Received: 31 January 2022 | Accepted: 15 February 2023

DOI: 10.1111/mafi.12389

ORIGINAL ARTICLE

MATHEMATICAL FINANCE An International Journal of Mathematics, Statistics and Financial Economics WILEY

# Designing universal causal deep learning models: The geometric (Hyper)transformer

**Beatrice Acciaio**[1] | **Anastasis Kratsios**[2] | **Gudmund Pammer**[1]

[1]Department of Mathematics, ETH Zürich, Zürich, Switzerland

[2]Department of Mathematics and Statistics, McMaster University, Hamilton, Ontario, Canada

**Correspondence**
Anastasis Kratsios, Department of Mathematics and Statistics, McMaster University, Hamilton, Ontario, Canada.
Email: kratsioa@mcmaster.ca

**Funding information**
H2020 European Research Council, Grant/Award Number: 852821; ETH Zürich Foundation; McMaster University and the NSERC Discovery, Grant/Award Number: RGPIN-2023-04482

**Abstract**

Several problems in stochastic analysis are defined through their geometry, and preserving that geometric structure is essential to generating meaningful predictions. Nevertheless, how to design principled deep learning (DL) models capable of encoding these geometric structures remains largely unknown. We address this open problem by introducing a universal causal geometric DL framework in which the user specifies a suitable pair of metric spaces $\mathcal{X}$ and $\mathcal{Y}$ and our framework returns a DL model capable of causally approximating any "regular" map sending time series in $\mathcal{X}^{\mathbb{Z}}$ to time series in $\mathcal{Y}^{\mathbb{Z}}$ while respecting their forward flow of information throughout time. Suitable geometries on $\mathcal{Y}$ include various (adapted) Wasserstein spaces arising in optimal stopping problems, a variety of statistical manifolds describing the conditional distribution of continuous-time finite state Markov chains, and all Fréchet spaces admitting a Schauder basis, for example, as in classical finance. Suitable spaces $\mathcal{X}$ are compact subsets of any Euclidean space. Our results all quantitatively express the number of parameters needed for our DL model to achieve a given approximation error as a function of the target map's regularity and the geometric

structure both of $\mathcal{X}$ and of $\mathcal{Y}$. Even when omitting any temporal structure, our universal approximation theorems are the first guarantees that Hölder functions, defined between such $\mathcal{X}$ and $\mathcal{Y}$ can be approximated by DL models.

**KEYWORDS**
adapted optimal transport, geometric deep learning, hypernetworks, metric geometry, random projection, stochastic processes, transformer networks, universal approximation

## 1 | INTRODUCTION

Due to breakthroughs in machine learning, optimization, and computing hardware, the last half-decade has seen a paradigm shift in many areas of applied mathematics, moving away from model-based approaches to model-free methods. This is most apparent in computational stochastic analysis and mathematical finance, where deep learning has unlocked previously intractable problems. Examples include computation of optimal hedges under market frictions and possibly rough volatility Buehler et al. (2019); Carbonneau and Godin (2021); Gierjatowicz et al. (2020); Horvath et al. (2021), numerical implementation of complicated local stochastic volatility models Cuchiero et al. (2020), numerical solutions to previously intractable principal-agent problems Campbell et al. (2021), pricing of derivatives relying on optimal stopping rules written on high-dimensional portfolios Becker et al. (2019, 2021); Herrera et al. (2021), data-driven prediction of price formation using ultra-high dimensional limit orderbook data Sirignano and Cont (2019); Zhang et al. (2019).

These deep learning-based methods are appealing not only because of their empirical success, but they are equally theoretically founded and are known to be able to approximately implement any "reasonable" function. This latter feature of deep neural networks is known as their *universal approximation property* and is the focal topic of this paper, in the context of non-anticipative functions between discrete-time path spaces. The universal approximation capabilities of classical neural network models are well-understood Hornik et al. (1990); Leshno et al. (1993); Kidger and Lyons (2020); Kratsios and Bilokopytov (2020). Nevertheless, little is known about whether or not neural network-based models can be used to approximate general stochastic processes or how to design a deep neural model which could.

The approximate implementation of "any" stochastic process' evolution, conditioned on its realized trajectory, is, of course, central to various areas of applied probability since this would help bridge the gap between abstract theoretical models and algorithmically deployed models. Though our examples are framed in the context of adapted optimal transport and mathematical finance, various other intersectional areas of machine learning and applied probability theory can utilize the results obtained in the present paper, such as computational signal processing, numerical weather predictions, and many others.

At this point, we make the leap towards abstraction and observe that the universal approximation of a stochastic process' evolution, conditioned on its realized trajectory, is only a special case of a broader phenomenon which we call a *causal map*. Briefly, given two metric spaces $\mathcal{X}$ and $\mathcal{Y}$, a causal map $F : \mathcal{X}^{\mathbb{Z}} \to \mathcal{Y}^{\mathbb{Z}}$ is a function which maps discrete-time paths in $\mathcal{X}$ to discrete-time paths in $\mathcal{Y}$ while respecting the *causal forward-flow of information* in time. We refer to this as

the *dynamic case* which incorporates a temporal flow into the *static case* $f : \mathcal{X} \to \mathcal{Y}$. In the case of stochastic processes, $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y}$ can be thought of as a space of laws of a process on a prespecified number of future steps, such as the Adapted Wasserstein space of Rüschendorf (1985), or a Fréchet space of random-vectors such as a local $L^p$-space. Nevertheless, the analysis we develop here is general enough to cover a broad range of discrete-time paths on suitable metric spaces $\mathcal{X}$ and $\mathcal{Y}$, where $\mathcal{X}$ is a subspace of an Euclidean space and $\mathcal{Y}$ is approximately representable by Euclidean information.

Concisely, our paper's main objective is to approximate any causal map between suitable discrete-time path spaces, which can have an arbitrarily long memory but which isn't overly reliant on the infinite past. We do so by proposing a new geometric deep learning model, called *geometric hypertransformers (GHTs)*, which naturally adapts to $\mathcal{Y}$'s (non-Euclidean) geometry, and our main quantitative result can be informally summarized as:

*GHTs can approximate any causal map $F : \mathcal{X}^{\mathbb{Z}} \to \mathcal{Y}^{\mathbb{Z}}$ over any time-horizon.*

We intentionally leave the description of what "any causal map" is and in which sense they are "approximated" vague at this point. This is to reflect the *modular nature* of our modeling framework, which can accommodate a broad range of different processes and modes of approximation. Examples include the approximation of any "integrable" stochastic process in a way which respects its adapted flow of information in the sense of Rüschendorf (1985); Backhoff-Veraguas et al. (2020b, 2020), any square-integrable martingale in the Martingale-Hardy sense (Weisz, 1994), a broad range of minimal parametric models in the sense of information geometry Amari (2016); Ay et al. (2015, 2017) but cast in the stochastic analysis setting, and the extreme but classical case of discrete-time dynamical systems (which we understand as deterministic processes).

Our neural network model, illustrated in Figure 1, emulates $F$ by only flowing information forward in time. The vertically-placed *black boxes* in Figure 1 are $\mathcal{Y}$-valued counterparts of the *transformer networks* of Vaswani et al. (2017), and variants of the probabilistic transformer networks of Kratsios et al. (2022); Kratsios (2023); Kratsios et al. (2022). Transformer networks are particularly appealing since, unlike recurrent neural networks, they can automatically process any input sequence without any recursion, which makes them much faster and more stable to train. Two advantages that transformers have over their recurrent neural network (RNN) Rumelhart et al. (1986) predecessors are that they avoid recursion and they learn how to encode any inputs before decoding them as predictions. This allows transformers to avoid lengthy and unstable training, and it gives them the flexibility to focus on different features of a path without overemphasizing its most recent movements. Thus, transformers have redefined the state of the art in sequential prediction tasks by effectively replacing LSTMs Hochreiter and Schmidhuber (1997).

Since we are not focusing on time-invariant $F$, that is $F$ which commute with time-shifts (i.e., $F(x_{\cdot+1})_{\cdot} = F(x_{\cdot})_{\cdot+1}$), then the parameters defining an GHT *efficiently* approximating $F$ must progressively update to accommodate changes in $F$. These updates, to our model are depicted in Figure 1 by horizontal arrows are implemented by a very small neural network known as a *hypernetwork* by Ha et al. (2017), acting on our model's parameter space. Its role is to interpolate the parameters of each of our *geometric transformer networks*, each acting on sequential segments of any input path from time $t_n$ to time $t_{n+1}$ across all $n \in \mathbb{Z}$ (depicted as vertical networks in Figure 1). The point here is a quantitative one, namely, it is known that neural networks require far fewer parameters to memorize (or interpolate) a finite set of input and output pairs

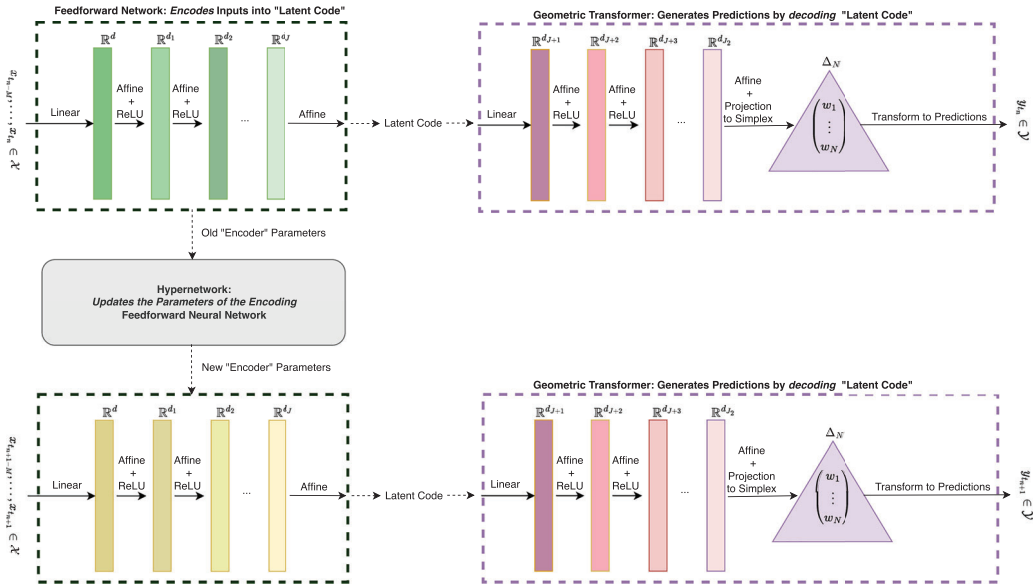



**FIGURE 1** Illustration of the Geometric Hypertransformer (GHT): At every time-step, a feedforward neural network, illustrated by the green and yellow boxes, maps the current time-series segment in $\mathcal{X}$ to a "latent code" in some Euclidean space $\mathbb{R}^{d_J}$. The geometric transformer, illustrated by the purple box, then transforms that "latent code" into the next prediction on $\mathcal{Y}$. Between each prediction the hypernetwork, illustrated by the gray box, updates the feedforward network's hidden weights so that the GHT architecture can adapt to changes in the time-series. [Color figure can be viewed at wileyonlinelibrary.com]

(Vershynin, 2020; Park et al., 2021; Yun et al., 2019) than what is required to uniformly approximate a function implementing this memorization.

In the context of stochastic processes, our model can be visualized using Figures 1 and 2. Briefly, given a (possibly infinite) $d$-dimensional path $\boldsymbol{x}$, our GHT model sequentially assimilates long segments of $\boldsymbol{x}$ before forecasting the process' distribution over a fixed number[1] $N_F \in \mathbb{N}_+$ of future steps. Then, once a new portion of the path is observed, our GHT model's internal parameters are updated, and the next process' law over the next $N_F$ increments, conditioned on the observed path, is predicted. We illustrate this sequential procedure in Figures 1 and 2 where the future set of steps of the process are color coded to indicate the network's corresponding weights.

From the deep learning side of this interdisciplinary story, and to the best of authors' knowledge, the results presented here are the first approximation theoretic result advocating for the effectiveness of *hypernetworks*. Thus, we complement the extensive emerging empirical studies on hypernetworks (see Ha et al. (2017); Zhang et al. (2019); von Oswald et al. (2020)).

The hypernetwork defining the GHT weaves together instances of our main model for the static case, which we call *geometric transformers (GTs)*. Even in the static case, we obtain a novel universal approximation theorem which can be summarized as

$$GTscanapproximateanyHölderfunctionf:\mathcal{X}\to\mathcal{Y},$$

where $\mathcal{Y}$ is as above, and $\mathcal{X}$ is a subspace of the Euclidean space $\mathbb{R}^d$. We emphasize that the output spaces covered by our results are much broader than the theoretically backed neural network architectures available in the literature. The closest principled sequential deep learning results to ours are the theory of reservoir computers Grigoryeva and Ortega (2019); Gonon and Ortega


**FIGURE 2** Illustration of a causal map which sends the sequence of realizations of a time-series to the law of its next few steps. In this depiction, each of the color represents the output of this causal map evaluated at different instances in time. [Color figure can be viewed at wileyonlinelibrary.com]

(2020), especially echo-state networks Grigoryeva and Ortega (2018a); Gonon and Ortega (2021), as well as single-layer recurrent neural networks Hutter et al. (2021), each of which is known to be able to approximate specific classes of generalized discrete-time dynamical systems evolving in Euclidean spaces.

To fully appreciate geometric hypertransformers, let us overview the current state of the art in theoretically-founded modeling of stochastic processes. When approximating jump diffusion processes, Gonon and Schwab (2021) showed that solutions to stochastic differential equations whose coefficients are parameterized by neural networks, called *neural SDEs* Chen et al. (2018); Jia and Benson (2019, ?); Kidger et al. (2021); Morrill et al. (2021), can approximate any *single* expected path-functional of a Markovian SDE with uniformly bounded Lipschitz coefficients. Nevertheless, the approximation of the law of any regular SDE remains an open problem. Alternatively, reservoir computers of Jaeger and Haas (2004); Maass et al. (2002) have been demonstrated to approximate time-invariant random dynamical systems with possibly infinitely long memory but *fading memory* Gonon et al. (2020); Grigoryeva & Ortega (2019) and in Cuchiero et al. (2021) using *random signature-based approaches* drawing from rough path theory Lyons (1994).

Circling back to the origins of recurrent models, we arrive at the extreme case of deterministic, stochastic processes, by which we mean discrete-time dynamical systems with an arbitrarily long memory. These are typically approximated by using RNNs with the LSTM architecture of Hochreiter and Schmidhuber (1997). Classical RNNs have long been known to have the capacity to approximate any *computable function* Siegelmann and Sontag (1995), and preliminary results indicating that they are universal approximators of any regular dynamical systems on $\mathbb{R}^d$ were suggested by Schäfer and Zimmermann (2006). Nevertheless, it has only recently been shown in Hutter et al. (2021) that classical RNNs, which are not reservoir computers, can approximate any time-inhomogeneous linear dynamical system on $\mathbb{R}^d$ with rapidly decaying memory.

Recently, various deep learning models mapping into infinite dimensional Banach spaces have been proposed such as the DeepONets of Lu et al. (2021); Liu et al. (2022), the Fourier Neural Operators of Kovachki et al. (2021), and the generalized feedforward model of Benth et al. (2023)

generalized to Fréchet spaces (in the special case where the space of processes is a Martingale-Hardy space or a similar linear space of semi-Martingales, see (Cohen & Elliott, 2015, Part IV)). Our models can cover approximation of functions taking values in Banach spaces, and, more generally in Fréchet spaces. Most notably our results are not limited to linear spaces and they cover a broad range of non-vectorial metric spaces: Wasserstein spaces, adapted Wasserstein spaces, many spaces arising from information geometry, and many others.

It is worth noting that simultaneous vertical and horizontal connections have very recently and successfully been proposed in the few-shot image classification literature in the hypertransformer model of Zhmoginov et al. (2022). In analogy with Figure 1, the vertical networks are the standard transformers of Vaswani et al. (2017) and the horizontal networks are convolutional neural networks LeCun et al. (1989).

## 1.1 | Deep learning in mathematical finance

To fully gauge the potential impact of our work in mathematical finance, we briefly highlight some of the recent advances made possible by the deep learning methods in finance, in relation to sequential learning. Recently, many novel neural network-based frameworks have been proposed for market simulation and data-driven model selection. The ability to accurately model time-series data is critical for numerous applications in the finance industry, and neural networks offer different ways to tackle these problems as alternatives to classical approaches. In particular, synthetic generation of time-series can be performed in a completely data-based fashion without imposing assumptions on the underlying stochastic dynamics. The generated data can be used, for example, to facilitate training and validation of models. Time-series generation has been successfully approached via autoencoders as, for example, in Buehler et al. (2020); Wiese et al. (2021), and via adversarial generation, as in Koshiyama et al. (2019) with classical GANs, in Acciaio and Krach (2022) with newly developed transport-theoretic techniques, and in Ni et al. (2021, 2020) in conjunction with signature method. Other than direct path generation, neural networks have been employed as function approximators for drift and diffusions of the modeled SDE system, for robust and data-driven model selection mechanisms, as for example, done in Arribas et al. (2020); Cuchiero et al. (2020); Gierjatowicz et al. (2020); Wiese et al. (2020); Cohen et al. (2021); Kidger et al. (2021); van Rhijn et al. (2021).

## 1.2 | Organization of paper

Our paper is organized as follows. Section 2 introduces the relevant mathematical and machine learning tools to formulate our geometric deep learning problem. Section 3 covers our results in the static case, where we introduce a broad class of non-Euclidean metric spaces $\mathcal{Y}$ as well as our geometric deep learning model, before showing that our model can approximate any Hölder function defined on a compact subset $\mathcal{X}$ and taking values in $\mathcal{Y}$. Section 4 then addresses the dynamic case by first introducing causal maps between the discrete-time path spaces $\mathcal{X}^{\mathbb{Z}}$ and $\mathcal{Y}^{\mathbb{Z}}$, and then describing how the recurrent extension of our static model can approximate any causal map whose memory asymptotically vanishes. Section 5 provides examples of spaces and causal maps covered by our framework, in particular relating to stochastic analysis and mathematical finance. Section 7 contains all the paper's proofs, auxiliary metric-theoretic definitions, and needed technical analytic lemmas.

## 2 | PRELIMINARIES

### 2.1 | Background

We recognize the interdisciplinary nature of this paper, and therefore, this section covers the necessary background in the areas of optimal transport, deep learning, and metric geometry. The reader familiar with any of these areas is encouraged to skip the corresponding introductory section.

#### 2.1.1 | Optimal transport

A central feature of optimal transport is its ability to lift the geometry of a base space onto the set of probabilities. We refer to Figalli (2010) for a detailed introduction to optimal transport. Typically for the field, we write, when $(\mathcal{X}, d_{\mathcal{X}})$ is a metric Polish space, $\mathcal{P}(\mathcal{X})$ for the set of all Borel probability measures on $\mathcal{X}$ which is then equipped with the topology of weak convergence of measures. When $1 \leq p < \infty$, we denote by $\mathcal{P}_p(\mathcal{X})$ the subset of probabilities that finitely integrate $x \mapsto d_{\mathcal{X}}(x, x_0)^p$ for some (and thus for any) $x_0 \in \mathcal{X}$. Similarly, we equip $\mathcal{P}_p(\mathcal{X})$ with the Wasserstein $p$-distance $\mathcal{W}_p$, that is, for $\mu, \nu \in \mathcal{P}_p(\mathcal{X})$, the metric defined by

$$\mathcal{W}_p(\mu,\nu) \overset{\text{def.}}{=} \inf_{\pi \in \mathrm{Cpl}(\mu,\nu)} \left\{ \int d_{\mathcal{X}}(x,y)^p \, \pi(dx,dy) \right\}^{1/p},$$

where $\mathrm{Cpl}(\mu,\nu) \overset{\text{def.}}{=} \{\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{X}) : \pi \text{ has first marginal } \mu, \text{ second marginal } \nu\}$.

In the context of mathematical finance, the *adapted Wasserstein distance*, that is a variation of the Wasserstein distance, can be used to obtain sharp quantitative results for sequential decision making problems and in robust finance, see Backhoff-Veraguas et al. (2020a); Bartl et al. (2021) among others. For $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^{dT})$, the adapted Wasserstein-$p$ distance is defined by

$$\mathcal{AW}_p(\mu,\nu) \overset{\text{def.}}{=} \inf_{\pi \in \mathrm{Cpl}_{\mathrm{bc}}(\mu,\nu)} \left\{ \int \sum_{t=1}^{T} |x_t - y_t|^p \pi(dx,dy) \right\}^{1/p},$$

where $\mathrm{Cpl}_{\mathrm{bc}}(\mu,\nu)$ is the set of probability measures $\pi \in \mathrm{Cpl}(\mu,\nu)$ satisfying the bi-causality constraint: for all $t = 1, \dots, T$,

$$\pi(dy_1, \dots, dy_t | x_{1:T}) = \pi(dy_1, \dots, dy_t | x_{1:t}) \quad \text{and} \quad \pi(dx_1, \dots, dx_t | y_{1:T}) = \pi(dx_1, \dots, dx_t | y_{1:t}),$$

where we use $x_{1:T}$ to denote the vector $(x_1, x_2, \dots, x_T)$. We note that if $T = 1$, $(\mathcal{P}_p(\mathbb{R}^{dT}), \mathcal{AW}_p)$ coincides precisely with the Wasserstein space $(\mathcal{P}_p(\mathbb{R}^d), \mathcal{W}_p)$.

#### 2.1.2 | Feedforward neural networks with parametric activation functions

In this section we recall the definition of *feedforward neural networks*. Introduced in the landmark paper Palm (1944) as a model for the cognition, a feedforward neural network is a function

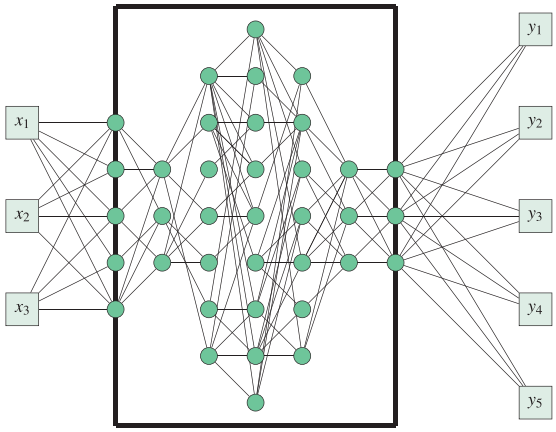



**FIGURE 3** A feedforward neural networks is a **black box** deep learning model which, given enough neurons, can be trained to approximate any function which continuously maps inputs in $\mathbb{R}^3$ to outputs in $\mathbb{R}^5$, uniformly on compact sets. [Color figure can be viewed at wileyonlinelibrary.com]

between Euclidean spaces which processes an input by iteratively applying affine maps with a fixed component-wise non-linear function $\sigma : \mathbb{R} \to \mathbb{R}$ called an *activation function*.

Feedforward neural networks (often abbreviated simply as neural networks) are often interpreted as *black boxes* and frequently visualized pictorially as in Figure 3, where each green circle represents an application of the activation function $\sigma : \mathbb{R} \to \mathbb{R}$ sometimes called a *neuron*, the sequentially black lines called *connections* represent univariate affine transformations of the preceding neuron's output until the network's outputs are produced. Note that the analogy with neurocomputing arises when considering $\sigma(x) = I_{\{x \geq 0\}}$ where the state 1 represents the "activation/firing" of a neuron upon receiving an input signal.

Recently, there has been significant theoretical work on activation functions which can adapt to their inputs, meaning that $\sigma$ is not a single function but rather a family of functions (Beknazaryan, 2021; Jiao et al., 2021; Ramachandran et al., 2018; Yarotsky & Zhevnerchuk, 2020; Yarotsky, 2021a). The advantage here is that an aptly chosen family of activation functions can "collaborate" to achieve superior approximation rates or help the original network exhibit desirable properties, such as being able to easily implement the identity and thus overcome overfitting Hardt and Ma (2016) and allow to represent the same function with fewer parameters Cheridito et al. (2021). Accordingly, our neural network models will depend on activation functions.

**Definition 2.1** (Trainable Activation Functions). A *trainable activation function* is any map $\sigma : \mathbb{R}^2 \times \mathbb{R} \ni (\alpha, x) \mapsto \sigma_\alpha(x) \in \mathbb{R}$.

We will consider neural network models built using either of three different types of activation's functions, "singular and trainable", "smooth non-polynomial and trainable", and classical continuous activation functions which are non-trainable and exhibit some basic regularity. Our quantitative approximation results depend on the choice of activation function.

Our interest in trainable activation functions, with discontinuities, stems from and adds to the emerging body of literature on the improved expressibility Yarotsky (2021b); Shen et al. (2021a); Park et al. (2021); Shen et al. (2021b); Kratsios and Zamanlooy (2022). Briefly, such activation functions allow us to build neural networks achieving exponential approximation rates. Thus, they provide an upper-bound for the trainable *singular* activation approximation rate which can be achieved by a practically deployed deep learning model.

In what follows, for any $x \in \mathbb{R}$, we denote $\lfloor x \rfloor \overset{\text{def.}}{=} \max\{n \in \mathbb{Z} : n \leq x\}$.

**Definition 2.2** (Trainable Activation Function: Singular-ReLU Type). A trainable activation function $\sigma$ is of *ReLU+Step type* if

$$\sigma_\alpha : \mathbb{R} \ni x \mapsto \alpha_1 \max\{x, \alpha_2 x\} + (1-\alpha_1)\lfloor x \rfloor \in \mathbb{R}.$$

For most $\alpha \in \mathbb{R}^2$, the singular-ReLU type activation function $\sigma_\alpha$ is discontinuous. We juxtapose the rates we derive with these singular types of trainable activation functions, with neural networks built using non-singular (i.e., smooth) and trainable activation functions; popular are the Swish activation function (Ramachandran et al., 2018), the analytic and periodic activation functions used in SIRENs Sitzmann et al. (2020) and studied in Siegel and Xu (2020), and GeLU Hendrycks and Gimpel (2016), and the classical sigmoid activation function. By the quantitative results of Kidger & Lyons (2020), we know that polynomial activation functions do generate universal approximators in the deep regime, unlike the shallow regime of Pinkus (1999). However, the quantitative results of Kratsios (2023) show that such activation functions are strictly less expressive than their non-polynomial counterparts; thus we do not consider such activation functions as a distinguished special case.

**Definition 2.3** (Trainable Activation Function: Smooth-ReLU-Type). A trainable activation function $\sigma$ is of *smooth non-polynomial type* if there is a non-polynomial $\sigma^\star \in C^\infty(\mathbb{R})$, for which

$$\sigma_\alpha : \mathbb{R} \ni x \mapsto \alpha_1 \max\{x, \alpha_2 x\} + (1-\alpha_1)\sigma^\star(x) \in \mathbb{R}.$$

For completeness, we also consider the general case consisting of activation functions which do not fall into either of these two classes. The most common example of such an activation function is the ReLU unit of Fukushima and Miyake (1982). We only require the regularity condition of Kidger & Lyons (2020).

**Definition 2.4** (Classical Activation Function). Let $\sigma^\star \in C(\mathbb{R})$ be non-affine and such that there is some $x \in \mathbb{R}$ at which $\sigma$ is differentiable and has non-zero derivative. Then, $\sigma$ is a *classical regular activation function* if, for every $\alpha \in \mathbb{R}^2$, $\sigma_\alpha = \sigma^\star$.

In what follows we will often be applying our trainable activation functions component-wise. For positive integers $n, m$, we denote the set of $n \times m$ matrices by $\mathbb{R}^{n\times m}$. More precisely, we mean the following operation defined for any $N \in \mathbb{N}_+$, $\bar{\alpha} \in \mathbb{R}^{N\times 2}$ with $i^{th}$ row denoted as $\bar{\alpha}_i$, and $x \in \mathbb{R}^N$, by

$$\sigma_{\bar{\alpha}} \bullet x \overset{\text{def.}}{=} \left(\sigma_{\bar{\alpha}_i}(x_i)\right)_{i=1}^N.$$

We may now formally define feedforward neural networks with trainable activation functions.

**Feedforward Neural Networks with Trainable Activation Functions:** Fix $J \in \mathbb{N}_+$ and a multi-index $[d] \overset{\text{def.}}{=} (d_0, \dots, d_{J+1})$, with $d_0, \dots, d_{J+1} \in \mathbb{N}_+$, and let $P([d]) = \sum_{j=0}^J d_{j+1}(d_j+3) - 2d_{J+1}$. We identify any vector $\theta \in \mathbb{R}^{P([d])}$ with

$$\begin{aligned} \theta \leftrightarrow \left((A^{(j)}, b^{(j)}, \bar{\alpha}^{(j)})_{j=0}^{J-1}, (A, c)\right), \\ (A^{(j)}, b^{(j)}, \bar{\alpha}^{(j)}) \in \mathbb{R}^{d_{j+1}\times d_j} \times \mathbb{R}^{d_{j+1}} \times \mathbb{R}^{d_{j+1}\times 2},\ A \in \mathbb{R}^{d_{J+1}\times d_J}, c \in \mathbb{R}^{d_{J+1}}. \end{aligned} \tag{1}$$

With the identification in (1), and similarly to Gribonval et al. (2021), we recursively define the representation function of a $[d]$-dimensional deep feedforward network by

$$\begin{aligned} \mathbb{R}^{P([d])} \times \mathbb{R}^{d_0} \ni (\theta, x) \mapsto \hat{f}_\theta(x) &\overset{\text{def.}}{=} A\, x^{(J)} + c, \\ x^{(j+1)} &\overset{\text{def.}}{=} \sigma_{\bar{\alpha}^{(j)}} \bullet (A^{(j)} x^{(j)} + b^{(j)}) \qquad \text{for } j = 0, \dots, J-1, \\ x^{(0)} &\overset{\text{def.}}{=} x. \end{aligned} \tag{2}$$

We denote by $\mathcal{NN}^{\sigma}_{[d]}$ the family of $[d]$-dimensional deep feedforward networks $\{\hat{f}_\theta\}_{\theta \in \mathbb{R}^{P([d])}}$ described by (2). The subset of $\mathcal{NN}^{\sigma}_{[d]}$ consisting of networks $\hat{f}_\theta$ with each $\bar{\alpha}_i^{(j)} = (1, 0)$ in (2) is denoted by $\mathcal{NN}^{ReLU}_{[d]}$ and consists of the familiar deep ReLU networks. The value $\max_{0 \le j \le J+1} d_j$ is $\hat{f}$'s *width*, $J$ is called $\hat{f}$'s *depth*[2], and $J + 1$ is the number of $\hat{f}$'s *layers*.

### 2.1.3 | Metric capacity

Many quantitative universal approximation results for classical feedforward networks exhibit the same approximation rate for any compact subset of the input space of the same *diameter*, defined for any $K \subseteq \mathcal{X}$ by $\operatorname{diam}(K) \overset{\text{def.}}{=} \sup_{x,y \in K} d(x, y)$; see Yarotsky and Zhevnerchuk (2020). However, one would expect that for "simpler compacts" approximation requires less complicated networks. Indeed, this is a feature of our main quantitative approximation result, which also relates the approximation quality to the complexity of the compact set on which the approximation holds. This complexity is expressed in two ways, first in terms of the regularity of each path $\boldsymbol{x} \in \mathcal{X}^{\mathbb{Z}}$, and second in terms of the size of the compact set in which each $x_t$ is to be approximated.

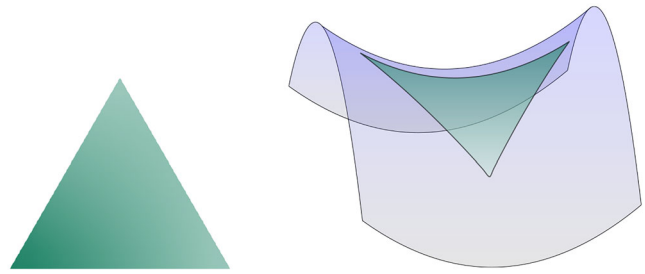
For the latter, we use a refinement of the usual notion of dimension, for smooth manifolds, which can distinguish between compact subsets of different intermediate "fractal dimensions". More specifically, we consider the *covering dimension* of a metric spaces, in the sense of (Heinonen, 2001, Section 10), which we control by the metric space's *metric capacity*, as defined in (Bruè et al., 2021a, Definition 1.6). Following Bruè et al. (2021a), we define the metric capacity of a *subset* $K \subseteq \mathbb{R}^d$ as the map $\kappa_K : (0, 1] \to \mathbb{N} \cup \{\infty\}$ which sends any $\delta > 0$ to

$$\kappa_K(\delta) \overset{\text{def.}}{=} \sup \left\{ k \in \mathbb{N} : \exists x_0, \dots, x_k,\ \exists r > 0 \text{ s.t. } \sqcup_{i=1}^{k} \operatorname{Ball}_K(x_i, \delta r) \subseteq \operatorname{Ball}_K(x_0, r) \right\}, \tag{3}$$

where, as usual, for any $x \in K$ and any $r > 0$, we set $\operatorname{Ball}_K(x, r) \overset{\text{def.}}{=} \{z \in K : \|z - x\| < r\}$, $\|\cdot\|$ the Euclidean norm, and $\sqcup$ denotes the union of disjoint sets. We make things concrete with the following two examples.

In what follows, we say that a function $f : \mathbb{N} \to \mathbb{R}$ *asymptotically grows at a linear rate*, written $f(n) \in \Theta(n)$, if there are constants $c, C > 0$ and $N \in \mathbb{N}$ such that $cn \le f(n) \le Cn$, for every $n \ge N$.

**Example 2.5** (Euclidean Spaces). For the $d$-dimensional Euclidean space, the discussions on (Heinonen, 2001, page 82) and (Bruè et al., 2021a, Proposition 1.7) imply that $\log_2(\kappa_{\mathbb{R}^d}(5^{-1})) \in \Theta(d)$.



(a) Inputs: $N$-Simplex (b) Outputs: Inscribed $N$-simplex in $\mathscr{Y}$.

**FIGURE 4** The mixing function $\eta$ inscribes any $N$-simplex into the approximately simplicial space $\mathscr{Y}$, sending any "weight" $w$ in the $N$-simplex $\Delta_N$ to the point $\eta(w, \mathcal{Y})$ in $\mathscr{Y}$ whose vertices are the set of $N$ points $\mathcal{Y}$ in $\mathscr{Y}$. [Color figure can be viewed at wileyonlinelibrary.com]

Similarly, when $K$ is a Riemannian submanifold of $\mathbb{R}^n$, the metric capacity is proportional to the intrinsic dimension of $K$, as defined in differential geometry. Thus, our approximation rates automatically adapt to paths which lie in low-dimensional submanifolds of our input space.

**Example 2.6** (Riemannian Submanifolds). Together, (Lang & Schlichenmaier, 2005, Proposition 2.7) and (Bruè et al., 2021a, Proposition 1.7) imply that, if $K$ is a $d$-dimensional compact Riemannian submanifold of $\mathbb{R}^n$, then $\log_2(\kappa_K(5^{-1})) \in \Theta(d)$.

Note that any subset $K \subseteq \mathbb{R}^d$ has finite metric capacity [3] , meaning that $\kappa_K(t) < \infty$ for all $t \in (0, 1]$.

## 3 | STATIC CASE: UNIVERSAL APPROXIMATION INTO QAS SPACES

We begin by presenting our results in the static case, where we find that a continuous function between suitable metric spaces can always be approximated by our geometric transformer model (Theorem 3.8). Our driving example is the approximation, in the (adapted) Wasserstein sense, of the transition kernels of higher order Markov processes; see Examples 3.6, 4.5, 4.8, and 5.11.

