

One model to solve them all: 2BSDE families via neural operators

Takashi FURUYA ^{*} Anastasis KRATSIOS [†] Dylan POSSAMAÏ [‡] Bogdan RAONIĆ [§]

November 4, 2025

Abstract

We introduce a mild generative variant of the classical neural operator model, which leverages Kolmogorov–Arnold networks to solve infinite families of second-order backward stochastic differential equations (2BSDEs) on regular bounded Euclidean domains with random terminal time. Our first main result shows that the solution operator associated with a broad range of 2BSDE families is approximable by appropriate neural operator models. We then identify a structured subclass of (infinite) families of 2BSDEs whose neural operator approximation requires only a polynomial number of parameters in the reciprocal approximation rate, as opposed to the exponential requirement in general worst-case neural operator guarantees.

Key words: Neural operators, solution operators, backward stochastic differential equations, exponential approximation rates.

1 Introduction

Fix a positive integer $d \in \mathbb{N}^*$. We work on a filtered probability space $(\Omega, \mathcal{F}, \mathbb{P} := (\mathcal{F}_t)_{t \in [0, \infty)}, \mathbb{P})$ carrying a d -dimensional $(\mathbb{F}, \mathbb{P})$ -Brownian motion W . Fix a sufficiently regular bounded open domain $\mathcal{D} \subset \mathbb{R}^d$, as well as maps $\mu : \mathbb{R}^d \rightarrow \mathbb{R}^d$, $\Sigma : \mathbb{R}^d \rightarrow \mathbb{R}^{d \times d}$, and $f : \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^{d \times d} \rightarrow \mathbb{R}$, as well as an initial point $x \in \mathcal{D}$. We are interested in *simultaneously* approximately solving each 2BSDE in the (non-empty) compact infinite family $\mathcal{B} \subseteq (X, Y^{g, f_0}, Z^{g, f_0}, \Upsilon^{g, f_0}, A^{g, f_0})_{(g, f_0) \in \mathfrak{W}}$ where $\mathfrak{W}$ is a suitable subset of the Sobolev spaces $W^{1,p}(\partial\mathcal{D}) \times W^{1,p}(\mathcal{D})$. These 2BSDEs are defined through the system

$$\begin{aligned} X_t &= x + \int_0^t \beta(X_s) ds + \int_0^t \Sigma(X_s) dW_s, \quad t \geq 0, \quad \mathbb{P}\text{-a.s.}, \quad \tau := \inf \{t \geq 0 : X_t \notin \mathcal{D}\}, \quad (\text{SDE}) \\ Y_t^{g, f_0} &= \underbrace{g(X_\tau)}_{\text{Perturbation}} + \int_{t \wedge \tau}^\tau \left(\underbrace{f(X_s, Y_s^{g, f_0}, Z_s^{g, f_0}, \Upsilon_s^{g, f_0})}_{\text{Reference generator}} + \underbrace{f_0(X_s)}_{\text{Perturbation}} - \frac{1}{2} \text{Tr}[\Sigma(X_s) \Sigma^\top(X_s) \Upsilon_s^{g, f_0}] \right) ds \\ &\quad - \int_{t \wedge \tau}^\tau Z_s^{g, f_0} \cdot dX_s, \quad t \in [0, \tau), \quad \mathbb{P}\text{-a.s.}, \quad (\text{FBSDE}) \\ Z_t^{g, f_0} &= z_0 + \int_0^t A_s^{g, f_0} ds + \int_0^t \Upsilon_s^{g, f_0} dX_s, \quad t \in [0, \tau), \quad \mathbb{P}\text{-a.s.} \quad (\text{2BSDE}) \end{aligned}$$

Using a variant (see [Section 3.1](#) below for the proof) of the results of [Cheridito, Soner, Touzi, and Victoir \[16\]](#) for 2BSDEs with random terminal time τ , as above, for each pair $(g, f_0) \in \mathfrak{W}$, if the following elliptic PDE

$$f(x, u(x), \nabla u(x), \nabla^2 u(x)) = -f_0(x), \quad x \in \mathcal{D} \quad u(x) = g(x), \quad x \in \partial\mathcal{D}, \quad (1.1)$$

admits a smooth enough solution, then the 2BSDE system [\(SDE\)](#), [\(FBSDE\)](#), [\(2BSDE\)](#) admits a solution of the form

$$Y_t^{g, f_0} = u(X_t), \quad Z_t^{g, f_0} = \nabla u(X_t), \quad \Upsilon_t^{g, f_0} = \nabla^2 u(X_t), \quad A_t^{g, f_0} = \mathcal{L} \nabla u(X_t), \quad t \in [0, \tau), \quad \mathbb{P}\text{-a.s.}, \quad (1.2)$$

^{*}Department of Biomedical Engineering, Doshisha University, Kyōto, Japan, takashi.furuya0101@gmail.com.

[†]Department of Mathematics, McMaster University, McMaster University, Hamilton, Canada, kratsiosa@mcmaster.ca.

[‡]ETH Zürich, Department of Mathematics, Zürich, Switzerland, dylan.possamai@math.ethz.ch.

[§]ETH Zürich, Seminar for Applied Mathematics and ETH AI Center, Zürich, Switzerland, braonic@ethz.ch.

where $\mathcal{L}$ denotes the generator associated to the forward process X (without the drift term), defined for any continuous bounded test function f on $\mathbb{R}^d$ by

$$\mathcal{L}(f) := \frac{1}{2} \text{Tr}[\Sigma(x)\Sigma(x)^\top \nabla^2 f(x)], \quad x \in \mathbb{R}^d,$$

see [16, Equations (2.9) and (2.11)] for a similar result in the parabolic case.

Our first main result, [Theorem 3.7](#), guarantees that the following solution map is approximable by a neural operator

$$\begin{aligned} \Gamma^+ : W^{1,\infty}(\mathcal{D}; \mathbb{R}) \times W^{1,\infty}(\mathcal{D}; \mathbb{R}) &\longrightarrow W^{1,\infty}(\mathcal{D}; \mathbb{R}) \\ (f_0, g) &\longmapsto u \end{aligned} \quad (1.3)$$

where f_0 and g are the source and boundary data of the PDE in (1.1), respectively; which equivalently perturb the generator and the terminal condition of the associated 2BSDEs with random terminal time τ in (FBSDE).

Consequently, the solution map associated to the family of second-order BSDEs is approximable by our stochastic neural operator model (which extends the neural operator model of [Furuya and Kratsios](#) in [31, Definition 4] from the classical BSDE setting to 2BSDEs). This result thus provides a 2BSDE analogue of neural operator approximability results, which typically follow a two-step strategy: first, establish a quantitative universal approximation theorem for general Hölder-continuous functions with the same source and target as the solution map (see *e.g.* [Lu, Jin, Pang, Zhang, and Karniadakis](#) [64], [Korolev](#) [46], [Galimberti, Kratsios, and Livieri](#) [33], [Yu, Becquey, Halikias, Mallory, and Townsend](#) [96], [Lanthaler, Mishra, and Karniadakis](#) [56], [Lu, Jin, and Karniadakis](#) [63], [Lanthaler, Li, and Stuart](#) [57], [Kratsios, Furuya, Benitez, Lassas, and de Hoop](#) [50], [Schwab, Stein, and Zech](#) [85], [Gödeke and Fernsel](#) [38], [Furuya, Taniguchi, and Okuda](#) [32], and [Adcock, Brugiapaglia, Dexter, and Moraga](#) [4]); second, show that the solution map is sufficiently continuous, for instance Hölder-continuous, often via a perturbation analysis, verifying in turn it is in the scope of the main theorem, see [Alvarez, Ekren, Kratsios, and Yang](#) [5], [Horvath, Kratsios, Limmer, and Yang](#) [43], [Lanthaler and Stuart](#) [55] or [Firouzi, Yang, and Kratsios](#) [30].

One may ask if favourable approximation rates are achievable if the reference generator f is simple enough, while still of course having a meaningful structure for several applications in optimal control and mathematical finance. Indeed, in [Theorem 3.11](#) we show that this is the case when the reference generator is of the simplified form

$$f(x, y, z, w) := -\text{Tr}[\gamma(x)w] - \text{div}(\gamma)(x) \cdot z + \mu(x) \cdot z + \lambda(x)y + \tilde{f}(x, y) \quad (1.4)$$

for some smooth enough maps $\lambda : \mathbb{R}^d \rightarrow \mathbb{R}$, $\gamma : \mathbb{R}^d \rightarrow \mathbb{R}^{d \times d}$, and $\mu : \mathbb{R}^d \rightarrow \mathbb{R}^d$ and where $\tilde{f} : \mathbb{R}^d \times \mathbb{R} \rightarrow \mathbb{R}$ is still sufficiently smooth. In this setting, we reduce the general fully non-linear elliptic PDE in (1.1) to the following semi-linear form

$$-\nabla \cdot \gamma \nabla u(x) + \mu(x) \cdot \nabla u(x) + \lambda(x)u(x) + \tilde{f}(x, u) = \underbrace{-f_0(x)}_{\text{Perturbation}}, \quad x \in \mathcal{D}, \quad u(x) = \underbrace{g(x)}_{\text{Perturbation}}, \quad x \in \partial\mathcal{D}. \quad (1.5)$$

[Theorem 3.11](#) both extends [31, Theorem 1] by allowing μ and λ to be non-zero and Σ to be non-constant and positive-definite, while no longer requiring any *a priori* knowledge of the PDE itself to be hard-coded into our design of the NO. This is because the latter authors use explicit knowledge of Green’s function associated with the PDE $-\nabla \cdot \gamma \nabla u(x) + \mu(x) \cdot \nabla u(x) + \lambda(x)u(x)$ to show that it admits a decomposition $\Phi(x - y) + \Psi(x, y)$, where Φ is a ‘difficult to approximate’ singular part and Ψ is an ‘easily approximated’ smooth part. The convolution with the singular component Φ is then hard-coded into each of their NO architectures by leveraging the explicit closed form for Φ obtained in [11]. In contrast, in our approach no such closed-form nor *a priori* knowledge of the PDE is required in our NO build. As should be expected, these extensions also come at the cost of devising an entirely different proof strategy.

The PDE in (1.5) can be connected back to the 2BSDE (FBSDE) either when the divergence of γ is absorbed into μ , or in the special case where γ is divergence-free, *i.e.* $\text{div}(\gamma)^\top = 0$, implying that $\nabla \cdot \gamma \nabla u = \text{Tr}[\gamma \nabla^2 u]$. In addition, when γ is valued in the set of semi-definite matrices, and we take for Σ any matrix square root of $e2\gamma$ (that is to say $\Sigma \Sigma^\top = 2\gamma$), then (1.5) reduces to the more standard Hamilton–Jacobi–Bellman–type semilinear equation

$$\tilde{f}(x, u) + \lambda(x)u(x) + \mu(x) \cdot \nabla u(x) - \frac{1}{2} \text{Tr}[\Sigma(x)\Sigma(x)^\top \nabla^2 u(x)] = -f_0(x), \quad x \in \mathcal{D}. \quad (1.6)$$

In dimension $d \geq 2$, there is a whole zoology of divergence-free γ ; thus this special case completely subsumes the case where γ is constant, as considered in [31]. For example, when $d = 2$, if γ is positive-definite-valued then

there exist a twice continuously differentiable potential $\varphi_\gamma : \mathbb{R}^2 \rightarrow \mathbb{R}$ (this is the so-called Airy potential) such that $\gamma(x) = R^\top (\nabla^2 \varphi_\gamma(x)) R$ for the symplectic matrix $R := e_1 e_2^\top - e_2 e_1^\top$ (where (e_1, e_2) is the canonical basis of $\mathbb{R}^2$). A simple non-constant example of such an Airy potential φ_γ which additionally yields a positive-definite γ is $\varphi_\gamma(x, y) := (x^2 + y^2)^2$.

Our first objective is, therefore, to *simultaneously* approximate the solution operator to general families of fully non-linear elliptic problems (1.1) and to obtain favourable rates for semi-linear special cases of the form (1.5). Our strategy will be to construct a neural operator (NO) model which directly approximates (Theorem 3.7 resp. Theorem 3.11) the coefficient-to-solution operator mapping any $(g, f_0) \in \mathfrak{W}$ to the elliptic PDE it defines via (1.1) (resp. (1.5)). Then, using the connections between elliptic PDEs and 2BSDEs with random terminal time in (1.2) formalised by our non-linear Feynman–Kac formula in Proposition 3.1, we construct an adapter transforming the functions output for our NO to tuples of stochastic processes approximating the solution to the family of associated 2BSDEs, see Theorem 3.13.

1.1 Related literature

There is a mature numerical literature on second-order BSDEs (2BSDEs), including weak approximation and time-discretisation schemes by Possamaï and Tan [80], Ren and Tan [82], Yang, Zhao, and Zhou [94], and the recent non-equidistant scheme of Pak, Hwang, and Kim [75]. Learning-based approaches have also appeared (*e.g.*, Beck, E, and Jentzen [7], Pereira, Wang, Chen, Reed, and Theodorou [77], Duong [24], Xiao, Qiu, and Nikan [93]), but these methods are essentially *per-instance*: they must be re-run (or re-trained) whenever coefficients or boundary data change. By contrast, we learn a *solution operator* that acts on the entire compact family of problems indexed by (g, f_0) , so a single trained model simultaneously solves all members of the family, both at the PDE and at the 2BSDE level via the PDE–(2)BSDE correspondence (Cheridito, Soner, Touzi, and Victoir [16]; see also Pardoux [76], Soner, Touzi, and Zhang [87]).

When it comes to finite-dimensional ML for non-linear PDEs, a large body of work trains a finite-dimensional network for each target PDE separately (*e.g.*, Nüsken and Richter [74], Pham, Warin, and Germain [78], Germain, Laurière, Pham, and Warin; Germain, Pham, and Warin; Germain, Pham, and Warin [34; 35; 36], Lefebvre, Loeper, and Pham [58], Zhou, Han, and Lu [97], Hu and Laurière [44], Nguwi, Penent, and Privault [73]). Provable *exponential* behaviour in this setting typically requires strong structure: either linear second-order elliptic operators (Marcati and Schwab [66; 67]) or analyticity of the *single* target solution, so that one may invoke classical exponential approximation of analytic functions by neural networks (Mhaskar and Micchelli [69], Mhaskar [68], E and Wang [25]).

On the other hand, neural operators (NOs) learn the infinite-dimensional coefficient-to-solution map and hence *simultaneously* solve all PDEs in a parametric class with a single model; see the early universality observation of Chen and Chen [15], the DeepONet/FNO line (Lu, Jin, and Karniadakis [63], Kovachki, Li, Liu, Azizzadenesheli, Bhattacharya, Stuart, and Anandkumar [47]), the CNO universality Raonić, Molinaro, de Ryck, Rohner, Bartolucci, Alaifari, Mishra, and de Bézenac [81], and a large set of abstract guarantees in Banach/Besov/Sobolev and non-linear metric settings (Yu, Becquey, Halikias, Mallory, and Townsend [96], Lu, Jin, Pang, Zhang, and Karniadakis [64], Lanthaler, Mishra, and Karniadakis [56], Adcock, Brugiapaglia, Dexter, and Moraga [2], Korolev [46], Cuchiero, Schmock, and Teichmann [17], Neufeld and Schmock [72], Kratsios, Furuya, Benitez, Lassas, and de Hoop [50], Adcock, Brugiapaglia, Dexter, and Moraga [4], Gödeke and Fernsel [38], Lanthaler and Stuart [55], Schwab, Stein, and Zech [85], de Hoop, Lassas, and Wong [20], Furuya, Taniguchi, and Okuda [32], Kratsios, Schmock, and Zimmermann [52], Acciaio, Kratsios, and Pammer [1], Kratsios, Liu, Lassas, de Hoop, and Dokmanic [49]). Within this line, *exponential* (sometimes ‘exponential-in-depth’) expression rates are known for holomorphic operator classes (Adcock, Dexter, and Moraga Scheuermann [3]), for certain linear elliptic PDEs (including polytopal domains) (Marcati and Schwab [66; 67]), and for specific semilinear elliptic equations on smooth domains (Furuya and Kratsios [31]). Other exponential statements rely either on super-expressive activations with effectively infinite pseudo-dimension (Shen, Yang, and Zhang [86], Pollard [79], Alvarez, Ekren, Kratsios, and Yang [5]) or on implicit/equilibrium-layer constructions exploiting convex variational structure (Kratsios, Neufeld, and Schmock [51]).

Our contribution in this landscape is that we design a NO that *simultaneously* (i) approximates the solution operator of a broad class of second-order elliptic PDEs/2BSDEs and (ii) retains *exponential-in-depth* rates in a substantially more general semi-linear regime than in the closest prior work. Concretely

- (i) *family-level learning*. We approximate the coefficient-to-solution map Γ^+ on a compact infinite family indexed by (f_0, g) , hence a single training phase serves the whole family (PDEs and the associated 2BSDEs). For fully

non-linear elliptic equations we obtain general operator-level approximability (algebraic rates) by combining quantitative NO universality on Besov/Sobolev scales (Yu, Becquey, Halikias, Mallory, and Townsend [96], Lu, Jin, Pang, Zhang, and Karniadakis [64], Lanthaler, Mishra, and Karniadakis [56], Adcock, Brugiapaglia, Dexter, and Moraga [2], Korolev [46], Galimberti, Kratsios, and Livieri [33]) with stability of the solution map (Krylov-type assumptions; cf. Krylov [54]).

- (ii) *Exponential rates for semi-linear equations under general forward dynamics.* In the semi-linear case

$$-\nabla \cdot \gamma(x) \nabla u + \mu(x) \cdot \nabla u + \lambda(x) u + \tilde{f}(x, u) = -f_0(x), \quad u|_{\partial \mathcal{D}} = g,$$

with smooth, uniformly elliptic γ and smooth μ, λ , we implement the classical fixed-point map by a non-local NO layer built from (approximated) Green kernels; existence/regularity of Green functions for variable-coefficient operators is standard (Kim and Sakellaris [45]). This yields accuracy ε with *logarithmic* depth $L = O(\log(1/\varepsilon))$, *constant* width, and a finite non-local rank that scales polynomially in $1/\varepsilon$. Compared to Furuya and Kratsios [31], which hard-codes the singular part of the Green's kernel and effectively assumes a driftless, constant-diffusion forward (so that the singular Φ is known in closed form), our construction does *not* require a closed-form kernel split and therefore covers far more general, state-dependent Itô diffusions in the forward process and variable-coefficient elliptic operators, while preserving exponential depth-rates.

- (iii) *From PDE to (2)BSDE at the operator level.* Because each 2BSDE in the family admits the PDE representation, our NO for the elliptic map transfers directly to a NO for the (Y, Z, Υ, A) -processes associated with the *entire* 2BSDE family.

Building upon the finite-dimensional lower bounds of Yarotsky [95], it was recently shown in Lanthaler and Stuart [55] that arbitrary continuous—or even several times continuously Fréchet differentiable—non-linear operators between Sobolev spaces cannot be uniformly approximated on compact sets by NOs without requiring an exponential number of trainable parameters in the reciprocal approximation error. Consequently, without additional structure beyond simple smoothness, there are insurmountable obstructions to operator learning. Thus, even if one could establish Hölder-continuity of the coefficient-to-solution operator in the fully non-linear setting (*e.g.* using results of Taylor [88], which we do show in Section A.4) the solution operator would still not be regular enough to permit meaningful approximation rates. In such cases, any quantitative result is practically no more meaningful than an existential statement on the approximability of the coefficient-to-solution operator (see Theorem 3.7), akin to the qualitative (rate-free) universal abstract approximation results of Chen and Chen [14], Benth, Detering, and Galimberti [8], or Bilokopytov and Xanthos [9] for other NO architectures.

When it comes to the closest exponential-rate results available in the literature, relative to linear/holomorphic NO rates (Marcati and Schwab [66; 67], Adcock, Dexter, and Moraga Scheuermann [3]), we require neither analyticity nor specialised domains; and unlike exponential claims relying on super-expressive activations or implicit/equilibrium layers (Shen, Yang, and Zhang [86], Pollard [79], Alvarez, Ekren, Kratsios, and Yang [5], Kratsios, Neufeld, and Schmocker [51]), our architecture maintains finite capacity per layer with explicit depth/width/rank scaling. Crucially, compared to Furuya and Kratsios [31], our exponential regime permits markedly *more general* forward dynamics and variable-coefficient elliptic operators, because the Green-kernel is learned/approximated rather than injected in closed form.

2 Preliminaries

2.1 Notation

Let $p \in (1, \infty)$. We denote by $p' \in (1, \infty)$ the conjugate of p such that $1/p + 1/p' = 1$. We let $\mathbb{N}$ be set of non-negative integers, $\mathbb{N}^*$ the set of positive integers, and $\mathbb{Z}$ the set of all negative and non-negative integers. We henceforth fix an ambient dimension¹ $d \in \mathbb{N}^*$; and let $\mathbb{S}_d^+$ denote the set of $d \times d$ (real) positive-definite matrices. Recall that, every symmetric positive definite matrix $A \in \mathbb{S}_d^+$ has a unique well-defined square-root given by $\sqrt{A} := \log(\exp(A)/2)$ where $\exp$ is the matrix exponential and $\log$ is its (unique) inverse on $\mathbb{S}_d^+$, see *e.g.* Arabpour, Armstrong, Galimberti, Kratsios, and Livieri [6, Lemma C.5]. For any $d \in \mathbb{N}^*$ denote the Fröbenius norm of any $d \times d$ matrix A by $\|A\|_F$.

¹In [31] an explicit expression for the singular part of the Green's function associated to the stopped forward process's induced elliptic PDE was required, which additionally constrained $d \geq 3$ there, but not herein.

Given any metric space $(\mathcal{X}, \rho)$, any $x \in \mathcal{X}$, and any radius $r \geq 0$, we define the open ball $B_{(\mathcal{X}, \rho)}(x, r) := \{u \in \mathcal{X} : \rho(x, u) < r\}$. Given any two vector spaces V and W , and any $x \in V$ and $y \in W$, we write $x \oplus y := (x, y) = V \times W$. For any $p \geq 1$, we let $\ell^p(\mathbb{Z})$ be the set of real-valued sequences $(u_n)_{n \in \mathbb{Z}}$ indexed by $\mathbb{Z}$ such that

$$\sum_{n \in \mathbb{Z}} |u_n|^p < \infty.$$

We also let $\mathbb{L}^p(\mathbb{R})$ be the set of p -integrable Lebesgue-measurable functions on $\mathbb{R}$.

For any $I \in \mathbb{N}$, we use $C^I(\mathbb{R})$ to denote the vector space of real-valued at-least I -times continuously differentiable functions on $\mathbb{R}$, and $C_c^I(\mathbb{R})$ for the subset thereof consisting of those compactly supported functions therein. For any $(s, d, D) \in (\mathbb{N}^*)^3$, we write $C^s(\mathcal{D}, \mathbb{R}^D)$ (resp. $C^\infty(\mathcal{D}, \mathbb{R}^D)$) for set of functions from $\mathbb{R}^d$ to $\mathbb{R}^D$ which are at-least s -times (resp. smooth) continuously differentiable when restricted to $\mathcal{D}$. We refer the reader to Appendix A.2.3 for wavelet-centric definitions of Besov, and thus Sobolev, spaces.

