Neural Operators Can Play Dynamic Stackelberg Games

Guillermo A. Alvarez[*]   GUIALV@UMICH.EDU
University of Michigan
Department of Mathematics
2074 East Hall, 530 Church Street Ann Arbor, Michigan, USA

Ibrahim Ekren[*]   IEKREN@UMICH.EDU
University of Michigan
Department of Mathematics
2074 East Hall, 530 Church Street Ann Arbor, Michigan, USA

Anastasis Kratsios[*†]   KRATSIOA@MCMASTER.CA
McMaster University and the Vector Institute
Department of Mathematics
1280 Main Street West, Hamilton, Ontario, L8S 4K1, Canada

Xuwei Yang[*]   HENRYYANGXUWEI@GMAIL.COM
McMaster University
Department of Mathematics
1280 Main Street West, Hamilton, Ontario, L8S 4K1, Canada

Abstract

Dynamic Stackelberg games are a broad class of two-player games in which the leader acts first, and the follower chooses a response strategy to the leader's strategy. Unfortunately, only stylized Stackelberg games are explicitly solvable since the follower's best-response operator (as a function of the control of the leader) is typically analytically intractable. This paper addresses this issue by showing that the *follower's best-response operator* can be approximately implemented by an *attention-based neural operator*, uniformly on compact subsets of adapted open-loop controls for the leader. We further show that the value of the Stackelberg game where the follower uses the approximate best-response operator approximates the value of the original Stackelberg game. Our main result is obtained using our universal approximation theorem for attention-based neural operators between spaces of square-integrable adapted stochastic processes, as well as stability results for a general class of Stackelberg games.

Keywords: attention, neural operators, dynamic games, Stackelberg equilibria, universal approximation, fast rates

1. Introduction

In many areas of the social sciences and economics, interactions among decision-makers can be modeled as games. For example, firms may compete by setting prices or production levels to maximize profits. Each player seeks to maximize their own gain (or equivalently, minimize

---

*. Equal contribution, all authors are listed in alphabetic order.

†. Corresponding author.

costs), but the outcome depends not only on their own actions but also on the actions of others. Game theory provides a mathematical framework for analyzing such strategic interactions, where an equilibrium represents a stable state in which no player can improve their payoff by unilaterally deviating. In the classical formulation of Stackelberg games, there are generally two players: a leader (major) who moves first and a follower (minor) who then reacts. One is typically interested in studying the equilibrium of these games, in which both players cannot increase their utilities by (re)acting differently. The generic structure of these games has led their equilibria to become a powerful mathematical tool to describe the evolution of incentives in complex environments with economic applications ranging from control of disease transmission in epidemiology (Aurell et al., 2022; Hubert et al., 2022), contract design (Conitzer and Sandholm, 2006; Elie et al., 2019; Keppo et al., 2024; Hernández and Possamaï, 2024; Hernández et al., 2024), advertising (He et al., 2008), mobile network planning (Zheng et al., 2018), economic behaviour in oligopolies (Carmona and Dayanıklı, 2021), brokerage (Alvarez et al., 2023), (re)insurance (Chen and Shen, 2018; Cao et al., 2022; Kroell et al., 2023; Ghossoub and Zhu, 2024), risk management (Bensalem et al., 2020; Li et al., 2022), algorithmic auction/mechanism design (Conitzer and Sandholm, 2006; Konrad and Leininger, 2007; Dierks and Seuken, 2022), security (Kar et al., 2017; An et al., 2011), and green investments (Zhang et al., 2023). Most of these applications consider *dynamic* Stackelberg games, where the game is played continuously in several rounds. Even though these games provide powerfully descriptive theoretical vehicles, the highly intertwined structure of Stackelberg equilibria can be challenging both numerically and analytically.