### 3.1 | The metric geometry of $\mathscr{Y}$

One cannot expect to have a universal approximation result for any metric space due to topological obstructions; see (Kratsios & Papon, 2022, Theorem 7). However, there is a rich class of output metric spaces $(\mathscr{Y}, d_{\mathscr{Y}})$ (abbreviated by $\mathscr{Y}$), encompassing most spaces encountered in stochastic analysis, and specified by exactly two simple conditions for which the approximation results presented in this paper apply.

The first condition on $\mathscr{Y}$, generalizes the existence of a *geodesic bicombing*, as studied by Lang and Plaut (2001); Descombes and Lang (2015); Basso (2018); Miesch (2018); Basso and Miesch (2019), and is analogous in spirit to the idea of a *simplicial topological space* from homotopy theory (Stacks project authors, 2021, Chapter 83.2) and from fuzzy set theory (Barr, 1986), and the peaked partitions of unity in metric space theory Semadeni (2006). Essentially, we require that any *inscribed simplex* in $\mathscr{Y}$ formed by joining any number of points in $\mathscr{Y}$ (illustrated by Figure 4(b)) is nothing else but a deformation of the standard Euclidean simplex with the same number of vertices (as illustrated by Figure 4(a)). In particular, if $\mathscr{Y}$ is a geodesic space, then the edges of the inscribed simplex in Figure 4(b) can be geodesics. However, this is by no means a requirement.

We denote the $N$-simplex by $\Delta_N \overset{\text{def.}}{=} \left\{ w \in [0,1]^N : \sum_{n=1}^N w_n = 1 \right\}$, and we set $\widehat{\mathcal{Y}} \overset{\text{def.}}{=} \bigcup_{N \in \mathbb{N}_+} (\Delta_N \times \mathcal{Y}^N)$.

**Definition 3.1** (Approximately Simplicial). A metric space $(\mathcal{Y}, d_{\mathcal{Y}})$ is said to be *approximately simplicial* if there is a function $\eta : \widehat{\mathcal{Y}} \to \mathcal{Y}$, called a mixing function, and constants $C_\eta \geq 1$ and $p \in \mathbb{N}_+$, such that for every $N \in \mathbb{N}_+$, $w = (w_1, \dots, w_N) \in \Delta_N$ and $\mathcal{Y} = (y_1, \dots, y_N) \in \mathcal{Y}^N$, one has

$$d_{\mathcal{Y}}(\eta(w, \mathcal{Y}), y_i) \leq C_\eta \left( \sum_{j=1}^N d_{\mathcal{Y}}(y_i, y_j)^p w_j \right)^{1/p}.$$

Note that, in particular, the function $\eta$ satisfies $\eta(e_i, \mathcal{Y}) = y_i$, for all $i = 1, \dots, N$, where $\{e_i\}_{i=1}^N$ denotes the standard basis of $\mathbb{R}^N$.

Only assuming that $\mathcal{Y}$ is approximately simplicial is not enough, since $\mathcal{Y}$ can be "too large" and universal approximation of any continuous function to $\mathcal{Y}$ by a "reasonable" model may be impossible. For example, if $\mathcal{Y}$ is non-separable, such as the family of càdlàg paths maps from $[0,1]$ to $\mathbb{R}$ with the uniform topology, then no sequence of models depending continuously on a finite number of real parameters can even approximate constant functions therein.

Our second condition on the output space's geometry requires that we can parameterize a dense subset of $\mathcal{Y}$. Moreover, we organized these parameterized subsets into a nested sequence of $\varepsilon$-nets which can be implemented by any idealized computer, capable of processing finite-dimensional real-vectors. In order for these $\varepsilon$-nets to be implementable, we require that each $\varepsilon$-net is explicitly parameterized by some Euclidean space. The sequence of the parameterizations is called a *quantization of* $\mathcal{Y}$; with the name drawing from the fact that our notion of quantization generalizes the quantization of probability measures (see Graf and Luschgy (2007); Liu and Pagès (2020)).

**Definition 3.2** (Quantizability). Let $(\mathcal{Y}, d_{\mathcal{Y}})$ be a metric space. If there is a sequence $Q \overset{\text{def.}}{=} (Q_q)_{q \in \mathbb{N}_+}$ of functions $Q_q : \mathbb{R}^{D_q} \to \mathcal{Y}$, with $D_q \in \mathbb{N}_+$, such that

(1) for any $q \in \mathbb{N}_+$ and $z \in \mathbb{R}^{D_q}$, there exists $\tilde{z} \in \mathbb{R}^{D_{q+1}}$ such that $Q_q(z) = Q_{q+1}(\tilde{z})$;
(2) for every $y \in \mathcal{Y}$ and $\varepsilon > 0$, there exist $q \in \mathbb{N}_+$ and $z \in \mathbb{R}^{D_q}$ satisfying

$$d_{\mathcal{Y}}\left(y, Q_q(z)\right) < \varepsilon,$$

then $(\mathcal{Y}, d_{\mathcal{Y}})$ is said to be *quantizable*. Furthermore, $Q$ is said to quantize $(\mathcal{Y}, d_{\mathcal{Y}})$.

We remark that it is advantageous not to require that the vectorial parameterization of these $\varepsilon$-nets be continuous, especially when $\mathcal{Y}$ is a discrete metric space . However, under the additional assumption that our parameterizations of these $\varepsilon$-nets by Euclidean inputs is continuous, then the quantizability of $\mathcal{Y}$ implies separability of its topology. Therefore, quantizability can be interpreted as a *quantitative metric analogue* of separability, which itself is an otherwise *qualitative topological property*.

Since quantizability is a quantitative property of $\mathcal{Y}$, then we measure the complexity required to parameterize an $\varepsilon$-net collection in any compact subset of $\mathcal{Y}$ by simply counting the minimum dimension of a Euclidean space required to parameterize that $\varepsilon$-net. The map sending an $\varepsilon$ and a

compact subset of $\mathcal{Y}$ to this minimum dimension is called a "quantization modulus". The role of the quantization modulus is therefore analogous to other moduli appearing in quantitative metric theories; examples include, moduli of continuity which quantifying metric distortion; moduli of smoothness in (Draganov & Ivanov, 2014; Heikkinen et al., 2016) quantifying higher-order distortions, or the modulus of padded decomposability quantifying the complexity of a random partition (Krauthgamer et al., 2004, 2005).

**Definition 3.3** (Modulus of Quantizability). In the notation of Definition 3.2. Let $(\mathcal{Y}, d_{\mathcal{Y}})$ be a metric space quantized by $Q = (Q_q)_{q\in\mathbb{N}_+}$. For any given compact $\mathcal{K} \subseteq \mathcal{Y}$, the *quantization modulus* of $Q$ on $\mathcal{K}$ is the map $\mathcal{Q}_{\mathcal{K}} : \mathbb{R}_+ \to \mathbb{N}$ that sends any $\varepsilon > 0$ to $\mathcal{Q}_{\mathcal{K}}(\varepsilon) \overset{\text{def.}}{=} D_{q_{\mathcal{K},\varepsilon}}$ where

$$q_{\mathcal{K},\varepsilon} \overset{\text{def.}}{=} \inf \left\{ q \in \mathbb{N}_+ : \forall y \in \mathcal{K}, \exists z \in \mathbb{R}^{D_q} \text{ s.t. } d_{\mathcal{Y}}(y, Q_q(z)) < \varepsilon \right\}.$$

The map $Q$ is called *regular* if $\mathcal{Q}_{\mathcal{K}}(\varepsilon)$ is finite for all $\varepsilon > 0$ and all compact subsets $\mathcal{K} \subseteq \mathcal{Y}$.

As we will see in Section 5, most metric spaces encountered when working with stochastic processes are both quantizable and approximately simplicial. Since we only consider output spaces with precisely these two properties, we name metric spaces carrying this additional structure.

**Definition 3.4** (QAS Space). A metric space $(\mathcal{Y}, d_{\mathcal{Y}})$ together with a function $\eta$ satisfying Definition 3.1 and a sequence $Q = (Q_q)_{q\in\mathbb{N}_+}$ satisfying Definition 3.2 will be called a *Quantizable and Approximately Simplicial space* (QAS space). We denote QAS spaces by the tuple $(\mathcal{Y}, d_{\mathcal{Y}}, \eta, Q)$.

The additional structure carried by QAS spaces allows us to approximately parameterized the inscribed simplices (as in Figure 4(b)) using Euclidean data projected onto the corresponding standard simplex (as in Figure 4(a)). This is because every QAS space naturally defines a neural network layer which approximately encodes $\mathcal{Y}$'s geometry, by mixing the attention payed to $N$ particles in $\mathcal{Y}$ (represented by the vertices of the inscribed simplex in Figure 4(b)) via the mixing function $\eta$.

**Definition 3.5** (Geometric Attention Mechanism). Let $(\mathcal{Y}, d_{\mathcal{Y}}, \eta, Q)$ be a QAS space. Then the *geometric attention mechanism* on $\mathcal{Y}$ is the family of functions $(\text{attention}_{N,q})_{N,q\in\mathbb{N}_+}$ defined by

$$\text{attention}_{N,q} : \mathbb{R}^N \times \mathbb{R}^{N\times D_q} \ni (u, (z_n)_{n=1}^N) \to \eta\left(\Pi_{\Delta_N}(u), (Q_q(z_n))_{n=1}^N\right) \in \mathcal{Y},$$

where $\Pi_{\Delta_N}$ is the projection of $\mathbb{R}^N$ onto the $N$-simplex $\Delta_N$.

The Wasserstein space $(\mathcal{P}_1(\mathbb{R}^2), \mathcal{W}_1)$ can be made into a QAS space (see Examples 5.5 and 5.6 below for details); in which case, a variant of the probabilistic attention mechanism of Kratsios et al. (2022) is a special case of our geometric attention mechanism; illustrated visually in Figure 5[4]. For instance, given three points $z_1, z_2, z_3$ in $\mathbb{R}^2$ which we wish to charge with mass, Figure 5 illustrates the map

$$\text{attention}_{3,1}(u, (z_1, z_2, z_3)) \overset{\text{def.}}{=} \sum_{n=1}^{3} w_n \delta_{z_n},$$

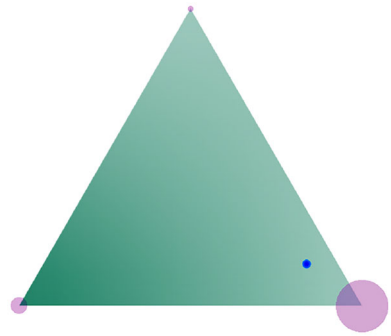


**FIGURE 5** Simplex inscribed in the Wasserstein Space $\mathcal{P}_1(\mathbb{R}^2)$. Here, the blue dot represents the normalized weight $w$ in the 3-simplex $\Delta_3$, the vertices of the simplex represent the points $z_1, z_2, z_3 \in \mathbb{R}^2$ to be charged with mass, and the size of the violet bubble around each vertex represents the measure $\sum_{n=1}^{3} w_n \delta_{z_n}$ charging the points $z_1, z_2, z_3$. [Color figure can be viewed at wileyonlinelibrary.com]

which sends any vector $u \in \mathbb{R}^3$ to the nearest normalized weight $w \overset{\text{def.}}{=} \Pi_{\Delta_3}(u)$ in the 3-simplex $\Delta_3$, as visualized by the $\bullet$. Then it distributes the weights in $\bullet$ amongst the three points $z_1, z_2, z_3$ in $\mathbb{R}^2$, with the amount of mass illustrated by the radius of the their respective violet bubbles.

More generally, given $N$ probability measures $y_1, \dots, y_N \in \mathcal{P}_1(\mathbb{R}^d)$, one can find empirical probability measures $\hat{y}_n \overset{\text{def.}}{=} \frac{1}{D_q} \sum_{q=1}^{D_q} \delta_{z_{n,q}}$ with $\{z_{n,q}\}_{n,q=1}^{N,D_q}$ in $\mathbb{R}^d$ which quantize $\{y_n\}_{n=1}^{N}$, that is, $\mathcal{W}_1(y_n, \hat{y}_n) \approx 0$; see (Chevallier, 2018, Corollary 3). In this case, the geometric attention becomes

$$\text{attention}_{N,q}(u, z) \overset{\text{def.}}{=} \sum_{n=1}^{N} \frac{w_n}{D_q} \sum_{q=1}^{D_q} \delta_{z_{n,q}} = \sum_{n=1}^{N} w_n \hat{y}_n, \tag{4}$$

where $w \overset{\text{def.}}{=} \Pi_{\Delta_N}(u)$ and $z \overset{\text{def.}}{=} (z_{n,q})_{n,q=1}^{N,D_q}$. In the next example, we extend this framework to the case of distributions on path spaces, and replace the Wasserstein distance with its *adapted counterpart* Rüschendorf (1985), which respects the flow of information.

**Example 3.6** (Adapted Wasserstein Space with Convex Combinations). We consider $(\mathcal{Y}, d_{\mathcal{Y}}) \overset{\text{def.}}{=} (\mathcal{P}_p([0,1]^{dT}), \mathcal{AW}_p)$. In difference to the Wasserstein distance case, empirical distributions are not consistent estimators with respect to the adapted Wasserstein distance. Thus, to quantize $(\mathcal{P}_p([0,1]^{dT}), \mathcal{AW}_p)$, we can not take functions as in (4). Instead, Backhoff et al. (2020) suggests the use of an *adapted empirical distribution*. Let $r = (T+1)^{-1}$ for $d = 1$, and $r = (dT)^{-1}$ for $d \geq 2$. For all $q \geq 1$, partition the cube $[0,1]^d$ into the disjoint union of $q^{rd}$ cubes with edges of length $q^{-r}$, and let $\varphi^q : [0,1]^d \to [0,1]^d$ map each small cube to its center. By (Backhoff et al., 2020, Theorem 1.3), the following family $(Q_q)_{q \in \mathbb{N}_+}$ of functions quantizes $(\mathcal{P}_p([0,1]^{dT}), \mathcal{AW}_p)$:

$$Q_q : \mathbb{R}^{dT \times q} \ni z = (z^1, \dots, z^q) \mapsto \frac{1}{q} \sum_{s=1}^{q} \delta_{(\varphi^q(z_1^s), \dots, \varphi^q(z_T^s))} \in \mathcal{P}_p([0,1]^{dT}).$$

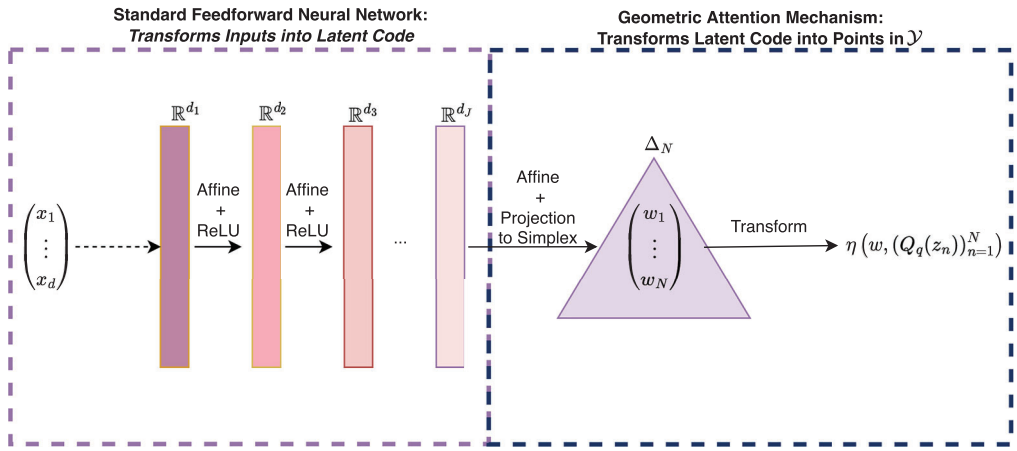



**FIGURE 6** An illustration of the geometric transformer architecture (see Definition 3.7). Euclidean inputs are first mapped to latent vectorial outputs by a feedforward neural network, these latent vectors are then transformed into predictions on $\mathcal{Y}$ by the geometric attention mechanism. [Color figure can be viewed at wileyonlinelibrary.com]

Thus, $(\mathcal{P}_p([0,1]^{dT}), \mathcal{AW}_p, \eta, Q)$ is a QAS space associated to the geometric attention mechanism

$$\text{attention}_{N,q}\left(u,(z^n)_{n=1}^N\right) \overset{\text{def.}}{=} \sum_{n=1}^{N} \frac{\Pi_{\Delta_N}(u)_n}{q} \sum_{s=1}^{q} \delta_{(\varphi^q((z^n)_1^s),\ldots,\varphi^q((z^n)_T^s))}.$$

Below, we introduce our static models, which we call *geometric transformers*. These work like the standard transformers of Vaswani et al. (2017), by decomposing the approximation of a suitable function $f : \mathbb{R}^d \to \mathcal{Y}$ into two steps. In the first step, a deep feedforward network learns how to encode any input $x \in \mathcal{X} \subseteq \mathbb{R}^d$ into a vector in a deep feature space $\mathbb{R}^N$. This step is illustrated in Figure 6 by the feedforward network sending the input $(x_1, x_2, x_3)$ to the deep features $(u_1, \ldots, u_5)$. In the second step, illustrated by the rectangular purple node in Figure 6, the deep features in $\mathbb{R}^N$ are decoded into $\mathcal{Y}$-valued predictions (illustrated in Figure 4(b)).

The decoding step is implemented by our geometric reinterpretation of the attention mechanism of Bahdanau et al. (2014), designed for Natural Language Processing tasks, and of its probabilistic counterpart of Kratsios (2023); Kratsios et al. (2022). The critical difference between our *geometric attention* and the above attention mechanisms is that the geometric attention is customized for $\mathcal{Y}$'s geometry, whereas the others are suited to their respective spaces.

**Definition 3.7** (Geometric Transformer). Let $(\mathcal{Y}, d_{\mathcal{Y}}, \eta, Q)$ be a QAS space and $d \in \mathbb{N}$. Fix a trainable activation function $\sigma$, constants $N, q \in \mathbb{N}_+$, and a multi-index $[d]$ with $d_0 = d$ and $d_J = N$. A *geometric transformer* (GT) from $\mathbb{R}^d$ to $\mathcal{Y}$ is a function $\hat{\rho} : \mathbb{R}^d \to \mathcal{Y}$ with representation

$$\hat{\rho} = \text{attention}_{N,q}\left(\hat{f}_\theta(\cdot), Y\right), \tag{5}$$

where $\hat{f}_\theta \in \mathcal{NN}^\sigma_{[d]}$ and $Y \in \mathbb{R}^{N \times D_q}$. The set $\mathcal{GT}^\sigma_{[d],N,q}(\mathbb{R}^d, \mathcal{Y})$ of geometric transformers from $\mathbb{R}^d$ to $\mathcal{Y}$ with activation function $\sigma$ of complexity $([d], N, q)$ consists of all $\hat{\rho}$ with representation (5).

In what follows, if already clear from the context, we will omit the specification "from $\mathbb{R}^d$ to $\mathcal{Y}$ with activation function $\sigma$ of complexity $([d], N, q)$" when talking of a geometric transformer.

**TABLE 1** Upper bounds on the model complexity of the geometric transformer network, $\hat{\rho} \overset{\text{def.}}{=} \text{attention}_{N,q}(\hat{f}_\theta(\cdot), Y)$ of Theorem 3.8. $N$ is the same for all activation functions.

| Activation ($\sigma$) | Singular (2.2) | Smooth (2.3) | Classical (2.4) |
|---|---|---|---|
| Depth ($J$) | $(N-1)(1+(2^6nD+3))$ | $\mathcal{O}((N-1)(1+\tilde{\varepsilon}^{-2n/\alpha}L_f^{2n/\alpha}(1+n/4)^{2n/\alpha}))$ | – |
| Width | $n(N-2)+\max\{n, 5W+13\}$ | $n(N-1)+3$ | n + N + 1 |
| # Parameters P([d]) | $(\frac{11}{4}n^2(N-1)^2-1)(N-1)\max\{n+3,5W+16\}^2(2^6nD+4)$ | $\mathcal{O}((\frac{11}{4}n^2(N-1)^2-1)(N-1)(n+6)^2(\tilde{\varepsilon}^{-4n/\alpha}L_f^{4n/\alpha}(1+n/4)^{4n/\alpha}+1))$ | $(N+n+1)^2$ (Depth + 1) |
| Implicit Parameter (D) | $\varepsilon_A = \sqrt{N}n^{\frac{\alpha}{2}}W^{-\sqrt{D}}(W^{(1-\alpha)\sqrt{D}}+2)$ | – | – |
| ln($N$) | $\ln(\kappa_{\mathcal{X}}(5^{-1}))\lceil\alpha^{-1}(\log_2(\text{diam}(\mathcal{X}))-\log_2(\varepsilon_A/3L_f)+\log_2((C_\eta 2c\lceil 1/\alpha\rceil\cdot\log_2(\kappa_{\mathcal{X}}(5^{-1})))_{++}))\rceil$ | | |
| q | $Q_{f(\mathcal{X})}(\varepsilon_Q)$ | | |

## 3.2 | Static case – Universal approximation into QAS spaces

We now present our first *universal approximation theorem* (Theorem 3.8). This result is both a refinement and a generalization of the classical universal approximation theorem, where we use geometric transformers in place of feedforward networks.

Following Kovachki et al. (2021); Liang et al. (2021), we decompose the total approximation error incurred by approximating a target function $f : \mathcal{X} \to \mathcal{Y}$ by a geometric transformer into three components: $\varepsilon_A, \varepsilon_Q, \varepsilon_{NN} > 0$. The first term $\varepsilon_A$ represents the *intrinsic error* incurred by approximately representing $f$ by a map defined on a specific simplex inscribed in $\mathcal{Y}$ (as in Figure 4(b)). Since this inscribed simplex' endpoints may not be representable by Euclidean features, we need to perturb them using the quantization maps $Q_q$, and the *quantization error* $\varepsilon_Q$ captures this. Then, $\varepsilon_{NN}$ represents the *encoding error* originated from the approximation of $f$ by a feedforward network, which approximately encodes the elements of $\mathcal{X}$ into deep features which are then passed to the geometric attention layer to be decoded into predictions on $\mathcal{Y}$. For $0 < \alpha \leq 1$, we denote the set of $\alpha$-Hölder functions from $\mathcal{X}$ to $\mathcal{Y}$ by $C^\alpha(\mathcal{X}, \mathcal{Y})$.

**Theorem 3.8** (Metric Transformers are Universal Approximators of QAS-Space-Valued Functions). *Let $0 < \alpha \leq 1$, $\mathcal{X} \subseteq \mathbb{R}^d$ be compact, $(\mathcal{Y}, d_{\mathcal{Y}}, \eta, Q)$ be a QAS space, and let $\sigma$ be a trainable activation function as in Definitions 2.2, 2.3, or 2.4. Then, for every $f \in C^\alpha(\mathcal{X}, \mathcal{Y})$, every "intrinsic error" $\varepsilon_A > 0$, "quantization error" $\varepsilon_Q > 0$, and "encoding error" $\varepsilon_{\mathcal{N}\mathcal{N}} > 0$, there exist positive integers $N, q \in \mathbb{N}_+$, a matrix $Y \in \mathbb{R}^{N \times D_q}$, and a geometric transformer $\hat{\rho} \in \mathcal{GT}^{\sigma}_{[d],N,q}(\mathbb{R}^d, \mathcal{Y})$ satisfying*

$$\sup_{x \in \mathcal{X}} d_{\mathcal{Y}}(f(x), \hat{\rho}(x)) \leq \varepsilon_A + \varepsilon_Q + \varepsilon_{\mathcal{N}\mathcal{N}},$$

*with $\hat{\rho}$ as in (5), and where the number of parameters determining $\hat{\rho}$ are recorded in Table 1.*

Table 3 reports the "space complexity" of the geometric transformer built in Theorem 3.8, with each column describing a different aspect of the universal approximation capabilities of our neural

**TABLE 2** Upper bounds on the model complexity of the feedforward network $\hat{f}_\theta$ in Proposition 3.10.

| Activation $\sigma$ | Singular (2.2) | Smooth (2.3) | Classical (2.4) (Kidger et al., 2021) |
|---|---|---|---|
| Depth ($J$) | $m(1+(2^6nD+3))$ | $\mathcal{O}(m(1+\tilde{\varepsilon}^{-2n/\alpha}L_f^{2n/\alpha}(1+n/4)^{2n/\alpha}))$ | – |
| Width | $n(m-1)+\max\{n, 5W+13\}$ | $nm+3$ | n + m + 2 |
| # Parameters P([d]) | $(\frac{11}{4}n^2m^2-1)m\max\{n+3, 5W+16\}^2(2^6nD+4)$ | $\mathcal{O}((\frac{11}{4}n^2m^2-1)m(n+6)^2(\tilde{\varepsilon}^{-4n/\alpha}L_f^{4n/\alpha}(1+n/4)^{4n/\alpha}+1))$ | $(n+m+2)^2$ (Depth + 1) |
| Implicit Parameter ($D$) | $\varepsilon = n^{\frac{\alpha}{2}}W^{-\sqrt{D}}(W^{(1-\alpha)\sqrt{D}}+2)$ | – | – |

network model. Moving from right to left, the third column in Table 1 confirms that "narrow" geometric transformers with classical (non-trainable) activation functions, such as ReLU, sigmoid, or the swish activation function of Ramachandran et al. (2018), have the capacity to approximate any function in $C^\alpha(\mathcal{X},\mathcal{Y})$.

The second column gives precise quantitative approximation rates for geometric transformers with trainable activation functions and smooth-ReLU-type. The deep feedforward used in building these models thus has the capacity to precisely implement the identity and can be trained using stochastic gradient descent methods, of course, depending on $\mathcal{Y}$. For details see Backhoff-Veraguas et al. (2022) for the case where $\mathcal{Y}$ is the Wasserstein space, Gallego et al. (2015) when $\mathcal{Y}$ is a suitable Banach space, and Bonnabel (2013); Tripuraneni et al. (2018); Alimisis et al. (2021) for the case where $\mathcal{Y}$ is a complete Riemannian manifold with bounded sectional curvature.

The first column of Table 3 covers the efficiency of transformer networks if one relaxes the need to have continuous activation functions. Though such neural network models cannot be trained using conventional stochastic gradient descent-type algorithms, they can be implemented using randomized approaches such as *random neural network* (or *extreme learning machine*); see Louart et al. (2018); Gonon et al. (2020); Kratsios and Zamanlooy (2022). Not only is this column most pertinent to transformer networks trained with randomized methods, but it also highlights the potential of transformer approaches to geometric deep learning, applications to stochastic analysis, and mathematical finance.

In what follows, for any $x\in\mathbb{R}$, we denote $\lceil x\rceil \overset{\text{def.}}{=} \min\{n\in\mathbb{Z} : n\geq x\}$ and $x_{++}\overset{\text{def.}}{=}\max\{1,x\}$.

*Remark* 3.9 (The Width Parameter in Tables 1 and 2). The width parameter $W$ in Proposition 3.10 only concerns the case where the approximating feedforward networks are wide and utilizes a trainable activation function of singular-ReLU-type (first column of Table 2). For example, one can take $W=\lceil\varepsilon^{-1}\rceil$.

A key step in the derivation of Theorem 3.8, is the following refinement of the central universal approximation theorem for deep feedforward networks with (possibly) trainable activation function. Briefly, the following result is a universal approximation theorem, which reflects not only the complexity of the target function being approximated and the size of the compact subset $K\subseteq\mathbb{R}^n$ on which the approximation holds, but also $K$'s *fractal dimension*. This refines many universal approximation theorems in the literature. For instance, it refines Yarotsky (2018) which

**TABLE 3** Compression rate required outside the prescribed time-horizon, in the universal approximation theorem of Theorem 4.11, as a function of the path space.

| Compact Path-space | Compression Rate $c_\varepsilon(n)$ for $\lvert n\rvert > N_T$ |
|---|---|
| $K^w$ | $4\varepsilon^{-1}\,\omega_{\rho_{\varepsilon/4}}\circ\omega_{f_{\varepsilon/4}}\left((\lvert n\rvert - N_T)\delta_+ + dm(\varepsilon/4)\,(\mathrm{diam}(K) + w(\lvert n\rvert) + w(N_T))\right)$ |
| $K^{\exp}_{\mathbb{C},C^\star,\varepsilon}$ | $4\varepsilon^{-1}\,\omega_{\rho_{\varepsilon/4}}\circ\omega_{f_{\varepsilon/4}}\left((\lvert n\rvert - N_T)\delta_+ + dm(\varepsilon/4)\,(\mathrm{diam}(K) + w_{\boldsymbol{C},C^\star}(n) + w_{\boldsymbol{C},C^\star}(N_T))\right)$ |
| $K^{\infty}_{C,p}$ | $4\varepsilon^{-1}L_{\alpha,\rho_{\varepsilon/4}}L^{\alpha}_{\alpha,f_{\varepsilon/4}}((\lvert n\rvert - N_T)\delta_+[1 + (dm(\varepsilon/4)+1)C^{\frac{1}{p}}\delta_+^{\frac{1-p}{p}}])^{\alpha^2}$ |
| $K^{\alpha}_{C,p}$ | $4\varepsilon^{-1}L_{\alpha,\rho_{\varepsilon/4}}L^{\alpha}_{\alpha,f_{\varepsilon/4}}((\lvert n\rvert - N_T)\delta_+ + (dm(\varepsilon/4)+1)C^{\frac{1}{p}}\delta_+^{\frac{1}{p}}(1 + 2\zeta(\frac{\alpha}{p-1}))^{\frac{p-1}{p}})^{\alpha^2}$ |
| $K^{\mathbb{Z}}$ | $4\varepsilon^{-1}L_{\alpha,\rho_{\varepsilon/4}}L^{\alpha}_{\alpha,f_{\varepsilon/4}}((\lvert n\rvert - N_T)\delta_+ + (dm(\varepsilon/4)+1)\mathrm{diam}(K))^{\alpha^2}$ |
| "Worst-Case" Arbitrary $\mathscr{K}$ | $\max_{\boldsymbol{x}\in\mathscr{K},\,k\le\lvert n\rvert}\ 4\varepsilon^{-1}\max\{1, d_{\mathscr{Y}}(F^{\rho_\varepsilon,f_\varepsilon}(\boldsymbol{x})_{t_k}, F^{\rho_\varepsilon,f_\varepsilon}(\boldsymbol{x})_{t_n})\}$ |

Here, $K\subseteq\mathbb{R}^d$ is compact, $C>0$, $\alpha<1-p$, $p\ge 1$ when $\mathscr{K}=K^{\alpha}_{C,p}$, and $p>0$ when $\mathscr{K}=K^{\infty}_{C,p}$, $L_{\alpha,\rho_\varepsilon}$, and $L_{\alpha,f_\varepsilon}$ are the $\alpha$-Hölder constants of $\rho_\varepsilon$ and $f_\varepsilon$ in (8), and $\zeta$ is the Riemann zeta function. When $\mathscr{K}=K^{\exp}_{\mathbb{C},C^\star,\varepsilon}$, $w_{\boldsymbol{C},C^\star}(n)\overset{\text{def.}}{=}\max\{C_0,\varepsilon^{-1/2}\,C^\star C_n^{1/2}e^{-nC_n\delta_-/2}\}$, for each $n\in\mathbb{N}_+$, the positive constants $\boldsymbol{C}=(C_n)_{n=0}^{\infty}$ and $C^\star$ are as in Proposition 4.2 and $\varepsilon>0$ can be taken to be the approximation error from Theorem 4.11.

concerns functions defined on the unit cube, it provides a quantitative version of Kidger et al. (2021), and it parallels the findings of Shaham et al. (2018) beyond the case where $K$ is a differentiable sub-manifold of $\mathbb{R}^n$; all while allowing for trainable activation functions. Note that in the special case where $\mathbb{R}^m=\mathbb{R}$, $K=[0,1]^n$, and $\sigma$ is of singular-ReLU type, we recover (Shen et al., 2021b, Theorem 1).

**Proposition 3.10** ($\mathcal{NN}^{\cdot}_{[d]}$-Networks are Efficient Universal Approximators). *Let $n,m\in\mathbb{N}_+$, $K\subseteq\mathbb{R}^n$ be a compact set with at least two points, $0<\alpha\le 1$, $f\in C^{\alpha}(K,\mathbb{R}^m)$, and let $\sigma$ be an activation function as in Definitions 2.2, 2.3 or 2.4. For every approximation error $\varepsilon>0$ and any width parameter $W\in\mathbb{N}_+$, there is a feedforward neural network $\hat{f}_\theta$ satisfying*

$$\sup_{x\in K}\|f(x)-\hat{f}_\theta(x)\|\le C_K\varepsilon,$$

*where the constant $C_K>0$ encodes the "complexity" of the input space $K$ and is defined by*

$$C_K\overset{def.}{=}c\sqrt{m}\lceil\alpha^{-1}\rceil\underbrace{\log_2\left(\kappa_K(5^{-1})\right)}_{\text{Dimension of }K}\underbrace{\mathrm{diam}(K)^{\alpha}}_{\text{Size of }K},\tag{6}$$

*where $c>0$ is an absolute constant[5] and $\kappa_K$ is defined in (3). Furthermore, the space complexity of $\hat{f}_\theta$ is recorded in Table 2.*

## 4 | DYNAMIC CASE – UNIVERSAL APPROXIMATION OF CAUSAL MAPS

This section's main result (Theorem 4.11) states that any function between discrete-time path spaces which flow information forward can be approximated by the dynamic extension of our geometric transformer neural network architecture.

## 4.1 | Compact subsets in discrete-time path spaces

We are interested in compact subsets $\mathcal{K}$ of the path space $\mathcal{X}^{\mathbb{Z}}$. In order to detail some different classes of paths which are relevant to our analysis, we need to fix the rate at which time flows, by specifying a time-grid $\{t_n\}_{n\in\mathbb{Z}} \subseteq \mathbb{R}$. Any $t_n < 0$ represents the past, $t_0 = 0$ is the present, and any $t_n$ are future times $t_n > 0$.

On the time-grid we only assume that it spans all time, and that it is not overly sparse nor overly clustered at any instance in time. These requirements are formalized by the following assumption. Given any $\boldsymbol{x} \in \mathcal{X}^{\mathbb{Z}}$, we define $\Delta_n \boldsymbol{x} \overset{\text{def.}}{=} x_{t_n} - x_{t_{n-1}}$, and set $\Delta t_n \overset{\text{def.}}{=} t_{n+1} - t_n$.

**Assumption 4.1** (Non-Degenerate Time-Grid). The time-grid $\mathbb{T} \overset{\text{def.}}{=} \{t_n\}_{n\in\mathbb{Z}} \subseteq \mathbb{R}$ satisfies: $t_0 = 0$ and

(i) **Spans all time:** $\inf_{n\in\mathbb{Z}} t_n = -\infty$ and $\sup_{n\in\mathbb{Z}} t_n = \infty$,
(ii) **Not overly sparse nor overly clustered:**

$$0 < \delta_- \overset{\text{def.}}{=} \inf_{n\in\mathbb{Z}} \Delta t_n \leq \sup_{n\in\mathbb{Z}} \Delta t_n \overset{\text{def.}}{=} \delta_+ < \infty.$$

Our dynamic universal approximation theorem will hold on compact subsets $\mathcal{K}$ of the path-space $\mathcal{X}^{\mathbb{Z}}$, which, by Tychonoff's Product Theorem (Munkres, 2000, Theorem 37.3), are of the form $\mathcal{K} = \prod_{n\in\mathbb{Z}} K_n$, where each $K_n \subseteq \mathcal{X}$ is compact. Our worst-case model complexity result concerns compact subsets of the path space at this level of generality.