Throughout this paper, $(\Omega, \mathcal{F}, \mathbb{P} := (\mathcal{F}_t)_{t \geq 0}, \mathbb{P})$ will denote a filtered probability space satisfying the usual conditions. For any $T > 0$ we use $\mathcal{H}_T^2$ to denote the class of square-integrable predictable processes $X : [0, T] \times \Omega \rightarrow \mathbb{R}$.

2.2 Deep learning

Neural operators (NOs) extend deep learning from finite-dimensional vector spaces to infinite-dimensional Banach spaces, with standard NOs specialising in function-to-function mappings. Broadly speaking, there are three types of NO builds between function spaces: the Fourier neural operator-type builds (FNO), which iteratively use finitely parametrised integral-kernel affine transformations between their non-linearities; DeepONet-type architectures (see [Lu, Jin, and Karniadakis \[63\]](#)) which learn to adaptively regress against learnable bases; and encoder—processor—decoder-type models, such as PCA-Net (see [Chan, Jia, Gao, Lu, Zeng, and Ma \[12\]](#)) which project infinite-dimensional data using a Schauder basis before processing it via a standard finite-dimensional neural network, and then reassembles finite-dimensional basis functions using the network’s outputs as coefficients.

The first and last of these models tend to be more numerically stable, the middle construction can exhibit advantageous approximation rates, and the third model is more readily generalisable to non—function space settings (see *e.g.* [Galimberti, Kratsios, and Livieri \[33\]](#)) by directly lifting the approximation guarantees for classical neural networks (see *e.g.* [Yarotsky](#); [Bolcskei, Grohs, Kutyniok, and Petersen](#); [DeVore, Hanin, and Petrova](#); [Gribonval, Kutyniok, Nielsen, and Voigtlaender](#); [Kratsios and Zamanlooy](#); [Shen, Yang, and Zhang](#); [Hong and Kratsios](#); [Schneider, Ullrich, and Vybiral](#)) to infinite dimensions. Our neural-operator build combines the best of the first two models using a two-branch structure: the top branch of an FNO-type, the bottom branch inspired by DeepONets, with coefficients shared between layers. Moreover, we map into non—function space targets when applying our deep-learning model in the 2BSDE setting by transforming its function space-valued outputs into processes via a ‘Feynman—Kac adapter’, that is to say a custom non-trainable readout layer encoding our nonlinear Feynman—Kac representation ([Proposition 3.1](#)). Finally, we allow the non-linearities injecting structure at each layer of our NO to be adaptive rather than fixed, as in classical NO builds, thereby maximizing their flexibility, for instance granting them the ability to exactly perform multiplication, a property not shared by classical piecewise-linear ReLU activation functions.

2.2.1 Residual Kolmogorov–Arnold networks (Res-KANs)

The key idea behind Kolmogorov–Arnold networks (KANs) is to make the activation function itself trainable. In KANs, one typically focuses on the spline part of the following definition [Liu, Wang, Vaidya, Ruehle, Halverson, Soljagic, Hou, and Tegmark \[62\]](#), with the role of the remaining part of the activation function being an afterthought, normally taken to some standard non-linearity such as the Swish or Sigmoid functions. In this paper, we explicitly exploit both parts of KANs activation functions, and as such, we add some basic structural requirements to the ‘non-spline’ part of the activation function (below in (2.1)) which serves a pointed role in our approximation theory in connection with the multi-resolution analysis (MRA); see *e.g.* [Mallat \[65\]](#).

Specifically, the activation $\sigma_{\beta:I} : \mathbb{R} \rightarrow \mathbb{R}$ maps any $x \in \mathbb{R}$ to a mixture of spline basis functions of varying degrees

$$\sigma_{\beta:I}(x) := \underbrace{\beta_{-1}\sigma_s(x) + \beta_0\sigma_w(x)}_{\text{Spectral structure}} + \underbrace{\sum_{i=1}^I \beta_i \mathcal{N}_i(x)}_{\text{Local structure}} \quad (2.1)$$

where $I \in \mathbb{N}$, $\beta = (\beta_{-1}, \beta_0, \cdot, \beta_I)^\top \in \mathbb{R}^{I+2}$ is a trainable vector of coefficients, and where for $i \in \{1, \dots, I\}$, $\mathcal{N}_i : \mathbb{R} \rightarrow \mathbb{R}$ are the cardinal B-splines which, following [Mhaskar and Micchelli \[70, Equation \(4.28\)\]](#), can be defined by $\mathcal{N}_0(x) := \mathbf{1}_{[0,1]}$ and for any $i \in \mathbb{N}^*$

$$\mathcal{N}_I(x) := \sum_{j=0}^{I+1} \frac{(-1)^j \binom{I+1}{j}}{I!} \text{ReLU}(x - j)^I, \quad x \in \mathbb{R}. \quad (2.2)$$

Furthermore, $\sigma_s : \mathbb{R} \rightarrow \mathbb{R}$ as well as $\sigma_w : \mathbb{R} \rightarrow \mathbb{R}$ and satisfy the spectral properties in [Assumption 2.1](#) below. However, before turning to the properties, we elucidate the first few wavelets in [Figure 1](#).



Figure 1: The cardinal B-splines of orders $I = 0, 1$, and 2.

Assumption 2.1 (Daubechies properties of order I). *Fix $I \in \mathbb{N}$. The respective ‘scale’ and ‘wavelet’ activation function σ_s and σ_w both belong to $C_c^I(\mathbb{R})$ if $I > 0$ (resp. $L^2(\mathbb{R})$ when $I = 0$ with compact essential support) and satisfy the refinement equation of [Daubechies \[18, Equation \(3.47\)\]](#), that is to say that there is a sequence of low-pass filters $(h_k)_{k \in \mathbb{Z}} \in \ell^2(\mathbb{Z})$ summing to $\sqrt{2}$, satisfying the orthogonality condition²*

$$\sum_{k \in \mathbb{Z}} h_{k-2i} h_{k-2j} = \mathbf{1}_{\{i=j\}}, \quad \forall (i, j) \in \mathbb{Z}^2,$$

and such that

$$\sigma_s(x) = \sqrt{2} \sum_{k \in \mathbb{Z}} h_k \sigma_s(2x - k), \quad x \in \mathbb{R}, \quad (2.3)$$

$$\sigma_w(x) = \sqrt{2} \sum_{k \in \mathbb{Z}} (-1)^k h_{1-k} \sigma_s(2x - k), \quad x \in \mathbb{R}. \quad (2.4)$$

The existence of such activation functions (called Daubechies father and mother wavelets respectively), for arbitrary I , is guaranteed by [Triebel \[89, Theorem 1.61.\(ii\)\]](#), while algorithmic constructions can be found in [Daubechies \[19, Chapter 6.4\]](#), and are standard in modern signal processing. Nevertheless, in the very low regularity regime where $I = 0$, the Haar system and the indicator function is a transparent example where [Assumption 2.1](#) holds.

Example 2.2 (Haar wavelets and indicator function for discontinuous regularity). *If $I = 0$ then, the indicator function of the unit interval $\sigma_s := \mathbf{1}_{[0,1]}$ and the Haar wavelet $\sigma_M := \mathbf{1}_{[0,1/2)} - \mathbf{1}_{[1/2,1]}$ satisfy [Assumption 2.1](#) with $h_0 = h_1 = \frac{1}{\sqrt{2}}$ and $h_k = 0$ whenever $|k| \geq 2$. Thus, σ_M and σ_s belong to $L^2(\mathbb{R})$ as expected since $I = 0$.*

In a KAN, this activation operates component-wise, with parameters tailored to each neuron. That is, for any integer k , any $x \in \mathbb{R}^k$, and $\beta := (\beta_1, \dots, \beta_k) \in \mathbb{R}^{(I+2) \times k}$, we define

$$\begin{aligned} \sigma_{\beta:I} \bullet : \mathbb{R}^k &\rightarrow \mathbb{R}^k \\ x = (x_1, \dots, x_k)^\top &\mapsto (\sigma_{\beta_1:I}(x_1), \dots, \sigma_{\beta_k:I}(x_k))^\top. \end{aligned} \quad (2.5)$$

We now introduce the core idea of *residual* KAN networks. These networks incorporate an additional *residual connection*, ensuring that signal is preserved during activation. Residual connections, standard in modern deep

²See [\[18, Equation \(3.18\)\]](#)

learning architectures, help stabilise training by preserving gradient flow and regularising the loss landscape, see [Riedi, Balestrierio, and Baraniuk \[83\]](#). They also mitigate vanishing gradients that can be caused by normalisation layers. Following [Acciaio, Kratsios, and Pammer \[1\]](#), we allow for flexible use of these residual paths, potentially modulated by a trainable gating mechanism.

More precisely, we fix positive integers d_{out} and d_{in} , matrices $(A, G) \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}} \times \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, with G being diagonal (i.e. $G_{i,j} = 0$ for $(i, j) \in \{1, \dots, d_{\text{out}}\} \times \{1, \dots, d_{\text{in}}\}$ with $i \neq j$), as well as $b \in \mathbb{R}^{d_{\text{out}}}$, and $\beta \in \mathbb{R}^{(I+2) \times d_{\text{out}}}$, a matrix of trainable coefficients. We then define for $x \in \mathbb{R}^{d_{\text{in}}}$

$$\mathcal{L}(x|A, b, \beta, G : I) := \underbrace{\sigma_{\beta:I} \bullet (Ax + b)}_{\text{KAN layer}} + \underbrace{Gx}_{\text{Residual connection}} \quad (2.6)$$

Although compositions of such KAN layers define valid functions, these may lack higher-order smoothness—an issue for applications such as PDE solving that require high regularity. There are two ways to address this: (1) enforce that $\beta_i = 0$ for small i , or (2) apply a smoothing layer at the output. We adopt the first strategy to ensure that the functions realised by our *smoothed residual* KANs are infinitely differentiable.

Definition 2.3 (Residual KANs (Res-KANs)). *Let D and I be positive integers, and let $\alpha > 0$. A residual Kolmogorov–Arnold network (Res-KAN) is a function $\hat{f} : \mathbb{R}^d \rightarrow \mathbb{R}^D$ with representation, for some $L \in \mathbb{N}^*$*

$$\hat{f} = A^{(L)} f^{(L)} + b^{(L)}, \quad (2.7)$$

with

$$f^{(0)}(x) = x, \quad x \in \mathbb{R}^d, \quad f^{(\ell)} = \mathcal{L}(f^{(\ell-1)}|A^{(\ell)}, b^{(\ell)}, \beta^{(\ell)}, G^{(\ell)} : I), \quad \ell \in \{1, \dots, L\},$$

where, for $\ell \in \{1, \dots, L\}$, $A^{(\ell)}$ and $G^{(\ell)}$ are $d_{\ell+1} \times d_\ell$ matrices with $G^{(\ell)}$ diagonal, $\beta^{(\ell)}$ is a $(I+2) \times d_{\ell+1}$ matrix, $b \in \mathbb{R}^{d_{\ell+1}}$, for given positive integers $(d_0, \dots, d_{L+1})$ satisfying $d_0 = d$ and $d_{L+1} = D$. In addition, for any $\ell \in \{1, \dots, L\}$, $\beta^{(\ell)}$ satisfies the sparsity pattern ensuring smoothness³

$$\beta_{i,j}^{(\ell)} = 0, \quad i < \lceil \alpha \rceil \text{ and } j \in \{1, \dots, d_{\ell+1}\}. \quad (2.8)$$

We denote the class of all Res-KANs with L hidden layers, width $W := \max_{\ell \in \{1, \dots, L+1\}} d_\ell$, adaptivity parameter I , and smoothness parameter α , by $\text{Res-KAN}_{L,W}^{I,\alpha}(\mathbb{R}^d, \mathbb{R}^D)$.

2.2.2 Neural operator architectures

We recall that we have fixed a constant $1 < p < \infty$ and $\mathcal{D} \subset \mathbb{R}^d$, a bounded open domain. The classical neural operators are defined in, e.g., [Kovachki, Li, Liu, Azizzadenesheli, Bhattacharya, Stuart, and Anandkumar \[47\]](#) or [Lanthaler, Li, and Stuart \[57\]](#).

Importantly, our NO architecture (see [Figure 2](#)) contains both encoder–processor–decoder (EPD) type and Fourier neural operator (FNO)-type ‘branches’ at each layer, whereby spectral features and physical features are iteratively processed in parallel, and then combined together using the adaptively activated neurons spearheaded by the KAN paradigm [\[48\]](#), rather than the static activation strategy of classical MLPs. The resulting architecture thus exhibits beneficial properties both of FNO-type models and EPD-type models.

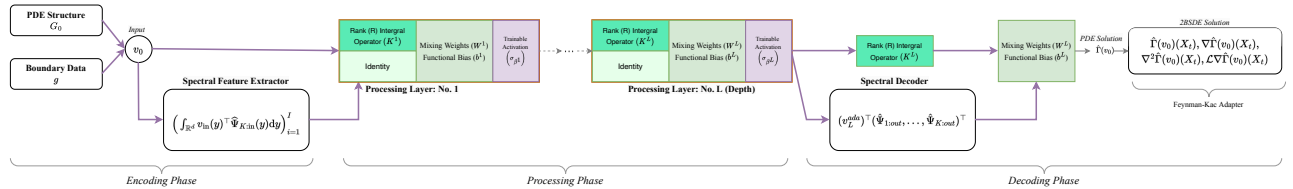


Figure 2: The KANO ([Definition 2.4](#)) pipeline.

What is illustrated in [Figure 2](#) is as follows.

- 0) First boundary data (g) and the PDE structure (G_0) are concatenated into an input v_0 .

³The $\lceil \alpha \rceil$ -time continuous differentiability of $\hat{f}$ follows from that of B-splines (see [DeVore and Sharpley \[22\]](#)), and the chain rule.

- 1) Learnable spectral features akin to FNOs are then extracted from v_0 and concatenated thereto.
- 2) At each processing iteration, the top NO branch applies a finite rank (R) integral operator, then all features are mixed and adaptively activated.
- 3) Finally the predictions are decoded via two branches: one applying another finite rank integral operator together as with to the FNO and the other leveraging a (trainable) spectral feature decoding akin to EPD, before both branches are mixed together to obtain the final prediction $\hat{\Gamma}$.

In the 2FBNO variant (**Definition 2.5**): $\hat{\Gamma}(v_0)$ is passed through the *Feynman–Kac adapter* (see **Proposition 3.1**).

This being said, we can now proceed with the definition. In the remainder of the paper $d_{\text{in}} = 2$, any tuple $v_{\text{out}} \in W^{1,\infty}(\mathcal{D}; \mathbb{R})^{d_{\text{in}}}$ will correspond to a pair of boundary and source data (g, f_0) , and $d_{\text{out}} = 1$. However, since many of these result can be use in more general approximation theory of solutions operators to other PDEs, we keep the definition of our KANO model general enough to accomodate other applications.

Definition 2.4 (Kolmogorov–Arnold neural operator (KANO)). *Fix positive integers $d_{\text{in}}, d_{\text{out}}, L, W, \hat{L}, \hat{W}, D_{\text{ada}}, W_{\text{ada}}$, as well as smoothness parameters $\alpha > 0$ and $I \in \mathbb{N}^*$ with $3 \leq \alpha \leq I$. We define a Kolmogorov–Arnold neural operator (KANO) $\hat{\Gamma} : W^{1,\infty}(\mathcal{D}; \mathbb{R})^{d_{\text{in}}} \rightarrow W^{1,\infty}(\mathcal{D}; \mathbb{R})^{d_{\text{out}}}$ to be any map sending any $v_{\text{in}} \in W^{1,\infty}(\mathcal{D}; \mathbb{R})^{d_{\text{in}}}$ to some $v_{L+1} \in W^{1,\infty}(\mathcal{D}; \mathbb{R})^{d_{\text{out}}}$ where v_{L+1} is defined iteratively by*

$$v_0(x) := \begin{pmatrix} v_0^{\text{crs}}(x) \\ v_0^{\text{ada}}(x) \end{pmatrix} := \begin{pmatrix} v_{\text{in}}(x) \\ \int_{\mathbb{R}^d} v_{\text{in}}(y)^\top \hat{\Psi}_{1:\text{in}}(y) dy \\ \vdots \\ \int_{\mathbb{R}^d} v_{\text{in}}(y)^\top \hat{\Psi}_{K:\text{in}}(y) dy \end{pmatrix}, \quad x \in \mathcal{D}, \quad (2.9)$$

$$v_{\ell+1}(x) := \begin{pmatrix} v_{\ell+1}^{\text{crs}}(x) \\ v_{\ell+1}^{\text{ada}}(x) \end{pmatrix} := \sigma_{\beta_\ell : I} \bullet \left(W^\ell \begin{pmatrix} v_\ell^{\text{crs}}(x) + (K^{(\ell)} v_\ell)(x) \\ v_\ell^{\text{ada}}(x) \end{pmatrix} + b^\ell(x) \right), \quad \ell \in \{0, \dots, L-1\}, \quad x \in \mathcal{D}, \quad (2.10)$$

$$v_{L+1}(x) := W^{(L)} \begin{pmatrix} v_L^{\text{crs}}(x) + (K^{(L)} v_L)(x) \\ (v_L^{\text{ada}})^\top(x) \begin{pmatrix} \hat{\Psi}_{1:\text{out}}(x) \\ \vdots \\ \hat{\Psi}_{K:\text{out}}(x) \end{pmatrix} \end{pmatrix} + b^{(L)}(x), \quad x \in \mathcal{D}, \quad (2.11)$$

where $\sigma_{\beta:I}$ is as in **Equation (2.5)** and acts as in **(2.1)**. In particular, $\beta_\ell \in \mathbb{R}^{(I+2) \times d_{\ell+1}}$, each $(\hat{\Psi}_{k:\text{in}})_{k \in \{1, \dots, K\}}$ and $(\hat{\Psi}_{k:\text{out}})_{k \in \{1, \dots, K\}}$ are Res-KANs of depth D_{ada} and width W_{ada} , and for any $\ell \in \{0, \dots, L+1\}$, we have $W^{(\ell)} \in \mathbb{R}^{d_{\ell+1} \times d_\ell}$

$$(K^{(\ell)} v)(x) := \int_{\mathcal{D}} k_{NN}^{(\ell)}(x, y) v(y) dy, \quad x \in \mathcal{D}, \quad v \in L^p(\mathcal{D}; \mathbb{R})^{d_\ell}, \quad b^{(\ell)}(x) := b_{NN}^{(\ell)}(x), \quad x \in \mathcal{D},$$

where $k_{NN}^{(\ell)} \in \text{Res-KAN}_{\hat{L}, \hat{W}}^{I, \alpha}(\mathbb{R}^{d \times d}, \mathbb{R}^{d_{\ell+1} \times d_\ell})$ and $b_{NN}^{(\ell)} \in \text{Res-KAN}_{\hat{L}, \hat{W}}^{I, \alpha}(\mathbb{R}^d, \mathbb{R}^{d_\ell})$ are Res-KANs of depth $\hat{L}$ and width $\hat{W}$. We denote the above class of KANOs by

$$\mathcal{NO}_{\hat{L}, \hat{W}}^{L, W, I, \alpha}(W^{1,\infty}(\mathcal{D}; \mathbb{R})^{d_{\text{in}}}, W^{1,\infty}(\mathcal{D}; \mathbb{R})^{d_{\text{out}}}),$$

which we abbreviate to $\mathcal{NO}_{\hat{L}, \hat{W}}^{L, W, I, \alpha}$ when the dimensions and domains are contextually evident.

For $I := \lceil s \rceil$, we henceforth abbreviate

$$\mathcal{NO}_{I, \alpha} := \bigcup_{(L, \hat{L}, W, \hat{W}, \alpha) \in (\mathbb{N}^*)^4 \times (0, 1)} \mathcal{NO}_{\hat{L}, \hat{W}}^{L, W, I, \alpha}, \quad (2.12)$$

Motivated by the PDE representation of the solutions to each member of our family of second-order BSDEs, given in (1.2), and due to [16], we extend the (semi-)classical class of neural operators above to the following stochastic model as follows.

Definition 2.5 (2Generative neural operators (2FBNO)). *Fix dimensions d , and d_{in} , as well as smoothness parameters $3 \leq \alpha \leq I$, with $I \in \mathbb{N}^*$, and fix depths $L \in \mathbb{N}^*$, $\widehat{L} \in \mathbb{N}^*$, and widths $W \in \mathbb{N}^*$, $\widehat{W} \in \mathbb{N}^*$. The class of forward-backward KANOs (2FBNOs) $\mathcal{FB}_{\widehat{L}, \widehat{W}, X}^{L, W, I, \alpha}$ consists of all*

$$\begin{aligned} \widehat{\Gamma} : W^{1, \infty}(\mathcal{D}, \mathbb{R})^{d_{\text{in}}} &\longrightarrow (\mathcal{H}_T^2)^4 := \prod_{i=1}^4 \mathcal{H}_T^2 \\ f := (f_1, \dots, f_{d_{\text{in}}}) &\longmapsto (\widehat{Y}^f, \widehat{Z}^f, \widehat{\Upsilon}^f, \widehat{A}^f), \end{aligned}$$

for which there is a $\Gamma \in \mathcal{NO}_{\widehat{L}, \widehat{W}}^{L, W, I, \alpha}(W^{1, \infty}(\mathcal{D}; \mathbb{R})^{d_{\text{in}}}, W^{1, \infty}(\mathcal{D}; \mathbb{R}))$ satisfying the representation

$$Y_t^f = \Gamma(f)(X_t), \quad Z_t^f = (\nabla \Gamma(f))(X_t), \quad \Upsilon_t^f = (\nabla^2 \Gamma(f))(X_t), \quad \text{and} \quad A_t^f = (\mathcal{L} \nabla \Gamma(f))(X_t),$$

where, as before, $\mathcal{L}$ denotes the generator of X , without the drift.

3 Main results

3.1 Elliptic PDE representation of the 2BSDE system

For the reader's convenience, we repeat the PDE in (1.1).

$$f(x, u(x), \nabla u(x), \nabla^2 u(x)) = -f_0(x), \quad x \in \mathcal{D}, \quad u(x) = g(x), \quad x \in \partial \mathcal{D}, \quad (3.1)$$

Proposition 3.1 (Non-linear Feynman–Kac's formula). *Let u be a classical solution to the PDE (1.1), such that all the quantities below are defined and continuous in time*

$$Y_t = u(X_t), \quad Z_t = \nabla u(X_t), \quad \Upsilon_t = \nabla^2 u(X_t), \quad A_t = \mathcal{L} \nabla u(X_t), \quad t \in [0, \tau], \quad \mathbb{P}\text{-a.s.},$$

where

$$X_t = x + \int_0^t \beta(X_s) ds + \int_0^t \gamma(X_s) dW_s, \quad t \geq 0, \quad \mathbb{P}\text{-a.s.}, \quad \tau := \inf \{t \geq 0 : X_t \notin \mathcal{D}\}.$$

Then (Y, Z, Υ, A) is a solution to (FBSDE)–(2BSDE).