This paper shows that deep learning can provide a viable and generic computational vehicle by which dynamic Stackelberg games can be computationally solved. We exhibit a class of *neural operators* leveraging an attention mechanism which can approximately implement the follower's best response map to arbitrary precision, uniformly on compact subsets of the leader's strategies (defined below). Unlike most neural operator models, which focus on learning the solution map of PDEs (Kovachki et al., 2021; Lanthaler et al., 2022a; Lee et al., 2023; Lanthaler and Stuart, 2023; Kovachki et al., 2023; Benitez et al., 2023; Raonic et al., 2024; Bartolucci et al., 2024; Fanaskov and Oseledets, 2024) or inverse-problems (Calderon-Macias, 1997; Molinaro et al., 2023; de Hoop et al., 2022, 2024), our neural operators are not defined between function spaces but between spaces of stochastic controls. Our attention mechanism mimics that of Vaswani et al. (2017) used in transformers (Bahdanau et al., 2015) while reflecting the geometry of the input and output spaces of stochastic processes, and it extends the attention mechanisms of Acciaio et al. (2023). We note that, it is natural to consider compact sets of controls not only from approximation-theoretic vantage point but also from the control-theoretic perspective; this is because this guarantees the existence of a Stackelberg equilibrium under only minimal assumptions.

Here, we consider the following class of dynamic Stackelberg games, with stochastic effects, where both players (re)act in continuous time according to the following general dynamics

$$dX_t = f(X_t, u_t^0, u_t^1)dt + \sigma(X_t, u_t^0, u_t^1)dW_t$$

where $W. \overset{\text{def.}}{=} (W_t)_{t \geq 0}$ is d-dimensional standard Brownian motions and $u_\cdot^i \overset{\text{def.}}{=} (u_t^i)_{t \geq 0}$ are the (re)actions/strategies of each player, where $i = 0$ is the leader and $i = 1$ indexes

the follower. Each player seeks to optimize their respective objective functions, one in which we impose minimal continuity requirements since we are not interested in analytic expression, which would require highly stylized assumptions on the dynamics and objective functions of all involved players. Rather, our goal is to show that a deep learning solution via neural operators is possible for a broad class of Stackelberg games lying outside the score of these classical stylized settings. Our first main result (Theorem 4) shows under enough strong-convexity requirements on the utility of the follower the best response maps depend continuously on each leader's actions, then there is a neural operator which can approximate the follower's best response map, uniformly over any compact set of actions of the leader, to any given precision.

In general, it is well-known that the approximation of non-linear maps/operators between infinite-dimensional Hilbert spaces by deep learning models may be practically challenging due to necessarily slow convergence rates; see e.g. Lanthaler and Stuart (2023), which is effectively an exacerbated version of the curse of dimensionality known in the finite-dimensional setting, see e.g. Shen et al. (2022). Unlike the finite-dimensional setting, sufficient smoothness is insufficient to obtain reasonable convergence rates in general, (Galimberti et al., 2022), and typically, one has to hope for favorable structures which can be exploited by the neural operator; see e.g. Marcati and Schwab (2023), to obtain fast convergence rates. Fortunately, we identify a set of structures that can be exploited by our neural operator for a class of Stackelberg games that encompass analytically-solvable linear-quadratic games. Our second main result (Theorem 7) shows that if the compact set of controls is compatible with the best-response map, then one may guarantee efficient convergence rates for neural operator approximations of the best response map for the follower. We conclude that neural operators can efficiently approximate the solutions to a wide class of Stackelberg games, encompassing those games which are solvable via classical analytic means.

Additionally, we identify an *unsupervised* objective function which provides a heuristic helping to detect the suitability of a neural operator approximating the best response map of the follower (Theorem 5). Importantly, this heuristic objective function does not require observations of the true optimal response (which would be an irrealistic supervised problem).

We use control-theoretic arguments to prove that the leader and the follower have optimal strategies with enough continuous dependence on one another to be *uniformly* approximable on compact sets of (re)actions. We also provide an illustrative counterexample showing that the best-response map might fail to be continuous without these strong-convexity requirements. Thus, without the strong-convexity assumption, discontinuous functions cannot be uniformly approximated by any continuous models due to the Uniform Limit Theorem; see e.g. Munkres (2000, Theorem 21.6).