Although we take inspiration for considering compact path spaces of this form from the machine learning for dynamical systems and reservoir computing literature Grigoryeva & Ortega (2019), we are equally motivated by stochastic analysis; specifically from the high-probability behavior of paths of solutions of stochastic differential equations (SDEs). The next result gives a precise statement motivating our first compact class of paths. We begin by considering the link between general stochastic differential equations, and the following compact path space

$$K^{\exp}_{\mathbb{C},C^\star,\varepsilon} \overset{\text{def.}}{=} \left\{ \boldsymbol{x} \in \mathcal{X}^{\mathbb{Z}} : \|x_0\| \leq C_0 \text{ and } (\forall n \in \mathbb{N}_+) \, \|x_{t_n}\| \leq \frac{C^\star}{\varepsilon^{1/2}} C_n^{1/2} e^{-n C_n \delta_-/2} \right\}, \tag{7}$$

which in addition to the time-grid $\mathbb{T}$ is defined by the following hyperparameters: a sequence of positive constants $\mathbb{C} \overset{\text{def.}}{=} (C_n)_{n\in\mathbb{Z}}$, $C^\star > 0$, and a fixed $0 < \varepsilon \leq 1$. We illustrate this phenomenon for SDEs with deterministic initial condition and a broad range of SDEs with random initial condition.

**Proposition 4.2** (Common Paths of SDEs Satisfy our Compactness Conditions). *Fix a time grid $\mathbb{T} \overset{\text{def.}}{=} \{t_n\}_{n\in\mathbb{Z}}$ satisfying Assumption 4.1 and a sequence of positive constants $\mathbb{L} \overset{\text{def.}}{=} (L_n)_{n\in\mathbb{Z}}$. Let $(\Omega, \mathcal{F}, \mathbb{F} \overset{\text{def.}}{=} (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$ be any filtered probability space supporting an $n$-dimensional Brownian motion $(W_t)_{t\geq 0}$, and any pair of $\mathbb{F}$-progressively measurable processes $(\alpha_t)_{t\geq 0}$ and $(\beta_t)_{t\geq 0}$ from $[0,\infty)\times\Omega\times\mathbb{R}^d$ to $\mathbb{R}^d$ and to the set of $d \times n$-matrices respectively, satisfying*

*(i) **Integrability:** For every $n \in \mathbb{N}_+$, $\mathbb{E}[\int_0^{t_n} \|\alpha_{t_n}(0)\|^2 + \|\beta_{t_n}(0)\|^2 dt] < \infty$;*

(ii) ***Local Lipschitz Regularity:*** *For every integer $n \in \mathbb{N}$, every $t \in [0, t_n]$, and every pair of points $x, y \in \mathbb{R}^d$, it holds that $\|\beta_t(x) - \beta_t(y)\| + \|\alpha_t(x) - \alpha_t(y)\| \leq L_n \|x - y\|$;*

(iii) ***Initial Condition****: Assume either that $X_0$ is a constant $x \in \mathbb{R}^d$ or $X_0$ is a $\mathcal{F}_0$-measurable sub-Gaussian random vector and is independent of $W_0$.*

*Let $(X_t)_{t\geq 0}$ be a solution to the SDE*

$$X_t = X_0 + \int_0^t \alpha_s(X_s)ds + \int_0^t \beta_s(X_s)\, dW_s, \qquad t \geq 0.$$

*Then, there exists a sequence of positive constants $(C_n)_{n\in\mathbb{Z}}$, depending only on the time-grid $\mathbb{T}$ and on the local Lipschitz constants $\mathbb{L}$, such that, for every $0 < \varepsilon \leq 1$, the discrete-time stochastic process $(X_{t_n})_{n\in\mathbb{N}}$ belongs to the compact subset $K^{\exp}_{\mathbb{C},C_{(\delta_-,\mathbb{L})},\varepsilon} \subseteq (\mathbb{R}^d)^{\mathbb{N}}$ with high probability:*

$$\mathbb{P}\Big((X_{t_n})_{n\in\mathbb{Z}} \in K^{\exp}_{\mathbb{C},C_{(\delta_-,\mathbb{L})},\varepsilon}\Big) \gtrsim (1-\varepsilon),$$

*where $C_{(\delta_-,\mathbb{L})} > 0$ depends only on $\delta_-$ and on $\mathbb{L}$, and $\gtrsim$ hides a constant depending only on $X_0$'s law.*

*Remark* 4.3. The constant suppressed by $\gtrsim$ in Proposition 4.2 is 1 if $X_0$ is deterministic, and when $X_0$ is sub-Gaussian, it is positive for large enough $L_0$.

Proposition 4.2 guarantees that, with high probability, the paths of an SDE with a random sub-Gaussian initial state belongs to a path space of the form (7). However, one can easily construct other types of stochastic processes whose typical paths lie in more general compact subsets of the product space $(\mathbb{R}^d)^{\mathbb{Z}}$. Similarly to the weighted approximation literature Prolla (1971); Schmocker (2022), a broad family of path-spaces can be defined by a compact subset $K \subseteq \mathbb{R}^d$ and a monotone increasing map $w : \mathbb{Z} \to [0, \infty)$, which in analogy with the reservoir computing Grigoryeva and Ortega (2018b); Grigoryeva & Ortega (2019) and approximation theory Prolla (1971); Schmocker (2022) literature, we call a *weighting function*. Together, the pair $(K, w)$ define the following path-space comprised of all paths in $(\mathbb{R}^d)^{\mathbb{Z}}$ that are $w(i)$-close to the compact set $K$ at time $t_i \in \mathbb{Z}$:

$$K^w \overset{\text{def.}}{=} \left\{\boldsymbol{x} \in (\mathbb{R}^d)^{\mathbb{Z}} : \forall i\, \exists y \in K \text{ s.t. } \|x_{t_i} - y\| \leq w(|i|)\right\}.$$

More broadly, regular classes of paths will yield smaller networks. In particular, we consider three classes of paths, $K^{\mathbb{Z}}$, $K^{\infty}_{C,p}$, and $K^{\alpha}_{C,p}$, induced by a fixed compact $K \subseteq \mathbb{R}^d$, and defined as follows. The first notable case considers uniformly bounded paths for all time. This case is typical within the reservoir computing literature, see for example, Grigoryeva and Ortega (2018b), and is formalized by setting $\mathcal{K} = K^{\mathbb{Z}}$.

The next case consists of paths that pass through $K$ at the present time, and whose time-increments uniformly control the $p$-variation. For any $C, p > 0$, we define

$$K^{\infty}_{C,p} \overset{\text{def.}}{=} \left\{\boldsymbol{x} \in (\mathbb{R}^d)^{\mathbb{Z}} : x_0 \in K \text{ and } (\forall n \in \mathbb{Z})\, \|\Delta_n \boldsymbol{x}\|^p \leq C|\Delta t_n|\right\}.$$

Our last distinguished compact class of paths in $(\mathbb{R}^d)^{\mathbb{Z}}$ mimics the approximation spaces of DeVore and Lorentz (1993), recently studied in the neural network context in Gribonval et al.

(2021), and describes paths whose $p$-variation rapidly converges to 0. This set consists of paths that pass through $K$ at the present time, and whose weighted $p$-variation converges when re-weighted by a factor of $|n|_{++}^{\alpha}$. Formally, for any fixed $C > 0$, $p \geq 1$, and any $0 < \alpha < 1 - p$, we define

$$K_{C,p}^{\alpha} \overset{\text{def.}}{=} \left\{ \boldsymbol{x} \in (\mathbb{R}^d)^{\mathbb{Z}} : x_0 \in K \text{ and } \sum_{n \in \mathbb{Z}} \frac{\|\Delta_n \boldsymbol{x}\|^p}{|\Delta t_n||n|_{++}^{\alpha}} \leq C \right\}.$$

Morally, $|n|_{++}$ is $|n|$ with the added technical point being that we avoid division by 0 when $n = 0$.

Next, we describe the functions between suitable discrete-time path spaces.

## 4.2 | Causal map of approximable complexity

We build on the ideas of the reservoir computing literature (Boyd & Chua, 1985; Jaeger, 2001, 2001; Grigoryeva and Ortega, 2019; Manjunath & Jaeger, 2013), and on that of non-anticipative functionals of Cont and Fournie (2010); Cont and Fournié (2013). In our setting this translates to maps $F : \mathcal{X}^{\mathbb{Z}} \to \mathcal{Y}^{\mathbb{Z}}$ between discrete-time path spaces which are *causal* in the following sense.

**Definition 4.4** (Causal Map). Given any two metric spaces $\mathcal{X}, \mathcal{Y}$, a map $F : \mathcal{X}^{\mathbb{Z}} \to \mathcal{Y}^{\mathbb{Z}}$ is called a *causal map* if, for every $t \in \mathbb{Z}$ and $x, x' \in \mathcal{X}^{\mathbb{Z}}$ with $x_s = x'_s$ for $s \leq t$, we have $F(x)_s = F(x')_s$.

Let us consider our first, and possibly most familiar, broad class of stochastic processes. This example frames solutions to discrete-time SDEs in the language of causal maps taking values in paths in Wasserstein spaces.

The crucial point here is that the framework of causal maps encompasses all classical discretized diffusion processes, such as any standard neural SDE model Cuchiero et al. (2020); Gierjatowicz et al. (2020); Kidger et al. (2021). We employ independent non-Gaussian noise in the SDEs in our example to illustrate that causal maps can comfortably describe much more complicated structures than what is describable with any diffusion model. Similarly, we highlight that the drift and diffusion coefficients of the SDE are of much lower regularity than what can be handled with the classical theory of diffusion (without resorting to stochastic differential inclusions Kisielewicz (2013)).

**Example 4.5** (Discrete-time SDEs (with Non-Gaussian Noise)). Let $p \geq 1$ and $(\mathcal{Y}, d_{\mathcal{Y}}) = (\mathcal{P}_p(\mathbb{R}^d), \mathcal{W}_p)$. Let $(W_n)_{n \in \mathbb{N}}$ be a sequence of independent standard Gaussians, and let $(Z_n)_{n \in \mathbb{Z}}$ be a sequence of i.i.d. random vectors in $\mathbb{R}^d$, independent of $(W_n)_{n \in \mathbb{N}}$, with $Z_0 \sim \nu \in \mathcal{P}_p(\mathbb{R}^d)$. For simplicity, we assume that $\delta \overset{\text{def.}}{=} \delta_+ = \delta_-$, where $\delta_+$ and $\delta_-$ are as in Assumption 4.1. Let the Hölder coefficients of the drift coefficients $\mu(t_n, \cdot) \in C^{\alpha}(\mathbb{R}^d, \mathbb{R}^d)$ and diffusion coefficients $\sigma(t_n, \cdot) \in C^{\alpha}(\mathbb{R}^d, \mathbb{R}^{d \times d})$ be uniformly bounded in $(t_n)_{n \in \mathbb{N}}$, and define at time $t_n$, for initial condition $X_{t_n} = x_{t_n} \in \mathbb{R}^d$, $X_{t_{n+1}}$ by

$$X_{t_{n+1}} = x_{t_n} + \delta \mu(t_n, x_{t_n}) + \sqrt{\delta} \sigma(t_n, x_{t_n}) W_n + Z_n.$$

This discrete-time SDE induces the causal map

$$F(\boldsymbol{x})_{t_n} \overset{\text{def.}}{=} \mathcal{L}(X_{t_{n+1}} | X_{t_n} = x_{t_n}),$$

where $\mathscr{L}(Y|X = x)$ denotes the (a.s. well-defined) conditional law of a random variable $Y$ given $X = x$.

*Remark* 4.6 (Higher Order Markovian SDEs). By extending the state space, Example 4.5 also covers higher order Markovian SDEs, that is, when drift and diffusion coefficients depend on finitely many past states. In order to ease reading, the example was presented in the Markovian case.

In principle, a causal map's memory and internal structure may be infinitely complicated since one can easily construct such maps which depend on the arbitrarily distant past. It would be surprising if these pathological causal maps can be approximated in any reasonable variant of the "uniform on compact sets" sense. Therefore, in analogy with Boyd and Chua (1985); Gonon et al. (2020), we also exclude such maps. Conversely, one would expect that any interesting universal approximation theorem for causal maps must encompass all causal maps that depend only on a finite (but potentially long) memory and process any path-segment using finite (but potentially large) number of Euclidean features.

**Definition 4.7** (Causal Map of Finite Complexity). Let $L, m \in \mathbb{N}_+$, $\alpha \in (0, 1]$, $f \in C^\alpha(\mathbb{R} \times \mathcal{X}^m, \mathbb{R}^L)$ (called the *encoding map*), and $\rho \in C^\alpha(\mathbb{R}^L, \mathcal{Y})$ (called the *decoding map*). We associate to $f$ and $\rho$ the system $F^{f,\rho} : \mathcal{X}^{\mathbb{Z}} \ni \boldsymbol{x} \mapsto (\rho(f(t_n, x_{t_{n-m}:t_n})))_{n\in\mathbb{Z}} \in \mathcal{Y}^{\mathbb{Z}}$, and call it a *causal map of finite complexity*. $F^{f,\rho}$ is said to be *time-homogeneous* if $f$ has no explicit dependence on the first (time) component.

By using the parabolic PDE representation of an SDE, offered by the Feynman-Kac formula, Hutzenthaler et al. (2020); Grohs et al. (2022) have shown that feedforward neural networks can efficiently approximate most SDE's associated PDE. Likewise, regular path-functionals of a jump-diffusion process with Lipschitz coefficients can efficiently be approximated by neural SDEs Gonon and Schwab (2021). In a similar spirit, we find that the causal maps of Example 4.5 are not only of finite complexity but, a fortiori, they admit a simple representation specified by only a small number of parameters.

**Example 4.8** (Discrete-time SDEs (with Non-Gaussian Noise)). Continuing Example 4.5, with the simplified assumption that $\delta = 1$, we show that the causal map $F$ is of finite complexity, since it can be expressed as $F(\boldsymbol{x})_{t_n} = \rho(f(t_n, x_{t_n}))$, where the encoding and decoding maps are defined by

$$f(t, x) \overset{\text{def.}}{=} (x + \mu(t, x), \sigma(t, x)) \quad \text{and} \quad \rho(\mu, \sigma) \overset{\text{def.}}{=} N_d(\mu, \sigma\sigma^\top) \star \nu,$$

where $\star$ denotes the convolution, and $N_d(\mu, \sigma\sigma^\top)$ a $d$-dimensional normal distribution with mean $\mu$ and covariance matrix $\sigma\sigma^\top$. Clearly, $f(t_n, \cdot) \in C^\alpha(\mathbb{R}^d, \mathbb{R}^d \times \mathbb{R}^{d\times d})$, and moreover we have

$$\begin{aligned}\mathcal{W}_p(\rho(\mu_1, \sigma_1), \rho(\mu_2, \sigma_2)) &\le \mathcal{W}_p\left(N_d\left(\mu_1, \sigma_1\sigma_1^\top\right), N_d\left(\mu_2, \sigma_2\sigma_2^\top\right)\right)\\ &\le \mathbb{E}[\|\mu_1 - \mu_2 + (\sigma_1 - \sigma_2) \cdot W_n\|^p]^{1/p}\\ &\le \|\mu_1 - \mu_2\| + \|\sigma_1 - \sigma_2\|\, \mathbb{E}[\|W_n\|^p]^{1/p},\end{aligned}$$

so that $\rho$ belongs to $C^1(\mathbb{R}^d \times \mathbb{R}^{d\times d}, (\mathcal{P}_p(\mathbb{R}^d), \mathcal{W}_p))$.

In general, causal maps of finite complexity extend Example 4.5 in a number of directions. For instance, they can describe stochastic processes which become Markovian in an extended state-space, encoding finitely many previous states realized by the process. Furthermore, the map $f$ in Example 4.8 is a very particular case of an encoding map, and it does not have to be an affine function of the process' current state, drift, and volatility. In general, $f$ can be replaced by any Hölder function of time and current state, as in the theory of (non-linear) random dynamical systems (see (Freidlin & Wentzell, 1984, Section 4.5)). Moreover, the decoding map does not need to take values in the subspace of the Wasserstein space $(\mathcal{P}_1(\mathbb{R}^d), \mathcal{W}_1)$ of $d$-dimensional Gaussian measures convoluted with $\nu$, representing the possible process' next step. Rather, $\rho$ can be any Hölder map into $(\mathcal{P}_1(\mathbb{R}^d), \mathcal{W}_1)$, or more broadly, $\rho$ can map into the *adapted Wasserstein space* $(\mathcal{P}_1(\mathbb{R}^{dN_F}), \mathcal{AW}_1)$ of Rüschendorf (1985) which robustly describes the process' conditional law on the next $N_F$ future steps. Note that these two generalizations coincide when $N_F = 1$. This is detailed in Example 5.11 in Section 5.2 below. Finally, $f$ does not have to have closed-form expression depending only on a finite segment of the process' history nor relating that data to a finite number of latent parameters decoded by $\rho$. Analogously, $\rho$ can be highly complicated and does not need to be expressible in closed-form as a decoding map depending on finitely many parameters.

These considerations naturally lead to our reinterpretation of the *fading memory property*, which was first formalized by Boyd and Chua (1985) but whose origins date back to ideas of Volterra and Wiener Wiener (1958). We note that, since its introduction, the fading memory property has been a central tool for deriving approximation theorems for dynamical systems between Euclidean spaces (see Lukoševičius and Jaeger (2009); Grigoryeva & Ortega (2019); Gonon et al. (2020); Manjunath (2020)) and is closely linked to the Echo State Property Jaeger (2001) that is key to reservoir computing Manjunath and Jaeger (2013); Gonon and Ortega (2021); Grigoryeva and Ortega (2018a).

**Definition 4.9** (Approximable Complexity). A causal map $F : \mathcal{X}^{\mathbb{Z}} \to \mathcal{Y}^{\mathbb{Z}}$ is of *approximable complexity* (AC Map) if there exist functions $m, L : (0,\infty) \to \mathbb{N}_+$ and $c_{AC} : \mathbb{Z} \times (0,\infty) \to [1,\infty)$ such that, for each $\varepsilon > 0$ and each compact $\mathcal{K} \subseteq \mathcal{X}^{\mathbb{Z}}$, there is $\alpha \in (0,1]$ and there exist an approximate encoding map $f_\varepsilon \in C^\alpha(\mathbb{R} \times \mathcal{X}^{m(\varepsilon)}, \mathbb{R}^{L(\varepsilon)})$ and an approximate decoding map $\rho_\varepsilon \in C^\alpha(\mathbb{R}^{L(\varepsilon)}, \mathcal{Y})$ such that the associated causal map of finite complexity satisfies

$$\sup_{n\in\mathbb{Z}} \sup_{\boldsymbol{x}\in\mathcal{K}} \frac{d_{\mathcal{Y}}(F^{f_\varepsilon,\rho_\varepsilon}(\boldsymbol{x})_{t_n}, F(\boldsymbol{x})_{t_n})}{c_{AC}(n,\varepsilon)} < \varepsilon. \tag{8}$$

An AC map is said to be *time-homogeneous* if there is a family $(F^{f_\varepsilon,\rho_\varepsilon})_{\varepsilon>0}$ of time-homogeneous systems of finite complexity satisfying (8).

Intuitively, an AC map $F$ is very close to some causal map of finite complexity $F^{f_\varepsilon,\rho_\varepsilon}$ on some (possibly large) time window around the current time, after which the two may begin to drift apart. Figure 7 illustrates a typical output of $F(\boldsymbol{x})$ (in violet) and $F^{f_\varepsilon,\rho_\varepsilon}$ (in orange) evaluated on some path $\boldsymbol{x} \in \mathcal{X}^{\mathbb{Z}}$. The rate at which this drifting apart occurs is expressed by the compression rate $c_{AC}$. The value of $c_{AC}$ is illustrated by the width of the turquoise region in Figure 7.

In the case of $\mathcal{Y} = \mathbb{R}$, for any $n \in \mathbb{Z}$ and any $\boldsymbol{x} \in \mathcal{X}^{\mathbb{Z}}$ the condition in (8) is equivalent to

$$F(\boldsymbol{x})_{t_n} \in \left[F^{f_\varepsilon,\rho_\varepsilon}(\boldsymbol{x})_{t_n} - c_{AC}(n,\varepsilon)\varepsilon, F^{f_\varepsilon,\rho_\varepsilon}(\boldsymbol{x})_{t_n} + c_{AC}(n,\varepsilon)\varepsilon\right].$$

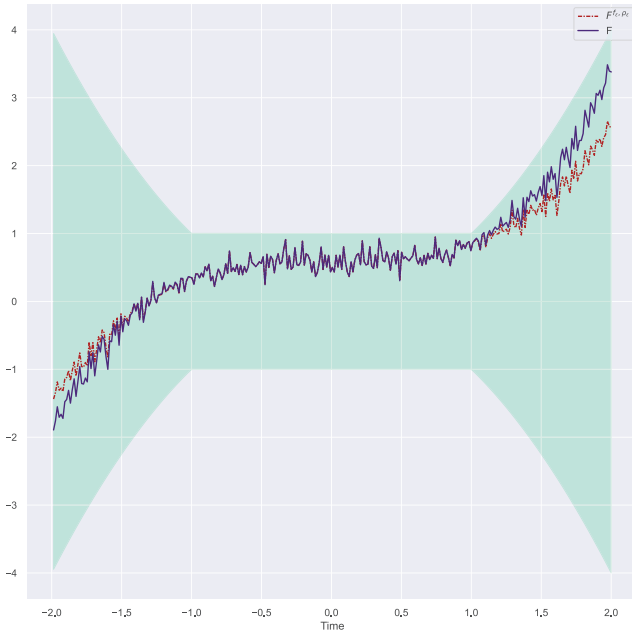



**FIGURE 7** An AC map $F$'s compression rate $c_{AC}$, in Definition 4.9, quantifies the rate at which a finitely-parameterized approximation of $F$ decays. [Color figure can be viewed at wileyonlinelibrary.com]

Therefore, the outputs of the (possibly intractable) causal map $F$ each belongs to a region defined by the (tractable) causal map of finite complexity $F^{f_\varepsilon,\rho_\varepsilon}$. An example of an AC map, but which is not of finite complexity, is given in Section 5.

## 4.3 | The geometric hypertransformer (GHT) model

We now introduce our geometric deep learning model, which we will use to approximate AC maps. The idea is to approximate AC maps by sequences of geometric transformers, which we sew together using a small auxiliary hypernetwork. Just as in Figure 1, the role of this hypernetwork is to sequentially generate the next geometric transformer (given the current geometric transformer's parameters). From the lens of stochastic analysis, the hypernetwork encoding the dynamic version of our geometric transformer is akin to a Feller process' infinitesimal generator (Engel & Nagel, 2000, Theorem II.3.8).

**Definition 4.10** (Geometric Hypertransformer). Let $(\mathcal{Y}, d_{\mathcal{Y}}, \eta, Q)$ be a QAS space, and fix positive integers $m, L, d$, a multi-index $[d]$ where $d_0 = (m+1)d$ and $d_J = L$, and an activation function $\sigma$. Let $\hat{\rho} : \mathbb{R}^L \to \mathcal{Y}$ be a geometric transformer, $h \in \mathcal{NN}_{\cdot}^{ReLU}$ a network mapping $\mathbb{R}^{P([d])}$ to itself, $\theta \in \mathbb{R}^{P([d])}$, and let $N \in \mathbb{N}_+$. The *geometric hypertransformer* (GHT) generated by $(\hat{\rho}, h, \theta, N)$ is the causal map $F^{(\hat{\rho},h,\theta,N)} : (\mathbb{R}^d)^{\mathbb{Z}} \to \mathcal{Y}^{\mathbb{Z}}$ defined by

$$F^{(\hat{\rho},h,\theta,N)}(\boldsymbol{x})_{t_n} \overset{\text{def.}}{=} \hat{\rho} \circ \hat{f}_{\theta_n}(x_{t_{n-m}:t_n}), \qquad n \in \mathbb{Z},$$

where $\hat{f}_{\theta_n} \in \mathcal{NN}^{\sigma}_{[d]}$ for all $n \in \mathbb{Z}$, and the parameters $(\theta_n)_{n\in\mathbb{Z}}$ are defined recursively by

$$\theta_n \overset{\text{def.}}{=} \theta \text{ for } n \leq -N \quad \text{and} \quad \theta_{n+1} \overset{\text{def.}}{=} \begin{cases} h(\theta_n) & : -N < n < N \\ \theta_n & : n \geq N. \end{cases} \tag{9}$$

Transformer networks typically carry an encoder–decoder structure, which means that each $\hat{f}$ can be thought of as the composition of two neural networks. The role of the first network is to encode the incoming input information into some deep latent features designed to optimize the prediction of the second, decoder, network, whose role is to generate the predictions at each time-step. In our context of the AC map $F$, the role of each of these networks becomes explicit. Namely, the encoder network's role will be to approximate the $f_\varepsilon$ given by its approximable complexity, and the role of the decoder network is then to approximate the measure-valued map $\rho_\varepsilon$, also given by $F$'s approximable complexity.

GHTs can approximate AC maps on an arbitrarily long but finite-time horizon without suffering from performance degradation. However, this is not the case when approximating an AC map across an infinite-time horizon, as in this case the GHT's approximation quality will eventually begin to degrade past a prespecified moment in time. Given a discrete path space $\mathcal{K} \subseteq \mathcal{X}^{\mathbb{Z}}$, the rate at which the performance of a GHT $\hat{F} : \mathcal{X}^{\mathbb{Z}} \to \mathcal{Y}^{\mathbb{Z}}$ degrades beyond a finite-time horizon $N_T \in \mathbb{N}_+$, with hyperparameter $\lambda > 0$, is quantified by the map $c^{\hat{F}}_{\mathcal{K},N_T,\lambda} : \mathbb{Z} \to [1,\infty)$ defined on any $n \in \mathbb{N}_+$ by

$$c^{\hat{F}}_{\mathcal{K},N_T,\lambda}(n) \overset{\text{def.}}{=} \sup_{x\in\mathcal{K}} \max\left\{1, \lambda\, d_{\mathcal{Y}}\left(\hat{F}(x)_{t_n}, \hat{F}(x)_{t_{\text{sgn}(n)\cdot N_T}}\right)\right\}, \tag{10}$$

where $\text{sgn}(n) \overset{\text{def.}}{=} -1$ if $n$ is a negative integer and $\text{sgn}(n) = 1$ otherwise. Outside the time-window $\{-N_T, \dots, N_T\}$, the map in (10) plays an analogous re-normalizing role to an AC maps' compression rate $c_{AC}$.

## 4.4 | Main result – GHTs are universal causal maps

In what follows, we will use the notation

$$N_T \overset{\text{def.}}{=} \min\left\{ \min\{n \in \mathbb{N}_+ \,:\, t_n \geq T\}, |\max\{n \in \mathbb{N}_- \,:\, t_n \leq -T\}|\right\}. \tag{11}$$

Moreover, in Theorem 4.11, for a fixed compact set $\mathcal{K} \subseteq \mathcal{X}^{\mathbb{Z}}$, $\mathcal{X} \subseteq \mathbb{R}^d$, and an AC map $F : \mathcal{X}^{\mathbb{Z}} \to \mathcal{Y}^{\mathbb{Z}}$, we will denote by $L_{\alpha,\rho_\varepsilon}$ and $L_{\alpha,f_\varepsilon}$ the $\alpha$-Hölder constant of $\rho_\varepsilon$ and $f_\varepsilon$ in (8), and set $K_n \overset{\text{def.}}{=} \text{pj}_n(\mathcal{K})$, where $\text{pj}_n$ is the projection into the $n$-th coordinate, $\text{pj}_n : (\mathbb{R}^d)^{\mathbb{Z}} \ni (x_{t_u})_{u\in\mathbb{Z}} \mapsto x_{t_n} \in \mathbb{R}^d$.

**Theorem 4.11** (Adapted Universal Approximation via GHTs). *Let $\mathcal{X}$ be a subset of $\mathbb{R}^d$ and $\mathcal{H} \subseteq \mathcal{X}^{\mathbb{Z}}$ a compact subset. Fix any AC map $F : \mathcal{X}^{\mathbb{Z}} \to \mathcal{Y}^{\mathbb{Z}}$, and a "time span" $T > 0$, with $T = t_{\bar{n}}$ for some $\bar{n} \in \mathbb{N}_+$. Then, for every $\varepsilon > 0$, there is a "compression rate" $c_\varepsilon : \mathbb{Z} \to (0,\infty)$, with $c_\varepsilon(n) = 1$ if $|n| \leq N_T$ and otherwise recorded in Table 3 (depending on $\mathcal{K}$), such that the following holds: there are a multi-index $[d]$, a geometric transformer $\hat{\rho} \in \mathcal{GT}^{\sigma}_{\cdot,N,q}(\mathbb{R}^d, \mathcal{Y})$, a hypernetwork $h \in \mathcal{NN}^{ReLU}_{\cdot}$ mapping $\mathbb{R}^{P([d])}$ to itself, $\theta \in \mathbb{R}^{P([d])}$, and $N \in \mathbb{N}_+$, such that the GHT $F^{(\hat{\rho},h,\theta,N)}$ generated by $(\hat{\rho}, h, \theta, N)$*

*satisfies*

$$\sup_{n\in\mathbb{Z}}\sup_{\boldsymbol{x}\in\mathscr{K}}\frac{d_{\mathscr{Y}}\left(F(\boldsymbol{x})_{t_n},F^{(\hat{\rho},h,\theta,N)}(\boldsymbol{x})_{t_n}\right)}{\max\left\{c_{AC}(n,\varepsilon),c_{\varepsilon}(n),c^{F^{(\hat{\rho},h,\theta,N)}}_{\mathscr{K},N_T,8/\varepsilon}(n)\right\}}<\varepsilon.$$

*Moreover, $F^{(\hat{\rho},h,\theta,N)}$'s model complexity can be estimated by:*

(1) ***Encoder Complexity:*** *As in Table 2, for* $\varepsilon=\min_{n=-N_T,\dots,N_T}\frac{1}{C_{K_{t_n-dm(\varepsilon/4):t_n}}}(\frac{\varepsilon}{8L_{\alpha,\rho_{\varepsilon/4}}})^{1/\alpha}$, *where* $C_{K_{t_n-dm(\varepsilon/4):t_n}}$ *is the constant defined in (6), and* $\alpha$ *is the regularity of* $f_{\varepsilon/4}$ *as in (8).*
(2) ***Decoder Complexity:*** *As in Table 1*
(3) ***Hypernetwork Complexity:*** $[d]=[P,M,M,P]$, *where P is the number of trainable parameters of the encoder network, as described in Table 2, and M is given by:*

$$M=\min\left\{\tilde{M}\in\mathbb{N}:\ 2\left\lfloor\frac{\tilde{M}}{2}\right\rfloor\left\lfloor\frac{\tilde{M}}{4P}\right\rfloor\geq N_T\right\}.$$

*Remark* 4.12. From Table 3, one can see that only in the case of general paths $\mathscr{K}$ there is explicit dependence on $F$, whereas the other cases depend only on its regularity.

*Remark* 4.13. When applying Theorem 4.11 to the sample paths of the solution to an SDE, as in Proposition 4.2, the statement should be interpreted as a high probability guarantee, depending on the parameter $0<\varepsilon\leq 1$ from Proposition 4.2.

The new compression rate defined in Theorem 4.11 by $\max\{c_{AC}(\cdot,\varepsilon),c_{\varepsilon}(\cdot),c^{F^{(\hat{\rho},h,\theta,N)}}_{\mathscr{K},N_T,8/\varepsilon}(\cdot)\}$ estimates the width of the turquoise region in Figure 7 when approximating $F$ on an infinite-time horizon. We emphasize that, in the interval $[-T,T]$, $c_{\varepsilon}=1$ and $c^{F^{(\hat{\rho},h,\theta,N)}}_{\mathscr{K},N_T,8/\varepsilon}=1$ and therefore the compression rate $\max\{c_{AC}(\cdot,\varepsilon),c_{\varepsilon}(\cdot),c^{F^{(\hat{\rho},h,\theta,N)}}_{\mathscr{K},N_T,8/\varepsilon}(\cdot)\}$ coincides with $c_{AC}(\cdot,\varepsilon)$ therein.

We also point out that, if $F$ is a causal map of finite complexity, then $c_{AC}(n,\varepsilon)=1$ for all $n\in\mathbb{Z}$. Therefore, in this case $\max\{c_{AC}(\cdot,\varepsilon),c_{\varepsilon}(\cdot),c^{F^{(\hat{\rho},h,\theta,N)}}_{\mathscr{K},N_T,8/\varepsilon}(\cdot)\}=1$ in the time-interval $[-T,T]$. Naturally, $c$ depends on the regularity of the paths on which the causal map $F$ is being approximated, as reflected by our rates recorded in Table 3.

Furthermore, in the case where $F$ is time-homogeneous and $\mathscr{K}=K^{\mathbb{Z}}$, we recover the following analogue of the reservoir-computing-type results of Grigoryeva and Ortega (2018b); Gonon et al. (2020).

**Corollary 4.14** (Approximation of Time-Homogeneous Causal Maps). *Assume the setting of Theorem 4.11 and suppose further that F is time-homogeneous and* $\mathscr{K}=K^{\mathbb{Z}}$ *for some compact set* $K\subseteq\mathbb{R}^d$. *Then, the compression rate* $c_{\varepsilon}$ *is* 1 *for* $|n|\leq N_T$, *while for* $|n|>N_T$ *it is*

$$c_{\varepsilon}(n)=4\varepsilon^{-1}L_{\alpha,\rho_{\varepsilon/4}}L^{\alpha}_{\alpha,f_{\varepsilon/4}}((dm(\varepsilon/4)+1)\mathrm{diam}(K))^{\alpha^2}.$$

# 5 | EXAMPLES OF GEOMETRIC ATTENTION MECHANISMS

## 5.1 | Examples of QAS spaces

We illustrate the scope of our causal geometric deep learning framework with various examples of the spaces covered by our theory. In each case, we work out exactly what the geometric attention mechanism in *closed-form* is and, therefore, the form of the geometric transformer in that context. We begin with the infinite-dimensional linear case. Then, we progress to non-linear examples arising from optimal transport theory, and finally we transition to the finite dimensional but non-Euclidean setting within the context of information geometry.

### 5.1.1 | Linear spaces

Our framework easily encompasses a broad range of topological vector spaces relevant throughout much of applied probability theory and, in particular, to mathematical finance. These include all Euclidean spaces and any "well-behaved" Banach space, such as all $L^p$-spaces on $\sigma$-finite measure spaces where $1 \leq p < \infty$.