Proof. Since u is smooth enough, we can apply Itô's formula to obtain for any $t \in [0, \tau]$

$$u(X_t) = u(X_\tau) - \int_t^\tau \frac{1}{2} \text{Tr}[\gamma(X_s) \gamma^\top(X_s) \nabla^2 u(X_s)] ds - \int_t^\tau \nabla u(X_s) \cdot dX_s,$$

as well as

$$\nabla u(X_t) = \nabla u(x) + \int_0^t \nabla^2 u(X_s) dX_s + \int_0^t \mathcal{L} \nabla u(X_s) ds = \nabla u(x) + \int_0^t \Upsilon_s dX_s + \int_0^t A_s ds.$$

it follows by the PDE satisfied by u that

$$u(X_t) = g(X_\tau) + \int_t^\tau \left(f(X_s, Y_s, Z_s, \Upsilon_s) + f_0(X_s) - \frac{1}{2} \text{Tr}[\gamma(X_s) \gamma^\top(X_s) \Upsilon_s] \right) ds - \int_t^\tau Z_s^\top dX_s,$$

as desired. $\square$

3.2 General approximability guarantee

Let $0 < \delta \leq 1$ and let $\mathbb{S}_d^\delta$ denote the subset of $\mathbb{S}_d^+$ consisting of matrices satisfying the following near-norm preserving property: for every $x \in \mathbb{R}^d$

$$\delta \|x\|^2 \leq xA^\top x \leq \frac{1}{\delta} \|x\|^2.$$

We write generically u' for $(x_0, \dots, x_d) \in \mathbb{R}^{1+d}$, u'' for any element of $\mathbb{S}_d^\delta$, and $u := (u', u'')$.

Setting 3.2. and let $\bar{G} : \mathbb{R}^d \rightarrow [0, \infty)$ be Borel measurable. Fix constants $K_0, K_F \geq 0$, $L_F, C_g \geq 0$, and $0 < \delta \leq 1$. We require the following of the domain $\mathcal{D}$.

Assumption 3.3 (Domain Regularity). *The domain $\mathcal{D} \subseteq \mathbb{R}^d$ is a non-empty bounded domain with $C^{1,1}$ -boundary satisfying the exterior ball condition.*

Our general approximability result, for which favourable rates cannot generally be guaranteed, considers families of fully non-linear elliptic PDEs

$$f(x, u(x), \nabla u(x), \nabla^2 u(x)) = 0, \quad x \in \mathcal{D}, \quad u(x) = g(x), \quad x \in \partial\mathcal{D},$$

where the boundary data $g \in W^{k,p}(\partial\mathcal{D})$ is assumed to be sufficiently smooth, i.e. $k \geq 2$.

Following Krylov [54, Chapter 14], our PDEs will have sufficiently regular solutions under the following conditions.

Assumption 3.4. Assume that $p > d$, and fix constants $(c_1, c_2, R_0) \in (0, 1]^3$, $L_F \geq 0$, a function $\omega_F : [0, \infty) \rightarrow [0, \infty)$ with $\omega_F(0) = 0$, a Borel measurable function $\bar{G} : \mathbb{R}^d \rightarrow [0, \infty)$, and Borel measurable functions F and G of the variables (u_0, u', x) and (u, x) respectively. We have

(i) $f = F + G$, and for all $u'' \in \mathbb{S}_d^+$, $u' \in \mathbb{R}^{1+d}$, and $x \in \mathbb{R}^d$, we have

$$|G(u, x)| \leq c_1 \|u''\|_F + c_2 \|u'\| + \bar{G}(x), \quad F(0, x) = 0; \quad (3.2)$$

(ii) F is L_F -Lipschitz continuous with respect to u'' ;

(iii) for any $v \in \mathbb{R}$, $0 < r \leq R_0$, and $x \in \mathcal{D}$, there exists a convex function $\bar{F} : \mathbb{S}_d \rightarrow [0, \infty)$ such that

(a) $\bar{F}(0, x) = 0$, and $\nabla_{u''} \bar{F}$ has range in $\mathbb{S}_d^\delta$ at every point of twice differentiability of $\bar{F}$;

(b) for every $u'' \in \mathbb{S}_d^+$ with $\|u''\|_F = 1$, we have

$$\inf_{B(r, x) \cap \mathcal{D}} \sup_{\bar{r} > 0} \frac{|\bar{F}(u'_0, ru'', u) - \bar{F}(\tau u'')|}{r} \leq c_2 \text{Vol}(\mathcal{D} \cap B(r, x)), \quad (3.3)$$

where $\text{Vol}(A)$ denotes the d -dimensional Lebesgue measure of a Lebesgue-measurable set $A \subseteq \mathbb{R}^d$;

(c) for any $(u, v) \in \mathbb{R}^2$, $x \in \mathcal{D}$, and $u' \in \mathbb{S}_d^+$, we have

$$|F(u, u'', x) - F(v, u'', x)| \leq \omega_F(|u - v|) \|u''\|_F. \quad (3.4)$$

The next definition introduces appropriate perturbations of the original PDE we consider, and uses notations from Assumption 3.4.

Definition 3.5 (PDE perturbation space $\mathcal{X}_k(r)$). Fix $r > 0$, $k \in \mathbb{N}^*$ and let $\mathcal{X}_k(r)$ consist of all pairs $(\bar{G}_0, g) \in W^{2,p}(\mathcal{D}) \times W^{k,p}(\mathcal{D})$ with $\|g\|_{W^{k,p}(\mathcal{D})} \leq r$. Define $G_0 := G + \bar{G}_0$ and, for every pair $(G_0, g) \in \mathcal{X}_k(r)$ denote the solution to the following associated fully non-linear elliptic PDE by $u_{\bar{G}_0, g}$

$$\left(\underbrace{F + G}_{\text{Structure}} + \underbrace{\bar{G}_0}_{\text{Perturbation}} \right) (x, u(x), \nabla u(x), \nabla^2 u(x)) = 0, \quad \forall x \in \mathcal{D}, \quad u(x) = \underbrace{g(x)}_{\text{Perturbation}}, \quad \forall x \in \partial\mathcal{D}. \quad (3.5)$$

Example 3.6 (Source perturbations only). We can, of course, restrict ourselves to perturbations of the source condition itself only, in which case we may restrict our attention to $\bar{G}_0$ which are constant in their first argument; i.e. $\bar{G}_0(u, x) = f_0(x)$ for some $f_0 \in W^{k,p}(\mathcal{D})$, similarly to the special case in (1.5).

Theorem 3.7 (Approximability of the perturbation-to-solution map). *Fix $q \in [1, +\infty)$, let $\mathcal{D}$ be a bounded exterior-thick domain in $\mathbb{R}^d$ with $C^{1,1}$ -boundary, let $r > 0$, $k > 1 + \max\{1, \frac{d}{p}\}$, and $\mathcal{X} \subseteq \mathcal{X}_k(r)$ be compact.*

Suppose [Assumptions 3.3](#) and [3.4](#) hold and that both σ_s and σ_w satisfy [Assumption 2.1](#). Then, for every approximation error $\varepsilon > 0$, there exists some neural operator $\hat{\Gamma} \in \mathcal{NO}_{[k],1}$, cf. [\(2.12\)](#), satisfying the uniform estimate

$$\sup_{(\tilde{G}_0, g) \in \mathcal{X}} \|u_{\tilde{G}_0, g} - \hat{\Gamma}(\tilde{G}_0, g)\|_{W^{2,p}(\mathcal{D})} < \varepsilon. \quad (3.6)$$

The proof of [Theorem 3.7](#) is based on two ingredients. First, we establish the local-Lipschitz regularity of the coefficient-to-solution map associated to our family of fully non-linear elliptic PDEs ([Lemma A.15](#)) verifying the only necessary condition for approximability by continuous models classes; such as our NO, namely continuity—a property which need not be immediate for arbitrary coefficient-to-solution maps. Next, we rely on [Proposition A.14](#) which establishes a general universal approximation theorem for operators between Besov spaces.

In this sense, [Proposition A.14](#) for our NO architecture which, among other things, can be regarded as a generalisation of [Kovachki, Li, Liu, Aizzadenesheli, Bhattacharya, Stuart, and Anandkumar](#) [47, Theorem 11], which does not cover Besov spaces $B_{q,r}^s(\mathcal{D})$ for finite values of q and r (recall that $W^{s,p}(\mathcal{D}) = B_{q,r}^s(\mathcal{D})$ [90, Remark 1.2]). We emphasise that here, the case of finite q (and r) is necessary since $W^{s,\infty}(\mathcal{D})$ -spaces are automatically excluded from both [Proposition A.14](#) and [47, Theorem 11], as well as any encoder-decoder-type model using basis expansions (e.g. [33]), since $W^{s,\infty}$ is not separable and thus cannot admit a Schauder basis. Additionally, since this space is non-separable and any realistic NO model must be parameterised by finitely many parameters and depend continuously on them, any realistic NO model defines a separable space. As such, it cannot be dense/universal in spaces of continuous functions between non-separable spaces—again by elementary topological considerations.

We now consider the approximation of a specialized family of elliptic PDEs, whose solution operator exhibits enough structure so that it (not all continuous functions) can be approximated on non-separable space $W^{1,\infty}(\mathcal{D})$.

3.3 Feasible rates

3.3.1 Semi-linear elliptic PDE

In what follows, we will make use of the map $S_{\gamma,\mu,\lambda} : W^{(d+3)/2,2}(\partial\mathcal{D}; \mathbb{R}) \longrightarrow W^{1,\infty}(\mathcal{D}; \mathbb{R})$ sending boundary data to domain data, and defined for each $g \in W^{(d+3)/2,2}(\partial\mathcal{D}; \mathbb{R})$ by

$$S_{\gamma,\mu,\lambda}(g) := w_g, \quad (3.7)$$

where $w_g \in W^{(d+4)/2,2}(\mathcal{D}; \mathbb{R}) \subset W^{1,\infty}(\mathcal{D}; \mathbb{R})$. is the unique solution of

$$-\nabla \cdot \gamma \nabla w_g + \mu \cdot \nabla w_g + \lambda w_g = 0 \text{ in } \mathcal{D}, \quad w_g = g \text{ on } \partial\mathcal{D}.$$

We assume the following on the maps γ , μ and λ .

Assumption 3.8. *The maps $\gamma : \mathcal{D} \longrightarrow \mathbb{R}^{d \times d}$, $\mu : \mathcal{D} \longrightarrow \mathbb{R}^d$, and $\lambda : \mathcal{D} \longrightarrow \mathbb{R}$ satisfy the following conditions*

(i) $\gamma \in C^\infty(\bar{\mathcal{D}}; \mathbb{R}^{d \times d})$, $\mu \in C^\infty(\bar{\mathcal{D}}; \mathbb{R}^d)$, and $\lambda \in C^\infty(\bar{\mathcal{D}}; \mathbb{R})$ where $C^\infty(\bar{\mathcal{D}}; \mathbb{R}^d)$ and $C^\infty(\bar{\mathcal{D}}; \mathbb{R}^{d \times d})$ denote the spaces of all d -dimensional vector-valued and $d \times d$ matrix-valued functions that are infinitely differentiable on $\mathcal{D}$ and whose derivatives admit continuous extensions to the closure $\bar{\mathcal{D}}$;

(ii) γ is uniformly elliptic and bounded in the sense that there are positive constants γ_0 and γ_1 such that

$$\gamma_0 \|\xi\|^2 \leq \xi^\top \gamma(x) \xi \leq \gamma_1 \|\xi\|^2, \quad \forall (x, \xi) \in \mathcal{D} \times \mathbb{R}^d;$$

(iii) μ and λ are such that

$$\lambda \geq 0, \text{ and } \lambda \geq \nabla \cdot \mu \sum_{i=1}^d \partial_{x_i} \mu.$$

Next, we summaries our main assumptions on $\tilde{f}$.

Assumption 3.9. *The map $\tilde{f} : \mathcal{D} \times \mathbb{R} \longrightarrow \mathbb{R}$ satisfies*

(i) there exists $\delta_0 > 0$ and $H \in \mathbb{N}^* \setminus \{1, 2\}$ such that

$$\tilde{f}(x, z) = \sum_{h=0}^H \frac{\partial_z^h \tilde{f}(x, 0)}{h!} z^h, \text{ for } \|z\| < \delta_0, \text{ and } x \in \mathcal{D};$$

(ii) $\tilde{f}(\cdot, 0) = \partial_z^1 \tilde{f}(\cdot, 0) = 0$;

(iii) $\partial_z^h \tilde{f}(\cdot, 0) \in C^\infty(\bar{\mathcal{D}}; \mathbb{R})$ for all $h \in \{2, \dots, H\}$.

Assumption (i) posits that $\tilde{f}(x, z)$ is analytic at $z = 0$ and represented by a finite power series truncated at order H . Assumption (ii) removes the zeroth- and first-order terms, which are already captured by $f_0(x)$ and $\lambda(x)u(x)$ in (1.5). Assumption (iii) requires all coefficient functions to be smooth, ensuring a well-posed setting for the subsequent analysis.

Finally, we formulate a smallness assumption.

Assumption 3.10. We take $0 < \delta < \delta_0$ (where δ_0 comes from Assumption 3.9.(i)) so that

$$C_1 \delta < 1, \rho := C_2 \delta < 1, C_3 \delta < 1,$$

where the positive constants C_1, C_2, C_3 will appear in (A.6), (A.7), and (A.13), and depend only $p, d, \mathcal{D}, \tilde{f}, \gamma$, and μ .

Under the above assumptions, we have the following approximation guarantee for the *solution operator* of the PDE associated with our randomly stopped second-order BSDE system (SDE), (FBSDE), (2BSDE).

Theorem 3.11 (Exponential approximation rates: solution operator to the elliptic problem). *Let⁴ $d \geq 3$. Let Assumptions 3.8 to 3.10 hold. Suppose that $\mathcal{D}$ is a bounded open set with Lipschitz boundary in $\mathbb{R}^d$. Let $1 < s < 2$ and $1 \leq p < \frac{d}{d-1}$. Then, for any $0 < \varepsilon < 1$, there are positive integers $L, W, \hat{L}, \hat{W}$, and $\Gamma \in \mathcal{NO}_{\hat{L}, \hat{W}}^{L, W, \Gamma, \alpha}(W^{1, \infty}(\mathcal{D}; \mathbb{R})^2, W^{1, \infty}(\mathcal{D}; \mathbb{R}))$ such that*

$$\sup_{(f_0, g) \in \mathcal{B}} \left\| \Gamma^+(f_0, g) - \Gamma(f_0, S_{\gamma, \mu, \lambda}(g)) \right\|_{W^{1, \infty}(\mathcal{D}; \mathbb{R})} \leq \varepsilon.$$

where the supremum is taken over the set

$$\mathcal{B} := B_{W^{1, \infty}(\mathcal{D}; \mathbb{R})}(0, \delta^2) \times B_{W^{(d+3)/2, 2}(\partial \mathcal{D}; \mathbb{R})}(0, \delta^2).$$

Moreover, we have the following estimates for parameters $L = L(\Gamma)$, $W = W(\Gamma)$, $\hat{L} = \hat{L}(\Gamma)$, and $\hat{W} = \hat{W}(\Gamma)$,

$$L \leq C \log(\varepsilon^{-1}), W \leq C, \hat{L} \leq C, \hat{W} \leq C \varepsilon^{-\frac{1}{(s-1)p}},$$

where $C > 0$ depends only on $s, p, d, \mathcal{D}, \tilde{f}, \gamma$, and μ .

Our quantitative approximation rates are available because the family of elliptic PDEs considered here is well structured. In the fully general setting, however, since our NOs are continuous, one should not expect rates, as the solution operator should not even be expected to be continuous (let alone locally-Lipschitz continuous) which is necessary for approximability by the elementary uniform limit theorem from point-set topology, see Munkres [71, Theorem 21.6]. In that case—even if the solution operator is only continuous for general fully non-linear families—the best achievable rates are no better than worst-case bounds for approximating non-linear locally-Lipschitz continuous operators, see [55], which require an exponential increase in trainable neurons to achieve a linear decrease in error. Thus, even when approximability holds, any such ‘rate’ would be scarcely more informative than a simple existence statement.

Consequently, the principal obstacle is approximability, which is twofold:

(i) the relevant solution operator in the fully non-linear elliptic case must be regular enough to be approximable by some universal deep-learning class;

(ii) our models must be universal on the specific function spaces on which this solution map acts.

(i) requires a stability analysis of our PDE family under coefficient perturbations, while (ii) calls for a universal approximation theorem for our architecture, proved via basis-expansion techniques as in Proposition A.14, akin in spirit to [47, Theorem 11], that holds on more general Besov spaces over regular Euclidean domains. This two-step scheme was introduced for deep learning in stochastic filtering [42] and refined for differential games in [5; 29].

⁴This is need as our proof relies on the approximation results of [45] for the relevant Green’s function associated to our PDEs.

3.3.2 Solutions to the family of second-order BSDEs

We now derive the stochastic version of the above (deterministic) approximation theorem. We additionally require the following regularity conditions.

Assumption 3.12 (Regularity of forward process). *There is some $x_0 \in \mathcal{D}$ such that for each $R > 0$*

- (i) (local smoothness): $(\beta, \gamma) \in C_b^\infty(B_{\mathbb{R}^d}(x_0, 5R); \mathbb{R}^d \times \mathbb{S}_d^+)^2$;
- (ii) (local ellipticity): $\gamma(x)\gamma(x)^\top \geq c_{x_0, R} \mathbf{I}_d$, for every $x \in B_{\mathbb{R}^d}(x_0, 3R)$, for some $0 < c_{x_0, R} < 1$;
- (iii) there exists a unique strong solution to (SDE).

Theorem 3.13. *Let Assumptions 3.8 to 3.10 and 3.12 hold, then, for any $0 < \varepsilon < 1$ and any time-window $0 < T_- < T_+$, there are integers L, W, Δ, H , and $\widehat{\Gamma} \in \mathcal{FB}_{\widehat{L}, \widehat{W}, \widehat{\sigma}}^{L, W, \text{ReQU}}$ satisfying*

$$\sup_{(f, g) \in \mathcal{B}} \mathbb{E}^\mathbb{P} \left[\sup_{\tau \wedge T_- \leq t \leq T_+ \wedge \tau} \left| \widehat{\Gamma}(f, g)_t - (Y_t^x, Z_t^x) \right| \right] \lesssim \varepsilon,$$

where the supremum is taken over the set

$$\mathcal{B} := B_{W_0^{1, \infty}(\mathcal{D}; \mathbb{R})}(0, \delta^2) \times B_{H^{1+(d+1)/2}(\partial \mathcal{D}; \mathbb{R})}(0, \delta^2).$$

We have the same estimates for the parameters $L = L(\Gamma)$, $W = W(\Gamma)$, $\widehat{L} = \widehat{L}(\Gamma)$, and $\widehat{W} = \widehat{W}(\Gamma)$ as in Theorem 3.11.

4 Experimental results

In this section, we empirically validate our theoretical findings on two canonical benchmarks in the 2BSDE literature: the periodic semi-linear example of Chassagneux, Chen, Frikha, and Zhou [13] and the linear-quadratic control example of Pham, Warin, and Germain [78]. We deploy the KANO architecture with a slight modification in the kernel layer (see C.3 for details). Specifically, rather than jointly learning both the kernel basis and its coefficients, we fix the basis to a Fourier system, obtained via uniform discretisation of the spatial domain, while retaining trainable, Res-KAN-parametrised coefficients. Furthermore, skip connections parametrised by additional Res-KAN layers are introduced on top of the learnable Fourier kernel coefficients. The resulting spectral layer follows the kernel introduced in Li, Kovachki, Azizzadenesheli, Liu, Bhattacharya, Stuart, and Anandkumar [60].

4.1 Periodic semi-linear case

In this experiment, we study the periodic semi-linear benchmark of [13] in dimension $d = 5$. This benchmark consists of trigonometric drift-diffusion and has a closed-form solution $u(t, x)$ depending on $\sum_{i=1}^5 x_i$. This enables exact supervision of $u, \nabla u, \nabla^2 u$ and pathwise validation under periodic boundary conditions. The forward-backward SDE system and its closed-form solution are detailed in Section C.1.

A KANO model is trained on 4096 samples drawn according to the procedure in Section C.4, and subsequently evaluated along independently generated trajectories using the Euler-Maruyama sampler described in Section C.5. Figures 3 and 4 display the projections of two randomly selected trajectories onto the (x_1, x_2) -plane, together with the corresponding ground-truth solutions u , first and second partial derivatives $\partial u / \partial x_1$ and $\partial^2 u / \partial x_1^2$, and the respective predictions produced by the trained model along these trajectories. We observe that the model is generally able to accurately capture the solution, as well as the first and second partial derivatives along the entire trajectories, with only minor discrepancies in the second derivatives.

4.2 Linear-quadratic case

We next consider the LQ/Hamilton-Jacobi-Bellman benchmark proposed in [78] in $d = 5$ (see Section C.2 for details). It represents a HJB-type problem with quadratic cost, whose value function remains quadratic $u(t, x) = x^\top K(t)x$, and where $K(t)$ satisfies a Riccati ODE. It offers analytic targets for $u, \nabla u, \nabla^2 u$ and a clean test of learning constant-in-space Hessians and optimal-feedback structure.

The same training and inference pipeline as described in the semi-linear case is used, with a KANO network trained on 4096 samples. **Figure 5** presents two random trajectories projected onto the (x_1, x_2) -plane. The figure also compares the analytic solution u , its gradient components $\partial u / \partial x_1$, and the diagonal Hessian entries $\partial^2 u / \partial x_1^2$ with the corresponding model predictions along these paths. The predicted values of u closely follow the analytical solution. The derivatives are recovered with satisfactory accuracy, and the Hessian, which is expected to remain constant in space, is also well captured. Although the estimated derivatives show some deviations from the smooth exact values, their overall accuracy remains high. In summary, the network effectively learns and reproduces the solution u and its derivatives along the generated trajectories.

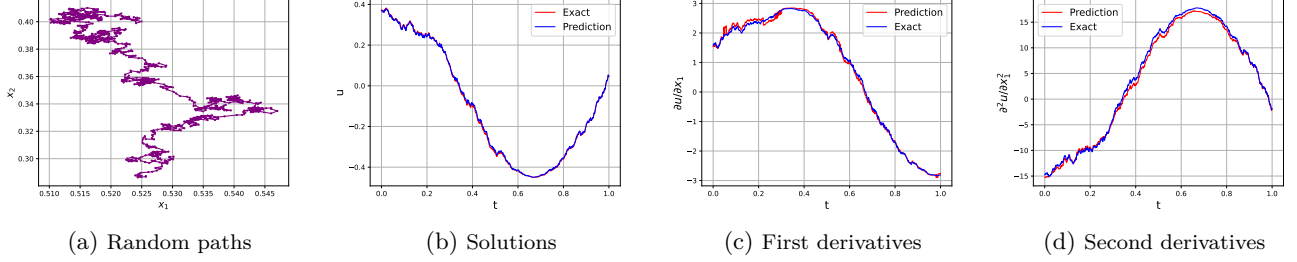


Figure 3: Ground-truth and KANO-predicted solutions for the first randomly selected trajectory of the periodic semilinear example from [13]. Each panel shows the projection onto the (x_1, x_2) -plane with u , $\partial u / \partial x_1$, and $\partial^2 u / \partial x_1^2$ along this path.