On the technical front: our main control-theoretic contributions (Lemma 9 together with 13) show that generically the optimal response of the follower is $1/2$-Hölder continuous in the leader's control if the problem of the follower is strongly convex. Our main approximation-theoretic contribution is a quantitative universal approximation theorem (Theorem 8) showing that neural operators are capable of approximating Hölder continuous non-linear operators between spaces of square-integrable $\mathbb{F}$-adapted processes open-loop controls. Generally speaking, an open loop control is a stochastic process adapted to the natural filtration generated by a Brownian motion $W_.$. We refer the reader to Sec-

tion 2 for a more detailed description of the concept. Together, these control-theoretic and approximation-theoretic contributions are enough to show that the best-response map is approximable by neural operators. Moreover, for suitable sets of controls, our precise analysis both of the regularity of the best response map and the dependence of the neural operator complexity on the set of controls, allow us to conclude that polynomial approximation rates are possible (Theorem 7).

Our neural operators leverage an attention-like decoding layer, similar to transformers (Vaswani et al., 2017), which allows for nonlinear decoding, unlike PCA-net (Lanthaler, 2023), the encoder-decoder models of Galimberti et al. (2022), and several others. The relationship between our infinite-dimensional analogue of the classical attention of Bahdanau et al. (2015) is also discussed. Attention mechanisms in operator learning, by now, have found common use in implementations; see e.g. the Galerkin transformers of Cao (2021) or the Continuum Attention Mechanism of Calvello et al. (2024). Interestingly, our architectural choices allow us to identify regimes in which efficient approximation is possible; which mirrors the recent results of Furuya and Kratsios (2024); Kratsios et al. (2025) which identify architectural tweaks allowing for efficient approximation rates of control problems distinct from the dynamic games considered in this paper.

1.1 Organization of Paper

Following the literature review in Section 2, the remainder of our paper is organized as follows:

(i) Section 2 and 3, respectively, review the literature and the necessary background in stochastic analysis and in deep approximation theory to formulate our main results.

(ii) Section 4 contains our main results.

(iii) Section 4.2 explains why our main results work by overviewing our proof strategy, during which we showcase our supporting technical results of independent technical interest.

(iv) Section 5.1 showcases examples of Stackelberg games satisfying our convexity requirements.

All technical derivations are relegated to appendix A.

2. Related Literature

2.1 Stackelberg Games in Machine Learning

In Stackelberg games, both players seek to maximize their gain while being fully rational. This characteristic "conditional" sequential structure is the hallmark challenge rendering Stackelberg games analytically intractable and the reason motivating significant attention from the machine learning community (Reisinger and Zhang, 2020; Ito et al., 2021; Gao et al., 2022; Haghtalab et al., 2023; Harris et al., 2023; Gerstgrasser and Parkes, 2023; Dayanikli and Lauriere, 2023), and their related FBSDEs (Forward-Backward Stochastic Differential Equations) (Furuya and Kratsios, 2024), looking for new algorithmic tools capable of solving this class of differential games. Nevertheless, there is currently no available

deep learning model which is guaranteed to solve a Stackelberg game, much less in continuous time with stochastic effects.

2.2 Neural Operators

The power of deep learning to solve previously intractable high-dimensional computational problems has motivated the deep learning community to extend these tools to the infinite-dimensional setting with models such as DeepONets (Lu et al., 2019, 2021; Goswami et al., 2022) and a variety of *neural operator* architectures; e.g. Fourier Neural Operators (Li et al., 2020a; Kovachki et al., 2021; Li et al., 2023), graph neural operators (Anandkumar et al., 2020), causal neural operators (Galimberti et al., 2022), neural operator analogues of transformers (Hao et al., 2023), convolutional neural operators (Raonic et al., 2024), encoder-decoder models such as PCA-net (Lanthaler et al., 2022b), and a myriad of other models. The community has largely been motivated by demand in the scientific computing community, focusing on designing neural operators tailored which can learn to solve (i.e. learn the solution operator) to high-dimensional partial integro-differential equations (PIDEs) with applications ranging from physics and engineering (De Ryck et al., 2024) to quantitative finance (Acciaio et al., 2023). However, the full power of neural operators remains otherwise largely unexplored as the community has focused mainly on the approximation capacity of these models between function spaces with a view towards PIDEs. Here, we probe the limits of neural operators beyond PIDEs by showing that they can solve a broad class of problems at the intersection of game theory and stochastic analysis.