**Example 5.1** (Fréchet Spaces). Let $(\mathcal{Y}, d_{\mathcal{Y}})$ be a Fréchet space. For any $N \in \mathbb{N}_+$, we may define the mixing function $\eta$ to be the map sending any $w \in \Delta_N$ and any $(y_n)_{n=1}^N \in \mathcal{Y}^N$ to

$$\eta\left(w, (y_n)_{n=1}^N\right) \overset{\text{def.}}{=} \sum_{n=1}^N w_n y_n.$$

Let $\dim(\mathcal{Y})$ denote dimension of $\mathcal{Y}$ as a vector space. If $\mathcal{Y}$ admits a Schauder basis[6] $(\beta_s)_{s=0}^{q^\star-1}$, where $q^\star \overset{\text{def.}}{=} \min\{\dim(\mathcal{Y}), \#\mathbb{N}\}$, then $(\mathcal{Y}, d_{\mathcal{Y}})$ is quantized by the functions $(Q_q)_{q\in\mathbb{N}}$ defined via

$$Q_q : \mathbb{R}^{q^\star} \ni z \mapsto \sum_{s=0}^{\min\{q,q^\star\}-1} z_s \cdot \beta_s \in \mathcal{Y},$$

where $z_s$ denotes the $s$-th component of $z$. Therefore, Fréchet spaces with Schauder basis can be endowed with a QAS space structure whose associated geometric attention is

$$\text{attention}_{N,q}\left(u, (z^n)_{n=1}^N\right) \overset{\text{def.}}{=} \sum_{n=1}^N \sum_{s=0}^{\min\{q,q^\star\}-1} \Pi_{\Delta_N}(u)_n z_s^n \cdot \beta_s.$$

In particular, Example 5.1 shows that our static universal approximation result for our architecture can approximate more general output spaces than the DeepONets of Lu et al. (2021); Liu et al. (2022). Further, our results yield quantitative counterparts to the Banach space-valued results of Benth et al. (2023)'s qualitative universal approximation results for their feedforward-type architecture.

Next, we explicate Example 5.1 with a Fréchet space central to classical mathematical finance. Namely, we use it to design a universal deep neural approach to *term structure modeling* which is

compatible with the Heath-Jarrow-Morton (HJM) framework of Heath et al. (1992). We note that, machine learning models have recently found successful applications in HJM-type frameworks since one can approximately impose no-arbitrage restrictions into the learned model Gambara and Teichmann (2020); Kratsios and Hyndman (2020). We also note that a feedforward counterpart to our transformer approach for forward-rate curve modeling in the Forward rate curve space of Filipović (2001) has recently been considered in Benth et al. (2022).

**Example 5.2** (Transformers in a Space of Forward Rate Curves). In (Filipović, 2001, Section 5.1), the author introduces a class of Hilbert spaces of functions on $[0,\infty)$ for modeling the term structure of interest rates, which are both economically meaningful (see (Filipović, 2001, Sections 4.2-4.3)) and convenient to analyze within the HJM framework. For instance, if $\alpha > 3$, the author considers the space $\tilde{H}_\alpha$ of all *absolutely continuous functions* $y : [0,\infty) \to \mathbb{R}$ (representing yields curves) for which the norm $\|y\| \overset{\text{def.}}{=} \langle y, y\rangle_\alpha^{1/2}$ is finite, where

$$\langle y, \tilde{y}\rangle_\alpha \overset{\text{def.}}{=} |y(0)\tilde{y}(0)| + \int_0^\infty |y'(t)\tilde{y}(t)|\,|1+t|^\alpha dt,$$

where $y'$ is a weak derivative of $y$ on $(0,\infty)$ and $\tilde{y}$ is some absolutely continuous function on $[0,\infty)$. For simplicity of exposition, let us consider the subset $H_\alpha \subseteq \tilde{H}_\alpha$ consisting of all functions satisfying the boundary condition $y(0) = 0$. Then, (Saitoh & Sawano, 2016, Example 1.3) shows that this is also a Reproducing Kernel Hilbert Space (RKHS) with reproducing kernel $K$ defined on any $(t,\tilde{t}) \in [0,\infty)^2$ by

$$K(t,\tilde{t}) \overset{\text{def.}}{=} \int_0^{\min\{t,\tilde{t}\}} \frac{1}{(1+t)^\alpha} dt = \frac{(\min\{t,t'\}+1)^{-\alpha}((\min\{t,t'\}+1)^\alpha - \min\{t,t'\}-1)}{\alpha-1}.$$

Applying (Aronszajn, 1950, Moore-Aronzajins Theorem), we conclude that the span of $\{K(t,\cdot) : t \in [0,\infty)\}$ is dense in $H_\alpha$. Now, since $[0,\infty)$ is separable and $K(\cdot,\cdot)$ is continuous, then there is a countable subset $\{t_n^\star\}_{n\in\mathbb{N}} \subseteq [0,\infty)$ for which

$$y_n \overset{\text{def.}}{=} K\left(t_n^\star, \cdot\right)$$

is a Schauder basis of $H_\alpha$. Thus, the geometric attention of Example 5.1 simplifies to

$$\text{attention}_{N,q}(u,(z^n)_{n=1}^N) \overset{\text{def.}}{=} \sum_{n=1}^N \sum_{s=0}^{q-1} \Pi_{\Delta_N}(u)_n z_s^n \cdot \frac{(\min\{t_n^\star,\cdot\}+1)^{-\alpha}((\min\{t_n^\star,\cdot\}+1)^\alpha - \min\{t_n^\star,\cdot\}-1)}{\alpha-1},$$

for some $\{t_n^\star\}_{n=0}^\infty$ in $[0,\infty)$.

*Remark* 5.3 (Other HJM-Type Frameworks via Example 5.2). As an interesting future research direction, one can likely modify Example 5.2 to suit other HJM-type of models used in equity or credit markets Carmona (2007) or stocks Kallsen and Krühner (2015).

*Remark* 5.4 (Example 5.2 Translates to Separable RKHSs). Since RKHSs are a central object in machine learning, it is worth emphasizing that the analysis carried out in Example 5.2 translates, nearly identically, to geometric transformers mapping in any other separable RKHS. The

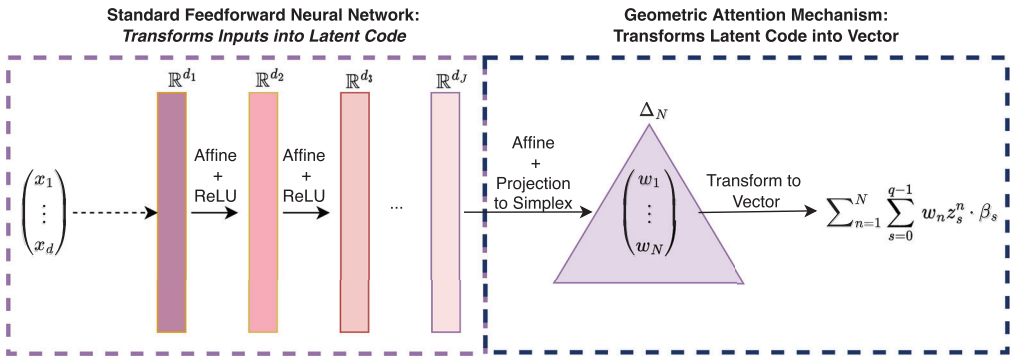



**FIGURE 8** The *geometric transformer* of Example 5.1 maps inputs in $\mathbb{R}^d$ to outputs in an *infinite-dimensional Fréchet space with a Schauder basis*. The outputs are convex combinations of $N$ vectors, in the Fréchet space, each of which can be exactly implemented as a linear combination of the first $q$ basis vectors $\{\beta_s\}_{s=0}^{q-1}$. [Color figure can be viewed at wileyonlinelibrary.com]

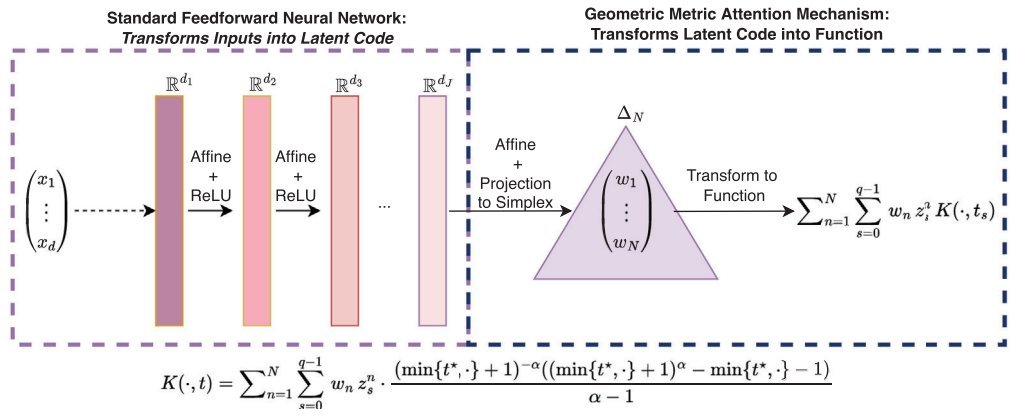



**FIGURE 9** The *geometric transformer* of Example 5.2 maps inputs in $\mathbb{R}^d$ to *forward rate curves in the Reproducing Kernel Hilbert space $H_\alpha$ of Filipović* (2001) which can be expressed as convex combinations of $H_\alpha$'s kernel function $K(\cdot,\cdot)$ with second component evaluated at specified times $t_0, \ldots, t_{q-1} \geq 0$. [Color figure can be viewed at wileyonlinelibrary.com]

geometric transformer in that context is the the same as in Figure 9 with only the kernel function swapped for the kernel of the new RKHs.

### 5.1.2 | Optimal transport spaces

Our results also apply to many spaces from optimal transport theory, adapted optimal transport, and consequentially robust finance.

**Example 5.5** (Wasserstein Space with Convex Combinations). Fix $p \geq 1$, let $Z \subseteq \mathbb{R}^d$ be closed, and either suppose that $Z = \mathbb{R}^d$ and $q > p$, or that $Z$ is bounded and $q \geq p$. Let $(\mathcal{Y}, d_{\mathcal{Y}}) \overset{\text{def.}}{=} (\mathcal{P}_q(Z), \mathcal{W}_p)$. Define the mixing function $\eta$ to be the map that, for each $N \in \mathbb{N}_+$, sends any $w \in \Delta_N$

and any $(y_n)_{n=1}^N \in \mathcal{Y}^N$ to

$$\eta(w,(y_n)_{n=1}^N) \overset{\text{def.}}{=} \sum_{n=1}^N w_n y_n. \tag{12}$$

Under our assumptions on $Z, p$, and $q$, we may apply (Chevallier, 2018, Theorem 2 or Corollary 3) (respectively) to conclude that $(\mathcal{P}_q(Z), \mathcal{W}_p)$ is quantized by the following family $(Q_q)_{q\in\mathbb{N}_+}$ of functions:

$$Q_q : \mathbb{R}^{d\times q} \ni z = (z_0,\dots,z_{q-1}) \mapsto \frac{1}{q}\sum_{s=0}^{q-1} \delta_{z_s} \in \mathcal{P}_q(Z). \tag{13}$$

Therefore, $(\mathcal{P}_q(Z), \mathcal{W}_p)$ can be endowed with a QAS space structure whose geometric attention mechanism, coincides with the probabilistic attention mechanism of Kratsios (2023); Kratsios et al. (2022), and is given by

$$\text{attention}_{N,q}(u,(z^n)_{n=1}^N) \overset{\text{def.}}{=} \frac{1}{q}\sum_{n=1}^N\sum_{s=1}^q \Pi_{\Delta_N}(u)_n \delta_{z_s^n}.$$

Still considering the Wasserstein space, for the next example, we show how the mixing function $\eta$ can be chosen to be a *non-linear averaging of distributions* unlike the linear mixing functions used to illustrate out theory thus far. We consider the notion of Wasserstein barycenters, as introduced in Agueh and Carlier (2011), which generalize McCann's interpolation problem originally defined only for two probability measures. This has a wide range of applications, for example when one wants to average features defined as distributions (as in computer vision). We refer to Cuturi and Doucet (2014) and Claici et al. (2018) for algorithms to efficiently compute Wasserstein barycenters and to Heinemann et al. (2022) for related statistical learning guarantees.

**Example 5.6** (Wasserstein barycenters). Let $(\mathcal{Y}, d_{\mathcal{Y}}) = (\mathcal{P}_p(\mathbb{R}^d), \mathcal{W}_p)$ be the set of measures with finite $p$-th moment on $\mathbb{R}^d$ equipped with the Wasserstein-$p$-distance. For each $N \in \mathbb{N}$ and fixed $(y_n)_{n=1}^N \in \mathcal{P}_p(\mathbb{R}^d)^N$, consider

$$S(y,w) \overset{\text{def.}}{=} \sum_{n=1}^N w_n \mathcal{W}_p(y,y_n)^p,$$

$$V(w) \overset{\text{def.}}{=} \inf_{y\in\mathcal{P}_p(\mathbb{R}^d)} S(y,w),$$

and define $\eta(w,(y_n)_{n=1}^N)$ as a measurable selection of optimizers of $V$. To see the existence of such a selection, first note that $V$ is continuous. Consequently,

$$B \overset{\text{def.}}{=} \left\{(y,w) \in \mathcal{P}_p(\mathbb{R}^d)\times\Delta_N \,:\, S(y,w) = V(w)\right\}$$

is a bounded subset of $\mathcal{P}_p(\mathbb{R}^d)\times\Delta_N$, and therefore relatively compact in $\mathcal{P}(\mathbb{R}^d)\times\Delta_N$. By (Figalli, 2010, Remark 6.12), $S$ is lower semicontinuous on $\mathcal{P}(\mathbb{R}^d)\times\Delta_N$, thus, $B$ is even compact in $\mathcal{P}(\mathbb{R}^d)\times$

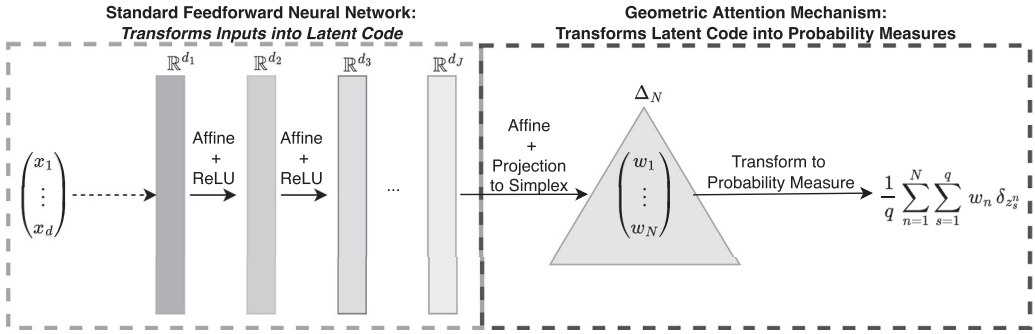



**FIGURE 10** The *geometric transformer* of Example 5.5 transforms inputs in $\mathbb{R}^d$ to outputs in the Wasserstein space $(\mathcal{P}_q(\mathbb{R}^d), \mathcal{W}_p)$. Every input in $\mathbb{R}^d$ is mapped to a weight $w$ in the $N$-simplex which is used to create a convex combination of $N$ empirical measures on $\mathbb{R}^d$, each of which charges at-most $q$ points with mass. [Color figure can be viewed at wileyonlinelibrary.com]

$\Delta_N$. Therefore, we find, for $z \in \mathcal{P}_p(\mathbb{R}^d)$ and $r > 0$, that

$$\left\{ w \in \Delta_N \,:\, \exists (y, w) \in B \text{ s.t. } y \in \overline{\text{Ball}}_{(\mathcal{P}_p(\mathbb{R}^d), \mathcal{W}_p)}(z, r) \right\}$$

is a closed subset of $\Delta_N$. Hence, (Bogachev, 2007, Theorem 6.9.3) to produces the desired measurable selection.

We remark that, for $p = 2$, by Agueh and Carlier (2011) the set of optimal couplings is a singleton if there is a $1 \le k \le N$ with $y_k$ absolutely continuous w.r.t. the Lebesgue measure. Hence, in this case the (unique) selection is even continuous. In any case, $\eta$ satisfies the following inequalities for all $i = 1, \dots, N$:

$$\mathcal{W}_p\left(\eta\left(w, (y_n)_{n=1}^N\right), y_i\right) \le \sum_{j=1}^{N} w_j \left(\mathcal{W}_p(y_j, y_i) + \mathcal{W}_p\left(\eta\left(w, (y_n)_{n=1}^N\right), y_j\right)\right) \le 2\left(\sum_{j=1}^{N} w_j \mathcal{W}_p(y_i, y_j)^p\right)^{1/p}.$$

Just as Example 3.6 provided an adapted counterpart to Example 5.5, so does the following construction provide an adapted counterpart to Example 5.6. The corresponding illustration is akin to Figure 11, mutatis mundanis.

**Example 5.7** (Adapted Wasserstein barycenters). Even though $(\mathcal{P}_p(\mathbb{R}^{dT}), \mathcal{AW}_p)$ on its own is not a geodesic space, its completion $(\mathcal{FP}_p, \mathcal{AW}_p)$ is geodesically complete; see Bartl et al. (2021), which also gives a probabilistic interpretation of the aforementioned space as a Wasserstein space of stochastic processes. Moreover, $(\mathcal{FP}_p, \mathcal{AW}_p)$ provides a way of taking barycenters of stochastic processes that take into account path properties as well as the arrow of time encoded in the underlying filtrations. Similar to Example 5.6, for fixed $N \in \mathbb{N}$ and $(y_n)_{n=1}^N$, we define the mixing function $\eta$ as a measurable selection of $\mathcal{AW}_p$-barycenters:

$$\eta(w, (y_n)_{n=1}^N) \in \arg\min \left\{ \sum_{n=1}^{N} w_n \mathcal{AW}_p(y, y_n)^p \,:\, y \in \mathcal{Y} \right\}.$$

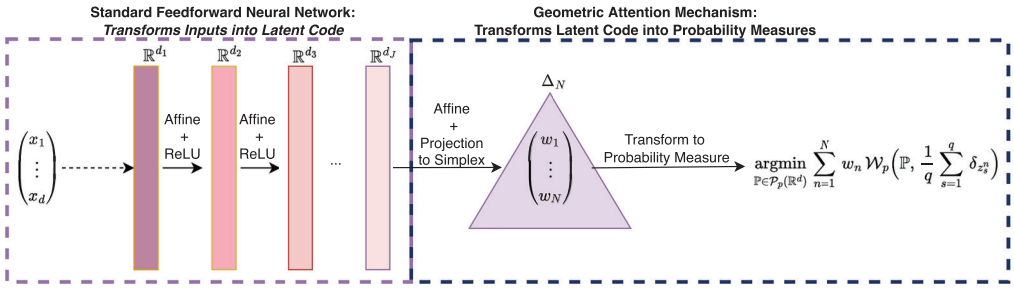



**FIGURE 11** The *geometric transformer* of Example 5.6 transforms inputs in $\mathbb{R}^d$ to outputs in the Wasserstein space $(\mathcal{P}_p(\mathbb{R}^d), \mathcal{W}_p)$. Every input in $\mathbb{R}^d$ is first mapped to a weight $w$ in the $N$-simplex. The weight $w$ together with $N$ different empirical measures $\sum_{s=1}^q \delta_{z_s^1}, \ldots, \sum_{s=1}^q \delta_{z_s^N}$ on $\mathbb{R}^d$ each of which charges at-most $q$ points with mass, defines a Wasserstein barycenter problem where the relative importance of each empirical measure is weighted according to $w$. A probability measure optimizing this Wasserstein barycenter problem is output by the GT. [Color figure can be viewed at wileyonlinelibrary.com]

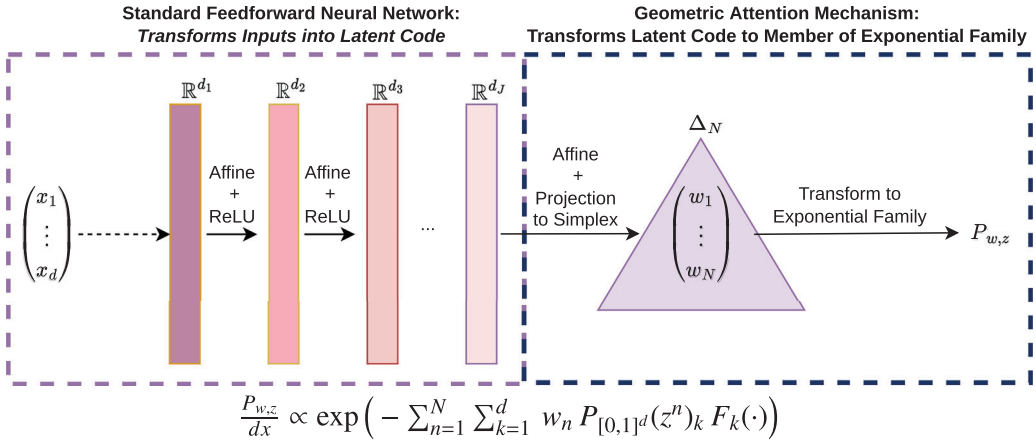



**FIGURE 12** The *geometric transformer* of Example 5.9 transforms inputs in $\mathbb{R}^d$ to the unique parameter associated to a probability measure in the natural parameter space $\Theta_{F_\cdot}$ associated to the statistical manifold $\mathcal{P}_{F_\cdot}$. The model outputs the unique probability measure in $\mathcal{P}_{F_\cdot}$ corresponding to this parameter in $\Theta_{F_\cdot}$. [Color figure can be viewed at wileyonlinelibrary.com]

The existence of $\eta$ can be shown along the lines of Example 5.6 due to the characterization of relative compact sets given in (Bartl et al., 2021, Theorem 1.7).

*Remark* 5.8 (Mixture Density Networks are Geometric Transformers). One can modify the quantization function from the previous section to output Gaussian mixtures with parameterized non-degenerate variance. Therefore, the mixture density network of Bishop (1994) is a particular case of our geometric transformer framework. Figures 8, 10, 12, and 13.

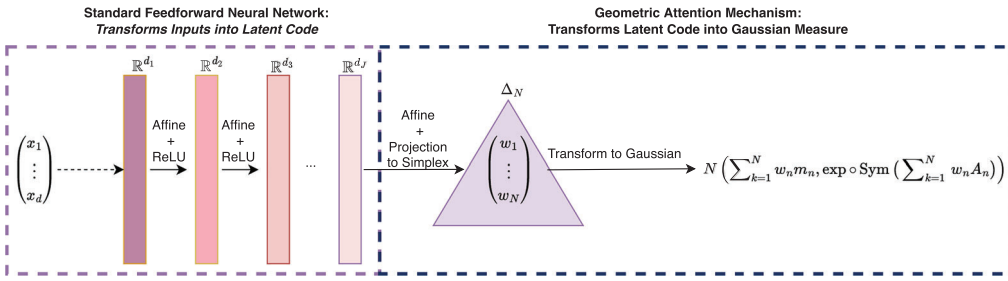



**FIGURE 13** The *geometric transformer* of Example 5.10 maps inputs in $\mathbb{R}^d$ to *non-degenerate Gaussian* probability measures on $\mathbb{R}^d$. The intrinsic distance function quantifying the dissimilarity between any such Gaussian measures is (15). [Color figure can be viewed at wileyonlinelibrary.com]

### 5.1.3 | Finitely parameterized families of probability distributions from information geometry

At times, more can be assumed of the probability distributions describing the target process. In particular, we are interested in the case where the process' law is characterized by a sufficient statistic and the conditional law of that process belongs to an exponential family of probability measures. Typical examples arise when working with continuous-time and finite-state Markov chains (see Jacobsen (1982); Küchler and Sø rensen (1997)).

This is not only of practical interest, but it provides a tangible class of non-linear examples, stemming in classical information geometry (Amari, 2016), where the mixing function $\eta$ is not a convex combination and where the maps $(Q_q)_{q\in\mathbb{N}_+}$ are all identical due to $\mathcal{Y}$'s finite dimensionality.

**Example 5.9** (Finite-Dimensional Exponential Families of Stochastic Processes). Let $F_1,\dots,F_d : (\mathbb{R}^n)^T \to \mathbb{R}$ be linearly independent continuous path functionals and let $\Theta_{F_\cdot}$ be the open convex subset of $\mathbb{R}^d$ defined by

$$\Theta_{F_\cdot} \overset{\text{def.}}{=} \left\{\theta\in\mathbb{R}^d \,:\, \int_{x\in\mathbb{R}^{nT}} \exp\left(-\sum_{k=1}^{d}\theta_k F_k(x)\right)dx<\infty\right\}.$$

The set $\Theta_{F_\cdot}$, is known as the *natural parameter space* of the statistical manifold

$$\mathcal{P}_{F_\cdot} \overset{\text{def.}}{=} \left\{ y_\theta\in\mathcal{P}\big((\mathbb{R}^n)^T\big) \,:\, (\exists\theta\in\Theta_{F_\cdot})\,\frac{dy_\theta}{dx}\propto\exp\left(-\sum_{k=1}^{d}\theta_k F_k(\cdot)\right) \text{ and } \mathbb{E}_{X\sim y_\theta}[\|X\|]<\infty\right\},$$

where the identification of $\Theta_{F_\cdot}$ with $\mathcal{P}_{F_\cdot}$ is given by the correspondence $\Theta_{F_\cdot}\ni\theta\leftrightarrow y_\theta\in\mathcal{P}_{F_\cdot}$; see (Amari, 2016, Chapter 2.1). Following Čencov (1982); Amari (2016), the family of Fisher information matrices $(I(\theta))_{\theta\in\Theta_{F_\cdot}}$, defined at each $\theta\in\Theta_{F_\cdot}$ by $I(\theta)\overset{\text{def.}}{=}(\mathbb{E}_{y_\theta}[\frac{\partial p_\theta}{\partial\theta_i}\frac{\partial p_\theta}{\partial\theta_j}])_{i,j=1}^d$, where $p_\theta\overset{\text{def.}}{=}\frac{dy_\theta}{dx}$, defines a Riemannian metric on $\Theta_{F_\cdot}$, called the *Fisher-Rao metric*. Appealing to the identification

$\varphi : \Theta_{F_\cdot} \ni \theta \to y_\theta \in \mathcal{P}_{F_\cdot}$, the Fisher-Rao metric induces an intrinsic metric on $\mathcal{P}_{F_\cdot}$ via

$$d_I(y_{\theta^1}, y_{\theta^2}) \overset{\text{def.}}{=} \inf_\gamma \int_0^1 \sqrt{\dot\gamma(t)^\top I(\theta)\dot\gamma(t)}dt, \tag{14}$$

where the infimum runs over all piecewise continuous $\gamma : [0,1] \to \Theta_{F_\cdot}$ with $\gamma(0) = \theta^1$ and $\gamma(1) = \theta^2$.

For simplicity, as in Abbas et al. (2021), let us assume that $[0,1]^d \subseteq \Theta_{F_\cdot}$ and let us set $(\mathcal{Y}, d_{\mathcal{Y}}) \overset{\text{def.}}{=} (\varphi([0,1]^d), d_I)$. Since $\varphi$ is a chart, it is smooth and therefore Lipschitz when restricted to $[0,1]^d$. Therefore, we may define $\eta$ such that, for every $N \in \mathbb{N}_+$, $w \in \Delta_N$ and $(y_{\theta^n})_{n=1}^N \in \mathcal{Y}^N$,

$$\frac{d\,\eta(w, (y_{\theta^n})_{n=1}^N)}{dx} \propto \exp\left(-\sum_{n=1}^N \sum_{k=1}^d w_n \theta_k^n F_k(\cdot)\right).$$

Since $\Theta_{F_\cdot}$ is a $d$-dimensional Riemannian manifold, then, similarly to Example 5.1, we may quantize it by the functions $(Q_q)_{q\in\mathbb{N}}$ defined via

$$\frac{dQ_q(z)}{dx} \propto \exp\left(-\sum_{k=1}^d [\Pi_{[0,1]^d}(z)]_k F_k(\cdot)\right),$$

where $\Pi_{[0,1]^d}$ is the metric projection of $\mathbb{R}^d$ onto the cube $[0,1]^d$, i.e. $\mathbb{R}^d \ni u \mapsto (\min\{\max\{0, u_i\}, 1\})_{i=1}^d \in [0,1]^d$. Thus, $(\mathcal{P}_{F_\cdot}, d_I)$ can be endowed with QAS space structure. For $q \geq d$, its associated geometric attention mechanism is

$$\frac{d\,\text{attention}_{N,q}\left(u, (z^n)_{n=1}^N\right)}{dx} \propto \exp\left(-\sum_{n=1}^N \sum_{k=1}^d [\Pi_{\Delta_N}(u)]_n\, [\Pi_{[0,1]^d}(z^n)]_k\, F_k(\cdot)\right).$$

We conclude our illustration of QAS spaces with an example relevant to Markovian SDEs and Gaussian processes.

**Example 5.10** (Non-Degenerate Gaussian Measures). Let $\mathcal{Y}$ be the set of probability measures $N(m, \Sigma)$ on $\mathbb{R}^d$ which are absolutely continuous with respect to the $d$-dimensional Lebesgue measure and whose Radon-Nikodym derivative can be written as

$$\frac{dN(m,\Sigma)}{dx} = (2\pi)^{-\frac{d}{2}} \det(\Sigma)^{-\frac{1}{2}} e^{-\frac{1}{2}(x-m)^\top \Sigma^{-1}(x-m)},$$

where $\Sigma$ is a $d \times d$ positive definite matrix and $m \in \mathbb{R}^d$. Accordingly, as in the information geometry literature Lovrić et al. (2000); Nielsen (2020); Malagò et al. (2018), we identify non-degenerate Gaussian distributions in $\mathcal{Y}$ with pairs $(\Sigma, \mu)$ parameterizing them. Instead of equipping $\mathcal{Y}$ with a Riemannian structure, we may simply equip it with the following complete metric, arising from the non-positive curvature geometry of the set of $d \times d$-symmetric positive-definite matrices (see

Meyer et al. (2011); Helgason (1979)):

$$d_{\mathcal{Y}}(N(m_1,\Sigma_1),N(m_2,\Sigma_2)) \overset{\text{def.}}{=} \sqrt{\|m_1-m_2\|^2+\left\|\sqrt{\Sigma_1}\log\left(\sqrt{\Sigma_1}^{-1}\sqrt{\Sigma_2}\sqrt{\Sigma_1}^{-1}\right)\sqrt{\Sigma_1}\right\|_F^2}, \quad (15)$$

where $\sqrt{\cdot}$ denotes the square-root of a symmetric positive-definite matrix, log denotes the matrix logarithm, and $\|\cdot\|_F$ denotes the Frobenius norm. Following Kratsios and Papon (2022), for each $q \in \mathbb{N}$ we set

$$Q_q : \mathbb{R}^d \times \mathbb{R}^{d(d+1)/2} \ni (m,A) \mapsto N(m, \exp\circ\mathrm{Sym}(A)),$$

where exp is the exponential matrix and for any $A=(A_{1,1},\dots,A_{1,d},\dots,A_{d,d}) \in \mathbb{R}^{d(d+1)/2}$ the matrix $\mathrm{Sym}(A)$ is defined by $\mathrm{Sym}(A)_{i,j} \overset{\text{def.}}{=} \mathrm{Sym}(A)_{j,i} \overset{\text{def.}}{=} A_{i,j}$, $i,j=1,\dots,d$. Thus, $(\mathcal{Y},d_{\mathcal{Y}})$ can be equipped with a QAS space structure. Its associated geometric attention mechanism is defined, the same way for each $q \in \mathbb{N}_+$, by sending any $(u,(m_n,A_n)_{n=1}^N)$ in $\mathbb{R}^N \times \mathbb{R}^{N\times d+d(d+1)/2}$ to

$$\mathrm{attention}_{N,q}(u,(m_n,A_n)_{n=1}^N) \overset{\text{def.}}{=} N\left(\sum_{k=1}^N \Pi_{\Delta_N}(u)_n m_n, \exp\circ\mathrm{Sym}\left(\sum_{k=1}^N \Pi_{\Delta_N}(u)_n A_n\right)\right).$$

## 5.2 | Examples of causal maps of approximable complexity

In this section, we present two examples of AC maps, which are not of finite complexity. The first illustrates how an AC map can fail to be of finite complexity while exhibiting a limited memory, because the decoding map $\rho$ depends on arbitrarily many parameters. The second example demonstrates an AC map where the converse is true. Namely, it has infinite memory but its decoding maps depend on a fixed and finite number of latent parameters.

In the following example, we modify the setting of Example 4.5, to show how an SDE's multi-step evolution defines an AC map into the adapted Wasserstein space. To keep technical details at a minimum, we consider a 1-dimensional SDE.

**Example 5.11** (Discrete-time SDEs (Adapted Wasserstein space)). Let $(\mathcal{Y},d_{\mathcal{Y}})=(\mathcal{P}_1(\mathbb{R}^2),\mathcal{AW}_1)$. Here, we assume that the Hölder coefficients of $\mu(t_n,\cdot)\in C^\alpha(\mathbb{R},\mathbb{R})$ and $\sigma(t_n,\cdot)\in C^\alpha(\mathbb{R},\mathbb{R})$ are uniformly bounded in $(t_n)_{n\in\mathbb{Z}}$, with constants $L_\mu$ and $L_\sigma$, respectively. For simplicity, we assume $\delta_-$ and $\delta_+$ in Assumption 4.1 to be $\delta_-=\delta_+=1$. For initial condition $X_{t_n}=x_{t_n}\in\mathbb{R}$, we define $X_{t_{n+1}}$ by

$$X_{t_{n+1}} = x_{t_n} + \mu(t_n,x_{t_n}) + \sigma(t_n,x_{t_n})W_n,$$

where $(W_n)_{n\in\mathbb{Z}}$ is a sequence of independent standard Gaussians. We consider the causal map that is, for $n\in\mathbb{Z}$ and $\boldsymbol{x}\in\mathbb{R}^{\mathbb{Z}}$, given by

$$F(\boldsymbol{x})_{t_n} \overset{\text{def.}}{=} \mathcal{L}(X_{t_{n+1}},X_{t_{n+2}}|X_{t_n}=x_{t_n}),$$

and proceed to show that it has approximable complexity. For this reason, let $\varepsilon > 0$ and $\mathcal{K} \subseteq \mathbb{R}^{\mathbb{Z}}$ be compact. We define, for $n \in \mathbb{N}$,

$$c_{AC}(n,\varepsilon) \overset{\text{def.}}{=} \sup_{\boldsymbol{x}\in\mathcal{K}} |\sigma(t_n, x_{t_n})|^{\alpha}.$$

For $m \in \mathbb{N}$ sufficiently large, we choose an increasing sequence of quantiles $(q_k)_{k=1}^m$ of a standard Gaussian $\gamma$, satisfying

$$\int_{\mathbb{R}} \sum_{k=0}^{m} \mathbb{1}_{I_k}(x)|x - q_k|^{\alpha}\,\gamma(dx) \le \frac{\varepsilon}{L_\mu + L_\sigma},$$

where we set $q_0 \overset{\text{def.}}{=} q_1, q_{m+1} \overset{\text{def.}}{=} \infty, I_k \overset{\text{def.}}{=} (q_k, q_{k+1}), k = 1, \dots, m$ and $I_0 \overset{\text{def.}}{=} (-\infty, q_1)$.