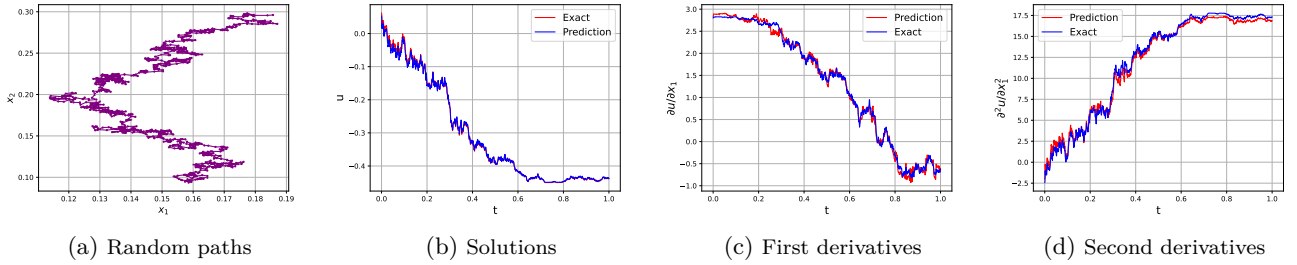


Figure 4: Continuation of **Figure 3**, showing the second randomly selected trajectory for the same semi-linear example.

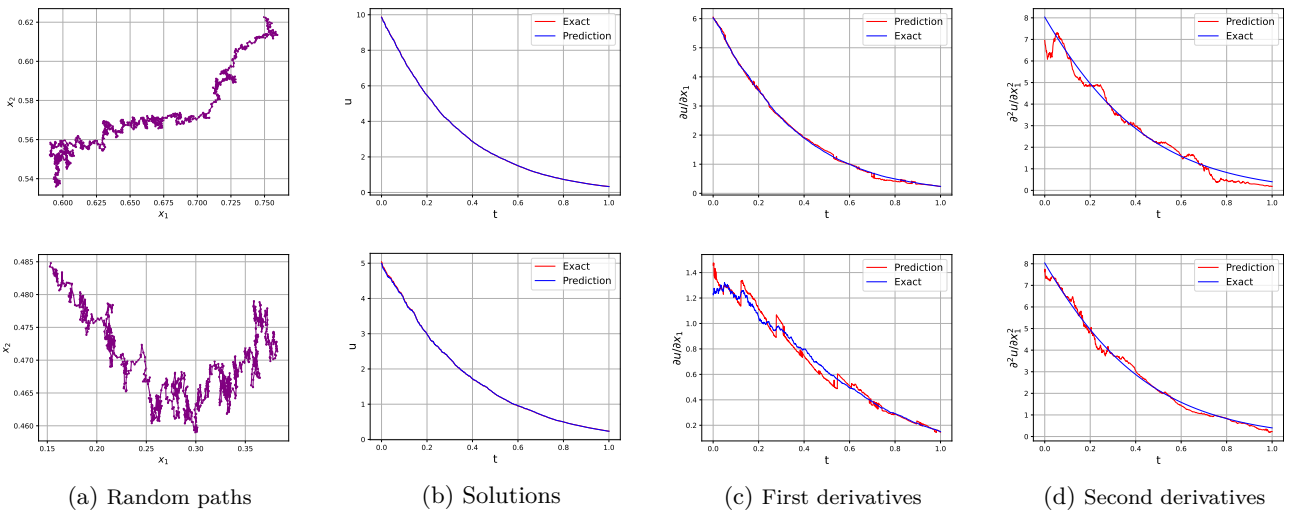


Figure 5: Comparison between the ground-truth and KANO-predicted solutions for the periodic linear-quadratic example of [78]. The figure shows two randomly selected trajectories projected onto the (x_1, x_2) -plane, together with the corresponding values of u , $\partial u / \partial x_1$, and $\partial^2 u / \partial x_1^2$ along these paths.

4.2.1 Ablation on the sample size

We next train a model using eight times fewer training samples than before *i.e.*, 512 samples) and evaluate it following the same procedure as in previous experiments. The resulting quantities of interest are shown in [Figure 6](#). We observe that in the vicinity of $t = 0$, the solution u is not well approximated, which in turn affects the accuracy of its first- and second-order partial derivatives. This behaviour is consistent with the theoretical discussion presented earlier: a sufficient number of training samples is required in the high-dimensional space $\mathbb{R}^d$ for the model to accurately capture the solution near $t = 0$.

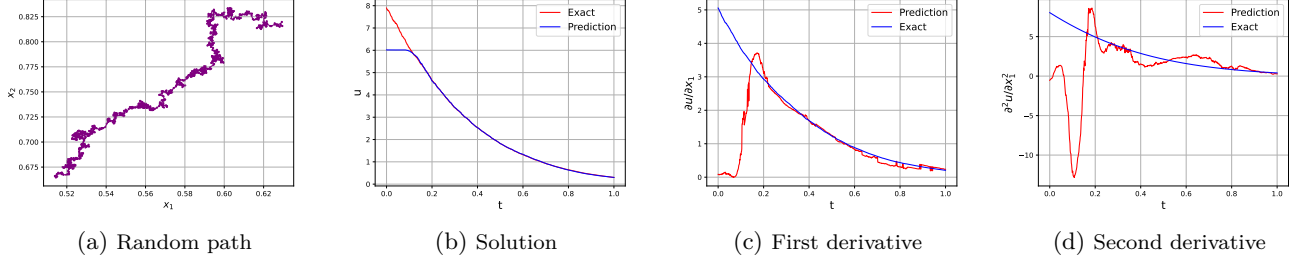


Figure 6: Comparison between the ground-truth and KANO-predicted solutions for the periodic linear-quadratic example of [\[78\]](#) in *low training data regime*. The figure shows two randomly selected trajectories projected onto the (x_1, x_2) -plane, together with the corresponding values of u , $\partial u / \partial x_1$, and $\partial^2 u / \partial x_1^2$ along these paths.

Acknowledgements

Takashi Furuya was supported by JSPS KAKENHI Grant Number JP24K16949, 25H01453, JST CREST JP-MJCR24Q5, JST ASPIRE JPMJAP2329. Anastasis Kratsios acknowledges financial support from an NSERC Discovery Grant No. RGPIN-2023-04482 and No. DGEGR-2023-00230, and they acknowledge that resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute⁵; they would also like to thank Behnoosh Zamanlooy for her support. Dylan Possamai gratefully acknowledges partial support by the SNF project MINT 205121-219818.

A Proof of PDE results

A.1 Proof of Theorem 3.11

This appendix contains the proofs of our paper’s main theoretical guarantees.

A.1.1 Well-posedness

Let $G_{\gamma, \mu, \lambda}(x, y)$ be a (real-valued) Green’s function for $-\nabla \cdot \gamma \nabla + \mu \cdot \nabla + \lambda$ with a Dirichlet boundary condition, *i.e.*, for $y \in \mathcal{D}$,

$$\begin{aligned} -\nabla \cdot \gamma \nabla G_{\gamma, \mu, \lambda}(\cdot, y) + \mu \cdot \nabla G_{\gamma, \mu, \lambda}(\cdot, y) + \lambda G_{\gamma, \mu, \lambda}(\cdot, y) &= -\delta(\cdot - y) \text{ in } \mathcal{D}, \\ G_{\gamma, \mu, \lambda}(\cdot, y) &= 0 \text{ on } \partial \mathcal{D}. \end{aligned}$$

Lemma A.1. *Let [Assumption 3.8](#) hold. Then, we have*

$$G_{\gamma, \mu, \lambda} \in W^{s, p}(\mathcal{D} \times \mathcal{D}; \mathbb{R}).$$

where $1 \leq p < \frac{d}{d-1}$ and $1 \leq s < 2$.

Proof. From [\[45, Theorem 8.1\]⁶](#), the Green function $G_{\gamma, \mu, \lambda}(x, y)$ for the operator $Lu := -\nabla \cdot \gamma \nabla u + \mu \cdot \nabla u + \lambda u$ can be estimated as for $\beta \in \mathbb{N}_0^d$ with $|\beta| \leq 1$

$$\|\partial_x^\beta G_{\gamma, \mu, \lambda}(x, y)\| \leq C_0 \|x - y\|^{1-d}, \quad (\text{A.1})$$

⁵<https://vectorinstitute.ai/partnerships/current-partners/>

⁶Note that our setting is that γ and μ are smooth. Thus, they are uniformly Dini continuous, which implies that they are of Dini mean oscillation.

where $C_0 > 0$ is a constant depending on $\mathcal{D}$, d , β , γ , μ , and λ . Also, applying [45, Theorem 8.1] to the Green function $g_{\gamma,\mu,\lambda}(y, x)$ for the adjoint operator $L^\top u = -\nabla \cdot (\gamma^\top \nabla u + \mu u) + \lambda$, the Green function $g_{\gamma,\mu,\lambda}(y, x)$ can be estimated, for $\beta \in (\mathbb{N}^*)^d$ with $\|\beta\| \leq 1$ by

$$\|\partial_y^\beta g_{\gamma,\mu,\lambda}(y, x)\| \leq C_0 \|y - x\|^{1-d}.$$

With [45, Proposition 6.13] and [Assumption 3.8.\(iii\)](#), we see that $G(x, y) = g(y, x)$ ($x \neq y$), which implies that

$$\|\partial_y^\beta G_{\gamma,\mu,\lambda}(x, y)\| \leq C_0 \|x - y\|^{1-d}. \quad (\text{A.2})$$

We now choose $R > 0$ such that $\mathcal{D} \subset B_{\mathbb{R}^d}(0, R)$. Using (A.2), we estimate that for $x \in \mathcal{D}$ and $\beta \in (\mathbb{N}^*)^d$ with $\|\beta\| \leq 1$

$$\begin{aligned} \int_{\mathcal{D}} \|\partial_x^\beta G_{\gamma,\mu,\lambda}(x, y)\|^p dy &\lesssim \int_{\mathcal{D}} \|x - y\|^{(1-d)p} dy = \int_{x-\mathcal{D}} \|z\|^{(1-d)p} dz \leq \int_{B_{\mathbb{R}^d}(0, 2R)} \|z\|^{(1-d)p} dz \\ &\lesssim \int_0^{2R} r^{(1-d)p} r^{d-1} dr = \int_0^{2R} r^{(d-1)(1-p)} dr \lesssim 1, \end{aligned} \quad (\text{A.3})$$

where we have used that $1 < p < \frac{d}{d-1}$. We can obtain the estimate for the derivative with respect to y similarly, using now (A.2). Note that we use the symbol $\lesssim$ to omit a multiplicative constant that is independent of x on the left-hand side. $\square$

Using the Green function $G_{\gamma,\mu,\lambda}(x, y)$, we define an integral operator encoding (1.5) by:

$$u(x) := \int_{\mathcal{D}} G_{\gamma,\mu,\lambda}(x, y) (\tilde{f}(y, u(y)) - f(y)) dy + w_g(x), \quad x \in \mathcal{D}, \quad (\text{A.4})$$

where $f_0 \in W^{1,\infty}(\mathcal{D}; \mathbb{R})$ and $w_g(x) \in W^{\frac{d+4}{2}, 2}(\mathcal{D}; \mathbb{R})$ is the unique solution of

$$-\nabla \cdot \gamma \nabla w_g + \mu \cdot \nabla w_g + \lambda w_g = 0, \text{ on } D, \quad w_g = g, \text{ on } \partial D.$$

where $g \in W^{\frac{d+3}{2}, 2}(\partial \mathcal{D})$. Note that, it is well known that a linear elliptic equation has the unique solution w_g (see, e.g., [37]). By the Sobolev embedding theorem (see, e.g., [Evans \[26, Section 5.6.3\]](#)) we have

$$W^{(d+4)/2, 2}(\mathcal{D}) \subset C^{(d+4)/2-d/2-1, \xi_0}(\overline{\mathcal{D}}) \subset W^{1,\infty}(\mathcal{D}),$$

where $0 < \xi_0 < 1$ is a constant. Hence, $w_g \in W^{1,\infty}(\mathcal{D})$. We define next the mapping T by

$$T(u)(x) := \int_{\mathcal{D}} G_{\gamma,\mu,\lambda}(x, y) (\tilde{f}(y, u(y)) - f_0(y)) dy + w_g(x), \quad x \in \mathcal{D},$$

We set

$$B_{W^{1,\infty}}(0, \delta) := \{u \in W^{1,\infty}(\mathcal{D}; \mathbb{R}) : \|u\|_{W^{1,\infty}(\mathcal{D}; \mathbb{R})} \leq \delta\},$$

$$B_{W^{(d+3)/2, 2}}(0, \delta) := \{g \in W^{(d+3)/2, 2}(\partial \mathcal{D}; \mathbb{R}) : \|g\|_{W^{(d+3)/2, 2}(\partial \mathcal{D}; \mathbb{R})} \leq \delta\}.$$

Then, $B_{W^{1,\infty}}(0, \delta)$ is a closed subset in $W^{1,\infty}(\mathcal{D}; \mathbb{R})$.

Lemma A.2. *Let [Assumptions 3.8](#) to [3.10](#) hold. Let $f \in B_{W^{1,\infty}(\mathcal{D}; \mathbb{R})}(0, \delta^2)$ and $g \in B_{W^{(d+3)/2, 2}(\partial \mathcal{D}; \mathbb{R})}(0, \delta^2)$. Then, the map $T : B_{W^{1,\infty}}(0, \delta) \longrightarrow B_{W^{1,\infty}}(0, \delta)$ is a ρ -contraction where $\rho \in (0, 1)$ is defined in [Assumption 3.10](#). In particular, there exists a unique solution of (A.4) in $B_{W^{1,\infty}}(0, \delta)$.*

Proof. We see that for $x \in \mathcal{D}$

$$\begin{aligned} T(w)(x) &:= \int_{\mathcal{D}} G_{\gamma,\mu,\lambda}(x, y) [\tilde{f}(y, w(y)) - f_0(y)] dy + w_g(x) \\ &= \int_{\mathcal{D}} G_{\gamma,\mu,\lambda}(x, y) \left(\sum_{h=2}^H \frac{\partial_z^h \tilde{f}(y, 0)}{h!} w(y)^h - f_0(y) \right) dy + w_g(x) \\ &= \sum_{h=2}^H \frac{1}{h!} \int_{\mathcal{D}} G_{\gamma,\mu,\lambda}(x, y) \partial_z^h \tilde{f}(y, 0) w(y)^h dy - \int_{\mathcal{D}} G_{\gamma,\mu,\lambda}(x, y) f_0(y) dy + w_g(x). \end{aligned}$$

First, we will show that $T : B_{W^{1,\infty}}(0, \delta) \longrightarrow B_{W^{1,\infty}}(0, \delta)$. Let $w \in B_{W^{1,\infty}}(0, \delta)$. Using this, that f_0 , and w_g are both in $B_{W^{1,\infty}}(0, \delta^2)$, and [Lemma A.1](#), we see that for any $\beta \in (\mathbb{N}^*)^d$ with $\|\beta\| \leq 1$, we have

$$\begin{aligned} \|\partial_x^\beta T(w)(x)\| &\lesssim \int_{\mathcal{D}} \|\partial_x^\beta G_{\gamma,\mu,\lambda}(x, y)\| \left(\sum_{h=2}^H \frac{1}{h!} |w(y)|^h + |f_0(y)| \right) dy + \|\partial_x^\beta w_g(x)\| \\ &\lesssim \delta^2 \int_{\mathcal{D}} \|\partial_x^\beta G_{\gamma,\mu,\lambda}(x, y)\| dy + \delta^2 \lesssim \delta^2. \end{aligned} \quad (\text{A.5})$$

This means that $T(w) \in W^{1,\infty}(\mathcal{D}; \mathbb{R})$. We also see that

$$\|T(w)\|_{W^{1,\infty}(\mathcal{D}; \mathbb{R})} \leq C_1 \delta^2, \quad (\text{A.6})$$

where $C_1 > 0$ is a constant depending on $p, d, \mathcal{D}, \tilde{f}, \gamma$, and μ . By choosing $\delta > 0$ in [Assumption 3.10](#), we have $Tw \in B_{W^{1,\infty}}(0, \delta)$.

Next, we will show that $T : B_{W^{1,\infty}}(0, \delta) \longrightarrow B_{W^{1,\infty}}(0, \delta)$ is a contraction mapping. Let $(w_1, w_2) \in B_{W^{1,\infty}}(0, \delta) \times B_{W^{1,\infty}}(0, \delta)$. Since

$$w_1(y)^h - w_2(y)^h = \left(\sum_{i=0}^{h-1} w_1(y)^{h-1-i} w_2(y)^i \right) (w_1(y) - w_2(y)),$$

we deduce that for any $\beta \in (\mathbb{N}^*)^d$ with $\|\beta\| \leq 1$, by Hölder's inequality and [Lemma A.1](#)

$$\begin{aligned} \|\partial_x^\beta T(w_1)(x) - \partial_x^\beta T(w_2)(x)\| &\lesssim \sum_{h=2}^H \frac{1}{h!} \int_{\mathcal{D}} \|\partial_x^\beta G_{\gamma,\mu,\lambda}(x, y)\| |w_1(y)^h - w_2(y)^h| dy \\ &\leq \sum_{h=2}^H \frac{1}{h!} \sum_{i=0}^{h-1} \int_{\mathcal{D}} \|\partial_x^\beta G_{\gamma,\mu,\lambda}(x, y)\| |w_1(y)^{h-1-i} w_2(y)^i| |w_1(y) - w_2(y)| dy \\ &\leq \sum_{h=2}^H \frac{h}{h!} \delta^{h-1} \int_{\mathcal{D}} \|\partial_x^\beta G_{\gamma,\mu,\lambda}(x, y)\| \lesssim \delta \|w_1 - w_2\|_{W^{1,\infty}(\mathcal{D}; \mathbb{R})}. \end{aligned}$$

Then, we have that

$$\|T(w_1) - T(w_2)\|_{W^{1,\infty}(\mathcal{D}; \mathbb{R})} \leq C_2 \delta \|w_1 - w_2\|_{W^{1,\infty}(\mathcal{D}; \mathbb{R})} = \rho \|w_1 - w_2\|_{W^{1,\infty}(\mathcal{D}; \mathbb{R})}, \quad (\text{A.7})$$

where $C_2 > 0$ is a constant depending on $p, d, \mathcal{D}, \tilde{f}, \gamma$, and μ . By choosing $\delta > 0$ as in [Assumption 3.10](#), we have that T is ρ -contraction mapping in $B_{W^{1,\infty}}(0, \delta)$. $\square$

Given the previous result, and using Banach's fixed-point theorem, the following solution operator is well-defined

$$\begin{aligned} \Gamma^+ : B_{W^{1,\infty}}(0, \delta^2) \times B_{W^{(d+3)/2,2}}(0, \delta^2) &\longrightarrow B_{W^{1,\infty}}(0, \delta) \\ (f_0, g) &\longmapsto u, \end{aligned}$$

where, u is the unique solution of [Equation \(A.4\)](#) in $B_{W^{1,\infty}}(0, \delta)$.

A.1.2 Proof of Theorem 3.11

We now prove Theorem 3.11 in a series of several steps. Throughout, the activation function applied component-wise to the neural operator layers in neural operator's neurons, *i.e.* in (2.9), will always be taken to be the squared-ReLU function, that is to say $\beta = (1, 0, \dots, 0)$ in (2.1) for the neural operator.

Let $(f_0, g) \in B_{W^{1,\infty}}(0, \delta^2) \times B_{W^{(d+3)/2,2}}(0, \delta^2)$ and let $u \in B_{W^{1,\infty}}(0, \delta)$ be a solution of (A.4), that is, $\Gamma^+(f, g) = u$. By [48, Theorem 1], for any $\varepsilon > 0$, there exist Res-KANs, with representation as in [Definition 2.3](#), $k_{nn}^h : \mathbb{R}^d \longrightarrow \mathbb{R}$, $h \in \{2, \dots, H\}$, and $k'_{nn} : \mathbb{R}^d \longrightarrow \mathbb{R}$ such that

$$\left\| k_{nn}^h(x, y) - \frac{1}{h!} G_{\gamma,\mu,\lambda}(x, y) \partial_z^h \tilde{f}(y, 0) \right\|_{W_{x,y}^{1,p}(\mathcal{D} \times \mathcal{D}; \mathbb{R})} \leq \varepsilon, \quad h \in \{2, \dots, H\}, \quad (\text{A.8})$$

and

$$\|k'_{nn}(x, y) - G_{\gamma, \mu, \lambda}(x, y)\|_{W_{x, y}^{1, p}(\mathcal{D} \times \mathcal{D}; \mathbb{R})} \leq \varepsilon, \quad (\text{A.9})$$

where depths $\widehat{L}(k_{nn}^h)$ and $\widehat{L}(k'_{nn})$ are of order $\mathcal{O}(1)$, while the width of $\widehat{W}(k_{nn}^h)$ and $\widehat{W}(k'_{nn})$ are of order $\mathcal{O}(\varepsilon^{-\frac{1}{(s-1)p}})$. Then, we define by

$$\widehat{L} := \widehat{L}(\Gamma) := \max\{\widehat{L}(k_{nn}^1), \dots, \widehat{L}(k_{nn}^H), \widehat{L}(k'_{nn})\}, \quad \widehat{W} := \widehat{W}(\Gamma) := \max\{\widehat{W}(k_{nn}^1), \dots, \widehat{W}(k_{nn}^H), \widehat{W}(k'_{nn})\},$$

Then, they are estimated by

$$\begin{cases} \widehat{L} \leq C, \\ \widehat{W} \leq C\varepsilon^{-\frac{1}{(s-1)p}}, \end{cases} \quad (\text{A.10})$$

where $C > 0$ is a constant depending on d, s, H , and p . We can then define the map T_{NN} by

$$T_{NN}(u)(x) := \sum_{h=2}^H \int_{\mathcal{D}} k_{nn}^h(x, y)(u(y))^h dy - \int_{\mathcal{D}} k'_{nn}(x, y)f(y)dy + w_g(x). \quad (\text{A.11})$$

Lemma A.3. *There exists a constant $C_4 > 0$ depending on $p, d, \mathcal{D}, \gamma, \mu$, and λ such that for any $u \in B_{W^{1, \infty}(\mathcal{D}; \mathbb{R})}(0, \delta)$*

$$\|T(u) - T_{NN}(u)\|_{W^{1, \infty}(\mathcal{D}; \mathbb{R})} \leq C_4 \varepsilon.$$