2.3 Open-Loop Stackelberg Equilibria

There are multiple notions of Stackelberg equilibria, depending on the information available to the players. In this work, we focus on the so-called open-loop equilibria, in which the players' strategies are stochastic processes adapted to the natural filtration generated by a common Brownian motion W.. The existing approaches to compute open-loop Stackelberg equilibria in continuous time (Bensoussan et al., 2015) are based on the Stochastic Maximum Principle (Yong and Zhou, 1999, Theorems 3.2 and 5.2). The previous results enable us to characterize, through the solution of an FBSDE, the follower's optimal response given any leader's strategy u_0. As a result, the leader's problem can be reduced to a stochastic control of FBSDEs, which can be solved by an extended version of the Stochastic Maximum Principle (Hu et al., 2018). The previous approaches present multiple challenges, and, in practice, very few problems beyond a linear-quadratic setting can be explicitly solved using the above method.

Notation

We use the following notation. For any $C \in \mathbb{N}_+$, we denote the C-simplex by $\Delta_C \stackrel{\text{def.}}{=} \{u \in [0,1]^C : \sum_{c=1}^{C} u_c = 1\}$ and we define the associated softmax function by $\text{softmax}_C : \mathbb{R}^C \ni w \mapsto \left(e^{w_c}/\sum_{c=1}^{C} e^{w_c}\right)_{c=1}^{C} \in \Delta_C$. Both in Δ_C and softmax_C, the subscript C will be suppressed when clear from the context. We will write $f \in \tilde{\mathcal{O}}(g)$ if $f \in \mathcal{O}(g \log^k(g))$ for some $k \in \mathbb{N}_+$.

3. Preliminaries

We now overview the background required to formulate the main results of our paper and formalize our neural operator model. We begin by formulating the dynamic Stackelberg game and the notion of Stackelberg equilibrium. Next, we overview the definition of a multilayer perceptron (MLP) from deep learning and the notion of square-integrable predictable processes from stochastic analysis. Additional details are included in Appendix A.3.

Fix a time horizon $T > 0$ and let $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ be a filtered probability space whose filtration $\mathbb{F} \stackrel{\text{def.}}{=} (\mathcal{F}_t)_{0 \leq t \leq T}$ is generated by a d-dimensional Brownian motion $W. \stackrel{\text{def.}}{=} (W_t)_{0 \leq t \leq T}$; for some $d \in \mathbb{N}_+$. We consider the space $\mathcal{H}_T^2$ of all square-integrable $\mathbb{F}$-*predictable processes* on $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ whose elements $H. \stackrel{\text{def.}}{=} (H_t)_{0 \leq t \leq T} \in \mathcal{H}_T^2$ consist of d-dimensional $\mathbb{F}$-predictable processes for which the norm

$$\|H\|_{\mathcal{H}_T^2}^2 \stackrel{\text{def.}}{=} \mathbb{E}\left[\int_0^T |H_s|^2 d s \right]$$

is finite. Furthermore, $\mathcal{H}_T^2$ is a separable infinite-dimensional Hilbert space whose inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_T^2}$, is given for each $H., \tilde{H}. \in \mathcal{H}_T^2$ by

$$\langle H., \tilde{H}. \rangle_{\mathcal{H}_T^2} \stackrel{\text{def.}}{=} \mathbb{E}\left[\int_0^T H_s^\top \tilde{H}_s \, ds \right].$$