To ease notation, we write

$$q_k^n(x) \overset{\text{def.}}{=} x + \mu(t_n, x) + \sigma(t_n, x)q_k,$$

$$\mu^n(x) \overset{\text{def.}}{=} \mu(t_n, x), \quad \mu_k^n(x) \overset{\text{def.}}{=} \mu(t_n, q_k^{n-1}(x)),$$

$$\sigma^n(x) \overset{\text{def.}}{=} \sigma(t_n, x), \quad \sigma_k^n(x) \overset{\text{def.}}{=} \sigma(t_n, q_k^{n-1}(x)).$$

Then, the encoding and decoding maps are given by

$$f_\varepsilon(t_n, x) \overset{\text{def.}}{=} (x + \mu^n(x), \sigma^n(x), (\mu_k^{n+1}(x))_{k=1}^m, (\sigma_k^{n+1}(x))_{k=1}^m),$$

$$\rho_\varepsilon\left((\mu_k)_{k=0}^m, (\sigma_k)_{k=0}^m\right) \overset{\text{def.}}{=} N\left(\mu_0, \sigma_0^2\right)(dx_1) \sum_{k=0}^{m} \mathbb{1}_{\mu_0+\sigma_0 I_k}(x_1) N\left(\mu_{k\vee 1}, \sigma_{k\vee 1}^2\right)(dx_2).$$

By construction, we have that $f_\varepsilon(t_n, \cdot) \in C^\alpha(\mathbb{R}, \mathbb{R}^{m+1} \times \mathbb{R}^{m+1})$. Next we show that $\rho_\varepsilon$ belongs to $C^1(\mathbb{R}^{m+1} \times \mathbb{R}^{m+1}, (\mathcal{P}_1(\mathbb{R}^2), \mathcal{AW}_1))$. We may estimate

$$\begin{aligned}\mathcal{AW}_1(\rho_\varepsilon((\mu_k)_{k=0}^m, (\sigma_k)_{k=0}^m), \rho_\varepsilon((\hat{\mu}_k)_{k=0}^m, (\hat{\sigma}_k)_{k=0}^m)) &\le \sum_{k=0}^{m} \mathcal{W}_1\left(N\left(\mu_k, \sigma_k^2\right), N\left(\hat{\mu}_k, \hat{\sigma}_k^2\right)\right) \\ &\le \sum_{k=0}^{m} |\mu_k - \hat{\mu}_k| + |\sigma_k - \hat{\sigma}_k|\,\mathbb{E}[|W_n|].\end{aligned}$$

Finally, we verify that the causal map of finite complexity associated to $(f_\varepsilon, \rho_\varepsilon)$ satisfies (8). For this reason, let $\gamma_{x_{t_n}}^n$ be the normal distribution with mean $\mu^n(x_{t_n})$ and variance $\sigma^n(x_{t_n})^2$. Thus, we get for $\boldsymbol{x} \in \mathcal{K}$

$$\begin{aligned}&\mathcal{AW}_1(F(\boldsymbol{x})_{t_n}, \rho_\varepsilon(f_\varepsilon(t_n, x_{t_n}))) \\ &\quad\le \int_{\mathbb{R}} \sum_{k=0}^{m} \mathbb{1}_{\mu^n(x_{t_n})+\sigma^n(x_{t_n})I_k}(x)\mathcal{W}_1\left(N(\mu^{n+1}(x), \sigma^{n+1}(x)), N(\mu_k^{n+1}(x), \sigma_k^{n+1}(x))\right)\gamma_{x_{t_n}}^n(dx)\end{aligned}$$

$$\leq \int_{\mathbb{R}} \sum_{k=0}^{m} \mathbb{1}_{I_k}(x) \left( |\mu^{n+1}(\mu^n(x_{t_n}) + \sigma^n(x_{t_n})x) - \mu^{n+1}(\mu^n(x_{t_n}) + \sigma^n(x_{t_n})q_k)| \right.$$

$$\left. + |\sigma^{n+1}(\mu^n(x_{t_n}) + \sigma^n(x_{t_n})x) - \sigma^{n+1}(\mu^n(x_{t_n}) + \sigma^n(x_{t_n})q_k)| \, \mathbb{E}[|W_n|] \right) \gamma(dx)$$

$$\leq \int_{\mathbb{R}} \sum_{k=0}^{m} \mathbb{1}_{I_k}(x) |\sigma^n(x_{t_n})|^\alpha |x - q_k|^\alpha (L_\mu + L_\sigma) \gamma(dx) \leq |\sigma^n(x_{t_n})|^\alpha \varepsilon.$$

Our next example of an AC map draws from the dynamical systems literature, such as Hutter et al. (2021). In particular, the example shows how the approximation of stochastic processes can equally be framed as a causal map into an $L^1$-space rather than a Wasserstein space. This illustrates the "modularity" of our framework as well as the ability to model stochastic processes with genuinely infinite memory.

**Example 5.12** (Infinite-Memory SDEs Inducing AC Maps). Let $\alpha \in (0,1]$, $(\Omega, \mathcal{F}, \mathcal{P}, \mathbb{F} \overset{\text{def.}}{=} (\mathcal{F}_{t_n})_{n\in\mathbb{N}})$ be a filtered probability space, and $(W_{t_n})_{n\in\mathbb{N}}$ be a sequence of i.i.d. standard $d$-dimensional Gaussian random variables. For simplicity, we assume that $\delta_- = \delta_+ = 1$ where, $\delta_+$ and $\delta_-$ are as in Assumption 4.1. Let $M \in C^\alpha([0,\infty) \times \mathbb{R}^d, [0,1]^d)$, $\Sigma \in C^\alpha([0,\infty) \times \mathbb{R}^d, [0,1]^{d\times d})$, and let $(k_n)_{n\in\mathbb{Z}_-}$ be an absolutely summable real-valued sequence. Define the map

$$\mu : [0,\infty) \times (\mathbb{R}^d)^{\mathbb{Z}_-} \ni \left(t, (z_n)_{n\in\mathbb{Z}_-}\right) \mapsto \sum_{-\infty<n\leq 0} k_n M(t, z_n).$$

Then the causal map $F$ given by

$$F(\boldsymbol{x})_{t_n} \overset{\text{def.}}{=} N\left(x_{t_n} + \mu(t_n, (x_{t_s})_{s\leq n}), \Sigma(t_n, x_{t_n})\right) \in \mathcal{P}_1(\mathbb{R}^d),$$

for $\boldsymbol{x} \in (\mathbb{R}^d)^{\mathbb{Z}}$ and $n \in \mathbb{Z}$, has "infinite memory". This causal map has approximable complexity since the encoding and decoding maps of Definition 4.9 can be taken to be

$$f_\varepsilon(t, (x_{-m(\varepsilon)}, \dots, x_0)) \overset{\text{def.}}{=} \left( x_0 + \sum_{-m(\varepsilon)\leq n\leq 0} k_n M(t, x_n), \Sigma(t, x_0) \right) \text{ and } \rho_\varepsilon(\mu, \Sigma) \overset{\text{def.}}{=} N(\mu, \Sigma\Sigma^\top),$$

for $m(\varepsilon) \in \mathbb{N}$ large enough.

## 6 | DISCUSSION

In the current manuscript, we considered the case where $\mathcal{X}$ is a compact subset of a Euclidean case. Indeed, one can consider more general "finite dimensional" input spaces by precomposing our geometric transformer and geometric hypertransformer, architectures with *feature maps*, which map inputs from more general metric spaces into Euclidean space. Moreover, by Kratsios (2023) any such continuous feature map is "'suitable", meaning that it preserves a model's universal approximation property upon pre-composition, if and only if it is injective.

The situation complicates whenever $\mathcal{X}$ is no longer assumed to be a compact subset of Euclidean space, and instead, it is assumed to be a compact subset of an infinite-dimensional

Banach space $B$. This is because our proof techniques rely on the extendability of Hölder functions defined on $\mathcal{X}$ into a QAS space $(\mathcal{Y}, d_{\mathcal{Y}}, \eta, Q)$ to Hölder functions defined on all of $B$ and that the Hölder constant of this extended function can be controlled in terms of the original map. This is an active area of contemporary analytic research which is subtle, even in seemingly familiar cases, such as when $\mathcal{X}$ and $\mathcal{Y}$ are both Lebesgue spaces on $\mathbb{R}$ (see Naor (2001)).

In the dynamic case, if we replace the uniform topology on the Fréchet space $(\mathbb{R}^d)^{\mathbb{Z}}$ with a weighted TVS topology in the sense of Prolla (1971); Schmocker (2022), then it may be possible to derive qualitative universal approximation theorems which are valid on a broader family of compact subsets of the path-space $(\mathbb{R}^d)^{\mathbb{Z}}$ with this alternative topology. We would like to explore this direction and the possible connection to the compression function $c_{AC}$ in Definition 4.7 in future work.

## 7 | PROOFS

This final section contains proofs of the paper's main results. We start by recalling concepts and introducing notations that we will use throughout the section.

Let $(\mathcal{X}, d)$ be a metric space. We will often make use of a homeomorphism of a metric space to itself, which "makes any $\alpha$-Hölder function Lipschitz" and, following (Weaver, 2018, Section 2.6), is defined as follows. For any $0 < \alpha \leq 1$ we define the *$\alpha$-snowflake* $(\mathcal{X}, d^\alpha)$ of $(\mathcal{X}, d)$ to be the metric space whose underlying set is $\mathcal{X}$ equipped with the metric $d^\alpha$ given by

$$d^\alpha(x_1, x_2) \overset{\text{def.}}{=} (d(x_1, x_2))^\alpha, \qquad x_1, x_2 \in \mathcal{X}.$$

The metric ball in $(\mathcal{X}, d)$ of radius $r > 0$ at $x \in \mathcal{X}$ is denoted by $\text{Ball}_{(\mathcal{X},d)}(x, r) \overset{\text{def.}}{=} \{z \in \mathcal{X} : d(x, z) < r\}$. The metric space $(\mathcal{X}, d)$ is called *doubling*, if there is $C \in \mathbb{N}_+$ for which every metric ball in $(\mathcal{X}, d)$ can be covered by at most $C$ metric balls of half its radius. The smallest such constant is called $(\mathcal{X}, d)$'s *doubling number*, and is here denoted by $C_{(\mathcal{X},d)}$, see (Heinonen, 2001, Section 10.13) for further details. We denote the Euclidean distance on $\mathbb{R}^n$ by $d_n$, and we write $d_n^\alpha$ for the metric of the $\alpha$-snowflake $(\mathbb{R}^n, d_n^\alpha)$ of $(\mathbb{R}^n, d_n)$.

We also recall that, for any $\alpha > 0$, the $\alpha$-Hölder norm of a function $g : [0,1]^n \to \mathbb{R}^m$ is defined as

$$\|g\|_\alpha \overset{\text{def.}}{=} \sup_{x \in [0,1]^n} \|g(x)\| + \sup_{x_1, x_2 \in [0,1]^n,\, x_1 \neq x_2} \frac{\|g(x_1) - g(x_2)\|}{\|x_1 - x_2\|^\alpha}.$$

When $\|g\|_\alpha < \infty$, we write $g \in C^\alpha([0,1]^n, \mathbb{R}^m)$.

### 7.1 | Proofs for the static case

**Lemma 7.1** (Doubling Number of Snowflakes and Covering Number). *Let $(\mathcal{X}, d)$ be a doubling metric space with doubling number $C_{(\mathcal{X},d)}$. Then, for $\alpha \in (0, 1]$, the doubling number $C_{(\mathcal{X},d^\alpha)}$ of its $\alpha$-snowflake $(\mathcal{X}, d^\alpha)$ is bounded as*

$$C_{(\mathcal{X},d^\alpha)} \leq C_{(\mathcal{X},d)}^{\lceil 1/\alpha \rceil}. \tag{16}$$

*Moreover, if $\mathcal{X}$ has finite diameter, then, for any $\delta > 0$, there exist $N \overset{def.}{=} C_{(\mathcal{X},d_\mathcal{X})}^{\lceil \log_2(\mathrm{diam}(\mathcal{X})) - \log_2(\delta)\rceil}$ balls of radius $\delta$ covering $\mathcal{X}$, i.e., there are $x_i \in \mathcal{X}$ for $i = 1, \dots, N$ such that*

$$\mathcal{X} = \bigcup_{i=1}^{N} \mathrm{Ball}_{(\mathcal{X},d)}(x_i, \delta). \tag{17}$$

*Proof.* Since $(\mathcal{X}, d)$ is doubling, there are, for each $x \in \mathcal{X}$ and $r > 0$, $x_i \in K, i = 1, \dots, C_d$ such that

$$\mathrm{Ball}_{(\mathcal{X},d)}(x, r) \subseteq \bigcup_{i=1}^{C_{(\mathcal{X},d)}} \mathrm{Ball}_{(\mathcal{X},d)}(x_i, r/2).$$

Iteratively applying this reasoning $j$-times for $j \in \mathbb{N}$, we find

$$\mathrm{Ball}_{(\mathcal{X},d)}(x, r) \subseteq \bigcup_{i_1,\dots,i_j=1}^{C_{(\mathcal{X},d)}} \mathrm{Ball}_{(\mathcal{X},d)}\left(x_{i_1,\dots,i_j}, r/2^j\right), \tag{18}$$

where $x_{i_1,\dots,i_j}$ are elements of $\mathcal{X}$.

To see the first statement, note that

$$\mathrm{Ball}_{(\mathcal{X},d)}(x, r) = \mathrm{Ball}_{(\mathcal{X},d^\alpha)}(x, r^\alpha), \text{ thus } \mathrm{Ball}_{(\mathcal{X},d)}(x, r/2^{1/\alpha}) = \mathrm{Ball}_{(\mathcal{X},d^\alpha)}(x, r^\alpha/2).$$

For this reason, we choose $j = \lceil 1/\alpha \rceil$ and derive from (18) that a ball in $(\mathcal{X}, d^\alpha)$ can be covered by $C_{(\mathcal{X},d)}^j$ balls half the radius, which yields (16).

To see the second statement, note that $\mathcal{X} \subseteq \mathrm{Ball}_{(\mathcal{X},d)}(x, \mathrm{diam}(\mathcal{X}))$ for any $x \in \mathcal{X}$. Choose $j = \lceil \log_2(\mathrm{diam}(\mathcal{X})) - \log_2(\delta)\rceil$ and $r = \mathrm{diam}(\mathcal{X})$. We have $r \le 2^j\delta$ and therefore, by (20), $\mathcal{X}$ can be covered by $C_d^j = N$ balls of radius $\delta$. □

**Lemma 7.2** (Extension of Hölder Functions on Compact Subsets of Euclidean Space). *Let $K \subseteq \mathbb{R}^n$ be compact, $\alpha \in (0, 1]$, and $f \in C^\alpha(K, \mathbb{R}^m)$ with $\alpha$-Hölder constant $L_f$. Then there exists an extension $F \in C^\alpha(\mathbb{R}^n, \mathbb{R}^m)$ of $f$ with constant $L_F$, such that*

$$L_F \le c \cdot \lceil \alpha^{-1} \rceil \log_2\left(\kappa_K(5^{-1})\right) L_f, \tag{19}$$

*for some universal constant $c > 0$.*

*Proof.* Since the identity $(\mathbb{R}^n, d_n) \ni x \mapsto x \in (\mathbb{R}^n, d_n^\alpha)$ is a quasisymmetry (Heinonen, 2001, page 78), it is by definition a homeomorphism. Thus, $K$ is closed in $(\mathbb{R}^n, d_n^\alpha)$ if and only if it is closed in $(\mathbb{R}^n, d_n)$. Moreover, a function $f$ is an element of $C^\alpha(K, \mathbb{R}^m)$ if and only if $f \in C^1((K, d_n^\alpha), (\mathbb{R}^m, d_m))$, since

$$\|f(x_1) - f(x_2)\| \le L_f \|x_1 - x_2\|^\alpha = L_f d_n^\alpha(x_1, x_2) \quad \text{for all } x_1, x_2 \in K.$$

In order to find an extension $F$ of $f$ with $\alpha$-Hölder constant bounded by (19), we want to apply (Bruè et al., 2021a, Theorem 4.1). It remains to show that the doubling number $C_{(K,d_n^\alpha)}$ of $(K, d_n^\alpha)$

is bounded as

$$C_{(K,d_n^\alpha)} \leq \kappa_K (5^{-1})^{\lceil \alpha^{-1} \rceil}. \tag{20}$$

For this reason, we relate $C_{(K,d_n^\alpha)}$ with the doubling number $C_{(K,d_n)}$ of $(K, d_n)$. By (Bruè et al., 2021a, Proposition 1.7 (i)) we have the bound

$$C_{(K,d_n)} \leq \kappa_K (5^{-1}). \tag{21}$$

Now, by Lemma 7.1 and combining (21) with (16), we obtain (20). Therefore, we can apply (Bruè et al., 2021a, Theorem 4.1) and find an extension $F \in C^1((\mathbb{R}^n, d_n^\alpha), (\mathbb{R}^m, d_m))$ of $f$ with Lipschitz constant $L_F$ satisfying (19). Clearly, $F \in C^\alpha(\mathbb{R}^n, \mathbb{R}^m)$ with $\alpha$-Hölder constant $L_F$ which completes the proof. □

*Proof of Proposition* 3.10. If $f(x) = c$ for some constant $c > 0$, then the statement holds with the neural network $\hat{f}(x) = c$, which can be represented as in (2) with $[d] = (n, m)$, where $A^j$ is the 0 matrix for all $j$, and the "$c$" in (2) is taken to be this constant $c$. Therefore, we henceforth only need to consider the case where $f$ is not constant. Let us observe that, if we pick some $x^\star \in K$, then for any multi-index $[d]$ and any neural network $\hat{f}_\theta \in \mathcal{NN}^\sigma_{[d]}$, $\hat{f}_\theta(x) - f(x^\star) \in \mathcal{NN}^\sigma_{[d]}$, since $\mathcal{NN}^\sigma_{[d]}$ is invariant to post-composition by affine functions. Thus, we represent $\hat{f}_\theta(x) - f(x^\star) = \hat{f}_{\theta^\star}(x)$, for some $\theta^\star \in \mathbb{R}^{P([d])}$. Consequently:

$$\sup_{x \in K} \left| \|(f(x) - f(x^\star)) - \hat{f}_{\theta^\star}(x)\| - \|f(x) - \hat{f}_\theta(x)\| \right| = 0.$$

Therefore, without loss of generality, we assume that $f(x^*) = 0$ for some $x^* \in K$. By Lemma 7.2 we can extend $f$ to $F \in C^\alpha(\mathbb{R}^n, \mathbb{R}^m)$ with $\alpha$-Hölder constant bounded by (19).

**Step 1 – Normalizing $\tilde{f}$ to the Unit Cube:** First, we identify a hypercube "nestling" $K$. To this end, let

$$r_K \overset{\text{def.}}{=} \operatorname{diam}(K) \sqrt{\frac{n}{2(n+1)}}. \tag{22}$$

By Jung's Theorem (see Jung (1901)), there exists $x_0 \in \mathbb{R}^n$ such that the closed Euclidean ball $\overline{\operatorname{Ball}_{(\mathbb{R}^n, d_n)}(x_0, r_K)}$ contains $K$. Therefore, by Hölder's inequality, we have that the $n$-dimensional hypercube[7] $[x_0 - r_K \bar{1}, x_0 + r_K \bar{1}]$ contains $\overline{\operatorname{Ball}_{(\mathbb{R}^n, d_n)}(x_0, r_K)}$, where $\bar{1} = (1, \ldots, 1) \in \mathbb{R}^n$. Consequently, $K \subseteq [x_0 - r_K \bar{1}, x_0 + r_K \bar{1}]$. Let $\tilde{f} \overset{\text{def.}}{=} F|_{[x_0 - r_K \bar{1}, x_0 + r_K \bar{1}]}$, then $\tilde{f} \in C^\alpha([x_0 - r_K \bar{1}, x_0 + r_K \bar{1}], \mathbb{R}^m)$ is an extension of $f$ with $\alpha$-Hölder constant $L_{\tilde{f}}$ bounded by (19).

Since $K$ has at least two points, then $r_K > 0$. Hence, the affine function

$$T : \mathbb{R}^n \ni x \mapsto (2r_K)^{-1}(x - x_0 + r_K \bar{1}) \in \mathbb{R}^n$$

is well-defined, invertible, not identically 0, and maps $[x_0 - r_K \bar{1}, x_0 - r_K \bar{1}]$ to $[0, 1]^n$. Note that the $\alpha$-Hölder norm of $g \overset{\text{def.}}{=} \tilde{f} \circ T^{-1}$ is finite, as $g$ is $\alpha$-Hölder continuous with constant $L_g = (2r_K)^\alpha L_{\tilde{f}}$.

More explicitly, writing $u^* \overset{\text{def.}}{=} T(x^*) \in [0,1]^n$, we have $g(u^*) = 0$ and find

$$\begin{aligned}\|g\|_\alpha &= \sup_{u\in[0,1]^n} \|g(u)\| + \sup_{u_1,u_2\in[0,1]^n, u_1\neq u_2} \frac{\|g(u_1)-g(u_2)\|}{\|u_1-u_2\|^\alpha} \\ &= \sup_{u\in[0,1]^n} \|g(u)-g(u^*)\| + L_g \leq \sup_{u_1,u_2\in[0,1]^n} \|g(u_1)-g(u_2)\| + L_g \leq 2L_g \\ &\leq c\, 2^{\alpha+1} r_K^\alpha \left(\lceil \alpha^{-1}\rceil \log_2\left(\kappa_K(5^{-1})\right)\right) L_f, \end{aligned} \tag{23}$$

where we used that $L_{\tilde f}$ is bounded by (19). We define $\tilde g \overset{\text{def.}}{=} \|g\|_\alpha^{-1} g$, and get that, for each $i = 1,\dots,m$, the function $\tilde g^{(i)} \overset{\text{def.}}{=} \mathrm{pj}_i \circ \tilde g$ belongs to the unit ball of $C^\alpha([0,1]^n, \mathbb{R})$, where, for $i = 1,\dots,m$, $\mathrm{pj}_i$ denotes the canonical projection $\mathrm{pj}_i : \mathbb{R}^m \ni (x_1,\dots,x_m) \mapsto x_i \in \mathbb{R}$.

**Step 2 – Constructing the Approximator:** For $i = 1,\dots,m$, let $\hat f_{\theta^{(i)}} \in \mathcal{NN}^\sigma_{d^{(i)}}$ for some multi-index $[d^{(i)}] = (d_0^{(i)},\dots,d_{J^{(i)}+1})$ with $n$-dimensional input layer and 1-dimensional output layer, i.e. $d_0^{(i)} = n$ and $d_{J^{(i)}+1} = 1$, and let $\theta^{(i)} \in \mathbb{R}^{P([d^{(i)}])}$ be the parameters defining $\hat f_{\theta^{(i)}}$. Since the pre-composition by affine functions and the post-composition by linear functions of neural networks in $\mathcal{NN}^\sigma_{d^{(i)}}$ are again neural networks in $\mathcal{NN}^\sigma_{d^{(i)}}$, we have that $g_{\theta^{(i)}} \overset{\text{def.}}{=} \hat f_{\theta^{(i)}} \circ T^{-1}$ and $\tilde g_{\theta^{(i)}} \overset{\text{def.}}{=} \|g\|_\alpha^{-1} g_{\theta^{(i)}}$ are neural networks in $\mathcal{NN}^\sigma_{d^{(i)}}$. Note that due to bijectivity of the maps $T$ and $y \mapsto \|g\|_\alpha\, y$, the correspondence $\hat f_{\theta^{(i)}} \mapsto \tilde g_{\theta^{(i)}}$ is one-to-one. Denote the standard basis of $\mathbb{R}^m$ by $\{e_i\}_{i=1}^m$. We compute

$$\begin{aligned}\sup_{x\in K}\left\|f(x)-\sum_{i=1}^m \hat f_{\theta^{(i)}}(x)e_i\right\| &= \sup_{x\in K}\left\|\tilde f(x)-\sum_{i=1}^m \hat f_{\theta^{(i)}}(x)e_i\right\| \leq \sup_{x\in[x_0-r_K\bar 1, x_0+r_K\bar 1]}\left\|\tilde f(x)-\sum_{i=1}^m \hat f_{\theta^{(i)}}(x)e_i\right\| \\ &= \sup_{x\in[x_0-r_K\bar 1, x_0+r_K\bar 1]}\left\|\tilde f\circ T^{-1}\circ T(x)-\sum_{i=1}^m \hat f_{\theta^{(i)}}\circ T^{-1}\circ T(x)e_i\right\| \\ &= \sup_{u\in[0,1]^n}\left\|g(u)-\sum_{i=1}^m g_{\theta^{(i)}}(u)e_i\right\| \leq \sqrt{m}\sup_{u\in[0,1]^n}\max_{1\leq i\leq m}\left\|\mathrm{pj}_i\circ g(u)-g_{\theta^{(i)}}(u)\right\| \\ &= C_0 \sup_{u\in[0,1]^n}\max_{1\leq i\leq m}\left\|\tilde g^{(i)}(u)-\tilde g_{\theta^{(i)}}(u)\right\|, \end{aligned} \tag{24}$$

where $C_0 \overset{\text{def.}}{=} \|g\|_\alpha\sqrt{m}$. Since, for each $i = 1,\dots,m$, $\tilde g^{(i)}$ belongs to the unit ball of $C(([0,1]^n, d_n^\alpha), \mathbb{R})$, for $\sigma$ as in Definition 2.2 (resp. as in Definition 2.3) we may apply (Shen et al., 2021a, Theorem 1) (resp. (Kratsios and Papon, 2022, Proposition 59))

to conclude that, for any $W, D \in \mathbb{N}_+$ (resp. any $\tilde\varepsilon > 0$), and each $i = 1,\dots,m$, there are $\hat f_{\hat\theta^{(i)}} \in \mathcal{NN}^\sigma_{d^{(n)}}$ where $[d^{(i)}] = (d_0^{(i)},\dots,d_{J^{(i)}+1}^{(i)})$ such that

$$\begin{cases} J^{(i)} \leq 2^6 nD + 3 \text{ and } \max_{1\leq j\leq J^{(i)}+1} d_j{}^{(i)} \leq \max\{n, 5W+13\} & : \sigma \text{ of Definition 2.2} \\ J^{(i)} \in \mathcal{O}\left(\tilde\varepsilon^{-2n/\alpha} L_f^{2n/\alpha}(1+n/4)^{2n/\alpha}\right) \text{ and } \max_{1\leq j\leq J^{(i)}+1} d_j{}^{(i)} \leq n+3 & : \sigma \text{ of Definition 2.3} \end{cases} \tag{25}$$

and

$$\begin{cases} \sup_{u\in[0,1]^n} \max_{1\le i\le m} \|\tilde{g}^{(i)}(u) - \hat{f}_{\hat{\theta}^{(i)}}(u)\| \le n^{\frac{\alpha}{2}} W^{-\alpha\sqrt{D}} + 2n^{\frac{\alpha}{2}} W^{-\sqrt{D}} & : \sigma \text{ of Definition 2.2} \\ \sup_{u\in[0,1]^n} \max_{1\le i\le m} \|\tilde{g}^{(i)}(u) - \hat{f}_{\hat{\theta}^{(i)}}(u)\| \le \tilde{\varepsilon} & : \sigma \text{ of Definition 2.3.} \end{cases} \tag{26}$$

Consequently, (26) implies that

$$\begin{cases} \sup_{x\in K} \left\| f(x) - \sum_{i=1}^m \hat{f}_{\hat{\theta}^{(i)}}(x) e_i \right\| \le C_0 n^{\frac{\alpha}{2}} W^{-\sqrt{D}} (W^{(1-\alpha)\sqrt{D}} + 2) & : \sigma \text{ of Definition 2.2} \\ \sup_{x\in K} \left\| f(x) - \sum_{i=1}^m \hat{f}_{\hat{\theta}^{(i)}}(x) e_i \right\| \le C_0 \tilde{\varepsilon} & : \sigma \text{ of Definition 2.3.} \end{cases} \tag{27}$$

We express the constant $C_0$ in terms of the given $K$-dependent data by (22) and (23), thus obtaining

$$C_0 \le 2^{1+\alpha/2} c\sqrt{m}\,\mathrm{diam}(K)^\alpha \left(\frac{n}{n+1}\right)^{\alpha/2} \left(\lceil\alpha\rceil^{-1} \log_2\left(\kappa_K(5^{-1})\right)\right) L_f. \tag{28}$$

Therefore, we introduce the universal constant $\tilde{C}_0 \overset{\text{def.}}{=} 2^{1+\alpha/2}(\frac{n}{n+1})^{\alpha/2} c L_f$. Rearranging and combining with (25), we find that

$$\begin{cases} \varepsilon \overset{\text{def.}}{=} \sqrt{m} n^{\frac{\alpha}{2}} W^{-\sqrt{D}} (W^{(1-\alpha)\sqrt{D}} + 2) & : \sigma \text{ of Definition 2.2} \\ \varepsilon \overset{\text{def.}}{=} \sqrt{m}\tilde{\varepsilon} & : \sigma \text{ of Definition 2.3.} \end{cases} \tag{29}$$

Observe that $P([d^{(i)}]) = \sum_{j=0}^{J^{(i)}} d_{j+1}^{(i)}(d_j^{(i)} + 3) - 2d_{J^{(i)}+1}^{(i)} \le \sum_{j=0}^{J^{(i)}} d_{j+1}^{(i)}(d_j^{(i)} + 3) \le (\text{Depth} + 1) \times (\text{Width} + 3)^2$, where we denote the network's depth and width respectively by Depth and Width. Thus, we deduce that the number of parameters defining each network is bounded-above by

$$\begin{cases} P([d^{(i)}]) \le \max\{n+3, 5W+16\}^2 (2^6 nD + 4) & : \sigma \text{ of Definition 2.2} \\ P([d^{(i)}]) \le \mathcal{O}\left((n+6)^2 \left(\tilde{\varepsilon}^{-4n/\alpha} L_f^{4n/\alpha} (1 + n/4)^{4n/\alpha} + 1\right)\right) & : \sigma \text{ of Definition 2.3.} \end{cases} \tag{30}$$

**Step 3 – Counting Parameters:** Let $g_1 \bullet g_2$ denotes the component-wise composition of a univariate function $g_1$ with a multivariate function $g_2$. By construction, for any $k \in \mathbb{N}_+$, if $I_k$ denotes the $k \times k$-identity matrix, then $I_k \sigma_{(1,1)} \bullet I_k \in \mathcal{NN}_{d_k}^\sigma$ with $P([d]) = 2k$, and $I_k \sigma_{(1,1)} \bullet I_k = 1_{\mathbb{R}^k}$. Therefore, mutatis mutandis, $\mathcal{NN}^\sigma$ satisfies (Cheridito et al., 2021, Definition 4); whence, mutatis mutandis, we may apply (Cheridito et al., 2021, Proposition 5). Thus, there is a multi-index $[d] = (d_0, \dots, d_{J+1})$ with $d_0 = n$ and $d_J = m$, and a network $\hat{f}_\theta \in \mathcal{NN}_{[d]}^\sigma$ implementing $\sum_{i=1}^m \hat{f}_{\theta^{(i)}}$, i.e.

$$\sum_{i=1}^m \hat{f}_{\theta^{(i)}} e_i = \hat{f}_\theta,$$

such that $\hat{f}_\theta$'s depth ($J$) is bounded above by

$$J \leq \begin{cases} m(2^6 nD + 4) & : \sigma \text{ of Definition 2.2} \\ \mathcal{O}\Big(m\Big(\tilde{\varepsilon}^{-2n/\alpha} L_f^{2n/\alpha}(1+n/4)^{2n/\alpha} + 1\Big)\Big) & : \sigma \text{ of Definition 2.3,} \end{cases} \tag{31}$$

its width can be upper-bounded by

$$\max_{0\leq j\leq J+1} d_j \leq \begin{cases} n(m-1) + \max\{n, 5W+13\} & : \sigma \text{ of Definition 2.2} \\ nm + 3 & : \sigma \text{ of Definition 2.3,} \end{cases} \tag{32}$$

and it total number of parameters can be upper-bounded by

$$P([d]) \leq \left(\frac{11}{16}\, 2^2\, n^2 m^2 - 1\right) \sum_{i=1}^{m} P([d^{(i)}]). \tag{33}$$

Therefore, depending on the trainable activation function $\sigma$, (33) implies that

$$P([d]) \leq \begin{cases} \Big(\frac{11}{4} n^2 m^2 - 1\Big) m \max\{n+3, 5W+16\}^2 (2^6 nD + 4) & : \sigma \text{ of Definition 2.2} \\ \mathcal{O}\Big(\Big(\frac{11}{4} n^2 m^2 - 1\Big) m (n+6)^2 \Big(\tilde{\varepsilon}^{-4n/\alpha} L_f^{4n/\alpha}(1+n/4)^{4n/\alpha} + 1\Big)\Big) & : \sigma \text{ of Definition 2.3.} \end{cases} \tag{34}$$

Combining (34) with (30) yields our result for the cases where $\sigma$ is as in Definition 2.2 and as in Definition 2.3. The case where $\sigma$ is as in Definition 2.4 follows from (Kidger et al., 2021, Theorem 3.2). □

## 7.2 | Proof of Theorem 3.8

*Proof of Theorem* 3.8. Let $\alpha \in (0,1]$ and fix $\varepsilon_A, \varepsilon_Q > 0$.

**Step 1 – The Random Projection:** Let $A \subseteq \mathcal{X}$ be closed and denote by $C_{(A,d_{\mathcal{X}}^\alpha)}$ the doubling number of the $\alpha$-snowflake of $A$, where $d_{\mathcal{X}}$ is the base metric on $A$ defining the Wasserstein distance on $\mathcal{P}_1(A)$. By (Bruè et al., 2021a, Theorem 3.2), there exists a random projection $\Pi : \mathcal{X} \to \mathcal{P}_1(A)$ with Lipschitz constant $L_\Pi$ such that

$$L_\Pi \leq c \cdot \log_2(C_{(A,d_{\mathcal{X}}^\alpha)}), \tag{35}$$

where $c > 0$ is a universal constant. To make the bound of $L_\Pi$ more explicit, we estimate $C_{(A,d_{\mathcal{X}}^\alpha)}$. For this reason, consider $x \in A$ and $r > 0$. We may externally cover $\mathrm{Ball}_{(A,d_{\mathcal{X}}^\alpha)}(x,r)$ by $C^2_{(\mathcal{X},d_{\mathcal{X}}^\alpha)}$ balls of radius $r/4$ centered at points in $\mathcal{X}$. Consider a ball centered at $\tilde{x} \in \mathcal{X}$ with

$$\mathrm{Ball}_{(\mathcal{X},d_{\mathcal{X}}^\alpha)}(\tilde{x}, r/4) \cap A \neq \emptyset,$$

then we can find a ball of radius $r/2$ centered in $A$ covering it. Therefore, $\mathrm{Ball}_{(\mathcal{X},d_{\mathcal{X}}^{\alpha})}(x,r)$ may be internally covered by at most $C^2_{(\mathcal{X},d_{\mathcal{X}}^{\alpha})}$-balls of radius $r/2$. By inserting this into (35) and applying Lemma 7.1, we thus get

$$L_\Pi \leq 2c \cdot \log_2(C_{(\mathcal{X},d_{\mathcal{X}}^{\alpha})}) \leq 2c\lceil 1/\alpha \rceil \cdot \log_2(C_{(\mathcal{X},d_{\mathcal{X}})}) =: C_\Pi. \tag{36}$$

Note that the right-hand side of (36) is independent of the choice of $A$, and we may assume from now on without loss of generality that $C_\Pi \geq 1$.