Proof. Let $u \in B_{W^{1, \infty}(\mathcal{D}; \mathbb{R})}(0, \delta)$. We see that for $\beta \in (\mathbb{N}^*)^d$ with $\|\beta\| \leq 1$,

$$\begin{aligned} |\partial_x^\beta T(u)(x) - \partial_x^\beta T_{NN}(u)(x)| &\leq \sum_{h=2}^H \left\| k_{nn}^h(x, y) - \frac{1}{h!} G_{\gamma, \mu, \lambda}(x, y) \partial_z^h f(y, 0) \right\|_{W_{x, y}^{1, p}(\mathcal{D}; \mathbb{R})} \left(\int_{\mathcal{D}} |u(y)^h|^{p'} dy \right)^{1/p'} \\ &\quad + \|k'_{nn}(x, y) - G_{\gamma, \mu, \lambda}(x, y)\|_{W_{x, y}^{1, p}(\mathcal{D}; \mathbb{R})} \left(\int_{\mathcal{D}} |f(y)|^{p'} dy \right)^{1/p'} \leq C_4 \delta^2 \varepsilon < \varepsilon, \end{aligned} \quad (\text{A.12})$$

which is exactly the desired result. $\square$

Lemma A.4. *T_{NN} maps $B_{W^{1, \infty}}(0, \delta)$ to itself.*

Proof. Fix $u \in B_{W^{1, \infty}(\mathcal{D}; \mathbb{R})}(0, \delta)$. Using (A.6) and (A.12), we see that

$$\|T_{NN}(u)\|_{W^{1, \infty}(\mathcal{D}; \mathbb{R})} \leq \|T_{NN}(u)\|_{W^{1, \infty}(\mathcal{D}; \mathbb{R})} + \|T(u) - T_{NN}(u)\|_{W^{1, \infty}(\mathcal{D}; \mathbb{R})} \lesssim \delta^2.$$

Thus, we have that

$$\|T_{NN}(u)\|_{W^{1, \infty}(\mathcal{D}; \mathbb{R})} \leq C_3 \delta^2, \quad (\text{A.13})$$

where $C_3 > 0$ is a constant depending on $s, p, d, \mathcal{D}$, and γ . By the choice of δ in [Assumption 3.10](#), we see that $\|T_{NN}(u)\|_{W^{1, \infty}(\mathcal{D}; \mathbb{R})} \leq \delta$. $\square$

We can now define for an arbitrary positive integer J , the map $\Gamma_J : B_{W^{1, \infty}}(0, \delta^2) \times B_{W^{1, \infty}}(0, \delta^2) \longrightarrow W^{1, \infty}(\mathcal{D}; \mathbb{R})$ by

$$\Gamma_J(f_0, w_g) := \underbrace{T_{NN} \circ \dots \circ T_{NN}}_{J \text{ times}}(0) =: T_{NN}^{[J]}(0).$$

Lemma A.5. *Let $J := \lceil \log(1/\varepsilon) / \log(1/\rho) \rceil \in \mathbb{N}$. Then, there exists a constant $C_5 > 0$ depending on $p, d, \mathcal{D}, \gamma, \mu$, and λ such that for all $(f_0, g) \in B_{W^{1, \infty}}(0, \delta^2) \times B_{W^{(d+3)/2, 2}}(0, \delta^2)$*

$$\|\Gamma^+(f_0, g) - \Gamma_J(f_0, w_g)\|_{W^{1, \infty}(\mathcal{D})} \leq C_5 \varepsilon.$$

Proof. From [Lemma A.2](#), $T : B_{W^{1, \infty}}(0, \delta) \longrightarrow B_{W^{1, \infty}}(0, \delta)$ is ρ -contraction mapping, which implies that

$$\|\Gamma^+(f_0, g) - T^{[J]}(0)\|_{W^{1, \infty}(\mathcal{D}; \mathbb{R})} = \|T^{[J]}(u) - T^{[J]}(0)\|_{W^{1, \infty}(\mathcal{D}; \mathbb{R})} \lesssim \rho^J \|u\|_{W^{1, \infty}(\mathcal{D}; \mathbb{R})} \leq \rho^J \delta \lesssim \varepsilon, \quad (\text{A.14})$$

where u is the unique solution of (A.4) in $B_{W^{1,\infty}}(0, \delta)$. Next, we see that

$$\begin{aligned}
\|T^{[J]}(0) - \Gamma(f_0, w_g)\|_{W^{1,\infty}(\mathcal{D}; \mathbb{R})} &= \|T^{[J]}(0) - T_{NN}^{[J]}(0)\|_{W^{1,\infty}(\mathcal{D}; \mathbb{R})} \\
&\leq \sum_{h=1}^J \left\| (T^{[J-h+1]} \circ T_{NN}^{[h-1]})(0) - (T^{[J-h]} \circ T_{NN}^{[h]})(0) \right\|_{W^{1,\infty}(\mathcal{D}; \mathbb{R})} \\
&\leq \sum_{h=1}^J \rho^{J-h} \left\| (T \circ T_{NN}^{[h-1]})(0) - (T_{NN} \circ T_{NN}^{[h-1]})(0) \right\|_{W^{1,\infty}(\mathcal{D}; \mathbb{R})} \\
&= \sum_{h=1}^J \rho^{J-h} \|T(u_h) - T_{NN}(u_h)\|_{W^{1,\infty}(\mathcal{D}; \mathbb{R})}, \tag{A.15}
\end{aligned}$$

where, we see that, by Lemma A.4

$$u_h := T_{NN}^{[h-1]}(0) \in B_{W^{1,\infty}}(0, \delta).$$

Note that we define $T_{NN}^{[0]} := \text{Id}$. By Lemma A.3, we see that

$$\|T(u) - T_{NN}(u)\|_{W^{1,\infty}(\mathcal{D}; \mathbb{R})} \leq C_4 \varepsilon,$$

which implies that with (A.15)

$$\|T^{[J]}(0) - \Gamma(f_0, w_g)\|_{W^{1,\infty}(\mathcal{D}; \mathbb{R})} \leq \sum_{h=1}^J \rho^{J-h} C_5 \varepsilon \leq \sum_{h=0}^{\infty} \rho^h C_5 \varepsilon = \frac{C_5}{1-\rho} \varepsilon \lesssim \varepsilon. \tag{A.16}$$

Thus, by Equations (A.14) and (A.16), we conclude that

$$\|\Gamma^+(f_0, g) - \Gamma(f_0, w_g)\|_{W^{1,\infty}(\mathcal{D}; \mathbb{R})} \leq \|\Gamma^+(f_0, g) - T^{[J]}(0)\|_{W^{1,\infty}(\mathcal{D}; \mathbb{R})} + \|T^{[J]}(0) - \Gamma(f_0, w_g)\|_{W^{1,\infty}(\mathcal{D}; \mathbb{R})} \lesssim \varepsilon.$$

□

Let us remind the reader that Γ_J is defined by

$$\Gamma_J(f_0, w_g) = \underbrace{T_{NN} \circ \cdots \circ T_{NN}}_{J \text{ times}}(0) = T_{NN}^{[J]}(0).$$

where the operator T_{NN} is defined by

$$T_{NN}(u)(x) = \sum_{h=2}^H \int_{\mathcal{D}} k_{nn}^h(x, y) (u(y))^h dy - \int_{\mathcal{D}} k'_{nn}(x, y) f_0(y) dy + w_g(x) = \sum_{h=2}^H \int_{\mathcal{D}} k_{nn}^h(x, y) (u(y))^h dy + v_{f_0, g}(x)$$

where

$$v_{f_0, g}(x) := - \int_{\mathcal{D}} k'_{nn}(x, y) f_0(y) dy + w_g(x)$$

We see that $\Gamma_J(f_0, w_g)(x) = v_J(x)$ where $v_0 := 0$ and

$$v_{j+1}(x) := \sum_{h=2}^H \int_{\mathcal{D}} k_{nn}^h(x, y) (v_j(y))^h dy + v_{f_0, g}(x), \quad j \in \{0, \dots, J-1\}.$$

We define

$$W^{(0)} := \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix} \in \mathbb{R}^{2 \times 2},$$

and let $K_N^{(0)} : W^{1,\infty}(\mathcal{D}; \mathbb{R})^2 \longrightarrow W^{1,\infty}(\mathcal{D}; \mathbb{R})^2$ be defined by

$$\left(K^{(0)} \begin{pmatrix} f_0 \\ w_g \end{pmatrix} \right)(x) := \int_{\mathcal{D}} k_{NN}^{(0)}(x, y) \begin{pmatrix} f(y) \\ w_g(y) \end{pmatrix} dy,$$

where

$$k_{NN}^{(0)}(x, y) := \begin{pmatrix} k'_{NN}(x, y) & 0 \\ k'_{NN}(x, y) & 0 \end{pmatrix} \in \mathbb{R}^{2 \times 2}.$$

We therefore compute

$$W^{(0)} \begin{pmatrix} f_0(x) \\ w_g(x) \end{pmatrix} + \left(K^{(0)} \begin{pmatrix} f_0 \\ w_g \end{pmatrix} \right)(x) = \begin{pmatrix} v_{f_0, g}(x) \\ v_{f_0, g}(x) \end{pmatrix} = \begin{pmatrix} v_{f_0, g}(x) \\ v_1(x) \end{pmatrix}.$$

Next, we define $F_{\text{ReQU}} : \mathbb{R}^2 \rightarrow \mathbb{R}^H$ by

$$F_{\text{ReQU}}(u) := \begin{pmatrix} u_1 \\ (u_2)^2 \\ \vdots \\ (u_2)^H \end{pmatrix}, \quad u = (u_1, u_2) \in \mathbb{R}^2,$$

which can have an exact implementation by a ReQU neural networks (see [Li, Tang, and Yu \[59, Theorem 3.1\]](#)).

We define

$$W = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \end{pmatrix} \in \mathbb{R}^{2 \times H},$$

and $K : W^{1, \infty}(\mathcal{D}; \mathbb{R})^H \rightarrow W^{1, \infty}(\mathcal{D}; \mathbb{R})^2$, for $u = (u_1, \dots, u_H) \in W^{1, \infty}(\mathcal{D}; \mathbb{R})^{H+1}$

$$(Ku)(x) := \int_{\mathcal{D}} k_{NN}(x, y)u(y)dy = \begin{pmatrix} 0 \\ \sum_{h=2}^H \int_{\mathcal{D}} k_{nn}^h(x, y)u_h(y)dy \end{pmatrix},$$

where

$$k_{NN}(x, y) := \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & k_{nn}^2(x, y) & \cdots & k_{nn}^H(x, y) \end{pmatrix} \in \mathbb{R}^{2 \times H},$$

Then, we have that for $j \in \{1, \dots, J-1\}$

$$\begin{aligned} \left((W + K) \circ F_{\text{ReQU}} \begin{pmatrix} v_{f_0, g} \\ v_j \end{pmatrix} \right)(x) &= W \begin{pmatrix} v_{f_0, g}(x) \\ (v_j(x))^2 \\ \vdots \\ (v_j(x))^H \end{pmatrix} + K \begin{pmatrix} v_{f_0, g} \\ (v_j)^2 \\ \vdots \\ (v_j)^H \end{pmatrix}(x) = \begin{pmatrix} v_{f_0, g}(x) \\ \sum_{h=2}^H \int_{\mathcal{D}} k_{nn}^h(x, y)(v_j(y))^h dy + v_{f_0, g}(x) \end{pmatrix} \\ &= \begin{pmatrix} v_{f_0, g}(x) \\ v_{j+1}(x) \end{pmatrix}. \end{aligned}$$

Denoting $W' := (0, 1) \in \mathbb{R}^{1 \times 2}$, we finally obtain that

$$\Gamma_J(f, w_g) = W' \circ \underbrace{\left((W + K) \circ F_{\text{ReQU}} \circ \cdots \circ (W + K) \circ F_{\text{ReQU}} \right)}_{J-1 \text{ times}} \circ (W^{(0)} + K^{(0)}) \begin{pmatrix} f \\ w_g \end{pmatrix}.$$

Since the ReQU network can be represented by the KANs network [\[92, Theorem 3.2\]](#), we have, by the above construction,

$$\Gamma \in \mathcal{NO}_{L, W}^{L, W, I, \alpha}(W^{1, \infty}(\mathcal{D}; \mathbb{R})^2, W^{1, \infty}(\mathcal{D}; \mathbb{R})).$$

Moreover, the depth $L = L(\Gamma)$ and width $W = W(\Gamma)$ of the neural operator Γ can be estimated via

$$L(\Gamma) \lesssim J \leq C \log(\varepsilon^{-1}), \quad W(\Gamma) \lesssim H \leq C.$$

This concludes our proof of [Theorem 3.11](#); where, again, $\alpha = s$ and $I := \lceil \alpha \rceil$.

A.2 Proof of Theorem 3.7

The proof of our second main result relies on some tools from multi-resolution analysis and the wavelet theory of Besov spaces. We, therefore, now overview the necessary material.

A.2.1 Additional background

In what follows, we use $S(\mathbb{R}^d)$ to denote the Schwartz space on $\mathbb{R}^d$ and consider the space of distributions defined as the topological dual $D(\mathcal{D})'$. We define the restriction operator sending any distribution $g \in S(\mathbb{R}^d)$ to $g|_{\mathcal{D}} \in D(\mathcal{D})'$ defined by restriction of its action to test functions $\varphi \in D(\mathcal{D})$ *i.e.*

$$g|_{\mathcal{D}}(\varphi) := g(\varphi).$$

A.2.2 From wavelet para-bases to Besov spaces on Euclidean spaces

Fix $u \in \mathbb{N}$ and $(\sigma_S, \sigma_W) \in C^u(\mathbb{R}) \times C^u(\mathbb{R})$ satisfy [Assumption 2.1](#); that is to say that σ_S and σ_W are Daubechies father (also known as scaling function) and mother wavelets (also known as wavelet function) respectively, in the sense of [\[18\]](#). For each $j \in \mathbb{N}$ define the sets

$$G^j := \begin{cases} \{S, W\}^d, & \text{if } j = 0, \\ \{S, W\}^{d*} := \{S, W\}^d \setminus \{(S, \dots, S)\}, & \text{if } j > 0. \end{cases}$$

Now, for each ‘scale’ $j \in \mathbb{N}$, location $m \in \mathbb{Z}^d$, and each $G \in G^j$, define the tensorised Daubechies wavelet by

$$\tilde{\Psi}_{G,m}^j(x) := 2^{jd/2} \prod_{i=1}^d \sigma_{G_i}(2^j x_i - m_i), \quad x \in \mathbb{R}^d, \quad (\text{A.17})$$

where $G := (G_1, \dots, G_d)$. Let $\mathcal{O} := \{(j, G, m) : j \in \mathbb{N}, G \in G^j, m \in \mathbb{Z}^d\}$ and for each $(j, G, m) \in \mathcal{O}$ let

$$\frac{1}{(\beta_{G,m}^j)^2} := \int_{\mathbb{R}^d} (\tilde{\Psi}_{G,m}^j(x))^2 dx, \text{ and } \Psi_{G,m}^j := \frac{1}{\beta_{G,m}^j} \tilde{\Psi}_{G,m}^j(x), \quad x \in \mathbb{R}^d.$$

Then, as discussed on [Triebel \[90, page 13\]](#), for any $u \in \mathbb{N}$ we have that $(\Psi_{G,m}^j)_{(j,G,m) \in \mathcal{O}}$ is an orthonormal basis of $L^2(\mathbb{R}^d)$, and for every $f \in L^2(\mathbb{R}^d)$

$$f = \sum_{j \in \mathbb{N}} \sum_{G \in G^j} \sum_{m \in \mathbb{Z}^d} \lambda_{G,m}^j 2^{-jd/2} \Psi_{G,m}^j, \text{ where } \lambda_{G,m}^j := 2^{jd/2} \int_{\mathbb{R}^d} f(x) \Psi_{G,m}^j(x) dx, \quad (\text{A.18})$$

where the series converge in $L^2(\mathbb{R}^d)$.

A key properties of Besov spaces, from the approximation theoretic lense, is that they are entirely determined by the decay/convergence rates of the sequences $(\lambda_{G,m}^j)_{(j,G,m) \in \mathcal{O}}$, defined in [\(A.18\)](#). Indeed, for $(q, r) \in (0, +\infty]^2$ and $s \in \mathbb{R}$, if

$$u > \max\{s, \sigma_q - s\}, \text{ where } \sigma_q := d \max\left\{0, \frac{1}{q} - 1\right\}, \quad (\text{A.19})$$

as shown in [\[90, Theorem 1.20\]](#), $f \in S(\mathbb{R}^d)'$ belongs to the Besov space $\overline{B}_{q,r}^s(\mathbb{R}^d)$ if and only if the sequence $\lambda := (\lambda_{G,m}^j)_{(j,G,m) \in \mathcal{O}}$, defined by [\(A.18\)](#), satisfies

$$\|\lambda\|_{b_{q,r}^s}^r := \sum_{j=0}^{\infty} 2^{jr(s-d/q)} \sum_{G \in G^j} \left(\sum_{m \in \mathbb{Z}^d} |\lambda_{G,m}^j|^q \right)^{r/q} < \infty, \quad (\text{A.20})$$

with the usual modifications to the left-hand side of [\(A.20\)](#) if q or r are infinite. Additionally, the map $f \mapsto (2^{jd/2} \langle f, \Psi_{G,m}^j \rangle_{L^2(\mathbb{R}^d)})_{(j,G,m) \in \mathcal{O}}$ is a bi-Lipschitz linear isomorphism between $B_{q,r}^s(\mathbb{R}^d)$ and the (quasi-)Banach space $b_{q,r}^s$ of all sequences for which the (quasi-)norm $\|\cdot\|_{b_{q,r}^s}$ is finite.

A.2.3 Besov spaces on domains

We begin with the definition of Besov spaces on any domain (proper open set with non-empty interior) $O \subset \mathbb{R}^d$, with closure $\overline{O}$. We write $\mathcal{D}(O)$ for the space of complex-valued compactly supported smooth (test) functions on O , topologized with the canonical (Limit of Fréchet) LF-topology. Its dual space $D'(O)$ is the space of *distributions* on O , and a distribution $f \in D'(O)$ is said to be supported on a set $A \subseteq O$ if $f(\varphi) = 0$ for every $\varphi \in \mathcal{D}(O)$ such that $\varphi(x) = 0$ for all $x \notin A$; the support $\text{supp}(f)$ is the smallest closed set K with this property. For instance, if $x \in O$ then the Dirac distribution $\delta_x(\varphi) := \varphi(x)$ has support $\text{supp}(\delta_x) = \{x\}$, see [\[90, Chapter 2, page 28\]](#) for further details and notation. We now define the Besov (quasi-Banach) spaces on $\mathcal{D}$.

Definition A.6 (Besov spaces on domains). Let $\mathcal{D}$ be a domain, $(q, r) \in (0, +\infty]^2$, and $s \in \mathbb{R}$. The Besov space $\tilde{B}_{q,r}^s(\overline{\mathcal{D}})$ consists of all $f \in B_{q,r}^s(\mathbb{R}^d)$ supported in the closure $\overline{\mathcal{D}}$ and $\tilde{B}_{q,r}^s(\mathcal{D})$ consists of all distributions $f \in D(\mathcal{D})'$ for which there exists some $g \in B_{q,r}^s(\overline{\mathcal{D}})$ such that $f = g|_{\mathcal{D}}$. In either case, $\mathfrak{D} \in \{\mathcal{D}, \overline{\mathcal{D}}\}$, we equip $\tilde{B}_{q,r}^s(\mathfrak{D})$ with the interpolation norm

$$\|f\|_{\tilde{B}_{q,r}^s(\mathfrak{D})} := \inf \left\{ \|g\|_{B_{q,r}^s(\mathbb{R}^d)} : g \in \tilde{B}_{q,r}^s(\overline{\mathcal{D}}), f = g|_{\mathfrak{D}} \right\}.$$

We define the Besov spaces $\overline{B}_{q,r}^s(\mathcal{D})$ as follows

$$\overline{B}_{q,r}^s(\mathcal{D}) := \begin{cases} \tilde{B}_{q,r}^s(\mathcal{D}), & \text{if } 0 < q \leq \infty, 0 < r \leq \infty, s > \sigma_q, \\ B_{q,r}^0(\mathcal{D}), & \text{if } 1 < q < \infty, 0 < r \leq \infty, s = 0, \\ B_{q,r}^s(\mathcal{D}), & \text{if } 0 < q \leq \infty, 0 < r \leq \infty, s < 0. \end{cases} \quad (\text{A.21})$$

Following [90, Section 3], we now construct wavelet systems on arbitrary domains (open subsets $\Omega \subset \mathbb{R}^n$) using Whitney decompositions; an object which acts almost as a leitmotif in analysis from our PDE problems to fundamental result in the geometry of functions spaces [27; 28]. The idea is to partition Ω into dyadic cubes whose sizes adapt to the distance from the boundary, and then build localized wavelet bases on these cubes—maintaining the regularity and cancellation properties of classical $\mathbb{R}^n$ wavelets while conforming to the geometry of Ω .

These spaces can themselves be characterized in a similar way using compactly supported Daubechies wavelets. We fix a so-called *approximate lattice* $\mathbb{Z}_{\mathcal{D}} \subset \mathcal{D}$ consisting of points $\mathbb{Z}_{\mathcal{D}} = (x_r^j)_{(j,k) \in \mathbb{N} \times \{1, \dots, N_j\}}$ where, for each $j \in \mathbb{N}$, $N_j \in \overline{\mathbb{N}} := \mathbb{N} \cup \{\infty\}$ for which there exist positive constants c_1, c_2, c_3 satisfying the approximate ‘lattice separation condition’ at any scale $j \in \mathbb{N}$

$$|x_r^j - x_{r'}^j| \geq \frac{c_1}{2^j} \quad (\text{A.22})$$

and the separation from the ‘boundary condition’ at scale $j \in \mathbb{N}$

$$\inf_{\{z \in \mathbb{R}^d : \|z - x_r^j\| \leq c_2/2^j\}} \inf_{u \in \partial \mathcal{D}} \|z - u\| \geq \frac{c_3}{2^j}. \quad (\text{A.23})$$

Clearly the constants c_1, c_2 , and c_3 may be chosen to guarantee the existence of such a $\mathbb{Z}_{\mathcal{D}}$ for any domain $\mathcal{D}$. Intuitively, $\mathbb{Z}_{\mathcal{D}}$ acts precisely as the dyadic lattices $\bigcup_{j \in \mathbb{N}} 2^{-j} \mathbb{Z}^d$ does in $\mathbb{R}^d$ but is contained entirely within $\mathcal{D}$ and condition (A.22) vacuously holds when $\mathcal{D}$ is replaced by the Euclidean space.

For any $L \in \mathbb{N}$, to be specified retroactively, we denote $\sigma_s^L(\cdot) := \sigma_s(2^L \cdot)$, $\sigma_w^L(\cdot) := \sigma_w(2^L \cdot)$, and $\Psi_{G,m}^{j,L} := \Psi_{G,m}^j(2^L \cdot)$ for each $(j, G, m) \in \mathcal{O}$. In other words, the factor L rescales our setup and we will choose it so that our problem is properly ‘shrunk’ within our domain and aligned to the approximate lattice $\mathbb{Z}_{\mathcal{D}}$.

We are now ready to define wavelet classes tailored to general domains; we follow the terminology in [90, Definition 2.4], the existence of which is known (see e.g. [90, Theorem 2.33]).