**Step 2 – Estimating the External Covering Number of $f(\mathcal{X})$**: By passing on to the snowflake $(\mathcal{X}, d_{\mathcal{X}}^{\alpha})$, we may apply the reasoning of Lemma 7.1 and cover $(\mathcal{X}, d_{\mathcal{X}}^{\alpha})$, for any $\delta > 0$, by

$$N_\delta \leq C_{(\mathcal{X},d_{\mathcal{X}})}^{\lceil \left(\log_2(\mathrm{diam}(\mathcal{X})) - \log_2(\delta)\right)/\alpha \rceil} \tag{37}$$

distinct balls of radius $\delta$ centered in $\{x_i\}_{i=1}^{N_\delta} \subseteq \mathcal{X}$. From now on, let

$$\delta \overset{\text{def.}}{=} \frac{\varepsilon_A}{3L_f C_\eta C_\Pi}, \tag{38}$$

where we recall that $C_\eta \geq 1$ is the constant of the mixing function, see Definition 3.1, and write $N \overset{\text{def.}}{=} N_\delta$ and $\mathcal{X}_N \overset{\text{def.}}{=} \{x_i\}_{i=1}^{N}$. Then

$$\max_{x\in\mathcal{X}} \min_{1\leq i\leq N} d_{\mathcal{Y}}(f(x), f(x_i)) \leq L_f \max_{x\in\mathcal{X}} \min_{1\leq i\leq N} d_{\mathcal{X}}^{\alpha}(x, x_i) \tag{39}$$

$$\leq L_f \delta \leq \varepsilon_A/3. \tag{40}$$

In particular, the $\varepsilon_A/3$-external covering number of $f(\mathcal{X})$ is at most $N$.

**Step 3 – $\mathcal{P}_1(\mathcal{X}_N)$ and the simplex $\Delta_N$**: Let $\Pi_{\mathcal{P}_1(\mathcal{X}_N)} : (\mathcal{X}, d_{\mathcal{X}}^{\alpha}) \to (\mathcal{P}_1(\mathcal{X}_N), \mathcal{W}_1)$ be a $C_\Pi$-Lipschitz random projection, which exists by what was recalled in Step 1. We find by the triangle inequality and (38) that

$$d_{\mathcal{X}}^{\alpha}(x_1, x_2) \leq \delta \text{ and } d_{\mathcal{X}}^{\alpha}(x_2, x_3) \leq \delta \Rightarrow d_{\mathcal{X}}^{\alpha}(x_1, x_3) \leq (2\varepsilon_A)/(3L_f).$$

Choose $\tilde{\mathcal{X}}_N \overset{\text{def.}}{=} \{\tilde{x}_i\}_{i=1}^{\tilde{N}}$, $\tilde{N} \leq N$, to be a subset of $\mathcal{X}_N$ with

$$\delta/2 \leq \min_{i\neq j} d_{\mathcal{X}}^{\alpha}(\tilde{x}_i, \tilde{x}_j) \leq \mathrm{diam}(\mathcal{X})^{\alpha}, \tag{41}$$

$$\max_{1\leq i\leq N} \min_{1\leq j\leq \tilde{N}} d_{\mathcal{X}}^{\alpha}(x_i, \tilde{x}_j) \leq \delta. \tag{42}$$

Then $f(\tilde{\mathcal{X}}_N)$ is a $(2\varepsilon_A)/3$-covering of $f(\mathcal{X})$ consisting of at most $N$ points, because, by (42),

$$\max_{x\in\mathcal{X}} \min_{1\leq i\leq \tilde{N}} d_{\mathcal{Y}}(f(x), f(\tilde{x}_i)) \leq \max_{x\in\mathcal{X}} \min_{1\leq i\leq N} d_{\mathcal{Y}}(f(x), f(x_i)) + \min_{1\leq j\leq \tilde{N}} d_{\mathcal{Y}}(f(x_i), f(\tilde{x}_i)) \tag{43}$$

$$\leq L_f(\delta + \delta) \leq (2\varepsilon_A)/3. \tag{44}$$

From now on we will solely use $\tilde{\mathcal{X}}_N$ and, to ease of notation, rename $\tilde{\mathcal{X}}_N$ as $\mathcal{X}_N$, $\tilde{N}$ as $N$, and $\{\tilde{x}_i\}_{i=1}^{\tilde{N}}$ as $\{x_i\}_{i=1}^{N}$. Clearly, the Wasserstein space $(\mathcal{P}_1(\mathcal{X}_N), \mathcal{W}_1)$ is homeomorphic to $(\mathcal{P}_1(\mathcal{X}_N), \mathrm{TV})$, where TV is the total variation distance on $\mathcal{P}_1(\mathcal{X}_N)$, and the two metrics are equivalent thereon, since

$$(\delta/2) \cdot \mathrm{TV}(\mu,\nu) \le \mathcal{W}_1(\mu,\nu) \le \mathrm{diam}(\mathcal{X})^\alpha \cdot \mathrm{TV}(\mu,\nu). \tag{45}$$

Let $\|\cdot\|_1$ be the $\ell^1$-norm and $\|\cdot\|$ the Euclidean norm. Note that $(\mathcal{P}_1(\mathcal{X}_N), \mathrm{TV})$ is isometric to the simplex $(\Delta_N, \|\cdot\|_1)$ and homeomorphic to $(\Delta_N, \|\cdot\|)$. Indeed, there is a map $\iota_N : \mathcal{P}_1(\mathcal{X}_N) \to \Delta_N$ with

$$\|\iota_N(\mu) - \iota_N(\nu)\| \le \mathrm{TV}(\mu,\nu) = \|\iota_N(\mu) - \iota_N(\nu)\|_1 \le \sqrt{N} \cdot \|\iota_N(\mu) - \iota_N(\nu)\|, \tag{46}$$

for all $\mu, \nu \in \mathcal{P}_1(\mathcal{X}_N)$. Combining (36) with (45) and (46) yields an estimate on the Lipschitz constant, denoted by $L_{f_N}$, of $f_N \overset{\text{def.}}{=} \iota_N \circ \Pi_{\mathcal{P}_1(\mathcal{X}_N)} : (\mathcal{X}, d_{\mathcal{X}}^\alpha) \to (\Delta_N, \|\cdot\|) \subseteq (\mathbb{R}^N, \|\cdot\|)$:

$$\begin{aligned}
\|f_N(x) - f_N(\tilde{x})\| &\le \frac{2}{\delta} \mathcal{W}_1(\Pi_{\mathcal{P}_1(\mathcal{X}_N)}(x), \Pi_{\mathcal{P}_1(\mathcal{X}_N)}(\tilde{x})) \\
&\le \frac{4c\lceil 1/\alpha \rceil}{\delta} \cdot \log_2(C_{(\mathcal{X},d_{\mathcal{X}})}) \cdot d_{\mathcal{X}}^\alpha(x,\tilde{x}) \\
&= \left(12c^2 \lceil 1/\alpha \rceil^2 L_f\right) \cdot \left(\frac{C_\eta}{\varepsilon_A}\right)^{1/r} \cdot \log_2(C_{(\mathcal{X},d_{\mathcal{X}})})^2 \cdot d_{\mathcal{X}}^\alpha(x,\tilde{x})
\end{aligned}$$

for all $x, \tilde{x} \in \mathcal{X}$.

**Step 4 – Approximation of** $f$: We denote the elements of $f(\mathcal{X}_N)$ by $y_i \overset{\text{def.}}{=} f(x_i), i = 1, \dots, N$. By Definition 3.3, we have that $f(\mathcal{X}_N)$ is quantizable, i.e. there exists $q \in \mathbb{N}_+$ with $D_q \overset{\text{def.}}{=} Q_{f(\mathcal{X}_N)}(\varepsilon_Q)$ and there are $\{\tilde{y}_i\}_{i=1}^N \subseteq \mathcal{Y}$ such that

$$d_{\mathcal{Y}}(y_i, \tilde{y}_i) \le \min\left\{\frac{\varepsilon_Q}{3C_\eta}, \frac{\varepsilon_Q}{3}\right\}, \quad 1 \le i \le N, \tag{47}$$

and $\tilde{\mathcal{Y}}_N \overset{\text{def.}}{=} \{\tilde{y}_i\}_{i=1}^N$ are representable with $Q_{f(\mathcal{X}_N)}(\varepsilon_Q)$-parameters, that is, there are $z_1, \dots, z_N \in \mathbb{R}^{Q_{f(\mathcal{X}_N)}(\varepsilon_Q)}$ such that $Q_q(z_i) = \tilde{y}_i$ for $i = 1, \dots, N$ and the estimate (47) holds. Since $\mathcal{Y}_N = (y_1, \dots, y_N)$ and $\tilde{\mathcal{Y}}_N = (\tilde{y}_1, \dots, \tilde{y}_N)$ are fixed, to ease the notation, from here and till the end of the proof we will use

$$\begin{aligned}
\eta_N &: \Delta_N \to \mathcal{Y} : w \mapsto \eta(w, \mathcal{Y}_N), \\
\tilde{\eta}_N &: \Delta_N \to \mathcal{Y} : w \mapsto \eta(w, \tilde{\mathcal{Y}}_N).
\end{aligned}$$

Let $1 \le i \le N$, $w \in \Delta_N$ and $p \ge 1$. Since, for every $x_i$, the function $x \mapsto d_{\mathcal{Y}}(f(x), f(x_i))$ is $L_f$-Lipschitz on $(\mathcal{X}, d^\alpha)$, then we have

$$\mathcal{W}_p^p(\delta_{x_i}, \iota_N^{-1}(w)) = \sum_{j=1}^N w_j d_{\mathcal{X}}^\alpha(x_i, x_j)^p \ge \frac{1}{L_f^p} \sum_{j=1}^N w_j d_{\mathcal{Y}}(f(x_i), f(x_j))^p = \frac{1}{L_f^p} \sum_{j=1}^N w_j d_{\mathcal{Y}}(y_i, y_j)^p.$$

Using this, together with Definition 3.1 and (47), yields

$$
\begin{aligned}
d_{\mathcal{Y}}(\widetilde{\eta}_N(e_i), \widetilde{\eta}_N(w)) &\leq C_\eta \left( \sum_{j=1}^N w_j d_{\mathcal{Y}}(\tilde{y}_i, \tilde{y}_j)^p \right)^{1/p} \\
&\leq C_\eta \left( \sum_{j=1}^N w_j \left( d_{\mathcal{Y}}(\tilde{y}_i, y_i) + d_{\mathcal{Y}}(y_i, y_j) + d_{\mathcal{Y}}(y_j, \tilde{y}_j) \right)^p \right)^{1/p} \\
&\leq C_\eta \left( \sum_{j=1}^N w_j \left( \frac{2\varepsilon_Q}{3C_\eta} + d_{\mathcal{Y}}(y_i, y_j) \right)^p \right)^{1/p} \\
&\leq 2\varepsilon_Q/3 + C_\eta \left( \sum_{j=1}^N w_j d_{\mathcal{Y}}(y_i, y_j)^p \right)^{1/p} \\
&\leq 2\varepsilon_Q/3 + C_\eta L_f \mathcal{W}_p(\Pi_{\mathcal{P}_1(\mathcal{X}_N)}(x_i), \iota_N^{-1}(w)), \qquad (48)
\end{aligned}
$$

where we used Minkowski's inequality in the second to last step. Now we want to obtain an estimate of $\mathcal{W}_p(\Pi_{\mathcal{P}_1(\mathcal{X}_N)}(x_i), \Pi_{\mathcal{P}_1(\mathcal{X}_N)}(x))$, for $1 \leq i \leq N$ and $x \in \mathcal{X}$. If $x \in \mathcal{X}_N$, then $\mathcal{W}_p(\Pi_{\mathcal{P}_1(\mathcal{X}_N)}(x_i), \Pi_{\mathcal{P}_1(\mathcal{X}_N)}(x)) = d_{\mathcal{X}}(x, x_i)$. For the case of $x \notin \mathcal{X}_N$, we note from the proof of (Bruè et al., 2021a, Theorem 3.2) that the projection $\Pi_{\mathcal{P}_1(\mathcal{X}_N)}$ even satisfies

$$
\mathcal{W}_p(\Pi_{\mathcal{P}_1(\mathcal{X}_N)}(x_i), \Pi_{\mathcal{P}_1(\mathcal{X}_N)}(x)) \leq C_\Pi d_{\mathcal{X}}(x_i, x), \qquad (49)
$$

where crucially enters $x_i \in \mathcal{X}_N$.

Since, for $1 \leq i \leq N$, $\eta_N(e_i) = y_i = f(x_i)$, $\widetilde{\eta}_N(e_i) = \tilde{y}_i$ and $f_N(x_i) = e_i$, we find by (44), (47), (48), (49) and Lipschitz continuity of $\Pi_{\mathcal{P}_1(\mathcal{X}_N)}$, that

$$
\begin{aligned}
\sup_{x \in \mathcal{X}} d_{\mathcal{Y}}(f(x), \widetilde{\eta}(f_N(x))) &\leq \sup_{x \in \mathcal{X}} \min_{1 \leq i \leq N} \left\{ d_{\mathcal{Y}}(f(x), y_i) + d_{\mathcal{Y}}(y_i, \tilde{y}_i) + d_{\mathcal{Y}}(\tilde{y}_i, \widetilde{\eta}_N(f_N(x))) \right\} \qquad (50) \\
&\leq (2\varepsilon_A)/3 + \varepsilon_Q/3 + \sup_{x \in \mathcal{X}} d_{\mathcal{Y}}(\widetilde{\eta}_N(e_{i(x)}), \widetilde{\eta}_N(f_N(x))) \\
&\leq (2\varepsilon_A)/3 + \varepsilon_Q + \sup_{x \in \mathcal{X}} C_\eta L_f \mathcal{W}_p(\Pi_{\mathcal{P}_1(\mathcal{X}_N)}(x_{i(x)}), \iota_N^{-1} \circ f_N(x)) \\
&= (2\varepsilon_A)/3 + \varepsilon_Q + \sup_{x \in \mathcal{X}} C_\eta L_f \mathcal{W}_p(\Pi_{\mathcal{P}_1(\mathcal{X}_N)}(x_{i(x)}), \Pi_{\mathcal{P}_1(\mathcal{X}_N)}(x)) \\
&\leq (2\varepsilon_A)/3 + \varepsilon_Q + C_\eta L_f C_\Pi \delta, \qquad (51)
\end{aligned}
$$

where $i(x)$ satisfies $i(x) \in \arg\min_{1 \leq i \leq N} d_{\mathcal{X}}(x, x_i)$ and $b$ is the universal constant in (49). Recalling the definition of $\delta$, see (38), yields

$$
\sup_{x \in \mathcal{X}} d_{\mathcal{Y}}(f(x), \widetilde{\eta}_N(f_N(x))) \leq \varepsilon_A + \varepsilon_Q.
$$

Hence, $\widetilde{\eta}_N \circ f_N$ is an $(\varepsilon_A + \varepsilon_Q)$-approximation of $f$.

**Step 5 – Approximating $f_N$ by neural networks:** Fix an approximation error $\varepsilon_{\mathcal{NN}} > 0$ of the neural network. Consider either scenario:

- **Scenario A (Trainable Activation Function: Singular-ReLU-Type)**: Under Assumption 2.2, Proposition 3.10 guarantees that there is a network $\hat{f}_\theta \in \mathcal{NN}^{\sigma}_{[d]}$ where, $d = (d_0, \dots, d_J)$, $d_0 = d$, and $d_J = N$ of depth and width as recorded in Table 1 satisfying (52) below.
- **Scenario B (Trainable Activation Function: Smooth-ReLU-Type)**: Under Assumption 2.3, we may apply (Kratsios, 2023, Proposition 58) to conclude that for each $i = 1, \dots, N$ there is a feedforward network with activation function $\sigma^\star$ of width $d + N + 2$ and of depth $\mathcal{O}((1 + \frac{d}{4})^{\frac{2d}{\alpha}} \varepsilon^{\frac{-2d}{\alpha}})$ approximating each $\mathrm{pj}_i \circ f_N$. Since $\sigma_{(1,1)}$ is the identity on $\mathbb{R}$, then we may (mutatis mutandis) apply the parallelization construction of (Cheridito et al., 2021, Proposition 5) to conclude that there is a network $\hat{f}_\theta \in \mathcal{NN}^{\sigma}_{[d]}$ where $d = (d_0, \dots, d_J)$, $d_0 = d$, $d_J = N$, and of depth and width as recorded in Table 1, such that (52) below holds.
- **Scenario C (Classical Activation Function)**: Under Assumption 2.4, (Kidger & Lyons, 2020, Theorem 3.2) applies, whence, there exists a network $\hat{f}_\theta \in \mathcal{NN}^{\sigma}_{[d]}$ where $d = (d_0, \dots, d_J)$, $d_0 = d$, $d_J = N$, of width at most $d + N + 2$ satisfying (52) below.

In either case, we may choose $\hat{f}$ such that

$$\sup_{x \in \mathcal{X}} \|f_N(x) - \hat{f}_\theta(x)\| \leq \frac{\varepsilon_{\mathcal{NN}}}{C_\eta L_f \mathrm{diam}(\mathcal{X})^\alpha \sqrt{N}}. \tag{52}$$

Since the projection $\Pi_{\Delta_N} : (\mathbb{R}^N, \|\cdot\|_2) \to (\Delta_N, \|\cdot\|_2)$ onto the convex subset $\Delta_N \subseteq \mathbb{R}^N$ is 1-Lipschitz, and noting that $\Pi_{\mathcal{P}_1(\mathcal{X}_N)}(x) = \iota_N^{-1} \circ f_N(x)$, we may compute

$$\sup_{x \in \mathcal{X}} \mathcal{W}_1(\iota_N^{-1} \circ \Pi_{\Delta_N} \circ \hat{f}_\theta(x), \iota_N^{-1} \circ f_N(x)) = \sup_{x \in \mathcal{X}} \mathcal{W}_1(\iota_N^{-1} \circ \Pi_{\Delta_N} \circ \hat{f}_\theta(x), \iota_N^{-1} \circ \Pi_{\Delta_N} \circ f_N(x)) \tag{53}$$

$$\leq \sup_{x \in \mathcal{X}} \mathrm{diam}(\mathcal{X})^\alpha \sqrt{N} \cdot \|\Pi_{\Delta_N} \circ \hat{f}_\theta(x) - \Pi_{\Delta_N} \circ f_N(x)\|$$

$$\leq \sup_{x \in \mathcal{X}} \mathrm{diam}(\mathcal{X})^\alpha \sqrt{N} \cdot \|\hat{f}_\theta(x) - f_N(x)\|$$

$$\leq \frac{\varepsilon_{\mathcal{NN}}}{C_\eta L_f}, \tag{54}$$

where we applied (45) and (46) to obtain the first, and (52) to obtain the last inequality.

Now let $x \in \mathcal{X}$ and $i = i(x)$. Then, analogously to Step 4, (44) and (48) give

$$d_{\mathcal{Y}}(f(x), \widetilde{\eta}_N \circ \Pi_{\Delta_N} \circ \hat{f}_\theta(x)) \leq d_{\mathcal{Y}}(f(x), \widetilde{\eta}_N(e_i))) + d_{\mathcal{Y}}(\widetilde{\eta}_N(e_i), \widetilde{\eta}_N \circ \Pi_{\Delta_N} \circ \hat{f}_\theta(x))$$

$$\leq (2\varepsilon_A)/3 + 2\varepsilon_Q/3 + C_\eta L_f \mathcal{W}_1(\Pi_{\mathcal{P}_1(\mathcal{X}_N)}(x_i), \iota_N^{-1} \circ \Pi_{\Delta_N} \circ \hat{f}_\theta(x)). \tag{55}$$

Note that in the last term we can use the estimate

$$\mathcal{W}_1(\Pi_{\mathcal{P}_1(\mathcal{X}_N)}(x_i), \iota_N^{-1} \circ \Pi_{\Delta_N} \circ \hat{f}_\theta(x)) \leq \mathcal{W}_1(\Pi_{\mathcal{P}_1(\mathcal{X}_N)}(x_i), \Pi_{\mathcal{P}_1(\mathcal{X}_N)}(x)) + \mathcal{W}_1(\Pi_{\mathcal{P}_1(\mathcal{X}_N)}(x), \iota_N^{-1} \circ \Pi_{\Delta_N} \circ \hat{f}_\theta(x))$$

$$\leq C_\Pi \delta + \frac{\varepsilon_{\mathcal{NN}}}{C_\eta L_f}.$$

Finally, plugging this into (55) yields

$$\sup_{x \in \mathcal{X}} d_{\mathcal{Y}}(f(x), \widetilde{\eta}_N \circ \Pi_{\Delta_N} \circ \hat{f}_\theta(x)) \leq (2\varepsilon_A)/3 + 2\varepsilon_Q/3 + C_\eta L_f C_\Pi \delta + \varepsilon_{\mathcal{NN}}$$

$$\leq \varepsilon_A + \varepsilon_Q + \varepsilon_{\mathcal{NN}}.$$

This completes the approximation. **Step 6 – Expressing $N$ and $Q$ using metric entropy:** It remains to derive simple expressions for $N$ and $Q$. Together, (36), (37) and (38) imply that

$$\begin{aligned} N \leq & C_{(\mathcal{X},d_{\mathcal{X}})}^{\left\lceil \left(\log_2(\operatorname{diam}(\mathcal{X}))-\log_2(\delta)\right)/\alpha \right\rceil} \\ = & C_{(\mathcal{X},d_{\mathcal{X}})}^{\left\lceil \left(\log_2(\operatorname{diam}(\mathcal{X}))-\log_2\left(\varepsilon_A/\left(3L_f C_\eta C_\Pi\right)\right)\right)/\alpha \right\rceil} \\ = & C_{(\mathcal{X},d_{\mathcal{X}})}^{\left\lceil \alpha^{-1}\left(\log_2(\operatorname{diam}(\mathcal{X}))-\log_2\left(\varepsilon_A/(3L_f)\right)+\log_2\left((C_\eta C_\Pi)\right)\right)\right\rceil} \\ = & C_{(\mathcal{X},d_{\mathcal{X}})}^{\left\lceil \alpha^{-1}\left(\log_2(\operatorname{diam}(\mathcal{X}))-\log_2\left(\varepsilon_A/(3L_f)\right)+\log_2\left(C_\eta 2c\lceil 1/\alpha\rceil\cdot\log_2(C_{(\mathcal{X},d_{\mathcal{X}})})\right)\right)\right\rceil}. \end{aligned} \tag{56}$$

By (Bruè et al., 2021b, Proposition 1.7 (i)), we have the estimate $C_{(\mathcal{X},d_{\mathcal{X}})} \leq \kappa_{\mathcal{X}}(5^{-1})$. Thus, (56) can be bounded above as

$$\ln(N) \leq \ln\left(\kappa_{\mathcal{X}}\left(5^{-1}\right)\right)\left\lceil \alpha^{-1}\left(\log_2(\operatorname{diam}(\mathcal{X})) - \log_2\left(\varepsilon_A/3L_f\right) + \log_2\left(C_\eta 2c\lceil 1/\alpha\rceil\cdot\log_2\left(\kappa_{\mathcal{X}}\left(5^{-1}\right)\right)\right)\right)\right\rceil.$$

This completes the simplified estimate on $N$.

For $Q$, following (47) we have that $Q = Q_{f(\mathcal{X}_N)}(\varepsilon_Q)$. Moreover, by definition, since $\mathcal{X}_N \subseteq \mathcal{X}$, we have that

$$Q = Q_{f(\mathcal{X}_N)}(\varepsilon_Q) \leq Q_{f(\mathcal{X})}(\varepsilon_Q).$$

This completes the estimate on $Q$ as well as our proof. □

## 7.3 | Proofs for the dynamic case

*Proof of Proposition* 4.2. Under the assumed condition on the $\mathbb{F}$-progressively measurable stochastic processes $(\alpha_t)_{t\geq 0}$ and $(\beta_t)_{t\geq 0}$, we can conclude from (Touzi, 2013, Theorem 2.2) that

$$\mathbb{E}\left[\sup_{t\in[0,t_n]} ||X_t||^2\right] \leq C_n e^{C_n t_n} \tag{57}$$

holds for every $n \in \mathbb{N}_+$, for some $C_n > 0$, where each $C_n$ depends only on the prescribed local-Lipschitz constant $L_n$, on $t_n$, and on the law of $X_0$. For every $n \in \mathbb{N}_+$, let $\tilde{C}_n \geq C_n$ be chosen large enough such that the exponential sum $\sum_{n=1}^{\infty} e^{(C_n-\tilde{C}_n)\,n\,\delta_+}$ converges. Now, we may apply Tonelli's theorem and deduce that

$$\begin{aligned} \mathbb{E}\left[\sum_{n=1}^{\infty}\sup_{t\in[0,t_n]} ||X_t||^2 \frac{e^{-\tilde{C}_n t_n}}{C_n}\right] &= \sum_{n=1}^{\infty}\mathbb{E}\left[\sup_{t\in[0,t_n]} ||X_t||^2\right]\frac{e^{-\tilde{C}_n t_n}}{C_n} \leq \sum_{n=1}^{\infty} C_n\, e^{C_n t_n}\,\frac{e^{-\tilde{C}_n t_n}}{C_n} \\ &= \sum_{n=1}^{\infty} e^{(C_n-\tilde{C}_n)t_n} \leq \sum_{n=1}^{\infty} e^{(C_n-\tilde{C}_n)\,n\,\delta_-} < \infty. \end{aligned} \tag{58}$$

Since (58) implies that $\sum_{n=1}^{\infty} \sup_{t\in[0,t_n]} ||X_t||^2 \frac{e^{-(C_n-\tilde{C}_n)t_n}}{C_n}$ is a.s. finite, and since this quantity is non-negative, we may apply the Markov inequality to deduce that, for any given positive constant $\gamma$, the following concentration bound holds

$$\mathbb{P}\left(\sum_{n=1}^{\infty} \sup_{t\in[0,t_n]} ||X_t||^2 \frac{e^{-\tilde{C}_n t_n}}{C_n} \geq \gamma\right) \leq \frac{1}{\gamma} \sum_{n=1}^{\infty} e^{(C_n-\tilde{C}_n)t_n} \leq \frac{1}{\gamma} \sum_{n=1}^{\infty} e^{(C_n-\tilde{C}_n)n\,\delta_-}. \tag{59}$$

In particular, (59) implies the looser bound

$$\begin{aligned}\mathbb{P}\left(\sup_{n\in\mathbb{N}_+} \frac{e^{-\tilde{C}_n t_n/2}}{C_n^{1/2}} \sup_{t\in[0,t_n]} ||X_t|| \geq \gamma^{\frac{1}{2}}\right) &= \mathbb{P}\left(\sup_{n\in\mathbb{N}_+} \frac{e^{-\tilde{C}_n t_n}}{C_n} \sup_{t\in[0,t_n]} ||X_t||^2 \geq \gamma\right) \\ &\leq \mathbb{P}\left(\sum_{n=1}^{\infty} \sup_{t\in[0,t_n]} ||X_t||^2 \frac{e^{-\tilde{C}_n t_n}}{C_n} \geq \gamma\right) \\ &\leq \frac{1}{\gamma} \sum_{n=1}^{\infty} e^{(C_n-\tilde{C}_n)n\,\delta_-}.\end{aligned} \tag{60}$$

Define the positive constant $C_{(\delta_-,\mathbb{L})} \overset{\text{def.}}{=} \left(\sum_{n=1}^{\infty} e^{(C_n-\tilde{C}_n)n\,\delta_-}\right)^{1/2}$, and note that $C_{(\delta_-,\mathbb{L})}$ only depends on the local-Lipschitz constant $\mathbb{L}$ and on the minimal time-grid spacing $\delta_-$ of the time-grid $\mathbb{T}$. Define $\gamma \overset{\text{def.}}{=} \frac{1}{\varepsilon} \sum_{n=1}^{\infty} e^{(C_n-\tilde{C}_n)n\,\delta_-}$ where $0 < \varepsilon \leq 1$. Rearranging (60) yields

$$\begin{aligned}\mathbb{P}\left((\forall n \in \mathbb{N}_+)\, \|X_{t_n}\| \leq \frac{C_{(\delta_-,\mathbb{L})}}{\varepsilon^{1/2}} C_n^{1/2} e^{-C_n \delta_- n/2}\right) &\geq \mathbb{P}\left(\sup_{n\in\mathbb{N}_+} \frac{e^{-\tilde{C}_n t_n/2}}{C_n^{1/2}} \sup_{t\in[0,t_n]} ||X_t|| \leq \frac{C_{(\delta_-,\mathbb{L})}}{\varepsilon^{1/2}}\right) \\ &\geq 1-\varepsilon.\end{aligned} \tag{61}$$

• **Case** 1 - **Deterministic** $X_0$: If $X_0$ is deterministic, i.e. $X_0 = x \in \mathbb{R}^d$, then the concentration inequality in (61) implies that the discretized stochastic process $(X_{t_n})_{n\in\mathbb{N}}$ belongs to the compact subset $K^{\exp}_{\mathbb{C},C^\star,\varepsilon}$ of the product space $(\mathbb{R}^d)^{\mathbb{N}}$ with probability at least $(1-\varepsilon)$, where $C^\star \overset{\text{def.}}{=} 1$ and $C_0 \overset{\text{def.}}{=} \|x\|$.

• **Case** 2 - **Sub-Gaussian** $X_0$: If $X_0$ is sub-Gaussian, then there are positive constants $c_0$ and $c_1$, both depending only on the law of $X_0$, such that

$$\mathbb{P}(\|X_0\| \leq L_0) \geq 1 - c_0\, e^{c_1 L_0^2}. \tag{62}$$

By the independence of $X_0$ and $W_0$, we may combine the concentration inequalities (61) and (62) to conclude that the discretized stochastic process $(X_{t_n})_{n\in\mathbb{N}}$ belongs to the compact subset $K^{\exp}_{\mathbb{C},C^\star,\varepsilon}$ of the product space $(\mathbb{R}^d)^{\mathbb{N}}$ with probability at least $C^\star(1-\varepsilon)$, where $C^\star \overset{\text{def.}}{=} (1 - c_0\, e^{c_1 L_0^2})$ and $C_0 \overset{\text{def.}}{=} L_0$. □

Next, before embarking on the proof of our main result, namely Theorem 4.11, we take a moment to explain the intuition behind it. The proof ultimately reduces to controlling the

distance between the target causal map $F$ against an aptly chosen geometric hypertransformer model $\hat{F}$ defined on the same input and output spaces as $F$. Our proof works by decomposing an upper-bound for the distance between $F$ and $\hat{F}$ into four terms, each of which is either controlled by our choice of model $\hat{F}$ or by the regularity of the causal map $F$.

Let us briefly explain each of the four terms appearing in the proof. The first term, (FC. Approx) within the proof, is given by $F$'s approximable complexity and reduces the approximation problem of an infinitely complicated causal map to a causal map of finite complexity $F^{\rho_\varepsilon,f_\varepsilon}$ (implied by Definition 4.9). The second term, (Window) within the proof, controls the error between our candidate geometric hypertransformer $\hat{F}$ and the causal map of finite complexity $F^{\rho_\varepsilon,f_\varepsilon}$ within a prespecified finite-time horizon $[-T,T]$. The third term, (Growth F.A) within the proof, constructs the extrapolation quality function $c$ in Table 3 to control the deviation of $F^{\rho_\varepsilon,f_\varepsilon}$ outside the time window $[-T,T]$ from the values it takes within the time window $[-T,T]$. Similarly, the fourth and last term, (Growth $\hat{F}$) within the proof, controls the deviation of our candidate geometric hypertransformer model $\hat{F}$ outside, from its behavior within, the time window $[-T,T]$.

*Proof of Theorem* 4.11. **Step 1 - Decomposing Approximation Error:**

Since $\mathcal{K}\subseteq(\mathbb{R}^d)^{\mathbb{Z}}$ is compact and $F$ has approximable complexity, then there exists a function $c_{AC}:\mathbb{Z}\times(0,\infty)\to[1,\infty)$ such that, for every $\varepsilon>0$, there exist $f_\varepsilon\in C^\alpha(\mathbb{R}\times\mathbb{R}^{dm(\varepsilon)},\mathbb{R}^{L(\varepsilon)})$ and $\rho_\varepsilon\in C^\alpha(\mathbb{R}^{L(\varepsilon)},\mathcal{Y})$ as in Definition 4.9, satisfying

$$\sup_{n\in\mathbb{Z}}\sup_{\boldsymbol{x}\in\mathcal{K}}\frac{d_{\mathcal{Y}}\left(F(\boldsymbol{x})_{t_n},F^{\rho_\varepsilon,f_\varepsilon}(\boldsymbol{x})_{t_n}\right)}{c_{AC}(n,\varepsilon)}<\frac{\varepsilon}{4}. \tag{63}$$

*With an abuse of notation, we opt to write $f,\rho,m,L$ depending on $\varepsilon$ rather than $\varepsilon/4$ throughout the proof to ease the reading.*

For any other causal map $\hat{F}:\mathcal{X}^{\mathbb{Z}}\to\mathcal{Y}^{\mathbb{Z}}$ and any function $c=c_\varepsilon:\mathbb{Z}\to[1,\infty)$, by letting

$$c'(n,\varepsilon)\overset{\text{def.}}{=}\max\{c_{AC}(n,\varepsilon/4),c(n)\},$$

the following estimate holds:

$$\begin{aligned}\sup_{n\in\mathbb{Z}}\sup_{\boldsymbol{x}\in\mathcal{K}}\frac{d_{\mathcal{Y}}\left(F(\boldsymbol{x})_{t_n},\hat{F}(\boldsymbol{x})_{t_n}\right)}{c'(n,\varepsilon)}&\leq\sup_{n\in\mathbb{Z}}\sup_{\boldsymbol{x}\in\mathcal{K}}\left(\frac{d_{\mathcal{Y}}\left(F(\boldsymbol{x})_{t_n},F^{\rho_\varepsilon,f_\varepsilon}(\boldsymbol{x})_{t_n}\right)}{c'(n,\varepsilon)}+\frac{d_{\mathcal{Y}}\left(F^{\rho_\varepsilon,f_\varepsilon}(\boldsymbol{x})_{t_n},\hat{F}(\boldsymbol{x})_{t_n}\right)}{c'(n,\varepsilon)}\right)\\&\leq\frac{\varepsilon}{4}+\sup_{n\in\mathbb{Z}}\sup_{\boldsymbol{x}\in\mathcal{K}}\frac{d_{\mathcal{Y}}\left(F^{\rho_\varepsilon,f_\varepsilon}(\boldsymbol{x})_{t_n},\hat{F}(\boldsymbol{x})_{t_n}\right)}{c'(n,\varepsilon)}.\end{aligned} \tag{64}$$

Therefore,

$$\sup_{n\in\mathbb{Z}}\sup_{\boldsymbol{x}\in\mathcal{K}}\frac{d_{\mathcal{Y}}\left(F(\boldsymbol{x})_{t_n},\hat{F}(\boldsymbol{x})_{t_n}\right)}{c'(n,\varepsilon)}\leq\frac{\varepsilon}{4}+ \tag{FC. Approx}$$

$$\max_{|n|\leq N_T}\sup_{\boldsymbol{x}\in\mathcal{K}}\frac{d_{\mathcal{Y}}\left(F^{\rho_\varepsilon,f_\varepsilon}(\boldsymbol{x})_{t_n},\hat{F}(\boldsymbol{x})_{t_n}\right)}{c'(n,\varepsilon)} \tag{Window}$$

$$+\sup_{n>N_T}\sup_{\boldsymbol{x}\in\mathcal{K}}\frac{d_{\mathcal{Y}}\left(F^{\rho_\varepsilon,f_\varepsilon}(\boldsymbol{x})_{t_n},\hat{F}(\boldsymbol{x})_{t_{N_T}}\right)}{c'(n,\varepsilon)}+\sup_{n<-N_T}\sup_{\boldsymbol{x}\in\mathcal{K}}\frac{d_{\mathcal{Y}}\left(F^{\rho_\varepsilon,f_\varepsilon}(\boldsymbol{x})_{t_n},\hat{F}(\boldsymbol{x})_{t_{-N_T}}\right)}{c'(n,\varepsilon)} \tag{Growth F.A}$$

$$+\sup_{n>N_T}\sup_{\boldsymbol{x}\in\mathcal{K}}\frac{d_{\mathcal{Y}}\left(\hat{F}(\boldsymbol{x})_{t_n},\hat{F}(\boldsymbol{x})_{t_{N_T}}\right)}{c'(n,\varepsilon)}+\sup_{n<-N_T}\sup_{\boldsymbol{x}\in\mathcal{K}}\frac{d_{\mathcal{Y}}\left(\hat{F}(\boldsymbol{x})_{t_n},\hat{F}(\boldsymbol{x})_{t_{-N_T}}\right)}{c'(n,\varepsilon)},\quad (Growth\hat{F})$$

for $N_T$ defined in (11). The remainder of the proof is devoted to controlling terms (Window), (Growth F.A) and (Growth $\hat{F}$) for a particular systems $\hat{F}$, implemented by a recurrent probabilistic transformer model that will be specified later. We start by noticing that, by the triangle inequality,

$$\sup_{n>N_T}\sup_{\boldsymbol{x}\in\mathcal{K}}\frac{d_{\mathcal{Y}}\left(F^{\rho_\varepsilon,f_\varepsilon}(\boldsymbol{x})_{t_n},\hat{F}(\boldsymbol{x})_{t_{N_T}}\right)}{c'(n,\varepsilon)}\leq\sup_{\boldsymbol{x}\in\mathcal{K}}\frac{d_{\mathcal{Y}}\left(F^{\rho_\varepsilon,f_\varepsilon}(\boldsymbol{x})_{t_{N_T}},\hat{F}(\boldsymbol{x})_{t_{N_T}}\right)}{c'(n,\varepsilon)}\tag{65}$$

$$+\sup_{n>N_T}\sup_{\boldsymbol{x}\in\mathcal{K}}\frac{d_{\mathcal{Y}}\left(F^{\rho_\varepsilon,f_\varepsilon}(\boldsymbol{x})_{t_n},F^{\rho_\varepsilon,f_\varepsilon}(\boldsymbol{x})_{t_{N_T}}\right)}{c'(n,\varepsilon)}.\quad \text{(Growth F.B)}$$

And mutatis mundanes for the case where $n<-N_T$. Since the term on the RHS of (65) is bounded by (Window), then the control of the latter (see Step 2) together with the control of (Growth F.B), will imply the control of (Growth F.A) (see Step 4). For the particular system $\hat{F}$ specified in Step 2, we will control the terms (Window) and (Growth F.B) each by $\varepsilon/4$, and show that (Growth $\hat{F}$) vanishes. This will conclude the proof.