Definition A.7 (*u-wavelets*). Let $\mathcal{D}$ be an arbitrary domain in $\mathbb{R}^n$ with $\mathcal{D} \neq \mathbb{R}^n$ and let $\mathbb{Z}_{\mathcal{D}}$ ads well as $L \in \mathbb{N}$ and $u \in \mathbb{N}$ be as above. Let $K \in \mathbb{N}$, $D > 0$ and $c_4 > 0$. Then, consider a sub-family of $\{\Psi_{G,m}^j : j \in \mathbb{N}_+, G \in G^j, m \in \mathbb{Z}_{\mathcal{D}}\}$

$$\{\Phi_r^j : j \in \mathbb{N}; r \in \{1, \dots, N_j\}\}, \text{ where } N_j \in \overline{\mathbb{N}}. \quad (\text{A.24})$$

satisfying: $\text{supp}(\Phi_r^j) \subset B_{\mathbb{R}^d}(x_r^j, c_2 2^{-j})$, $j \in \mathbb{N}$, is called a *u-wavelet system* (with respect to $\mathcal{D}$) if it consists of the following three possible types of functions

- (i) **basic wavelets:** $\Phi_r^0 = \Psi_{G,m}^{0,L}$ for some $G \in \{S, W\}^d$, and $m \in \mathbb{Z}^d$;
- (ii) **interior wavelets:** $\Phi_r^j = \Psi_{G,m}^{j,L}$ for each $j \in \mathbb{N}$, and $m \in \mathbb{Z}_{\mathcal{D}}$ such that $\text{dist}(x_r^j, \overline{\mathcal{D}}) \geq c_4 2^{-j}$, for some $G \in \{S, W\}^{d*}$;
- (iii) **boundary wavelets:** $\Phi_r^j = \sum_{\{m' \in \mathbb{Z}^d : \|m - m'\| \leq K\}} d_{m,m'}^j \Psi_{F,m'}^{j,L}$, for each $j \in \mathbb{N}$ for which $\text{dist}(x_r^j, \Gamma) < c_4 2^{-j}$, for some $m := m(j, r) \in \mathbb{Z}^d$ and $d_{m,m'}^j \in \mathbb{R}$ with

$$\sum_{\{m' \in \mathbb{Z}^d : \|m - m'\| \leq K\}} |d_{m,m'}^j| \leq D, \text{ and } \text{supp}(\Psi_{F,m'}^{j,L}) \subset B(x_r^j, c_2 2^{-j}). \quad (\text{A.25})$$

We may now adapt the definition of the sequence spaces $b_{q,r}^s$, given in (A.20), to suit the approximate lattice $\mathbb{Z}_{\mathcal{D}}$, and thus the domain $\mathcal{D}$.

Definition A.8 (Sequence space $b_{q,r}^s$). *Let $\mathcal{D}$ be an arbitrary domain in $\mathbb{R}^n$ with $\mathcal{D} \neq \mathbb{R}^n$, let $\mathbb{Z}_{\mathcal{D}}$ be as above, $s \in \mathbb{R}$, and $(q, r) \in (0, \infty]^2$. Then $b_{q,r}^s(\mathbb{Z}_{\mathcal{D}})$ is the collection of all sequences*

$$\lambda := \{\lambda_r^j \in \mathbb{C} : j \in \mathbb{N}, r \in \{1, \dots, N_j\}\}, \text{ for some } N_j \in \overline{\mathbb{N}}, \quad (\text{A.26})$$

such that

$$\|\lambda\|_{b_{q,r}^s(\mathbb{Z}_{\mathcal{D}})}^q := \sum_{j=0}^{\infty} 2^{j(s-n/q)r} \left(\sum_{k=1}^{N_j} |\lambda_k^j|^q \right)^{r/q} < \infty. \quad (\text{A.27})$$

As we will see shortly, the wavelet system in (A.24) is a Schauder basis for several Besov spaces on domains, provided these domains possess a basic level of generic ‘thickness’ and regularity of their boundaries. We begin by first noting the relationship between the Besov sequence and function spaces, with the same indices, if the domain has a regular enough boundary.

A domain $\mathcal{D} \subseteq \mathbb{R}^d$ is said to be *special Lipschitz* if there exists a Lipschitz-continuous map $\beta : \mathbb{R}^{d-1} \rightarrow \mathbb{R}$ such that

$$\mathcal{D} = \{(\tilde{x}, x_d) \in \mathbb{R}^{d-1} \times \mathbb{R} : \beta(x) < x_d\}.$$

A *bounded Lipschitz domain* $\mathcal{D} \subset \mathbb{R}^d$ is a bounded domain $\mathcal{D}$ for which there exists a finite number of open balls $(B_1, \dots, B_N)$, for some $N \in \mathbb{N}^*$, where for $n \in \{1, \dots, N\}$ we have

$$B_n := \{x \in \mathbb{R}^d : \|x - x^{(n)}\| < r^{(n)}\}, \text{ for some } x^{(n)} \in \partial\mathcal{D}, \text{ and some } r^{(n)} > 0,$$

such that $(B_n)_{n \in \{1, \dots, N\}}$ is a cover of $\partial\mathcal{D}$, and there exist rotations of special Lipschitz domains $(\mathcal{D}_1, \dots, \mathcal{D}_N) \subseteq (\mathbb{R}^d)^N$ for which

$$B_n \cap \mathcal{D} = B_n \cap \mathcal{D}_n, \quad n \in \{1, \dots, N\}.$$

Now, given any domain with Lipschitz boundary, we may characterise the inclusion of any square-integrable function into a wide array of Besov spaces depending on its associated sequence λ belonging to the ‘little Besov’ sequence space with the same indices. The following result is [89, Corollary 4.28].

Lemma A.9 (Wavelet para-bases in Besov and Triebel–Lizorkin spaces on bounded Lipschitz domains). *Fix $(q, r) \in (1, \infty)^2$. For $K > 0$ small enough, if $5d/2 < K$ and $s \in (-K, K)$ then $f \in \mathcal{D}(\mathcal{D})'$ belongs to $\overline{B}_{q,r}^s(\mathcal{D})$ (resp. $\overline{F}_{q,r}^s(\mathcal{D})$) if and only if it admits the representation*

$$f = \sum_{(j,G,m) \in S^{\mathcal{D}}} \lambda_{G,m}^j 2^{-jd/2} \Psi_{G,m}^j, \quad (\text{A.28})$$

and the following holds

$$\left\| \left(2^{j(s-d/q)} \left\| (\lambda_{G,m}^j)_{(G,m) \in S_j^{\mathcal{D}}} \right\|_{\ell^q} \right)_{j \in \mathbb{N}} \right\|_{\ell^p} < \infty.$$

In what follows, given any $f \in \overline{B}_{q,r}^s$ we write $\lambda(f) := (\lambda_{G,m}^j)_{j,G,m \in S^{\mathcal{D}}}$ for the sequence defined in (A.28); provided that it is unique. We denote the linear map $f \mapsto \lambda(f)$ by I .

Lemma A.9 does not guarantee that the wavelet expansions themselves are uniquely determined. In general, these wavelet ‘bases’ are only frames. However, the next result shows that this is not necessarily the case for E -thick domains.

We say that a domain is *exterior thick*, or E -thick for short, if there are constants $0 < c_L \leq c_U$ and $j_0 \geq 0$ such that for every $j \in \mathbb{N}$ with $j \geq j_0$, there is a d -dimensional ‘interior’ cube $Q \subset \mathcal{D}$ with side-length

$$c_L 2^{-j} \leq \max \left\{ \ell(Q), \sup_{z \in Q^i} \inf_{u \in \partial\mathcal{D}} \|z - u\| \right\} \leq c_U 2^{-j}$$

where Q^i denotes the interior of any cube Q in the norm relative topology on $\mathcal{D}$ and $\ell(Q)$ denotes its side-length; i.e. $\ell(Q) := \sup_{x,y \in Q} \|x - y\|_{\infty}$; where $\|\cdot\|_{\infty}$ denotes the ∞ -norm on $\mathbb{R}^d$. In the case of a thick exterior domain, we obtain a Schauder basis using our u -wavelet expansion, see [90, Theorem 3.13 (ii)].

Theorem A.10 (Wavelet-based Schauder bases). *Let $\mathcal{D}$ be an E -thick domain in $\mathbb{R}^d$. Define for $u \in \mathbb{N}^*$*

$$\{\Phi_r^j : j \in \mathbb{N}, r \in \{1, \dots, N_j\}\}, \text{ for some } N_j \in \mathbb{N},$$

an orthonormal u -wavelet basis in $L_2(\mathcal{D})$. Then let $\overline{B}_{q,r}^s(\mathcal{D})$ be the space in [89, Equation (3.46)] and let

$$u > \max \{s, \sigma_{q,r} - s\}, \quad s \neq 0.$$

Then $f \in \mathcal{D}'(\mathcal{D})$ is an element of $\overline{B}_{q,r}^s(\mathcal{D})$ if and only if it can be represented as

$$f = \sum_{j=0}^{\infty} \sum_{k=1}^{N_j} \lambda_k^j 2^{-jd/2} \Phi_k^j, \quad \lambda \in b_{q,r}^s(\mathbb{Z}_{\mathcal{D}}),$$

with convergence holding in $\mathcal{D}'(\mathcal{D})$ and locally in any spaces $\overline{B}_{q,r}^{\sigma}(\mathcal{D})$ with $\sigma_{q,r} < s$. Furthermore, if $f \in \overline{B}_{q,r}^s(\mathcal{D})$ then the representation is unique with $\lambda = \lambda(f)$ as in (A.28) and I the linear map in Lemma A.9 is a bi-Lipschitz isomorphism of Banach spaces mapping $\overline{B}_{q,r}^s(\mathcal{D})$ onto $b_{q,r}^s(\mathbb{Z}_{\mathcal{D}})$. If, in addition, $q < \infty, r < \infty$, then $(\Phi_k^j)_{\{(j,k) \in \mathbb{N}^2 : k \in \{1, \dots, N_j\}\}}$ is an unconditional basis in $\overline{B}_{q,r}^s(\mathcal{D})$.

Having covered the necessary background, we now prove our universal approximation result, see Proposition A.14 below.

A.3 Proof of universal approximation

We now express the previous result in terms of neural networks.

Lemma A.11 (Wavelet implementation on domains). *Let $\mathcal{D}$ be a bounded domain with Lipschitz boundary⁷, let σ_w and σ_s satisfy Assumption 2.1 and $s \geq 2$. Let $G \in \{S, W\}^{d^*}$, $j \in \mathbb{N}$, and $m \in \mathbb{Z}_{\mathcal{D}}$. Then there exists a Res-KAN $\widehat{\Psi}_{G,m}^j : \mathbb{R}^d \rightarrow \mathbb{R}$ of depth d , width at-most $2d+1$, and using at-most $(5d^2 + 25d + 2)/2$ non-zero parameters satisfying*

$$\Psi_{G,m}^j(x) = \widehat{\Psi}_{G,m}^j, \quad x \in \mathbb{R}^d.$$

Our proof will use a recent result, [48, Lemma 1], which shows that the d -ary multiplication operator can be *exactly* implemented using Res-KANs, but only locally. This is in contrast to ReLU MLPs, which can only *approximate* it locally.

Lemma A.12 (Exact multiplication on arbitrarily large hypercubes). *For every $d \in \mathbb{N}^*$ and each $M > 0$, there exists a Res-KAN $\times_d^2 : \mathbb{R}^d \rightarrow \mathbb{R}$ satisfying for each $x \in [-M, M]^d$*

$$\times_d^2(x) = \prod_{i=1}^d x_i.$$

Moreover $\times_d^2$ has depth d , width at-most $2d+1$, and at-most $(5d^2 + 21d)/2$ non-zero parameters.

We can now proceed with the

Proof of Lemma A.11. Recall that Assumption 2.1, implies that σ_s in (2.1) is a scaling function (father wavelet) and σ_w in (2.1) is the corresponding mother wavelet. In fact, by Assumption 2.1, both are Daubechies wavelets and are thus in $C^u(\mathbb{R})$ and *compactly supported*. By their continuity, they are thus bounded. Whence, there is some $M > 0$ such that $\sigma_G(\mathbb{R}) \subseteq [-M, M]$ for each $G \in \{S, W\}$.

Consequently, for every specification $G = (G_1, \dots, G_d) \in \{S, W\}^{d^*}$, for every $j \in \mathbb{Z}$, we may represent the (multivariate) Daubechies wavelet $\Psi_{G,m}^j$, defined by rescaling the associated un-normalised wavelet $\widetilde{\Psi}_{G,m}^j$ in (A.17), by

⁷The following result holds, more general on (ϵ, δ) -domains and thus on any Lipschitz domain; however, we will not need that level of generality in the remainder of our paper.

$$\begin{aligned}\Psi_{G,m}^j(\cdot) &= \prod_{i=1}^d \frac{2^{jd/2}}{\beta_{G,W}^j} \sigma_{G_i}(2^{jd/2} \cdot -m) = \left(\prod_{i=1}^d \frac{2^{jd/2}}{\beta_{G,W}^j} \right) \prod_{i=1}^d \sigma_{G_i}(W_0^j \cdot -m) =: \kappa_{G,W}^j \prod_{i=1}^d \sigma_{G_i}(W_0^j \cdot -m) \\ &= \kappa_{G,M}^j \times_d^2 \circ \sigma_{G_i}(W_0^j \cdot -m),\end{aligned}\quad (\text{A.29})$$

where $\beta_{G,W}^j := \|\Psi_{G,m}^j\|_{L^2(\mathbb{R})}$ where $W_0^j := 2^{jd/2} \mathbf{I}_d$, $m \in \mathbb{Z}^d$ and where (A.29) holds by [48, Lemma 1] (having chosen M large enough); where $\times_d^2 : \mathbb{R}^d \rightarrow \mathbb{R}$ is a Res-KAN with depth d , width at-most $2d+1$, and at-most $\frac{5d^2+21d}{2}$ non-zero parameters.

Now, making use of the chosen structure of the ‘non-spline’ factor in our trainable activation function $\sigma_{\beta:I}$ in Equation (2.1), for each $i \in \{1, \dots, d\}$, if $G_i = S$ we set $\beta_i = (1) \oplus 0_{I+1}$ and if $G_i = W$ we set $\beta_i = (0) \oplus (1) \oplus 0_I$. Then, (A.29) can be re-expressed as

$$\Psi_{G,m}^j := \kappa_{G,W}^j \times_d^2 \circ \sigma_{G_i}(W_0^j \cdot -m) \quad (\text{A.30})$$

Now by [48, Lemma 1], $\times_d^2$ can be implemented by a ReLU MLP of depth d , width $2d+1$, and using at-most $(5d^2 + 21d)/2$ non-zero parameters. Consequently, $\times_d^2$ is representable/implementable by a ReLU MLP with depth d , width at-most $2d+1$, and using at-most $(5d^2 + 25d + 2)/2$ non-zero parameters. $\square$

A direct consequence of the previous result is the following.

Proposition A.13 (Res-KAN basis of Besov spaces). *Let $\mathcal{D}$ be a bounded exterior-thick domain, $(q, r) \in (1, \infty)^2$, and $s \geq 2$. Then, there is a Schauder basis*

$$\{\widehat{\Phi}_k^j : j \in \mathbb{N}, k \in \{1, \dots, N_j\}\}, \text{ for some } N_j \in \overline{\mathbb{N}}, \quad (\text{A.31})$$

of $\overline{B}_{q,r}^s(\mathcal{D})$ consisting of u -wavelets. Moreover, for each such k, j , $\widehat{\Phi}_k^j$ is implementable by a Res-KAN of depth d , width at-most $2d+1$, and using at-most $(5d^2 + 25d + 2)/2$ non-zero parameters.

Proof. This is a direct consequence of Lemma A.11, Definition A.7, and of [90, Theorem 3.13 (ii)]. $\square$

We now prove the universality of our models in the class of Hölder continuous maps between Besov spaces; recall the notation (2.12). We write $\text{Hld}(\overline{B}_{q,r}^s(\mathcal{D}), \overline{B}_{q,r}^s(\mathcal{D}))$ for the set of all α -Hölder continuous maps from $\overline{B}_{q,r}^s(\mathcal{D})$ to itself, for some $0 < \alpha \leq 1$.

Proposition A.14 (Universal approximation). *Let $d \in \mathbb{N}_+$, $s > 0$, and $\mathcal{D}$ be a bounded exterior-thick domain in $\mathbb{R}^d$, $(q, r) \in (1, \infty)^2$ and $2 \leq s$, and let $I := \lceil s \rceil$. If σ_s and σ_w satisfy Assumption 2.1, then $\mathcal{NO}_{I,\alpha}$ is dense in $\text{Hld}(\overline{B}_{q,r}^s(\mathcal{D}), \overline{B}_{q,r}^2(\mathcal{D}))$ for the (relative) topology induced by the topology of uniform convergence on compact sets.*

Proof. Since $\mathcal{D}$ is exterior-thick, $s \geq 2$, $(q, r) \in (1, \infty)^2$, σ_s and σ_w satisfy Assumption 2.1, and we set $I := \lceil s \rceil$ then, Proposition A.13 guarantees that we may exhibit a Schauder basis of $\overline{B}_{q,r}^s(\mathcal{D})$ consisting only of Res-KANs, as in (A.31).

Pick an enumeration $(\widehat{\Psi}_{k_\ell}^{j_\ell})_{\ell \in \mathbb{N}}$ thereof. Now, let $\mathfrak{F}$ consist of all functions $\widehat{F} : \overline{B}_{q,r}^s(\mathcal{D}) \rightarrow \overline{B}_{q,r}^2(\mathcal{D})$ of the form in [33, Equation 16] and [33, Definition 6 (Neural filters)]

$$\widehat{F} := \left(\widehat{\Psi}_{k_1}^{j_1}, \dots, \widehat{\Psi}_{k_K}^{j_K} \right)^\top \widehat{f}_{\text{ReLU}} \circ \begin{pmatrix} \int_{\mathbb{R}^d} f(x) \widehat{\Psi}_{k_1}^{j_1} dx \\ \vdots \\ \int_{\mathbb{R}^d} f(x) \widehat{\Psi}_{k_K}^{j_K} dx \end{pmatrix} \quad (\text{A.32})$$

for some $K \in \mathbb{N}^*$, and where $\widehat{f}_{\text{ReLU}} : \mathbb{R}^K \rightarrow \mathbb{R}^K$ is a ReLU feed-forward neural network defined as iteratively mapping any $x \in \mathbb{R}^K$ to the vector $\widehat{f}_{\text{ReLU}}(x) := x_{L+1}$ defined recursively by

$$\begin{aligned}x_{L+1} &:= W_{L+1} x_L \in \mathbb{R}^{d_{L+1}} := \mathbb{R}^{d_K} \\ x_{\ell+1} &:= \text{ReLU}(W_\ell x_\ell + b_\ell) \in \mathbb{R}^{d_{\ell+1}}, \quad x \in \mathbb{R}^K, \quad L \in \mathbb{N}_+, \quad \text{for } \ell \in \{0, \dots, L\} \\ x_0 &:= x \in \mathbb{R}^{d_0} := \mathbb{R}^{d_K}.\end{aligned}\quad (\text{A.33})$$

where the layer widths are $(d_0, \dots, d_{L+1}) \in (\mathbb{N}_+)^{L+2}$, $K = d_0 = d_{L+1}$, and for each such ℓ , we have $W_\ell \in \mathbb{R}^{d_{\ell+1} \times d_\ell}$, as well as $b_\ell \in \mathbb{R}^{d_{\ell+1}}$.

Since $(\hat{\Psi}_{k_\ell}^{j_\ell})_{\ell \in \mathbb{N}}$ is a Schauder basis of the Banach space $\bar{B}_{q,r}^s(\mathcal{D})$ and of $\bar{B}_{q,r}^2(\mathcal{D})$ then [33, Theorem 1] implies that $\mathfrak{F}$ is dense in $\text{Hld}(\bar{B}_{q,r}^s(\mathcal{D}), \bar{B}_{q,r}^2(\mathcal{D}))$ for the (relative) topology induced by the topology of uniform convergence on compact sets. In other words, for every compact $\mathcal{K} \subseteq \bar{B}_{q,r}^s(\mathcal{D})$, every $\varepsilon > 0$, and $0 < \alpha \leq 1$, and every α -Hölder continuous map $f : \bar{B}_{q,r}^s(\mathcal{D}) \rightarrow \bar{B}_{q,r}^2(\mathcal{D})$, there is some $\hat{F} \in \mathfrak{F}$ satisfying

$$\sup_{u \in \mathcal{K}} \|F(u) - \hat{F}(u)\|_{W^{2,p}(\mathcal{D})} < \varepsilon. \quad (\text{A.34})$$

To deduce our claim, we will show that $\mathfrak{F} \subseteq \mathcal{NO}_{I,\alpha}$. Let $\hat{F}$ be an arbitrary element of $\mathfrak{F}$, which thus admits a representation as in (A.32).

Now, for every $\ell \in \{0, \dots, L-1\}$, let $b^\ell(x) := \mathbf{0}_{(d+d_{\ell+1}) \times (d+d_{\ell+1})}x + \mathbf{0}_d \oplus b_\ell$ be a constant Res-KAN, see Equation (2.7), where $\mathbf{0}_{(d+d_{\ell+1}) \times (d+d_{\ell+1})}x$ is the $(d+d_{\ell+1}) \times (d+d_{\ell+1})$ zero matrix and $\mathbf{0}_d \in \mathbb{R}^d$ is the zero vector therein. Now, for every $\ell \in \{1, \dots, L-1\}$ define the matrix $W^\ell := \mathbf{0}_{d \times d} \otimes W_\ell$, where $\otimes$ denotes the Kronecker product and let $W^L := (0_{K \times d} | W_L)$ denotes the column-wise concatenation of the matrix $0_{K \times d}$ with the matrix W_L . Now, for each $\ell \in \{1, \dots, L\}$ let $\beta_\ell := (0, 0, 1, 0, \dots, 0) \in \mathbb{R}^{d_{\ell+1}+2}$. With these specifications, we see that the KANO Γ with representation (2.4) (where $d_{\text{in}} = 1$ and $d_{\text{out}} = 1$) is exactly equal to $\hat{F}$. We have thus shown that $\mathfrak{F} \subseteq \mathcal{NO}_{I,\alpha}$, which concludes our proof. $\square$

A.4 Stability estimate of general solution operator

Lemma A.15 (Linear stability of perturbations to PDE). *Under Assumptions 3.3 and 3.4, if $r > 0$ and $k > 1 + \max\{1, d/p\}$ then there exists a constant $L_{2,k,\mathcal{D}} > 0$ such that the non-linear operator*

$$\begin{aligned} \Gamma_{\text{Gen}} : \mathcal{X}_k(r) &\longrightarrow W_p^2(\mathcal{D}) \\ (\bar{G}_0, g) &\longmapsto u_{\bar{G}_0, g}, \end{aligned} \quad (\text{A.35})$$

is $L_{2,k,\mathcal{D}}$ -Lipschitz continuous.

Proof. Under Assumptions 3.3 and 3.4 we may apply [54, Theorem 14.1.3] to deduce that for every $((\bar{G}_0, g), (\bar{G}'_0, g')) \in \mathcal{X} \times \mathcal{X}$ and the respective solutions $u_{\bar{G}_0, g}, u_{\bar{G}'_0, g'}$ (which exist by [54, Theorem 14.1.5]) to their elliptic PDE in (3.5) with $G + \bar{G}_0$ and $G + \bar{G}'_0$ respectively instead of G , we have the estimate

$$\|u_{\bar{G}_0, g} - u_{\bar{G}'_0, g'}\|_{W_p^2(\mathcal{D})} \lesssim \|\bar{G}_0 - \bar{G}'_0\|_{L^p(\mathcal{D})} + \|g - g'\|_{W^{2,p}(\mathcal{D})} + \|u_{\bar{G}_0, g} - u_{\bar{G}'_0, g'}\|_{C(\mathcal{D})}, \quad (\text{A.36})$$

where $\lesssim$ suppress a multiplicative constant depending only on $c_1, c_2, R_0, \delta, L_F, \omega_F$, and on the domain $\mathcal{D}$. Next, applying [54, Lemma 6.6.10] we deduce that there is an absolute constant $C > 0$ such that $\|u_{\bar{G}_0, g} - u_{\bar{G}'_0, g'}\|_{C(\mathcal{D})} \leq C \sup_{x \in \partial \mathcal{D}} |g(x) - g'(x)| = \|g - g'\|_{C(\partial \mathcal{D})}$. Consequently, (A.36) may be bounded above by

$$\begin{aligned} \|u_{\bar{G}_0, g} - u_{\bar{G}'_0, g'}\|_{W_p^2(\mathcal{D})} &\lesssim \|\bar{G}_0 - \bar{G}'_0\|_{L^p(\mathcal{D})} + \|g - g'\|_{W^{2,p}(\mathcal{D})} + \|g - g'\|_{C(\partial \mathcal{D})} \\ &\leq \|\bar{G}_0 - \bar{G}'_0\|_{W^{2,p}(\mathcal{D})} + \|g - g'\|_{W^{2,p}(\mathcal{D})} + \|g - g'\|_{C(\partial \mathcal{D})} \\ &\leq \|\bar{G}_0 - \bar{G}'_0\|_{W^{2,p}(\mathcal{D})} + \|g - g'\|_{W^{2,p}(\mathcal{D})} + \|g - g'\|_{C(\mathcal{D})} \\ &\leq \|\bar{G}_0 - \bar{G}'_0\|_{W^{2,p}(\mathcal{D})} + \|g - g'\|_{W^{2,p}(\mathcal{D})} + \|g - g'\|_{W^{k,p}(\mathcal{D})} \\ &\leq \tilde{C}_{2,k,\mathcal{D}} \|\bar{G}_0 - \bar{G}'_0\|_{W^{k,p}(\mathcal{D})} + \tilde{C}_{2,k,\mathcal{D}} \|g - g'\|_{W^{k,p}(\mathcal{D})} + \|g - g'\|_{W^{k,p}(\mathcal{D})} \\ &\leq L_{2,k,\mathcal{D}} \left(\|\bar{G}_0 - \bar{G}'_0\|_{W^{k,p}(\mathcal{D})} + \|g - g'\|_{W^{k,p}(\mathcal{D})} \right), \end{aligned}$$

where we used in the fourth line the Sobolev embedding Theorem [26, Section 5.6.3], which holds provided that $k \leq 1 + \lceil \frac{d}{p} \rceil$, where the existence of the constant $\tilde{C}_{2,k,\mathcal{D}} > 0$ (which only depends on 2, k , and on $\mathcal{D}$) as well as the validity of the fifth line are ensured since we have assumed that $2 < k$ so that the Rellich—Kondrachov Theorem [88, Proposition 4.4] implies that $W^{2,p}(\mathcal{D})$ is compactly embedded in $W^{k,p}(\mathcal{D})$, and $C := 2\tilde{C}_{2,k,\mathcal{D}} + 1 > 1$. $\square$

We are now ready to establish our approximability result for the solution operator corresponding to the more general class of fully non-linear elliptic PDEs.

Proof of Theorem 3.7. Under **Assumptions 3.3** and **3.4**, **Lemma A.15** applies and guarantees that the non-linear operator Γ_{Gen} , defined in (A.35), is $L_{2,k,\mathcal{D}}$ -Lipschitz continuous on $\mathcal{X}_k(r)$. Now, since $2 < k < \infty$ and σ_S and σ_W satisfy **Assumption 2.1**, we may apply **Proposition A.14** to deduce that for every $\varepsilon > 0$ and every non-empty compact subset $\mathcal{X} \subseteq \mathcal{X}_k(r)$ (in the relative topology induced by inclusion in $W^{2,p}(\mathcal{D}) \times W^{k,p}(\mathcal{D})$) equipped with the norm topology) there exists $\hat{\Gamma} \in \mathcal{N}_{\lceil k \rceil,1}$ satisfying the uniform estimate

$$\sup_{(\bar{G}_0, g) \in \mathcal{X}} \|\Gamma_{\text{Gen}}(\bar{G}_0, g) - \hat{\Gamma}(\bar{G}_0, g)\|_{W^{2,p}(\mathcal{D})} < \varepsilon. \quad (\text{A.37})$$

Noting that, by definition, $u_{\bar{G}_0, g} = \Gamma_{\text{Gen}}(\bar{G}_0, g)$ for each $(\bar{G}_0, g) \in \mathcal{X}$ concludes the proof. $\square$

B Proof of stochastic results

To derive the stochastic counterparts of our results, we emphasise that our approach does not rely on any unconventional lifting channels—such as those introduced in [31]—which are non-standard within the operator learning literature and were originally proposed to enforce additional smoothness. Instead, we are able to combine the Bernstein and Sobolev inequalities with Itô-type formulas in a compatible manner, without imposing excessive smoothness assumptions on the PDE solutions. This is achieved through the following transfer principle, which requires conditions we borrow from de Marco [21].

Assumption B.1 (Regularity of the forward process). *(i) there is $\eta \geq 0$ such that μ and γ in (SDE) are of class C^∞ on $\mathbb{R}^d \setminus \overline{B_{\mathbb{R}^d}(0, \eta)}$. Moreover, for every $R > 0$ and $x_0 \in \mathbb{R}^d$, μ and γ are smooth on $B_{\mathbb{R}^d}(x_0, 3R) \subset \mathbb{R}^d \setminus \overline{B_{\mathbb{R}^d}(0, \eta)}$;*

(ii) there exist positive exponents q and $\bar{q} > 0$, as well as constants $0 < C_0 < 1$, $C_k > 0$ (for every multi-index α with $|\alpha| = k \geq 1$) such that

$$|\partial_\alpha \mu^i(x)| + |\partial_\alpha \gamma^{i,j}(x)| \leq C_k(1 + \|x\|^q), \quad x \in \mathbb{R}^d, \quad (i, j) \in \{1, \dots, d\}^2, \quad (\text{B.1})$$

$$C_0 \|x\|^{-\bar{q}} \mathbf{I}_d \leq \gamma(x) \gamma(x)^\top, \quad \|x\| > \eta; \quad (\text{B.2})$$

(iii) for every $p > 0$, $\sup_{0 \leq s \leq t} \mathbb{E}^\mathbb{P}[\|X_s\|^p] < \infty$;

(iv) (SDE) admits a strong solution.

Under these conditions, the process X admits for every $t \in (0, T]$ a smooth density satisfying some Gaussian-type decay and derivative bounds, as shown in [21, Theorem 2.2]. In what follows, if it exists, for any time $t \geq 0$, we denote the density of the law X_t with respect to the Lebesgue measure on $B_R(y_0)$, for any $y_0 \in \mathcal{D}$ and $R > 0$, by $\rho_{t, y_0} \in L^1(B_R(y_0); [0, \infty))$, where

$$L^1(B_R(y_0); [0, \infty)) := \{u \in L^1(B_R(y_0)) : u(x) \geq 0, \text{ Lebesgue-a.e.}\}.$$

Lemma B.2 (Transfer trick). *Let $1 \leq s < \infty$, $1 \leq r \leq \infty$, $x_0 \in \mathcal{D}$ be such that $\mathcal{D} \subseteq B_R(x_0)$ be a compact domain, and $(u, \hat{u}) \in W^{s,r}(\mathcal{D}) \times W^{s,r}(\mathcal{D})$ be such that*

$$\|u - \hat{u}\|_{W^{s,r}(\mathcal{D})} \leq \varepsilon. \quad (\text{B.3})$$

Suppose that X satisfies (SDE) and Assumption B.1 and τ is the first exit time of X from $\mathcal{D}$. If r is finite, then additionally assume that there is some $0 < \delta_{\mathcal{D}}$ such that $d(0, \mathcal{D}) := \inf_{x \in \mathcal{D}} \|x\|_2 \geq \delta_{\mathcal{D}}$ and fix a time-window $0 < T_- < T_+$. Then

$$\begin{aligned} \mathbb{E}^\mathbb{P} \left[\int_{T_-}^{T_+} \sum_{|\beta| \leq s} \|D^\beta u(X_t) - D^\beta \hat{u}(X_t)\| dt \right] &\lesssim_{r, T_+, \mathcal{D}} \varepsilon \left(C_{T_+} + \frac{1}{T_-^{3d/2-1}} \right), \text{ if } 1 \leq r < \infty, \\ \text{esssup}^\mathbb{P} \left\{ \sup_{0 \leq t \leq \tau} \|D^\beta u(X_t(\omega)) - D^\beta \hat{u}(X_t(\omega))\| \right\} &\leq \varepsilon, \text{ if } r = \infty, \end{aligned} \quad (\text{B.4})$$

where $C_{T_+} > 0$ is a constant depending only on T_+ .

Proof. For the case where $r = \infty$, simply note that $X_{t \vee \tau} \in \mathcal{D}$. $\mathbb{P}$ -a.s. Thus, for $\mathbb{P}$ -almost every $\omega \in \Omega$ we have that

$$\sum_{|\beta| \leq s} \|D^\beta u(X_t(\omega)) - D^\beta \hat{u}(X_t(\omega))\| \leq \sup_{x \in \mathcal{D}} \|D^\beta(u - \hat{u})(x)\| = \|u - \hat{u}\|_{W^{s,r}(\mathcal{D})} \leq \varepsilon,$$

where the last inequality holds since $s \geq 1$. Consequently, (B.4) holds.

We now turn our attention to the case where $1 \leq r < \infty$. Define $\tau^* := T_+ \wedge (\tau \vee T_-)$. Note that, if $t \in [T_-, T_+]$ then $X_{t \wedge \tau^*} \in \mathcal{D}$, $\mathbb{P}$ -a.s. In particular, since $\mathcal{D}$ is bounded, then for any $t \geq 0$, $X_{t \wedge \tau^*} \in L^\infty([0, T_+] \times \Omega, \mathbb{R}^d)$; whence, we may apply the Fubini–Tonelli theorem to deduce that

$$\mathbb{E}^\mathbb{P} \left[\int_{T_-}^{T_+} \sum_{|\beta| \leq s} \|D^\beta u(X_t) - D^\beta \hat{u}(X_t)\| dt \right] = \int_{T_-}^{T_+} \mathbb{E}^\mathbb{P} \left[\sum_{|\beta| \leq s} \|D^\beta u(X_t) - D^\beta \hat{u}(X_t)\| \right] dt. \quad (\text{B.5})$$

Now, since we are operating under **Assumption B.1**, we may apply [21, Theorem 2.2] to show that $\rho_{t,x_0} \in L_+^1(B_R(x_0))$ exists and there is a constant $C_{r,T_+} > 0$, depending only on r and T_+ , such that for every $x \in B_R(x_0)$ we have

$$|\rho_{t,x_0}(x)| \leq C_{r,T_+} \left(1 + \frac{1}{t^{3d/2}} \right) \|x\|^{-r}. \quad (\text{B.6})$$

In particular, since $\mathcal{D} \subseteq B_R(x_0)$ then (B.6) holds for every $x \in \mathcal{D}$. Consequently, (B.5) and (B.6) imply that

$$\begin{aligned} \mathbb{E}^\mathbb{P} \left[\int_{T_-}^{T_+} \sum_{|\beta| \leq s} \|D^\beta u(X_t) - D^\beta \hat{u}(X_t)\| dt \right] &= \int_{T_-}^{T_+} \int_{\mathcal{D}} p_{t,x_0}(x) \sum_{|\beta| \leq s} \|D^\beta u(x) - D^\beta \hat{u}(x)\| dx dt \\ &\leq \int_{T_-}^{T_+} \left(\int_{\mathcal{D}} p_{t,x_0}(x)^{r'} dx \right)^{1/r'} \\ &\quad \times \left(\int_{\mathcal{D}} \sum_{|\beta| \leq s} \|D^\beta u(x) - D^\beta \hat{u}(x)\|^r dx \right)^{1/r} dt \\ &\leq \int_{T_-}^{T_+} \left(\int_{\mathcal{D}} C_{r,T_+}^{r'} \left(1 + \frac{1}{t^{3d/2}} \right)^{r'} \|x\|^{-(rr')} dx \right)^{1/r'} \\ &\quad \times \left(\int_{\mathcal{D}} \sum_{|\beta| \leq s} \|D^\beta u(x) - D^\beta \hat{u}(x)\|^r dx \right)^{1/r} dt, \end{aligned}$$

where the second line follows by Hölder's inequality with $\frac{1}{r} + \frac{1}{r'} = 1$ (since $1 < r < \infty$). Now, the term

$$\left(\int_{\mathcal{D}} \sum_{|\beta| \leq s} \|D^\beta u(x) - D^\beta \hat{u}(x)\|^r dx \right)^{1/r},$$

is precisely the $W^{[s],r}(\mathcal{D})$ norm of $(u - \hat{u})$, which is bounded above by the $W^{s,r}(\mathcal{D})$ -norm, which in turn is bounded above by ε , recall (B.3). Hence

$$\begin{aligned} \mathbb{E}^\mathbb{P} \left[\int_{T_-}^{T_+} \sum_{|\beta| \leq s} \|D^\beta u(X_t) - D^\beta \hat{u}(X_t)\| dt \right] &\leq \varepsilon \int_{T_-}^{T_+} \left(\int_{\mathcal{D}} C_{r,T_+}^{r'} \left(1 + \frac{1}{t^{3d/2}} \right)^{r'} \|x\|^{-(rr')} dx \right)^{1/r'} dt \\ &\leq C_{r,T_+} \varepsilon \frac{\text{Vol}(\mathcal{D})^{1/r'}}{\delta_{\mathcal{D}}^r} \int_{T_-}^{T_+} \left(1 + \frac{1}{t^{3d/2}} \right) dt \\ &\leq C_{r,T_+} \varepsilon \frac{\text{Vol}(\mathcal{D})^{1/r'}}{\delta_{\mathcal{D}}^r} \left(T_+ - T_- + \frac{T_-^{1-3d/2} - T_+^{1-3d/2}}{3d/2 - 1} \right) \\ &\leq \varepsilon C_{p,T_+, \mathcal{D}} \left(C_{T_+} + \frac{1}{T_-^{3d/2-1}} \right) \end{aligned}$$

where we used the assumption that $d(\mathcal{D}, 0) \geq \delta_{\mathcal{D}} > 0$ and a simple supremum-bound, and where we defined

$$C_{p, T_+, \mathcal{D}} := C_{p, T_+} \frac{2\text{Vol}(\mathcal{D})^{1/r'}}{(3d-2)\delta_{\mathcal{D}}^r}, \text{ and } C_{T_+} := \left(\frac{3d}{2} - 1\right)T_+.$$

□

C Experimental details

C.1 Periodic semi-linear case

We consider a periodic example from [13] in $d = 5$ dimension, with $T = 1$, in which the forward SDE is given by

$$dX_t^{(i)} = b_i(X_t^{(i)}) dt + \sigma_{i,i}(X_t^{(i)}) dW_t^{(i)}, \quad i \in \{1, \dots, d\},$$

and the coefficients of the SDE are given by

$$b_i(x) := 0.2 \sin(2\pi x_i), \quad \sigma_{i,j}(x) := \frac{1}{\sqrt{d}\pi} \left(0.25 + 0.1 \cos(2\pi x_i)\right) \mathbf{1}_{\{i=j\}}, \quad (i, j) \in \{1, \dots, d\}^2.$$

The coefficients of the backward SDE

$$dY_t = -f(t, X_t, Y_t, Z_t) dt + Z_t \cdot dW_t, \quad Y_T = g(X_T),$$

are given by

$$g(x) := \frac{1}{\pi} \left(\sin \left(2\pi \sum_{i=1}^d x_i \right) + \cos \left(2\pi \sum_{i=1}^d x_i \right) \right),$$

$$f(t, x, y, z) := 2\pi^2 y \sum_{i=1}^d \sigma_{i,i}(x)^2 - \sum_{i=1}^d \frac{b_i(x)}{\sigma_{i,i}(x)} z_i + h(t, x),$$

where

$$h(t, x) := 2 \left(\cos \left(2\pi \sum_{i=1}^d x_i + 2\pi(T-t) \right) - \sin \left(2\pi \sum_{i=1}^d x_i + 2\pi(T-t) \right) \right).$$

The explicit solution u is given by

$$u(t, x) = \frac{1}{\pi} (\sin(\theta(t, x)) + \cos(\theta(t, x))),$$

where

$$\theta(t, x) := 2\pi \left(\sum_{i=1}^d x_i + (T-t) \right).$$

The spatial derivatives of u are given by

$$\frac{\partial u}{\partial x_i}(t, x) = 2(\cos(\theta(t, x)) - \sin(\theta(t, x))), \quad i \in \{1, \dots, d\},$$

and

$$\frac{\partial^2 u}{\partial x_i \partial x_j}(t, x) = -4\pi (\sin(\theta(t, x)) + \cos(\theta(t, x))), \quad (i, j) \in \{1, \dots, d\}^2.$$

C.2 Linear–quadratic (LQ) case

We consider a linear–quadratic case from [78] in $d = 5$ dimension, with $T = 1$. The forward SDE is a controlled process X_t in $\mathbb{R}^d$, defined by

$$dX_t = (AX_t + B\alpha_t)dt + D\alpha_t dW_t,$$

where α_t is a control process in $\mathbb{R}$, $(B, D) \in \mathbb{R}^d \times \mathbb{R}^d$ and $A \in \mathbb{R}^{d \times d}$. The quadratic cost that is minimised is

$$J(\alpha) := \mathbb{E} \text{bigg} \left[\int_0^T (X_t^\top Q X_t + \alpha_t^2 N) dt + X_T^\top P X_T \right],$$

where P and Q are non-negative, symmetric $d \times d$ matrices and $N > 0$.

The Bellman PDE associated with this process admits an explicit solution given by a quadratic form

$$u(t, x) = x^\top K(t)x,$$

where $K(t)$ solves the Riccati equation

$$\dot{K} + A^\top K + KA + Q - \frac{KBB^\top K}{N + D^\top KD} = 0, \quad K(T) = P.$$

In all the simulations, we set

$$A = \mathbf{I}_d, \quad B = D = \mathbf{I}_d, \quad Q = P = \frac{1}{d}\mathbf{I}_d, \quad N = d.$$

The stochastic coefficients associated to the controlled process are set to

$$\sigma = \frac{1}{\sqrt{d}}\mathbf{I}_d, \quad \text{and } \mu(t, x) = x.$$

In our isotropic setup, the Riccati matrix remains proportional to the identity, *i.e.*

$$K(t) = k(t)\mathbf{I}_d.$$

Then, the explicit forms of the spatial derivatives of u are given by

$$\nabla_x u(t, x) = 2K(t)x = 2k(t)x, \quad D_x^2 u(t, x) = 2K(t) = 2k(t)\mathbf{I}_d.$$

To compute the solution u and its derivatives, we employ a fourth-order Runge–Kutta (RK4) scheme to numerically approximate $K(t)$ (the solution of the Riccati equation).

C.3 Architectural details

The KANO architecture follows a lift–process–project design. The input features are first lifted to a higher-dimensional latent space using a feed-forward network, producing an initial latent representation $v^{(0)}$.

After lifting, a composition of several KANO blocks is applied to iteratively refine this latent field:

$$v^{(\ell+1)} = \Phi^{(\ell)}(v^{(\ell)}, x), \quad \ell \in \{0, \dots, L-1\},$$

where each block $\Phi^{(\ell)}$ performs a structured operator update combining coordinate encoding, spectral convolution, and residual connection. Each KANO block consists of three main components

1. a **positional encoder** maps the spatial coordinates through a Res–KAN network, producing coordinate-dependent features

$$v_{\text{pos}} = b(x);$$

2. a **spectral kernel path** performs a spectral convolution in the frequency domain, analogous to the Fourier neural operator (FNO) [60]. Specifically, the feature field is transformed via a two-dimensional fast Fourier transform (FFT), filtered by learnable complex-valued multipliers, and then mapped back to the spatial domain

$$v_{\text{kf}}(x) = \mathcal{F}^{-1}(\hat{W}(k)\mathcal{F}[v_{\text{in}}](k)),$$

where $\mathcal{F}$ and $\mathcal{F}^{-1}$ denote the forward and inverse Fourier transforms, and $\hat{W}(k)$ are learnable complex weights restricted to a finite number of Fourier modes and parametrised as Res–KANs;

3. a **residual path** applies a Res-KAN transformation on the tensor obtained by concatenating $(v_{\text{pos}}, v_{\text{kf}}, v_{\text{in}})$.

After stacking L such KANO blocks, the resulting field $v^{(L)}$ is projected back to the target dimension through a final projection layer. This composition enables multiscale feature extraction, efficient global coupling through spectral convolution, and local adaptivity through Res-KAN-based non-linear transformations.

We restrict our training to a 2D uniform grid that spans the first two coordinates of the d -dimensional space, while conditioning the model pointwise on the remaining $d - 2$ coordinates. The procedure for generating random training samples is described in detail in [Section C.4](#). Our model is trained to approximate 2D slices of the solution along the (x_1, x_2) -coordinates in $\mathbb{R}_+ \times \mathbb{R}^d$. Once trained, the model can be evaluated at any point in time and space by approximating the solution over these 2D slices and querying the corresponding (x_1, x_2) values (see [Section C.5](#) for details). This type of restricted operator learning is efficient due to the following reasons.

- **Uniform grids enable efficient kernels.** During training, the coordinates (x_1, x_2) are placed on a uniform grid, enabling convolution-like kernel layers to be computed efficiently via FFTs. This reduces the per-layer complexity from dense $O(s^4)$ to $O(s^2 \log s)$, making spectral kernels both computationally efficient and numerically stable.
- **Learning high-dimensional maps through 2D evaluations.** The operator is evaluated over the full s^2 grid simultaneously, while the remaining coordinates $(x_3, \dots, x_d)$ and time t are provided as additional input channels. This setup allows the network to capture intrinsic symmetries in the problem and to perform restricted operator learning, approximating $u(t, x)$ across $\mathbb{R}^d$ by predicting values at multiple 2D locations in parallel.
- **2D offers the optimal balance; 3D becomes costly.** Extending the FFT-based grid to three dimensions increases computational and memory demands to $O(s^3 \log s)$ per pass and substantially raises activation and storage costs. In practice, 2D grids strike the best balance between expressivity (capturing many spatial query points per sample) and efficiency, while still encoding d -dimensional dependencies through the auxiliary input channels.