From this is clear that, in the statement of Theorem 4.11, we could separately specify the "AC approximation error" (dominating (FC. Approx)), the "approximation error within the time window $[-T,T]$" (dominating (Window)), the "extrapolation error" (dominating (Growth F.A)), and the "growth error" (dominating (Growth $\hat{F}$)), so that the total error is bounded by their sum.

**Step 2 - Construction of Model and Control of the finite-time horizon term** (Window):

We first control the term in (Window). For integers $m,n\in\mathbb{Z}$ with $m<n$ and $\boldsymbol{x}\in\mathcal{K}$ we use the shorthand notation $x_{t_m:t_n}$ to refer to the vector $(x_{t_m},x_{t_{m+1}},\dots,x_{t_n})$. For each $n\in\mathbb{Z}$ with $|n|\leq N_T$, the continuity of the projection maps $\mathrm{pj}_n:(\mathbb{R}^d)^{\mathbb{Z}}\ni(x_{t_n})_{n\in\mathbb{Z}}\mapsto x_{t_n}\in\mathbb{R}^d$ implies that, for every $\tilde{n}\leq n\in\mathbb{Z}$, the subset $K_{t_{\tilde{n}}:t_n}\overset{\text{def.}}{=}\prod_{i=\tilde{n}}^{n}\mathrm{pj}_i(\mathcal{K})$ is non-empty and compact in $\mathbb{R}^d$. Therefore, for

$$\varepsilon_{A,1}\overset{\text{def.}}{=}\min_{n=-N_T,\dots,N_T}\frac{1}{C_{K_{t_n-dm(\varepsilon):t_n}}}\left(\frac{\varepsilon}{8L_{\alpha,\rho_\varepsilon}}\right)^{1/\alpha},$$

where $C_{K_{t_n-dm(\varepsilon/4):t_n}}$ is defined as in (6), and for every $n\in\mathbb{Z}$ with $|n|\leq N_T$, Proposition 3.10 implies that there exist a multi-index $[d^{(n)}]$ and a neural network $\hat{f}_{\theta_n}:\mathbb{R}^{d\,m(\varepsilon)}\to\mathbb{R}^{L(\varepsilon)}$ satisfying

$$\sup_{\boldsymbol{x}\in K_{t_n-dm(\varepsilon):t_n}}\left\|f_\varepsilon(t_n,x_{t_n})-\hat{f}_{\theta_n}(x_{t_n})\right\|\leq C_{K_{t_n-dm(\varepsilon):t_n}}\varepsilon_{A,1}\leq\left(\frac{\varepsilon}{8L_{\alpha,\rho_{\varepsilon/4}}}\right)^{1/\alpha},\tag{66}$$

where the complexity of $\hat{f}$ is recorded in Table 2. Further, we may assume that $\{\theta_n\}_{-T\leq t_n\leq T}$ are all distinct. Indeed, let us first consider the case of a singular trainable activation function as in Definition 2.2. With the notation of (1), by setting $\theta'_n\overset{\text{def.}}{=}(\theta_n,(I_{L(\varepsilon)},b'_n,\alpha'))$, with $b'_n\in\mathbb{R}^{L(\varepsilon)}$ such that $\|b'_n\|<(\frac{\varepsilon}{8L_{\alpha,\rho_\varepsilon/4}})^{1/\alpha}$ and $b'_{-N_T},\dots,b'_{N_T}$ all distinct, and with $\alpha'_i=(0,1)$ for each $i=1,\dots,L(\varepsilon)$, then the $\{\theta'_n\}_{-T\leq t_n\leq T}$ are also all distinct and the estimate in (66) holds up to a factor of 2. In other words, we add an additional layer of depth which approximates the identity on $\mathbb{R}^{L(\varepsilon)}$, whilst

making all $\theta_n$'s distinct for $|n| \leq N_T$. In the case of $\sigma$ as in Definition 2.3 or as in Definition 2.4, by the continuity of $\sigma$, we may pick $\theta_n'$ arbitrarily close to $\theta_n$ such that the $\theta_{-N_T}, \dots, \theta_{N_T}$ are all distinct.

Since $\sigma_{(1,1)}(x) = x$ for all $x \in \mathbb{R}$, then, without loss of generality, we may assume that $d_{J_{(n)}} = d_{J_{(m)}}$ for all $n, m = -N_T, \dots, N_T$, by adding identity layers in (2). Similarly, by adding 0 rows to the matrices $A^{(j)}$ and the vectors $b^{(j)}$ in (2), we may without loss of generality assume that $d_j^{(n)} = d_j^{(0)}$ for all $j = 1, \dots, J_{(n)}$ and $n = -N_T, \dots, N_T$. Consequently, $P([d^{(n)}]) = P([d^{(0)}])$ for all $-N_T, \dots, N_T$.

Now, let $\mathcal{K} \subseteq \mathbb{R}^{L(\varepsilon)}$ be defined by

$$\mathcal{K} \overset{\text{def.}}{=} \bigcup_{n=-N_T}^{N_T} \left\{ z \in \mathbb{R}^{L(\varepsilon)} \,:\, \|z - f(t_n, K_{t_n - dm(\varepsilon):t_n})\| \leq \varepsilon_{A,1} \right\}.$$

Since each $f(t_n, \cdot)$ is continuous and each $K_{t-dm(\varepsilon):t}$ is compact, then, for each $n = -N_T, \dots, N_T$, the sets $\{z \in \mathbb{R}^{L(\varepsilon)} \,:\, \|z - f(t_n, K_{t-dm(\varepsilon):t})\| \leq \varepsilon_{A,1}\}$ are compact. Moreover, as the union of finitely many compacts is again compact, then $\mathcal{K} \subseteq \mathbb{R}^{L(\varepsilon)}$ is compact as well. Therefore, we may apply Theorem 3.8 to conclude the existence of a network $\hat{\rho} : \mathcal{K} \to \mathcal{P}_1(\mathbb{R}^m)$ with representation

$$\hat{\rho}(\cdot) = \text{attention}_{N,q}(\hat{r}_\theta(\cdot), Y),$$

where $Y$ is an $N \times q \times m$-array, $\hat{r} \in \mathcal{NN}^{\cdot}_{[d]}$ maps from $\mathbb{R}^d$ to $\mathbb{R}^N$, and estimates for the number of parameters defining both are given in Table 1, and such that $\hat{\rho}$ satisfies the estimate

$$\sup_{\lambda \in \mathcal{K}} d_{\mathcal{Y}}(\rho_\varepsilon(\lambda), \hat{\rho}(\lambda)) \leq \frac{\varepsilon}{8}. \tag{67}$$

Thus, together with (66), (67) and the monotonicity of the modulus of continuity $\omega_{\rho_\varepsilon}$, we obtain the following estimate for every $n \in \mathbb{N}_+$ with $|n| \leq N_T$ and any $\boldsymbol{x} \in K_{t_n - dm(\varepsilon):t_n}$:

$$\begin{aligned}
&d_{\mathcal{Y}}\left(\rho_\varepsilon \circ f_\varepsilon(t_n, x_{t_n-dm(\varepsilon):t_n}), \hat{\rho} \circ \hat{f}_{\theta_n}(x_{t_n-dm(\varepsilon):t_n})\right) \\
&\leq d_{\mathcal{Y}}\left(\rho_\varepsilon \circ f_\varepsilon(t_n, x_{t_n-dm(\varepsilon):t_n}), \rho_\varepsilon \circ \hat{f}_{\theta_n}(x_{t_n-dm(\varepsilon):t_n})\right) + d_{\mathcal{Y}}\left(\rho_\varepsilon \circ \hat{f}_{\theta_n}(x_{t_n-dm(\varepsilon):t_n}), \hat{\rho} \circ \hat{f}_{\theta_n}(x_{t_n-dm(\varepsilon):t_n})\right) \\
&\leq \omega_{\rho_\varepsilon}\left(\left\| f_\varepsilon(t_n, x_{t_n-dm(\varepsilon):t_n}) - \hat{f}_{\theta_n}(x_{t_n-dm(\varepsilon):t_n}) \right\|\right) + d_{\mathcal{Y}}\left(\rho_\varepsilon \circ \hat{f}_{\theta_n}(x_{t_n-dm(\varepsilon):t_n}), \hat{\rho} \circ \hat{f}_{\theta_n}(x_{t_n-dm(\varepsilon):t_n})\right) \\
&\leq \omega_{\rho_\varepsilon}\left(\left(\frac{\varepsilon}{8L_{\alpha,\rho_\varepsilon/4}}\right)^{1/\alpha}\right) + d_{\mathcal{Y}}\left(\rho_\varepsilon \circ \hat{f}_{\theta_n}(x_{t_n-dm(\varepsilon):t_n}), \hat{\rho} \circ \hat{f}_{\theta_n}(x_{t_n-dm(\varepsilon):t_n})\right) \\
&\leq \frac{\varepsilon}{8} + \frac{\varepsilon}{8} = \frac{\varepsilon}{4}.
\end{aligned} \tag{68}$$

Next, we build the hypernetwork $h$ which implements our transformer network's recursive structure. Let $P \overset{\text{def.}}{=} P([d^{(0)}])$ and, for each $n = -N_T, \dots, N_T$, define

$$\tilde{\theta}_n \overset{\text{def.}}{=} a\,\theta_n - b, \tag{69}$$

where $a > 0$ and $b \in \mathbb{R}^P$ are such that $\tilde{\theta}_n \in [0,1]^P$ for each $n = -N_T, \dots, N_T$, and where $\{\theta_n\}_{n=-N_T}^{N_T}$ are as in (66). We may therefore apply (Yun et al., 2019, Theorem 3.1 (ii)) to conclude that there is a feedforward neural network $\tilde{h} \in \mathcal{NN}^{\text{ReLU}}_{P,M,M,P}$ such that, for each $|n| \leq N_T$,

$$\tilde{h}(\tilde{\theta}_n) = \tilde{\theta}_{n+1},$$

and where $M$ is the smallest positive integer satisfying

$$2\left\lfloor \frac{M}{2} \right\rfloor \left\lfloor \frac{M}{4P} \right\rfloor \geq N_T.$$

Define $h(z) \overset{\text{def.}}{=} a^{-1}(\tilde{h}(az-b)+b)$. Then, we extend $\{\theta_n\}_{n=-N_T}^{N_T}$ to an infinite sequence $(\theta_n)_{n\in\mathbb{Z}}$ by

$$\theta_n \overset{\text{def.}}{=} \theta_{-N_T} \text{ for } n < -N_T \quad \text{and} \quad \theta_n \overset{\text{def.}}{=} \theta_{N_T} \text{ for } n > N_T.$$

We henceforth set $\hat{F} \overset{\text{def.}}{=} F^{(\hat{\rho},h,\theta_{-N_T},N_T)}$. By construction, then, (68) and the fact that $c(n)=1$ for $|n| \leq N_T$ imply that

$$\max_{|n|\leq N_T} \sup_{\boldsymbol{x}\in\mathcal{K}} \frac{d_{\mathcal{Y}}\left(F^{\rho_\varepsilon,f_\varepsilon}(\boldsymbol{x})_{t_n}, \hat{F}(\boldsymbol{x})_{t_n}\right)}{c'(n,\varepsilon)} \leq \max_{|n|\leq N_T} \sup_{\boldsymbol{x}\in\mathcal{K}} d_{\mathcal{Y}}\left(F^{\rho_\varepsilon,f_\varepsilon}(\boldsymbol{x})_{t_n}, \hat{F}(\boldsymbol{x})_{t_n}\right) \leq \frac{\varepsilon}{4}.$$

**Step 3 - Control of the RGT decay term (Growth $\hat{F}$):**

Observe that $H(t_n,\theta_n)=\theta_{-N_T}$ for all $n \leq -N_T$, and $H(t_n,\theta_n)=\theta_{N_T}$ for all $n \geq N_T$. Moreover, since we have defined $\hat{F}$, we may define the map $c^{\hat{F}}_{\mathcal{K},N_T,8/\varepsilon}(n)$ as in (10). Therefore, as long as $c:\mathbb{Z}\to[1,\infty)$ satisfies the following property for every $|n|>N_T$:

$$c(n) \geq c^{\hat{F}}_{\mathcal{K},N_T,8/\varepsilon}(n), \tag{70}$$

then we can control term (Growth $\hat{F}$). This is because, writing out $c^{\hat{F}}_{\mathcal{K},N_T,8/\varepsilon}$ we obtain

$$c^{\hat{F}}_{\mathcal{K},N_T,8/\varepsilon}(n) = \max\left\{1, \sup_{x\in\mathcal{K}} \frac{8}{\varepsilon} d_{\mathcal{Y}}\left(\hat{F}(x)_{t_n}, \hat{F}(x)_{t_{N_T}}\right) I_{\{n\geq N_T\}} + \frac{8}{\varepsilon} d_{\mathcal{Y}}\left(\hat{F}(x)_{t_n}, \hat{F}(x)_{t_{-N_T}}\right) I_{\{n\leq -N_T\}}\right\}.$$

Thus, we may now control term (Growth $\hat{F}$) by

$$\begin{aligned}
&\sup_{n>N_T}\sup_{\boldsymbol{x}\in\mathcal{K}} \frac{d_{\mathcal{Y}}\left(\hat{F}(\boldsymbol{x})_{t_n}, \hat{F}(\boldsymbol{x})_{t_{N_T}}\right)}{c'(n,\varepsilon)} + \sup_{n<-N_T}\sup_{\boldsymbol{x}\in\mathcal{K}} \frac{d_{\mathcal{Y}}\left(\hat{F}(\boldsymbol{x})_{t_n}, \hat{F}(\boldsymbol{x})_{t_{-N_T}}\right)}{c'(n,\varepsilon)} \\
\leq &\sup_{n>N_T}\sup_{\boldsymbol{x}\in\mathcal{K}} \frac{d_{\mathcal{Y}}\left(\hat{F}(\boldsymbol{x})_{t_n}, \hat{F}(\boldsymbol{x})_{t_{N_T}}\right)}{c^{\hat{F}}_{\mathcal{K},N_T,8/\varepsilon}(n)} + \sup_{n<-N_T}\sup_{\boldsymbol{x}\in\mathcal{K}} \frac{d_{\mathcal{Y}}\left(\hat{F}(\boldsymbol{x})_{t_n}, \hat{F}(\boldsymbol{x})_{t_{-N_T}}\right)}{c^{\hat{F}}_{\mathcal{K},N_T,8/\varepsilon}(n)} \\
= &\frac{\varepsilon}{8}\sup_{n>N_T}\sup_{\boldsymbol{x}\in\mathcal{K}} \frac{d_{\mathcal{Y}}\left(\hat{F}(\boldsymbol{x})_{t_n}, \hat{F}(\boldsymbol{x})_{t_{N_T}}\right)}{\max\left\{1, d_{\mathcal{Y}}\left(\hat{F}(\boldsymbol{x})_{t_n}, \hat{F}(\boldsymbol{x})_{t_{N_T}}\right)\right\}} + \frac{\varepsilon}{8}\sup_{n<-N_T}\sup_{\boldsymbol{x}\in\mathcal{K}} \frac{d_{\mathcal{Y}}\left(\hat{F}(\boldsymbol{x})_{t_n}, \hat{F}(\boldsymbol{x})_{t_{-N_T}}\right)}{\max\left\{1, d_{\mathcal{Y}}\left(\hat{F}(\boldsymbol{x})_{t_n}, \hat{F}(\boldsymbol{x})_{t_{-N_T}}\right)\right\}} \\
\leq &\frac{\varepsilon}{4}.
\end{aligned} \tag{71}$$

**Step 4 - Control of the infinite-time horizon term** (Growth F.A):

It remains to control the terms in (Growth F.A) by controlling (Growth F.B). This part is divided in different cases, as depending on the particular form of $\mathcal{K}$ we get a different compression rate

$c$. For notational simplicity, we may now define the compression rate $c$ by

$$c(n) \overset{\text{def.}}{=} \max\left\{ c^{\hat{F}}_{\mathscr{K},N_T,8/\varepsilon}, \tilde{c}(n) \right\},$$

for $n \in \mathbb{Z}$, where $\tilde{c}$ is defined on a case-by-case basis in the following Cases (1-5).

*Case* 1: Let $w : \mathbb{Z} \to [0,\infty)$ be a weighting function, $K \subseteq \mathbb{R}^d$, and consider the case where $\mathscr{K}$ is

$$\mathscr{K} = K^w \overset{\text{def.}}{=} \left\{ \boldsymbol{x} \in \mathscr{X}^{\mathbb{Z}} \,:\, \exists y \in K \text{ s.t. } \|x_{t_i} - y\| \le w(|i|) \right\}.$$

Since $F$ is an AC map, $f_\varepsilon$ and $\rho_\varepsilon$ are uniformly continuous with respective moduli of continuity $\omega_{f_\varepsilon}(u) = L_{\alpha,f_\varepsilon}|u|^\alpha$ and $\omega_{\rho_\varepsilon}(u) = L_{\alpha,\rho_\varepsilon}|u|^\alpha$ for some $L_{\alpha,f_\varepsilon}, L_{\alpha,\rho_\varepsilon} \ge 0$ and some $\alpha \in (0,1]$. Thus, for every $n, n' \in \mathbb{Z}$ with $n < n'$ and every $\boldsymbol{x} \in \mathscr{K}$, we compute the estimate

$$\begin{aligned} d_{\mathscr{Y}}\left(F^{\rho_\varepsilon,f_\varepsilon}(\boldsymbol{x})_{t_n}, F^{\rho_\varepsilon,f_\varepsilon}(\boldsymbol{x})_{t_{n'}}\right) &= d_{\mathscr{Y}}\left(\rho_\varepsilon \circ f_\varepsilon(t_n, x_{t_{n-dm(\varepsilon)}:t_n}), \rho_\varepsilon \circ f_\varepsilon(t_{n'}, x_{t_{n'-dm(\varepsilon)}:t_{n'}})\right) \\ &\le \omega_{\rho_\varepsilon} \circ \omega_{f_\varepsilon}\left(|t_n - t_{n'}| + \|x_{t_{n-dm(\varepsilon)}:t_n} - x_{t_{n'-dm(\varepsilon)}:t_{n'}}\|\right) \\ &\le \omega_{\rho_\varepsilon} \circ \omega_{f_\varepsilon}\left(|n-n'|\delta_+ + dm(\varepsilon) \max_{i=1,\dots,dm(\varepsilon)} \|[x_{t_{n-dm(\varepsilon)}:t_n}]_i - [x_{t_{n'-dm(\varepsilon)}:t_{n'}}]_i\|\right), \end{aligned} \tag{72}$$

where $[\cdot]_i$ denotes the $i^{th}$ canonical projection onto the Cartesian product $(\mathbb{R}^d)^{m(\varepsilon)} \to \mathbb{R}^d$. Since each $[x_{t_{n-dm(\varepsilon)}:t_n}]_i$ (resp. $[x_{t_{n'-dm(\varepsilon)}:t_{n'}}]_i$) belongs to $K^w$, there is some $y_{i,n}$ (resp. $y_{i,n'}$) in $K$ satisfying the bound $\|[x_{t_{n-dm(\varepsilon)}:t_n}]_i - y_{i,n}\| \le w(|n|)$ (resp. $\|[x_{t_{n'-dm(\varepsilon)}:t_{n'}}]_i - y_{i,n'}\| \le w(|n'|)$). Therefore, the right-hand side of (72) can be bounded from above as

$$\begin{aligned} d_{\mathscr{Y}}\left(F^{\rho_\varepsilon,f_\varepsilon}(\boldsymbol{x})_{t_n}, F^{\rho_\varepsilon,f_\varepsilon}(\boldsymbol{x})_{t_{n'}}\right) &\le \omega_{\rho_\varepsilon} \circ \omega_{f_\varepsilon}\Bigg(|n-n'|\delta_+ + dm(\varepsilon) \max_{i=1,\dots,dm(\varepsilon)} \Big(\|[x_{t_{n-dm(\varepsilon)}:t_n}]_i - y_{i,n}\| \\ &\quad + \|y_{i,n} - y_{i,n'}\| + \|[x_{t_{n'-dm(\varepsilon)}:t_{n'}}]_i - y_{i,n'}\|\Big)\Bigg) \\ &\le \omega_{\rho_\varepsilon} \circ \omega_{f_\varepsilon}\left(|n-n'|\delta_+ + dm(\varepsilon)\left(w(|n|) + \operatorname{diam}(K) + w(|n'|)\right)\right). \end{aligned} \tag{73}$$

Define

$$\tilde{c} : \mathbb{Z} \ni n \mapsto \frac{4}{\varepsilon} \omega_{\rho_\varepsilon} \circ \omega_{f_\varepsilon}\left((|n| - N_T)\delta_+ + dm(\varepsilon)\left(w(|n|) + \operatorname{diam}(K) + w(N_T)\right)\right).$$

Then (73) implies that

$$\sup_{|n|>N_T} \sup_{\boldsymbol{x}\in\mathscr{K}} \frac{d_{\mathscr{Y}}\left(F^{\rho_\varepsilon,f_\varepsilon}(\boldsymbol{x})_{t_n}, F^{\rho_\varepsilon,f_\varepsilon}(\boldsymbol{x})_{t_{N_T}}\right)}{c'(n,\varepsilon)} \le \frac{\varepsilon}{4}. \tag{74}$$

Thus, in this case, term (Growth $\hat{F}$) is controlled by $\varepsilon/4$.

*Case* 2: Let $K$ be a compact subset of $\mathbb{R}^d$, $C, p > 0$, and consider the case where

$$\mathscr{K} = K^\infty_{C,p} = \left\{ \boldsymbol{x} \in (\mathbb{R}^d)^{\mathbb{Z}} \,:\, x_0 \in K,\ (\forall n \in \mathbb{Z})\ \|\Delta_n \boldsymbol{x}\|^p \le C|\Delta t_n| \right\}.$$

If $\boldsymbol{x} \in \mathscr{K}$, then for every $n, n' \in \mathbb{Z}$ with $n < n'$ we have that

$$\begin{aligned}\|x_{t_n} - x_{t_{n'}}\| &\leq \sum_{k=n-1}^{n'} \|\Delta_k \boldsymbol{x}\| \\ &\leq C^{\frac{1}{p}} \sum_{k=n}^{n'-1} |\Delta t_k|^{\frac{1}{p}} \\ &\leq (n'-n) C^{\frac{1}{p}} \delta_+^{\frac{1}{p}} < \infty, \end{aligned} \tag{75}$$

where we have applied Assumption 4.1 to derive the last inequality.

Since $F$ is an AC map then, $f_\varepsilon$ and $\rho_\varepsilon$ are uniformly continuous, with respective moduli of continuity $\omega_{f_\varepsilon}(u) = L_{\alpha, f_\varepsilon} |u|^\alpha$ and $\omega_{\rho_\varepsilon}(u) = L_{\alpha, \rho_\varepsilon} |u|^\alpha$ for some $L_{\alpha, f_\varepsilon}, L_{\alpha, \rho_\varepsilon} \geq 0$ and some $\alpha \in (0, 1]$. Thus, for every $n, n' \in \mathbb{Z}$ with $n < n'$ and every $\boldsymbol{x} \in \mathscr{K}$, we estimate

$$\begin{aligned} d_{\mathscr{Y}}\left(F^{\rho_\varepsilon, f_\varepsilon}(\boldsymbol{x})_{t_n}, F^{\rho_\varepsilon, f_\varepsilon}(\boldsymbol{x})_{t_{n'}}\right) &= d_{\mathscr{Y}}\left(\rho_\varepsilon \circ f_\varepsilon(t_n, x_{t_{n-dm(\varepsilon)}:t_n}), \rho_\varepsilon \circ f_\varepsilon(t_{n'}, x_{t_{n'-dm(\varepsilon)}:t_{n'}})\right) \\ &\leq \omega_{\rho_\varepsilon} \circ \omega_{f_\varepsilon}\left(|t_{n'} - t_n| + \|x_{t_{n-dm(\varepsilon)}:t_n} - x_{t_{n'-dm(\varepsilon)}:t_{n'}}\|\right) \\ &\leq \omega_{\rho_\varepsilon} \circ \omega_{f_\varepsilon}\left(|t_{n'} - t_n| + \sum_{i=n-dm(\varepsilon)}^{n} \|x_{t_i} - x_{t_{i+n'-n}}\|\right) \end{aligned} \tag{76}$$

$$\leq \omega_{\rho_\varepsilon} \circ \omega_{f_\varepsilon}\left((n'-n)\delta_+ + (dm(\varepsilon)+1)(n'-n) C^{\frac{1}{p}} \delta_+^{\frac{1}{p}}\right) \tag{77}$$

$$= \omega_{\rho_\varepsilon} \circ \omega_{f_\varepsilon}\left(|n'-n|\delta_+ \left[1 + (dm(\varepsilon)+1) C^{\frac{1}{p}} \delta_+^{\frac{1-p}{p}}\right]\right), \tag{78}$$

where we used the monotonicity of the moduli of continuity $\omega_{f_\varepsilon}$ and $\omega_{\rho_\varepsilon}$ in (76), and the estimate in (75) to get (77). Now define

$$\tilde{c} : \mathbb{Z} \ni n \mapsto \frac{4}{\varepsilon} \omega_{\rho_\varepsilon} \circ \omega_{f_\varepsilon}\left((|n| - N_T)\delta_+ [1 + (dm(\varepsilon)+1) C^{\frac{1}{p}} \delta_+^{\frac{1-p}{p}}]\right).$$

Then (78) implies that

$$\sup_{|n| > N_T} \sup_{\boldsymbol{x} \in \mathscr{K}} \frac{d_{\mathscr{Y}}\left(F^{\rho_\varepsilon, f_\varepsilon}(\boldsymbol{x})_{t_n}, F^{\rho_\varepsilon, f_\varepsilon}(\boldsymbol{x})_{t_{N_T}}\right)}{c'(n, \varepsilon)} \leq \frac{\varepsilon}{4}. \tag{79}$$

Thus, in this case, term (Growth $\hat{F}$) is controlled by $\varepsilon/4$.

*Case* 3: Let $K$ be a compact subset of $\mathbb{R}^d$, $C > 0$, $p \geq 1$, and $\alpha < 1 - p$. Consider the case where

$$\mathscr{K} = K^\alpha_{C,p} = \left\{ \boldsymbol{x} \in (\mathbb{R}^d)^{\mathbb{Z}} : x_0 \in K, \sum_{n \in \mathbb{Z}} \frac{\|\Delta_n \boldsymbol{x}\|^p}{|\Delta t_n| |n|_{++}^{\alpha}} \leq C \right\}.$$

For $\boldsymbol{x} \in \mathscr{K}$, then, for every $n, n' \in \mathbb{Z}$ with $n < n'$, we have that

$$\begin{aligned}
\|x_{t_n} - x_{t_{n'}}\| &\leq \sum_{k=n-1}^{n'} \|\Delta_k \boldsymbol{x}\| \\
&= \sum_{k=n-1}^{n'} \frac{\|\Delta_k \boldsymbol{x}\|}{|\Delta t_k|^{\frac{1}{p}} |k|_{++}^{\frac{\alpha}{p}}} |\Delta t_k|^{\frac{1}{p}} |k|_{++}^{\frac{\alpha}{p}} \\
&\leq \left( \sum_{k=n-1}^{n'} \left( \frac{\|\Delta_k \boldsymbol{x}\|}{|\Delta t_k|^{\frac{1}{p}} |k|_{++}^{\frac{\alpha}{p}}} \right)^p \right)^{\frac{1}{p}} \left( \sum_{k=n-1}^{n'} \left( |\Delta t_k|^{\frac{1}{p}} |k|_{++}^{\frac{\alpha}{p}} \right)^{\frac{p}{p-1}} \right)^{\frac{p-1}{p}} \\
&= \left( \sum_{k=n-1}^{n'} \frac{\|\Delta_k \boldsymbol{x}\|^p}{|\Delta t_k| |k|_{++}^{\alpha}} \right)^{\frac{1}{p}} \left( \sum_{k=n-1}^{n'} |\Delta t_k|^{\frac{1}{p-1}} |k|_{++}^{\frac{\alpha}{p-1}} \right)^{\frac{p-1}{p}} \\
&\leq C^{\frac{1}{p}} \left( \sum_{k=n-1}^{n'} |\Delta t_k|^{\frac{1}{p-1}} |k|_{++}^{\frac{\alpha}{p-1}} \right)^{\frac{p-1}{p}} \\
&\leq C^{\frac{1}{p}} \delta_+^{\frac{1}{p}} \left( \sum_{k=n-1}^{n'} |k|_{++}^{\frac{\alpha}{p-1}} \right)^{\frac{p-1}{p}}. \qquad (80)
\end{aligned}$$

Since $\frac{\alpha}{p-1} < -1$, then $\sum_{z \in \mathbb{Z}} |z|_{++}^{\frac{\alpha}{p-1}}$ is a p-series and it converges to $1 + 2\zeta(\frac{\alpha}{p-1}) < \infty$, where $\zeta$ denotes the Riemann zeta-function.[8] Therefore, the right-hand side of (80) implies the estimate

$$\|x_{t_n} - x_{t_{n'}}\| \leq C^{\frac{1}{p}} \delta_+^{\frac{1}{p}} (1 + 2\zeta\alpha/(p-1))^{\frac{p-1}{p}}. \qquad (81)$$

Analogously to the previous case, we have that for every $n, n' \in \mathbb{Z}$ and every $\boldsymbol{x} \in \mathscr{K}$, (81) implies the estimate

$$d_{\mathscr{Y}}\left(F^{\rho_\varepsilon, f_\varepsilon}(\boldsymbol{x})_{t_n}, F^{\rho_\varepsilon, f_\varepsilon}(\boldsymbol{x})_{t_{n'}}\right) \leq \omega_{\rho_\varepsilon} \circ \omega_{f_\varepsilon}\left( |n' - n|\delta_+ + (dm(\varepsilon) + 1) C^{\frac{1}{p}} \delta_+^{\frac{1}{p}} (1 + 2\zeta\alpha/(p-1))^{\frac{p-1}{p}} \right). \qquad (82)$$

By defining the extrapolation function as

$$\tilde{c} : \mathbb{Z} \ni n \mapsto \frac{4}{\varepsilon} \omega_{\rho_\varepsilon} \circ \omega_{f_\varepsilon}\left( (|n| - N_T)\delta_+ + (dm(\varepsilon) + 1) C^{\frac{1}{p}} \delta_+^{\frac{1}{p}} (1 + 2\zeta\alpha/(p-1))^{\frac{p-1}{p}} \right),$$

we obtain the same estimate as in (79).

*Case* 4: Let $K$ be a compact subset of $\mathbb{R}^d$, and consider the case where $\mathscr{K} = K^{\mathbb{Z}}$. For $\boldsymbol{x} \in \mathscr{K}$ then, for every $n, n' \in \mathbb{Z}$, we have that

$$d_{\mathscr{Y}}\left(F^{\rho_\varepsilon, f_\varepsilon}(\boldsymbol{x})_{t_n}, F^{\rho_\varepsilon, f_\varepsilon}(\boldsymbol{x})_{t_{n'}}\right) \leq \omega_{\rho_\varepsilon} \circ \omega_{f_\varepsilon}\left( |t_n - t_{n'}| + \|x_{t_{n-dm(\varepsilon)}:t_n} - x_{t_{n'-dm(\varepsilon)}:t_{n'}}\| \right) \qquad (83)$$

$$\leq \omega_{\rho_\varepsilon} \circ \omega_{f_\varepsilon}\big(|n-n'|\delta_+ + (dm(\varepsilon)+1)\mathrm{diam}(K)\big). \tag{84}$$

Therefore, we may define the extrapolation function by

$$\tilde{c} : \mathbb{Z} \ni n \mapsto \frac{4}{\varepsilon}\omega_{\rho_\varepsilon} \circ \omega_{f_\varepsilon}((|n| - N_T)\delta_+ + (dm(\varepsilon)+1)\mathrm{diam}(K)),$$

and obtain the same estimate as in (79).

*Case* 5: Let $\mathcal{K} \subseteq (\mathbb{R}^d)^{\mathbb{Z}}$ be an arbitrary compact set. In this case we define the extrapolation function by

$$\tilde{c} : \mathbb{Z} \ni n \mapsto \max_{\boldsymbol{x}\in\mathcal{K},\, k\leq|n|} \frac{4}{\varepsilon} \max\left\{1, d_{\mathcal{Y}}\left(F^{\rho_\varepsilon,f_\varepsilon}(\boldsymbol{x})_{t_k}, F^{\rho_\varepsilon,f_\varepsilon}(\boldsymbol{x})_{t_n}\right)\right\},$$

and obtain the same estimate as in (79). □

*Proof of Corollary* 4.14. The proof is exactly as in Theorem 4.11 without the term in time in (84):

$$d_{\mathcal{Y}}\left(F^{\rho_\varepsilon,f_\varepsilon}(\boldsymbol{x})_{t_n}, F^{\rho_\varepsilon,f_\varepsilon}(\boldsymbol{x})_{t_{n'}}\right) \leq \omega_{\rho_\varepsilon} \circ \omega_{f_\varepsilon}((dm(\varepsilon)+1)\mathrm{diam}(K)).$$

□

## ACKNOWLEDGMENTS

This research was supported by the ETH Zürich Foundation and the European Research Council (ERC) Starting Grant 852821-SWING. The authors would like to thank Juan-Pablo Ortega for his helpful discussion surrounding the fading memory property, as well as Behnoosh Zamanlooy, Giulia Livieri, and Ivan Dokmanić for their valuable feedback. The authors would also like to thank Valentin Debarnot for his very helpful model-visualization advice.

## DATA AVAILABILITY STATEMENT

No data sets were used in this study.

## ORCID

*Beatrice Acciaio* https://orcid.org/0000-0002-8138-2578
*Anastasis Kratsios* https://orcid.org/0000-0001-6791-3371
*Gudmund Pammer* https://orcid.org/0000-0003-2494-8739

## ENDNOTES

[1] Here $\mathbb{N}$ denotes the non-negative integers and $\mathbb{N}_+$ denotes the positive integers.

[2] Some authors, for example, Bartlett et al. (2019), refer to $J$ as the number of *hidden layers* in $\hat{f}$.

[3] By (Heinonen, 2001, Theorem 12.1) every subset of $\mathbb{R}^d$ has the doubling property (see Section 7 for the definition and (Heinonen, 2001, Section 10.13) for details), and following the discussion on (Bruè et al., 2021a, p. 3) a metric space is doubling if and only if its metric capacity is finite for all $\delta \in (0,1]$.

[4] For the relationship between probabilistic attention and the (classical) attention mechanism of Vaswani et al. (2017) see the discussion following (Kratsios et al., 2022, Equation 4).

[5] By absolute constant, we mean that $c$ is independent of $W$, $\varepsilon$, $n$, $m$, $\alpha$, and of $K$.

[6] A Schauder basis for a Fréchet space $(\mathcal{Y}, d_{\mathcal{Y}})$ is a linearly independent set $(\beta_s)_{s=0}^{q^*-1} \subseteq \mathcal{Y}$ for which, given any $y \in \mathcal{Y}$ there exists a unique sequence $(z_s^y)_{s=0}^{q^*-1}$ in $\mathbb{R}$ satisfying $\lim_{q\uparrow q^*} d_{\mathcal{Y}}(y, \sum_{s=0}^{q-1} z_s^y \beta_s) = 0$.

[7] For $x, y \in \mathbb{R}^n$ we we denote by $[x,y]$ the hypercube defined by $\prod_{i=1}^n [x_i, y_i]$.

[8] The Riemann zeta-function is defined by $\zeta : \mathbb{R} \ni s \mapsto \sum_{n=1}^{\infty} n^{-s} \in [-\infty, \infty]$.

## REFERENCES


Abbas, A., Sutter, D., Zoufal, C., Lucchi, A., Figalli, A., & Woerner, S. (2021). The power of quantum neural networks. *Nature Computational Science*, *1*(6), 403–409. URL https://doi.org/10.1038/s43588-021-00084-1

Acciaio, B., & Krach, F. (2022). Market generation via adapted Wasserstein GANs. *preprint*.

Agueh, M., & Carlier, G. (2011). Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, *43*(2), 904–924.

Alimisis, F., Orvieto, A., Becigneul, G., & Lucchi, A. (2021). Momentum improves optimization on Riemannian manifolds. In Banerjee, A., & Fukumizu, K. (Eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research* (pp. 1351–1359). PMLR. URL https://proceedings.mlr.press/v130/alimisis21a.html

Amari, S.-i. (2016). *Information geometry and its applications*, volume 194 of *Applied Mathematical Sciences*. Springer. URL https://doi.org/10.1007/978-4-431-55978-8

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society*, *68*(3), 337–404.

Arribas, I. P., Salvi, C., & Szpruch, L. (2020). Sig-sdes model for quantitative finance. *arXiv preprint arXiv:2006.00218*.

Ay, N., Jost, J., Lê, H. V., & Schwachhöfer, L. (2015). Information geometry and sufficient statistics. *Probability Theory Related Fields*, *162*(1-2), 327–364. ISSN 0178-8051. URL https://doi.org/10.1007/s00440-014-0574-8

Ay, N., Jost, J., Lê, H. V., & Schwachhöfer, L. (2017). *Information geometry*, volume 64 of *Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics [Results in Mathematics and Related Areas. 3rd Series. A Series of Modern Surveys in Mathematics]*. Springer. URL https://doi.org/10.1007/978-3-319-56478-4

Backhoff, J., Bartl, D., Beiglböck, M., & Wiesel, J. (2020). Estimating processes in adapted Wasserstein distance. *arXiv preprint arXiv:2002.07261*.

Backhoff-Veraguas, J., Bartl, D., Beiglböck, M., & Eder, M. (2020). Adapted Wasserstein distances and stability in mathematical finance. *Finance and Stochastics*, *24*, 601–632.

Backhoff-Veraguas, J., Bartl, D., Beiglböck, M., & Eder, M. (2020a). Adapted Wasserstein distances and stability in mathematical finance. *Finance and Stochastics*, *24*(3), 601–632. ISSN 0949-2984. URL https://doi.org/10.1007/s00780-020-00426-3

Backhoff-Veraguas, J., Bartl, D., Beiglböck, M., & Eder, M. (2020b). All adapted topologies are equal. *Probability Theory Related Fields*, *178*(3-4), 1125–1172. URL https://doi.org/10.1007/s00440-020-00993-8

Backhoff-Veraguas, J., Fontbona, J., Rios, G., & Tobar, F. (2022). Stochastic gradient descent in Wasserstein space. *arXiv preprint arXiv:2201.04232*.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Barr, M. (1986). Fuzzy set theory and topos theory. *Canad. Math. Bull.*,*29*(4), 501–508. URL https://doi.org/10.4153/CMB-1986-079-9

Bartl, D., Beiglböck, M., & Pammer, G. (2021). The Wasserstein space of stochastic processes. *arXiv preprint arXiv:2104.14245*.

Bartlett, P. L., Harvey, N., Liaw, C., & Mehrabian, A. (2019). Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, *20*, Paper No. 63, 17.

Basso, G. (2018). Fixed point theorems for metric spaces with a conical geodesic bicombing. *Ergodic Theory and Dynamical Systems*, *38*(5), 1642–1657.

Basso, G., & Miesch, B. (2019). Conical geodesic bicombings on subsets of normed vector spaces. *Advances in Geometry*, *19*(2), 151–164.

Becker, S., Cheridito, P., & Jentzen, A. (2019). Deep optimal stopping. *Journal of Machine Learning Research*, *20*, Paper No. 74, 25.

Becker, S., Cheridito, P., Jentzen, A., & Welti, T. (2021). Solving high-dimensional optimal stopping problems using deep learning. *European Journal of Applied Mathematics*, *32*(3), 470–514. ISSN 0956-7925. URL https://doi.org/10.1017/S0956792521000073

Beknazaryan, A. (2021). Neural networks with superexpressive activations and integer weights. *arXiv preprint arXiv:2105.09917*.

Benth, F. E., Detering, N., & Galimberti, L. (2023). Neural networks infréchet spaces. *Annals of Mathematics and Artificial Intelligence*, *91*(1), 75–103.

Benth, F. E., Detering, N., & Galimberti, L. (2022). Pricing options on flow forwards by neural networks in hilbert space. *arXiv preprint arXiv:2202.11606*.

Bishop, C. M. (1994). Mixture density networks. *Aston University Library*.

Bogachev, V. I. (2007). *Measure theory. Vol. II*. Springer-Verlag, Berlin. URL https://doi.org/10.1007/978-3-540-34514-5

Bonnabel, S. (2013). Stochastic gradient descent on Riemannian manifolds. *IEEE Transactions on Automatic Control*, *58*(9), 2217–2229. doi: 10.1109/TAC.2013.2254619.

Boyd, S., & Chua, L. (1985). Fading memory and the problem of approximating nonlinear operators with volterra series. *IEEE Transactions on Circuits and Systems*, *32*(11), 1150–1161.

Bruè, E., Di Marino, S., & Stra, F. (2021a). Linear lipschitz and c1 extension operators through random projection. *Journal of Functional Analysis*, *280*(4), 108868. doi: https://doi.org/10.1016/j.jfa.2020.108868. URL https://www.sciencedirect.com/science/article/pii/S0022123620304110

Bruè, E., Di Marino, S., & Stra, F. (2021b). Linear Lipschitz and $C^1$ extension operators through random projection. *Journal of Functional Analysis*, *280*(4), 108868.

Buehler, H., Gonon, L., Teichmann, J., & Wood, B. (2019). Deep hedging. *Quant. Finance*, *19*(8), 1271–1291. URL https://doi.org/10.1080/14697688.2019.1571683

Buehler, H., Horvath, B., Lyons, T., Perez Arribas, I., & Wood, B. (2020). A data-driven market simulator for small data environments. *Available at SSRN 3632431*.

Campbell, S., Chen, Y., Shrivats, A., & Jaimungal, S. (2021). Deep learning for principal-agent mean field games. *CoRR*, URL https://arxiv.org/abs/2110.01127

Carbonneau, A., & Godin, F. (2021). Equal risk pricing of derivatives with deep hedging. *Quant. Finance*, *21*(4), 593–608. URL https://doi.org/10.1080/14697688.2020.1806343

Carmona, R. A. (2007). Hjm: A unified approach to dynamic models for fixed income, credit and equity markets. In *Paris-Princeton Lectures on Mathematical Finance 2004* (pp. 1–50). Springer.

Čencov, N. N. (1982). *Statistical decision rules and optimal inference*, volume 53 of *Translations of mathematical monographs*. American Mathematical Society. Translation from the Russian edited by Lev J. Leifman.

Chen, R. T. Q., Rubanova, Y., Bettencourt, J., & Duvenaud, D. K. (2018). Neural ordinary differential equations. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., & Garnett, R. (Eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper/2018/file/69386f6bb1dfed68692a24c8686939b9-Paper.pdf

Cheridito, P., Jentzen, A., & Rossmannek, F. (2021). Efficient approximation of high-dimensional functions with neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, *33*(7), 3079–3093.

Chevallier, J. (2018). Uniform decomposition of probability measures: quantization, clustering and rate of convergence. *J. Appl. Probab.*, *55*(4), 1037–1045. https://doi.org/10.1017/jpr.2018.69

Claici, S., Chien, E., & Solomon, J. (2018). Stochastic Wasserstein barycenters. In *International Conference on Machine Learning* (pp. 999–1008). PMLR.

Cohen, S. N., & Elliott, R. J. (2015). *Stochastic calculus and applications* (2nd ed.). Probability and its applications. Springer. URL https://doi.org/10.1007/978-1-4939-2867-5

Cohen, S. N., Reisinger, C., & Wang, S. (2021). Arbitrage-free neural-sde market models. *arXiv preprint arXiv:2105.11053*.

Cont, R., & Fournie, D. (2010). A functional extension of the Ito formula. *C. R. Math. Acad. Sci. Paris*, *348*(1-2), 57–61. URL https://doi.org/10.1016/j.crma.2009.11.013

Cont, R., & Fournié, D.-A. (2013). Functional Itô calculus and stochastic integral representation of martingales. *Ann. Probab.*, *41*(1), 109–133. URL https://doi.org/10.1214/11-AOP721

Cuchiero, C., Gonon, L., Grigoryeva, L., Ortega, J.-P., & Teichmann, J. (2021). Discrete-time signatures and randomness in reservoir computing. *IEEE Transactions on Neural Networks and Learning Systems*, 1–10. doi: 10.3390/risks8040101.

Cuchiero, C., Khosrawi, W., & Teichmann, J. (2020). A generative adversarial network approach to calibration of local stochastic volatility models. *Risks*, *8*(4), 101.

Cuturi, M., & Doucet, A. (2014). Fast computation of wasserstein barycenters. In *International Conference on Machine Learning* (pp. 685–693). PMLR.

Descombes, D., & Lang, U. (2015). Convex geodesic bicombings and hyperbolicity. *Geom. Dedicata*, *177*, 367–384. URL https://doi.org/10.1007/s10711-014-9994-y

DeVore, R. A., & Lorentz, G. G. (1993). *Constructive approximation*, volume 303 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin. URL https://doi.org/10.1007/978-3-662-02888-9

Draganov, B. R., & Ivanov, K. G. (2014). A generalized modulus of smoothness. *Proc. Amer. Math. Soc.*, *142*(5), 1577–1590. URL https://doi.org/10.1090/S0002-9939-2014-11884-3

Engel, K.-J., & Nagel, R. (2000). *One-parameter semigroups for linear evolution equations*, volume 194 of *Graduate Texts in Mathematics*. Springer-Verlag. With contributions by S. Brendle, M. Campiti, T. Hahn, G. Metafune, G. Nickel, D. Pallara, C. Perazzoli, A. Rhandi, S. Romanelli and R. Schnaubelt.

Figalli, A. (2010). *Optimal transport: Old and new* [book review of mr2459454]. *Bull. Amer. Math. Soc. (N.S.)*, *47*(4), 723–727. URL https://doi.org/10.1090/S0273-0979-10-01285-1

Filipović, D. (2001). *Consistency problems for Heath-Jarrow-Morton interest rate models*, volume 1760 of *Lecture Notes in Mathematics*. Springer-Verlag. URL https://doi.org/10.1007/b76888

Freidlin, M. I., & Wentzell, A. D. (1984). *Random perturbations of dynamical systems*, volume 260 of *Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag. (3rd ed.). Translated from the Russian by Joseph Szücs. URL https://doi.org/10.1007/978-1-4684-0176-9

Fukushima, K., & Miyake, S. (1982). Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and Cooperation in Neural Nets* (pp. 267–285). Springer.

Gallego, F. A., Quintero, J. J., & Riano, J. C. (2015). Convergence of the steepest descent method with line searches and uniformly convex objective in reflexive banach spaces. *Mathematical Communications*, *20*(2), 161–173.

Gambara, M., & Teichmann, J. (2020). Consistent recalibration models and deep calibration. *arXiv preprint arXiv:2006.09455*.

Gierjatowicz, P., Sabate-Vidales, M., Siska, D., Szpruch, L., & Zuric, Z. (2020). Robust pricing and hedging via neural sdes. *Available at SSRN 3646241*.

Gonon, L., Grigoryeva, L., & Ortega, J.-P. (2020). Risk bounds for reservoir computing. *Journal of Machine Learning Research*, *21*, Paper No. 240, 61.

Gonon, L., & Ortega, J.-P. (2020). Reservoir computing universality with stochastic inputs. *IEEE Trans. Neural Netw. Learn. Syst.*, *31*(1), 100–112.

Gonon, L., & Ortega, J.-P. (2021). Fading memory echo state networks are universal. *Neural Networks*, *138*, 10–13. URL https://www.sciencedirect.com/science/article/pii/S0893608021000332

Gonon, L., & Schwab, C. (2021). Deep relu neural network approximation for stochastic differential equations with jumps. *arXiv preprint arXiv:2102.11707*.

Graf, S., & Luschgy, H. (2007). *Foundations of quantization for probability distributions*. Springer.

Gribonval, R., Kutyniok, G., Nielsen, M., & Voigtlaender, F. (2021). Approximation spaces of deep neural networks. *Constructive Approximation*, 1–109.

Grigoryeva, L., & Ortega, J.-P. (2018a). Echo state networks are universal. *Neural Networks*, *108*, 495–508. doi: https://doi.org/10.1016/j.neunet.2018.08.025. URL https://www.sciencedirect.com/science/article/pii/S089360801830251X

Grigoryeva, L., & Ortega, J.-P. (2018b). Universal discrete-time reservoir computers with stochastic inputs and linear readouts using non-homogeneous state-affine systems. *Journal of Machine Learning Research*, *19*, Paper No. 24, 40.

Grigoryeva, L., & Ortega, J.-P. (2019). Differentiable reservoir computing. *Journal of Machine Learning Research*, *20*, Paper No. 179, 62.

Grohs, P., Hornung, F., Jentzen, A., & von Wurstemberger, P. (2022). A proof that artificial neural networks overcome the curse of dimensionality in the numerical approximation of Black-Scholes partial differential equations. *Mem. Amer. Math. Soc., (Forthcoming)*.

Ha, D., Dai, A., & Le, Q. V. (2017). Hypernetworks. In *International Conference on Learning Representations*. URL https://openreview.net/forum?id=rkpACe1lx

Hardt, M., & Ma, T. (2016). Identity matters in deep learning. In *International Conference on Learning Representations*. URL https://openreview.net/forum?id=ryxB0Rtxx

Heath, D., Jarrow, R., & Morton, A. (1992). Bond pricing and the term structure of interest rates: A new methodology for contingent claims valuation. *Econometrica: Journal of the Econometric Society*, 77–105.

Heikkinen, T., Ihnatsyeva, L., & Tuominen, H. (2016). Measure density and extension of Besov and Triebel-Lizorkin functions. *J. Fourier Anal. Appl.*, *22*(2), 334–382. URL https://doi.org/10.1007/s00041-015-9419-9

Heinemann, F., Munk, A., & Zemel, Y. (2022). Randomized Wasserstein barycenter computation: Resampling with statistical guarantees. *SIAM J. Math. Data Sci.*, *4*(1), 229–259. URL https://doi.org/10.1137/20M1385263

Heinonen, J. (2001). *Lectures on analysis on metric spaces*. Universitext. Springer-Verlag, New York. URL https://doi.org/10.1007/978-1-4613-0131-8

Helgason, S. (1979). *Differential geometry, Lie groups, and symmetric spaces*. Academic Press.

Hendrycks, D., & Gimpel, K. (2016). Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Herrera, C., Krach, F., Ruyssen, P., & Teichmann, J. (2021). Optimal stopping via randomized neural networks.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. URL https://doi.org/10.1162/neco.1997.9.8.1735

Hornik, K., Stinchcombe, M., & White, H. (1990). Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. *Neural Network*, *3*(5), 551–560.

Horvath, B., Teichmann, J., & Žurič, Ž. (2021). Deep hedging under rough volatility. *Risks*, *9*(7). URL https://www.mdpi.com/2227-9091/9/7/138

Hutter, C., Gül, R., & Bölcskei, H. (2021). Metric entropy limits on recurrent neural network learning of linear dynamical systems. *Applied and Computational Harmonic Analysis*. doi: https://doi.org/10.1016/j.acha.2021.12.004. URL https://www.sciencedirect.com/science/article/pii/S1063520321001068

Hutzenthaler, M., Jentzen, A., Kruse, T., & Nguyen, T. A. (2020). A proof that rectified deep neural networks overcome the curse of dimensionality in the numerical approximation of semilinear heat equations. *SN Partial Differential Equations and Applications*, *1*(2), 10. URL https://doi.org/10.1007/s42985-019-0006-9

Jacobsen, M. (1982). *Statistical analysis of counting processes*, volume 12 of *Lecture Notes in Statistics*. Springer-Verlag.

Jaeger, H. (2001). The "echo state" approach to analysing and training recurrent neural networks-with an erratum note. *German National Research Center for Information Technology GMD Technical Report*, *148*(34), 13.

Jaeger, H., & Haas, H. (2004). Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. *Science*, *304*(5667), 78–80. doi: https://doi.org/10.1126/science.1091277. URL https://www.science.org

Jia, J., & Benson, A. R. (2019). Neural jump stochastic differential equations. In Wallach, H., Larochelle, H., Beygelzimer, A., d' Alché-Buc, F., Fox, E., & Garnett, R. (Eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper/2019/file/59b1deff341edb0b76ace57820cef237-Paper.pdf

Jiao, Y., Lai, Y., Lu, X., & Yang, Z. (2021). Deep neural networks with relu-sine-exponential activations break curse of dimensionality on h\" older class. *arXiv preprint arXiv:2103.00542*.

Jung, H. (1901). Ueber die kleinste kugel, die eine räumliche figur einschliesst. *Journal für die reine und angewandte Mathematik*, *123*, 241–257. URL http://eudml.org/doc/149122

Kallsen, J., & Krühner, P. (2015). On a heath–jarrow–morton approach for stock options. *Finance and Stochastics*, *19*(3), 583–615. URL https://doi.org/10.1007/s00780-015-0263-1

Kidger, P., Foster, J., Li, X., & Lyons, T. J. (2021). Neural sdes as infinite-dimensional gans. In Meila, M., & Zhang, T. (Eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research* (pp. 5453–5463). PMLR.

Kidger, P., Foster, J., Li, X., Oberhauser, H., & Lyons, T. (2021). Neural sdes as infinite-dimensional gans. *arXiv preprint arXiv:2102.03657*.

Kidger, P., & Lyons, T. (2020). Universal Approximation with Deep Narrow Networks. In Abernethy, J., & Agarwal, S. (Eds.), *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research* (pp. 2306–2327). PMLR.

Kisielewicz, M. (2013). *Stochastic differential inclusions and applications*, volume 80 of *Springer Optimization and Its Applications*. Springer. URL https://doi.org/10.1007/978-1-4614-6756-4

Koshiyama, A., Firoozye, N., & Treleaven, P. (2019). Generative adversarial networks for financial trading strategies fine-tuning and combination. *arXiv preprint arXiv:1901.01751*.

Kovachki, N., Lanthaler, S., & Mishra, S. (2021). On universal approximation and error bounds for fourier neural operators. *Journal of Machine Learning Research*, *22*(290), 1–76. URL http://jmlr.org/papers/v22/21-0806.html

Kratsios, A. (2023). Universal regular conditional distributions. *Constructive Approximation*.

Kratsios, A., & Bilokopytov, I. (2020). Non-euclidean universal approximation. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., & Lin, H. (Eds.), *Advances in Neural Information Processing Systems*, volume 33 (pp. 10635–10646). Curran Associates, Inc. URL https://proceedings.neurips.cc/paper/2020/file/786ab8c4d7ee758f80d57e65582e609d-Paper.pdf

Kratsios, A., Debarnot, V., & Dokmanić, I. (2022). Small transformers compute universal metric embeddings. *arXiv preprint arXiv:2209.06788*.

Kratsios, A., & Hyndman, C. (2020). Deep arbitrage-free learning in a generalized HJM framework via arbitrage-regularization. *Risks*, *8*(2), 40.

Kratsios, A., & Papon, L. (2022). Universal approximation theorems for differentiable geometric deep learning. *Journal of Machine Learning Research*, *23*(196), 1–73.

Kratsios, A., & Zamanlooy, B. (2022). Learning sub-patterns in piecewise continuous functions. *Neurocomputing*. doi: https://doi.org/10.1016/j.neucom.2022.01.036. URL https://www.sciencedirect.com/science/article/pii/S092523122200056X

Kratsios, A., Zamanlooy, B., Liu, T., & Dokmanić, I. (2022). Universal approximation under constraints is possible with transformers. In *International Conference on Learning Representations*. URL https://openreview.net/forum?id=JGO8CvG5S9

Krauthgamer, R., Lee, J. R., Mendel, M., & Naor, A. (2004). Measured descent: A new embedding method for finite metrics. In *45th Annual IEEE Symposium on Foundations of Computer Science* (pp. 434–443). IEEE.

Krauthgamer, R., Lee, J. R., Mendel, M., & Naor, A. (2005). Measured descent: a new embedding method for finite metrics. *Geom. Funct. Anal.*, *15*(4), 839–858. URL https://doi.org/10.1007/s00039-005-0527-6

Küchler, U., & Sørensen, M. (1997). *Exponential families of stochastic processes*. Springer Series in Statistics. Springer-Verlag. URL https://doi.org/10.1007/b98954

Lang, U., & Plaut, C. (2001). Bilipschitz embeddings of metric spaces into space forms. *Geometriae Dedicata*, *87*(1), 285–307.

Lang, U., & Schlichenmaier, T. (2005). Nagata dimension, quasisymmetric embeddings, and Lipschitz extensions. *Int. Math. Res. Not.*, *2005*(58), 3625–3655. URL https://doi.org/10.1155/IMRN.2005.3625

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, *1*(4), 541–551. doi: https://doi.org/10.1162/neco.1989.1.4.541

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

Liang, S., Jiang, S. W., Harlim, J., & Yang, H. (2021). Solving PDEs on Unknown Manifolds with Machine Learning. *arXiv e-prints*, arXiv:2106.06682.

Liu, H., Yang, H., Chen, M., Zhao, T., & Liao, W. (2022). Deep Nonparametric Estimation of Operators between Infinite Dimensional Spaces. *arXiv e-prints*, arXiv:2201.00217.

Liu, Y., & Pagès, G. (2020). Convergence rate of optimal quantization and application to the clustering performance of the empirical measure. *Journal of Machine Learning Research*, *21*(86), 1–36. URL http://jmlr.org/papers/v21/18-804.html

Louart, C., Liao, Z., & Couillet, R. (2018). A random matrix approach to neural networks. *The Annals of Applied Probability*, *28*(2), 1190–1248.

Lovrić, M., Min-Oo, M., & Ruh, E. A. (2000). Multivariate normal distributions parametrized as a Riemannian symmetric space. *J. Multivariate Anal.*, *74*(1), 36–48.

Lu, L., Jin, P., Pang, G., Zhang, Z., & Karniadakis, G. E. (2021). Learning nonlinear operators via deeponet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, *3*(3), 218–229. URL https://doi.org/10.1038/s42256-021-00302-5

Lukoševičius, M., & Jaeger, H. (2009). Reservoir computing approaches to recurrent neural network training. *Computer Science Review*, *3*(3), 127–149.

Lyons, T. (1994). Differential equations driven by rough signals. I. An extension of an inequality of L. C. Young. *Math. Res. Lett.*, *1*(4), 451–464. URL https://doi.org/10.4310/MRL.1994.v1.n4.a5

Maass, W., Natschläger, T., & Markram, H. (2002). Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural Comput.*, *14*(11), 2531–2560. URL https://doi.org/10.1162/089976602760407955

Malagò, L., Montrucchio, L., & Pistone, G. (2018). Wasserstein Riemannian geometry of Gaussian densities. *Inf. Geom.*, *1*(2), 137–179.

Manjunath, G. (2020). Stability and memory-loss go hand-in-hand: three results in dynamics and computation. *Proc. A.*, *476*(2242), 20200563, 16.

Manjunath, G., & Jaeger, H. (2013). Echo state property linked to an input: exploring a fundamental characteristic of recurrent neural networks. *Neural Comput.*, *25*(3), 671–696. URL https://doi.org/10.1162/NECO_a_00411

Meyer, G., Bonnabel, S., & Sepulchre, R. (2011). Regression on fixed-rank positive semidefinite matrices: A Riemannian approach. *The Journal of Machine Learning Research*, *12*, 593–625.

Miesch, B. (2018). The Cartan–Hadamard theorem for metric spaces with local geodesic bicombings. *L'Enseignement Mathématique*, *63*(1), 233–247.

Morrill, J., Salvi, C., Kidger, P., & Foster, J. (2021). Neural rough differential equations for long time series. In Meila, M., & Zhang, T. (Eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research* (pp. 7829–7838). PMLR.

Munkres, J. R. (2000). Second edition of [MR0464128]. Prentice Hall, Inc., Upper Saddle River, NJ, 2000. xvi+537 pp. ISBN: 0-13-181629-2 54-01. Topology.

Naor, A. (2001). A phase transition phenomenon between the isometric and isomorphic extension problems for hölder functions between lp spaces. *Mathematika*, *48*(1-2), 253–271.

Ni, H., Szpruch, L., Sabate-Vidales, M., Xiao, B., Wiese, M., & Liao, S. (2021). Sig-Wasserstein GANs for time series generation. *arXiv preprint arXiv:2111.01207*.

Ni, H., Szpruch, L., Wiese, M., Liao, S., & Xiao, B. (2020). Conditional Sig-Wasserstein GANs for time series generation. *arXiv preprint arXiv:2006.05421*.

Nielsen, F. (2020). An elementary introduction to information geometry. *Entropy*, *22*(10).

Palm, G. (1944). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, *5*, 115–133.

Park, S., Lee, J., Yun, C., & Shin, J. (2021). Provable memorization via deep neural networks using sub-linear parameters. In Belkin, M., & Kpotufe, S. (Eds.), *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research* (pp. 3627–3661). PMLR.

Park, S., Yun, C., Lee, J., & Shin, J. (2021). Minimum width for universal approximation. In *International Conference on Learning Representations*. URL https://openreview.net/forum?id=O-XJwyoIF-k

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

Prolla, J. B. (1971). Bishop's generalized Stone-Weierstrass theorem for weighted spaces. *Mathematische Annalen*, *191*(4), 283–289.

Ramachandran, P., Zoph, B., & Le, Q. V. (2018). Searching for activation functions. URL https://openreview.net/forum?id=SkBYYyZRZ

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536. URL https://doi.org/10.1038/323533a0

Rüschendorf, L. (1985). The Wasserstein distance and approximation theorems. *Z. Wahrsch. Verw. Gebiete*, *70*(1), 117–129. URL https://doi.org/10.1007/BF00532240

Saitoh, S., & Sawano, Y. (2016). *Theory of reproducing kernels and applications*, volume 44 of *Developments in Mathematics*. Springer. URL https://doi.org/10.1007/978-981-10-0530-5

Schäfer, A. M., & Zimmermann, H. G. (2006). Recurrent neural networks are universal approximators. In *International Conference on Artificial Neural Networks*, (pp. 632–640). Springer.

Schmocker, P. (2022). *Universal approximation on path spaces and applications in finance*. PhD thesis, University of St. Gallen.

Semadeni, Z. (2006). *Schauder bases in Banach spaces of continuous functions*, volume 918. Springer.

Shaham, U., Cloninger, A., & Coifman, R. R. (2018). Provable approximation properties for deep neural networks. *Applied and Computational Harmonic Analysis*, *44*(3), 537–557. doi: https://doi.org/10.1016/j.acha.2016.04.003. URL https://www.sciencedirect.com/science/article/pii/S1063520316300033

Shen, Z., Yang, H., & Zhang, S. (2021a). Deep network with approximation error being reciprocal of width to power of square root of depth. *Neural Computation*, *33*(4), 1005–1036. doi: https://doi.org/10.1162/neco_a_01364

Shen, Z., Yang, H., & Zhang, S. (2021b). Neural network approximation: Three hidden layers are enough. *Neural Networks*, *141*, 160–173. doi: https://doi.org/10.1016/j.neunet.2021.04.011. URL https://www.sciencedirect.com/science/article/pii/S0893608021001465

Siegel, J. W., & Xu, J. (2020). Approximation rates for neural networks with general activation functions. *Neural Networks*, *128*, 313–321.

Siegelmann, H. T., & Sontag, E. D. (1995). On the computational power of neural nets. *Journal of Computer and System Sciences*, *50*(1), 132–150.

Sirignano, J., & Cont, R. (2019). Universal features of price formation in financial markets: perspectives from deep learning. *Quant. Finance*, *19*(9), 1449–1459. URL https://doi.org/10.1080/14697688.2019.1622295

Sitzmann, V., Martel, J., Bergman, A., Lindell, D., & Wetzstein, G. (2020). Implicit neural representations with periodic activation functions. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., & Lin, H. (Eds.), *Advances in neural information processing systems*, volume 33 (pp. 7462–7473). Curran Associates, Inc. URL https://proceedings.neurips.cc/paper/2020/file/53c04118df112c13a8c34b38343b9c10-Paper.pdf

Stacks project authors, T. (2021). The stacks project. https://stacks.math.columbia.edu

Touzi, N. (2013). *Optimal stochastic control, stochastic target problems, and backward SDE*, volume 29 of *Fields Institute Monographs*. Springer; Fields Institute for Research in Mathematical Sciences. With Chapter 13 by Angès Tourin. URL https://doi.org/10.1007/978-1-4614-4286-8

Tripuraneni, N., Flammarion, N., Bach, F., & Jordan, M. I. (2018). Averaging stochastic gradient descent on Riemannian manifolds. In Bubeck, S., Perchet, V., & Rigollet, P. (Eds.), *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, (pp. 650–687). PMLR. URL https://proceedings.mlr.press/v75/tripuraneni18a.html

van Rhijn, J., Oosterlee, C. W., Grzelak, L. A., & Liu, S. (2021). Monte carlo simulation of SDES using GANS. *arXiv preprint arXiv:2104.01437*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998–6008).

Vershynin, R. (2020). Memory capacity of neural networks with threshold and rectified linear unit activations. *SIAM J. Math. Data Sci.*, *2*(4), 1004–1033. URL https://doi.org/10.1137/20M1314884

von Oswald, J., Henning, C., Sacramento, J., & Grewe, B. F. (2020). Continual learning with hypernetworks. In *International Conference on Learning Representations*. URL https://openreview.net/forum?id=SJgwNerKvB

Weaver, N. (2018). *Lipschitz algebras*. World Scientific Publishing Co. Pte. Ltd.. Second edition of [MR1832645].

Weisz, F. (1994). *Martingale Hardy spaces and their applications in Fourier analysis*, volume 1568 of *Lecture notes in mathematics*. Springer-Verlag. URL https://doi.org/10.1007/BFb0073448

Wiener, N. (1958). *Nonlinear problems in random theory*. Technology Press Research Monographs. The Technology Press of The Massachusetts Institute of Technology and John Wiley & Sons, Inc.; Chapman & Hall, Ltd..

Wiese, M., Knobloch, R., Korn, R., & Kretschmer, P. (2020). Quant gans: Deep generation of financial time series. *Quantitative Finance*, *20*(9), 1419–1440.

Wiese, M., Wood, B., Pachoud, A., Korn, R., Buehler, H., Phillip, M., & Bai, L. (2021). Multi-asset spot and option market simulation. arXiv preprint arXiv:2112.06823.

Yarotsky, D. (2018). Optimal approximation of continuous functions by very deep relu networks. In Bubeck, S., Perchet, V., & Rigollet, P. (Eds.), *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research* (pp. 639–649). PMLR. URL https://proceedings.mlr.press/v75/yarotsky18a.html

Yarotsky, D. (2021a). Elementary superexpressive activations. In Meila, M., & Zhang, T. (Eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research* (pp. 11932–11940). PMLR.

Yarotsky, D. (2021b). Elementary superexpressive activations. In Meila, M., & Zhang, T. (Eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research* (pp. 11932–11940). PMLR. URL https://proceedings.mlr.press/v139/yarotsky21a.html

Yarotsky, D., & Zhevnerchuk, A. (2020). The phase diagram of approximation rates for deep neural networks. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., & Lin, H. (Eds.), *Advances in Neural Information Processing Systems*, volume 33, (pp. 13005–13015). Curran Associates, Inc.

Yun, C., Sra, S., & Jadbabaie, A. (2019). Small relu networks are powerful memorizers: a tight analysis of memorization capacity. In Wallach, H., Larochelle, H., Beygelzimer, A., d' Alché-Buc, F., Fox, E., & Garnett, R. (Eds.),

*Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper/2019/file/dbea3d0e2a17c170c412c74273778159-Paper.pdf

Zhang, C., Ren, M., & Urtasun, R. (2019). Graph hypernetworks for neural architecture search. In *International Conference on Learning Representations*. URL https://openreview.net/forum?id=rkgW0oA9FX

Zhang, Z., Zohren, S., & Roberts, S. (2019). Deeplob: Deep convolutional neural networks for limit order books. *IEEE Transactions on Signal Processing*, *67*(11), 3001–3012. doi: https://doi.org/10.1109/TSP.2019.2907260

Zhmoginov, A., Sandler, M., & Vladymyrov, M. (2022). HyperTransformer: Model Generation for Supervised and Semi-Supervised Few-Shot Learning. *arXiv e-prints*, arXiv:2201.04182.

**How to cite this article:** Acciaio, B., Kratsios, A., & Pammer, G. (2024). Designing universal causal deep learning models: The geometric (Hyper)transformer. *Mathematical Finance*, *34,* 671–735. https://doi.org/10.1111/mafi.12389