Note that spectral convolution on uniform grids is employed to improve the training efficiency of the model. In operator learning settings, various efficient kernel architectures exist, see [Kovachki, Li, Liu, Azizzadenesheli, Bhattacharya, Stuart, and Anandkumar \[47\]](#), including convolution-based kernels, see [Raonić, Molinaro, de Ryck, Rohner, Bartolucci, Alaifari, Mishra, and de Bézenac \[81\]](#), wavelet-based kernels, see [Tripura and Chakraborty \[91\]](#), and transformer-based kernels, see [Herde, Raonić, Rohner, Käppli, Molinaro, de Bézenac, and Mishra \[40\]](#) or [Li, Meidani, and Farimani \[61\]](#), among others. The choice of the spectral kernel here is made solely to demonstrate that training a neural operator in the 2BSDE setting is feasible.

C.4 Training pipeline

In all our experiments, we draw samples from the domain uniformly. To draw a random training sample, we first draw a random time, as well as random locations for the $d - 2$ dimensions (the first 2 dimensions (x_1, x_2) are already sampled on uniform grids),

$$t \in [0, T], \quad c = (x_3, \dots, x_d) \in [0, 1)^{d-2}.$$

To get the training samples, we evaluate the model on a *uniform* $s \times s$ grid for the first two coordinates

$$\mathcal{G} := \left\{ (x_1^p, x_2^q) : x_1^p = \frac{p}{s-1}, \quad x_2^q = \frac{q}{s-1}, \quad (p, q) \in \{0, \dots, s-1\}^2 \right\},$$

and denote $N := s^2$ and $X := ((x_{1n}, x_{2n}))_{n \in \{1, \dots, N\}}$ the grid.

At each grid node n , the model receives the feature vector

$$\phi_n := (t, X, x_3, \dots, x_d) \in \mathbb{R}^{1+2+(d-2)} = \mathbb{R}^{d+1},$$

i.e. time and the $(d - 2)$ extra coordinates are *channels* constant across the 2d grid. A neural operator F_θ maps these inputs to the $\mathbb{R}^{s \times s}$ field,

$$\hat{u}_\theta(t, X, x_3, \dots, x_d) = F_\theta(\phi_n) \in \mathbb{R}^{s \times s}.$$

C.5 Inference pipeline

At test time, the learned approximation $\hat{u}_\theta$ can be evaluated at any query (t, x) in the domain by either of the following.

- *Spectral/Fourier synthesis.* If the decoder is spectral, we evaluate the Fourier-like synthesis operator at the desired coordinates to obtain $\hat{u}_\theta(t, x)$ directly. This is naturally suited to periodic problems and preserves differentiability with respect to (t, x) , enabling gradients to be obtained by automatic differentiation.
- *Grid interpolation.* When the model outputs values on a uniform $s \times s$ grid in (x_1, x_2) at a given time t , we interpolate that grid to any (x_1, x_2) in the domain (*e.g.* bilinear/bicubic interpolation). This route is simple, fast, and it requires no change to the trained model.

To evaluate the models along random paths, we generate d -dimensional SDE trajectories using the Euler–Maruyama scheme,

$$X_{n+1}^{(i)} = X_n^{(i)} + b_i(X_n^{(i)})\Delta t + \sigma_{i,i}(X_n^{(i)})\sqrt{\Delta t}\xi_n^{(i)}, \quad \xi_n^{(i)} \sim \mathcal{N}(0, 1).$$

The trained model is then evaluated along these trajectories, and its predictions are compared against the exact solution u and its first- and second-order partial derivatives. Derivatives of the neural operator are approximated using first-order finite difference scheme. To obtain model outputs at arbitrary spatial locations, we employ bilinear interpolation over the (x_1, x_2) grid.

References

- [1] B. Acciaio, A. Kratsios, and G. Pammer. Designing universal causal deep learning models: the geometric (hyper) transformer. *Mathematical Finance*, 34(2):671–735, 2024.
- [2] B. Adcock, S. Brugiapaglia, N. Dexter, and S. Moraga. On efficient algorithms for computing near-best polynomial approximations to high-dimensional, Hilbert-valued functions from limited samples. *ArXiv preprint arXiv:2203.13908*, 2022.
- [3] B. Adcock, N. Dexter, and S. Moraga Scheuermann. Optimal deep learning of holomorphic operators between banach spaces. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Proceedings of the 38th conference on advances in neural information processing systems (NeurIPS 2024)*, December 10–15, 2024, Vancouver, British Columbia, Canada, volume 37, pages 27725–27789, 2024.
- [4] B. Adcock, S. Brugiapaglia, N. Dexter, and S. Moraga. Near-optimal learning of Banach-valued, high-dimensional functions via deep neural networks. *Neural Networks*, 181(106761):1–25, 2025.
- [5] G. Alvarez, I. Ekren, A. Kratsios, and X. Yang. Neural operators can play dynamic Stackelberg games. *ArXiv preprint arXiv:2411.09644*, 2024.
- [6] R. Arabpour, J. Armstrong, L. Galimberti, A. Kratsios, and G. Livieri. Low-dimensional approximations of the conditional law of Volterra processes: a non-positive curvature approach. *ArXiv preprint arXiv:2405.20094*, 2024.
- [7] C. Beck, W. E, and A. Jentzen. Machine learning approximation algorithms for high-dimensional fully nonlinear partial differential equations and second-order backward stochastic differential equations. *Journal of Nonlinear Science*, 29(4):1563–1619, 2019.
- [8] F. E. Benth, N. Detering, and L. Galimberti. Neural networks in Fréchet spaces. *Annals of Mathematics and Artificial Intelligence*, 91(1):75–103, 2023.
- [9] E. Bilokopytov and F. Xanthos. A universal approximation theorem and its applications to vector lattice theory. *ArXiv preprint arXiv:2507.20219*, 2025.
- [10] Helmut Bolcskei, Philipp Grohs, Gitta Kutyniok, and Philipp Petersen. Optimal approximation with sparsely connected deep neural networks. *SIAM Journal on Mathematics of Data Science*, 1(1):8–45, 2019.
- [11] D. Cao and J. Wan. Expansion of Green’s function and regularity of Robin’s function for elliptic operators in divergence form. *Annali della Scuola Normale Superiore di Pisa - Classe di Scienze*, to appear, 2022.

- [12] T.-H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, and Y. Ma. PCANet: simple deep learning baseline for image classification? *IEEE Transactions on Image Processing*, 24(12):5017–5032, 2015.
- [13] J.-F. Chassagneux, J. Chen, N. Frikha, and C. Zhou. A learning scheme by sparse grids and Picard approximations for semilinear parabolic pdes. *IMA Journal of Numerical Analysis*, 43(5):3109–3168, 2023.
- [14] T. Chen and H. Chen. Approximations of continuous functionals by neural networks with application to dynamic systems. *IEEE Transactions on Neural Networks*, 4(6):910–918, 1993.
- [15] T. Chen and H. Chen. Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its application to dynamical systems. *IEEE Transactions on Neural Networks*, 6(4):911–917, 1995.
- [16] P. Cheridito, H. M. Soner, N. Touzi, and N. Victoir. Second-order backward stochastic differential equations and fully nonlinear parabolic PDEs. *Communications on Pure and Applied Mathematics*, 60(7):1081–1110, 2007.
- [17] C. Cuchiero, P. Schmock, and J. Teichmann. Global universal approximation of functional input maps on weighted spaces. *ArXiv preprint arXiv:2306.03303*, 2023.
- [18] I. Daubechies. Orthonormal bases of compactly supported wavelets. *Communications on Pure and Applied Mathematics*, 41(7):909–996, 1988.
- [19] I. Daubechies. *Ten lectures on wavelets*, volume 61 of *CBMS–NSF regional conference series in applied mathematics*. Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania, 1992.
- [20] M. V. de Hoop, M. Lassas, and C. A. Wong. Deep learning architectures for nonlinear operator functions and nonlinear inverse problems. *Mathematical Statistics and Learning*, 4(1):1–86, 2022.
- [21] S. de Marco. Smoothness and asymptotic estimates of densities for SDEs with locally smooth coefficients and applications to square root–type diffusions. *The Annals of Applied Probability*, 21(4):1282–1321, 2011.
- [22] R. A. DeVore and R. C. Sharpley. Besov spaces on domains in $\mathbb{R}$. *Transactions of the American Mathematical Society*, 335(2):843–864, 1993.
- [23] Ronald DeVore, Boris Hanin, and Guergana Petrova. Neural network approximation. *Acta Numerica*, 30:327–444, 2021.
- [24] H. Duong. *Solving high-dimensional fully nonlinear convex partial differential equations using deep learning*. PhD thesis, Florida State University, 2023.
- [25] W. E and Q. Wang. Exponential convergence of the deep neural network approximation for analytic functions. *Science China Mathematics*, 61(10):1733–1740, 2018.
- [26] L. C. Evans. *Partial differential equations*, volume 19 of *Graduate studies in mathematics*. American Mathematical Society, 2nd edition, 2010.
- [27] C. Fefferman. Whitney’s extension problem for C^m . *Annals of Mathematics*, 164:313–359, 2006.
- [28] C. Fefferman, A. Israel, and G. Luli. Sobolev extension by linear operators. *Journal of the American Mathematical Society*, 27(1):69–145, 2014.
- [29] D. Firoozi, A. Kratsios, and X. Yang. Simultaneously solving infinitely many LQ mean field games in Hilbert spaces: the power of neural operators. *ArXiv preprint arXiv:2510.20017*, 2025.
- [30] D. Firouzi, X. Yang, and A. Kratsios. Simultaneously solving infinitely many LQ mean field games in Hilbert spaces: the power of neural operators. *In preparation*, 2025.
- [31] T. Furuya and A. Kratsios. Simultaneously solving FBSDEs with neural operators of logarithmic depth, constant width, and sub-linear rank. *ArXiv preprint arXiv:2410.14788*, 2024.
- [32] T. Furuya, K. Taniguchi, and S. Okuda. Quantitative approximation for neural operators in nonlinear parabolic equations. In *The thirteenth international conference on learning representations (ICLR 2025), April 24–28, 2025, Singapore*, pages 1–29, 2025.

- [33] L. Galimberti, A. Kratsios, and G. Livieri. Designing universal causal deep learning models: the case of infinite-dimensional dynamical systems from stochastic analysis. *Constructive Approximation*, to appear, 2025.
- [34] M. Germain, M. Laurière, H. Pham, and X. Warin. DeepSets and their derivative networks for solving symmetric PDEs. *Journal of Scientific Computing*, 91(63):1–33, 2022.
- [35] M. Germain, H. Pham, and X. Warin. Approximation error analysis of some deep backward schemes for nonlinear PDEs. *SIAM Journal on Scientific Computing*, 44(1):A28–A56, 2022.
- [36] M. Germain, H. Pham, and X. Warin. Neural networks-based algorithms for stochastic control and PDEs in finance. In A. Capponi and C.-A. Lehalle, editors, *Machine learning and data sciences for financial markets*, pages 426–452. Cambridge University Press, 2023.
- [37] D. Gilbarg and N. S. Trudinger. *Elliptic partial differential equations of second order*, volume 224 of *Classics in mathematics*. Springer Berlin, Heidelberg, second edition, 2001.
- [38] J. Gödecke and P. Fernsel. New universal operator approximation theorem for encoder–decoder architectures. *ArXiv preprint arXiv:2503.24092*, 2025.
- [39] R. Gribonval, G. Kutyniok, M. Nielsen, and F. Voigtlaender. Approximation spaces of deep neural networks. *Constructive Approximation*, 55(1):259–367, 2022.
- [40] Maximilian Herde, B. Raonić, T. Rohner, R. Käppeli, R. Molinaro, E. de Bézenac, and S. Mishra. Poseidon: efficient foundation models for PDEs. In *Proceedings of the 38th conference on advances in neural information processing systems (NeurIPS 2024), December 10–15, 2024, Vancouver, British Columbia, Canada*, volume 37, pages 72525–72624, 2024.
- [41] Ruiyang Hong and Anastasis Kratsios. Bridging the gap between approximation and learning via optimal approximation by relu mlps of maximal regularity. *arXiv preprint arXiv:2409.12335*, 2024.
- [42] B. Horvath, A. Kratsios, Y. Limmer, and X. Yang. Deep Kalman filters can filter. *SSRN preprint 4615215*, 2023.
- [43] B. Horvath, A. Kratsios, Y. Limmer, and X. Yang. Transformers can solve non-linear and non-Markovian filtering problems in continuous time for conditionally Gaussian signals. *ArXiv preprint arXiv:2310.19603*, 2025.
- [44] R. Hu and N. Laurière. Recent developments in machine learning methods for stochastic control and games. *Numerical Algebra, Control and Optimization*, 14(3):435–525, 2024.
- [45] S. Kim and G. Sakellaris. Green’s function for second order elliptic equations with singular lower order coefficients. *Communications in Partial Differential Equations*, 44(3):228–270, 2019.
- [46] Y. Korolev. Two-layer neural networks with values in a Banach space. *SIAM Journal on Mathematical Analysis*, 54(6):6358–6389, 2022.
- [47] N. B. Kovachki, Z. Li, B. Liu, K. Azizzadenesheli, K. Bhattacharya, A. M. Stuart, and A. Anandkumar. Neural operator: learning maps between function spaces with applications to PDEs. *Journal of Machine Learning Research*, 24(89):1–97, 2023.
- [48] A. Kratsios and T. Furuya. Kolmogorov–Arnold networks: approximation and learning guarantees for functions and their derivatives. *ArXiv preprint arXiv:2504.15110*, 2025.
- [49] A. Kratsios, C. Liu, M. Lassas, M. V. de Hoop, and I. Dokmanic. Universal geometric deep learning via geometric attention. *ArXiv preprint arXiv:2304.12231*, 2023.
- [50] A. Kratsios, T. Furuya, J. A. L. Benitez, M. Lassas, and M. de Hoop. Mixture of experts soften the curse of dimensionality in operator learning. *ArXiv preprint arXiv:2404.09101*, 2024.
- [51] A. Kratsios, A. Neufeld, and P. Schmocker. Generative neural operators of log-complexity can simultaneously solve infinitely many convex programs. *ArXiv preprint arXiv:2508.14995*, 2025.
- [52] A. Kratsios, P. Schmocker, and P. Zimmermann. Deep inverse problem for double phase equation. *In preparation*, 2025.

- [53] K. Kratsios and B. Zamanlooy. Do ReLU networks have an edge when approximating compactly-supported functions? *Transactions on Machine Learning Research*, August:1–22, 2022.
- [54] N. V. Krylov. *Sobolev and viscosity solutions for fully nonlinear elliptic and parabolic equations*, volume 233 of *Mathematical surveys and monographs*. American Mathematical Society, Providence, Rhode Island, 2018.
- [55] S. Lanthaler and A. M. Stuart. The parametric complexity of operator learning. *IMA Journal of Numerical Analysis*, to appear, 2025.
- [56] S. Lanthaler, S. Mishra, and G. E. Karniadakis. Error estimates for DeepONets: a deep learning framework in infinite dimensions. *Transactions of Mathematics and Its Applications*, 6(1):1–141, 2022.
- [57] S. Lanthaler, Z. Li, and A. M. Stuart. Nonlocality and nonlinearity implies universality in operator learning. *Constructive Approximation*, 62:261–303, 2025.
- [58] W. Lefebvre, G. Loeper, and H. Pham. Differential learning methods for solving fully nonlinear PDEs. *Digital Finance*, 5(1):183–229, 2023.
- [59] B. Li, S. Tang, and H. Yu. Better approximations of high dimensional smooth functions by deep neural networks with rectified power units. *Communications in Computational Physics*, 27:379–411, 2020.
- [60] Z. Li, N. Kovachki, K. Azizzadenesheli, B. Liu, K. Bhattacharya, A. Stuart, and A. Anandkumar. Fourier neural operator for parametric partial differential equations. In *International conference on learning representations (ICLR 2021)*, pages 1–16, 2021.
- [61] Z. Li, K. Meidani, and A. B. Farimani. Transformer for partial differential equations’ operator learning. *Transactions on Machine Learning Research*, April:1–34, 2023.
- [62] Z. Liu, Y. Wang, S. Vaidya, F. Ruehle, J. Halverson, M. Soljagic, T. Y. Hou, and M. Tegmark. KAN: Kolmogorov–Arnold networks. In *The thirteenth international conference on learning representations (ICLR 2025)*, pages 1–47, 2025.
- [63] L. Lu, P. Jin, and G. E. Karniadakis. DeepONet: learning nonlinear operators for identifying differential equations based on the universal approximation theorem of operators. *ArXiv preprint arXiv:1910.03193*, 2019.
- [64] L. Lu, P. Jin, G. Pang, Z. Zhang, and G. E. Karniadakis. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3:218–229, 2021.
- [65] S. G. Mallat. Multiresolution approximations and wavelet orthonormal bases of $L^2(\cdot)$. *Transactions of the American Mathematical Society*, 315(1):69–87, 1989.
- [66] C. Marcati and C. Schwab. Exponential convergence of deep operator networks for elliptic partial differential equations. *SIAM Journal on Numerical Analysis*, 61(3):1513–1545, 2023.
- [67] C. Marcati and C. Schwab. Expression rates of neural operators for linear elliptic PDEs in polytopes. *Foundations of Computational Mathematics*, to appear, 2025.
- [68] H. N. Mhaskar. Neural networks for optimal approximation of smooth and analytic functions. *Neural Computation*, 8(1):164–177, 1996.
- [69] H. N. Mhaskar and C. A. Micchelli. Degree of approximation by neural and translation networks with a single hidden layer. *Advances in Applied Mathematics*, 16(2):151–183, 1995.
- [70] H. N. Mhaskar and Charles A. Micchelli. Approximation by superposition of sigmoidal and radial basis functions. *Advances in Applied Mathematics*, 13(3):350–373, 1992.
- [71] J. R. Munkres. *Topology*. Prentice Hall, Inc., Upper Saddle River, NJ, second edition, 2000.
- [72] A. Neufeld and P. Schmock. Universal approximation property of Banach space-valued random feature models including random neural networks. *ArXiv preprint arXiv:2312.08410*, 2023.
- [73] J. Y. Nguwi, G. Penent, and N. Privault. A deep branching solver for fully nonlinear partial differential equations. *Journal of Computational Physics*, 499:112712, 2024.

- [74] N. Nüsken and L. Richter. Solving high-dimensional Hamilton–Jacobi–Bellman PDEs using neural networks: perspectives from the theory of controlled diffusions and measures on path space. *Partial Differential Equations and Applications*, 2(48):1–48, 2021.
- [75] C.-G. Pak, H.-J. Hwang, and M.-C. Kim. A nonequidistant multistep scheme for second order backward stochastic differential equations with applications to stochastic optimal control. *International Journal of Applied and Computational Mathematics*, 11(58):1–19, 2025.
- [76] É. Pardoux. Backward stochastic differential equations and viscosity solutions of systems of semilinear parabolic and elliptic PDEs of second order. In L. Decreusefond, B. Øksendal, J. Gjerde, and Üstünel A.S., editors, *Stochastic analysis and related topics VI. Proceedings of the sixth Oslo–Silivri workshop, Geilo, 1996*, volume 42 of *Progress in probability*, pages 79–127, 1998.
- [77] M. Pereira, Z. Wang, T. Chen, E. Reed, and E. Theodorou. Feynman–Kac neural network architectures for stochastic control using second-order FBSDE theory. In *Proceedings of the 2nd conference on learning for dynamics and control*, volume 120 of *Proceedings of machine learning research*, pages 728–738, 2020.
- [78] H. Pham, X. Warin, and M. Germain. Neural networks-based backward scheme for fully nonlinear PDEs. *SN Partial Differential Equations and Applications*, 2(16):1–24, 2021.
- [79] D. Pollard. *Convergence of stochastic processes*. Springer series in statistics. Springer New York, NY, 1984.
- [80] D. Possamaï and X. Tan. Weak approximation of second-order BSDEs. *The Annals of Applied Probability*, 25(5):2535–2562, 2015.
- [81] B. Raonić, R. Molinaro, T. de Ryck, T. Rohner, F. Bartolucci, R. Alaifari, S. Mishra, and E. de Bézenac. Convolutional neural operators for robust and accurate learning of PDEs. In *Proceedings of the 37th conference on advances in neural information processing systems (NeurIPS 2023), December 10–16, 2023, New Orleans, Louisiana, United States of America*, volume 36, pages 77187–77200, 2023.
- [82] Z. Ren and X. Tan. On the convergence of monotone schemes for path-dependent PDE. *Stochastic Processes and their Applications*, 127(6):1738–1762, 2017.
- [83] R. H. Riedi, R. Balestrieri, and R. G. Baraniuk. Singular value perturbation and deep network optimization. *Constructive Approximation*, 57(2):807–852, 2023.
- [84] Cornelia Schneider, Mario Ullrich, and Jan Vybiral. Nonlocal techniques for the analysis of deep relu neural network approximations. *arXiv preprint arXiv:2504.04847*, 2025.
- [85] C. Schwab, A. Stein, and J. Zech. Deep operator network approximation rates for Lipschitz operators. *Analysis and Applications*, to appear, 2025.
- [86] Z. Shen, H. Yang, and S. Zhang. Deep network approximation: achieving arbitrary accuracy with fixed number of neurons. *Journal of Machine Learning Research*, 23(276):1–60, 2022.
- [87] H. M. Soner, N. Touzi, and J. Zhang. Wellposedness of second order backward SDEs. *Probability Theory and Related Fields*, 153(1–2):149–190, 2012.
- [88] M. E. Taylor. *Partial differential equations I. Basic theory*, volume 115 of *Applied mathematical sciences*. Springer Cham, third edition, 2023.
- [89] H. Triebel. *Theory of function spaces. III*, volume 100 of *Monographs in mathematics*. Birkhäuser Basel, 2006.
- [90] H. Triebel. *Function spaces and wavelets on domains*, volume 7 of *Tracts in mathematics*. European Mathematical Society, Zürich, 2008.
- [91] T. Tripura and S. Chakraborty. Wavelet neural operator: a neural operator for parametric partial differential equations. *ArXiv preprint arXiv:2205.02191*, 2022.
- [92] Y. Wang, J. W. Siegel, Z. Liu, and T. Y. Hou. On the expressiveness and spectral bias of KANs. *ArXiv preprint arXiv:2410.01803*, 2024.

- [93] X. Xiao, W. Qiu, and O. Nikan. Numerical approximation based on deep convolutional neural network for high-dimensional fully nonlinear merged PDEs and 2BSDEs. *Mathematical Methods in the Applied Sciences*, 47(7):6184–6204, 2024.
- [94] J. Yang, W. Zhao, and T. Zhou. Explicit deferred correction methods for second-order forward backward stochastic differential equations. *Journal of Scientific Computing*, 79(3):1409–1432, 2019.
- [95] D. Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94:103–114, 2017.
- [96] A. Yu, C. Becquey, D. Halikias, M. E. Mallory, and A. Townsend. Arbitrary-depth universal approximation theorems for operator neural networks. *ArXiv preprint arXiv:2109.11354*, 2021.
- [97] M. Zhou, J. Han, and J. Lu. Actor–critic method for high dimensional static Hamilton–Jacobi–Bellman partial differential equations based on neural networks. *SIAM Journal on Scientific Computing*, 43(6):A4043–A4066, 2021.