3.1 The Dynamic Stackelberg Game

In the previously introduced probability space $(\Omega, \mathcal{F}, \{\mathcal{F}\}_{0 \leq t \leq T}, \mathbb{P})$, we consider a Stackelberg game with a leader indexed with $i = 0$ and a follower indexed with $i = 1$. The state process of the game is described by the stochastic differential equation

$$dX_t = f(X_t, u_t^0, u_t^1)dt + \sigma(X_t, u_t^0, u_t^1)dW_t, \tag{1}$$

and $u_t^0 \in \mathbb{R}^{d_0}$ and $u_t^1 \in \mathbb{R}^{d_1}$ are the controls of the leader and the follower, respectively. The exact set of admissible controls will be provided below. We assume that a deterministic initial $X_0 \in \mathbb{R}^d$ is fixed and is known to both agents. Thus, we omit the dependence of various parameters on X_0. The cost functionals of the two players are given by

$$J_0(u^0, u^1) = \mathbb{E}\left[\int_0^T L_0(X_t, u_t^0, u_t^1)dt + g_0(X_T) \right], \tag{2}$$

$$J_1(u^0, u^1) = \mathbb{E}\left[\int_0^T L_1(X_t, u_t^0, u_t^1)dt + g_1(X_T) \right], \tag{3}$$

where $L_i : \mathbb{R}^d \times \mathbb{R}^{d_0} \times \mathbb{R}^{d_1} \mapsto [0, \infty)$ and $g_i : \mathbb{R}^d \mapsto [0, \infty)$, $i = 0, 1$. We require the following regularity conditions of the involved functions.

Assumption 1 (Regularity Conditions) *There exists a constant $K > 0$ such that for* $h(x, u^0, u^1) = f(x, u^0, u^1)$, $\sigma(x, u^0, u^1)$, $L_i(x, u^0, u^1)$, *and* $g_i(x)$, $i = 0, 1$,

$$|h(x, u^0, u^1) - h(\tilde{x}, \tilde{u}^0, \tilde{u}^1)| \leq K(|x - \tilde{x}| + |u^0 - \tilde{u}^0| + |u^1 - \tilde{u}^1|). \tag{4}$$

We define

$$\mathcal{U}_i = \left\{ u : [0, T] \times \Omega \to \mathbb{R}^{d_i} \,|\, u(\cdot) \text{ is } \{\mathcal{F}\}_t\text{-adapted, } \mathbb{E} \int_0^T |u_t|^2 dt < \infty \right\}$$

and fix $\mathcal{K}_0 \subset \mathcal{U}_0$ so that $\mathcal{K}_0$ is the set of possible controls of the leader, $\mathcal{U}_1$ is the set of possible controls of the follower. The introduction of $\mathcal{K}_0$ is needed due to the fact that our operators in Subsection 3.2 will only perform optimization relative to a compact subset $\mathcal{K}_0$ of $\mathcal{U}_0$.

For each $u^0 \in \mathcal{U}_0$, the set of best responses for the follower is

$$\mathcal{R}(u^0) = \left\{ u \in \mathcal{U}_1 : J_1(u^0, u) \leq J_1(u^0, u^1), \, \forall u^1 \in \mathcal{U}_1 \right\}. \tag{5}$$

Following the definitions in Bensoussan et al. (2015), we define adapted open-loop (AOL) responses of the follower to the controls of the leader by

$$\bar{\mathcal{U}}_1 = \left\{ u : [0, T] \times \Omega \times \mathcal{K}_0 \to \mathbb{R}^{d_1} \,|\, \forall u^0 \in \mathcal{K}_0, \, u(\cdot, u^0) \in \mathcal{U}_1 \right\}. \tag{6}$$

We recall that implicitly we make the assumption that the initial point X_0 of X is fixed throughout the paper. If this initial condition is not fixed, one has to allow the elements of $\bar{\mathcal{U}}_1$ to also depend on this initial condition as it is the case in Bensoussan et al. (2015).

We study the Stackelberg equilibria for the leader-follower problem; that is, a set of leader-follower strategies wherein the follower optimally responds to the preemptive optimal action of the leader in such a way that neither player can gain utility by perturbing their strategy. Formally, a Stackelberg equilibrium is defined as follows.

Definition 1 (Stackelberg Equilibrium) *A Stackelberg equilibrium (relative to $\mathcal{K}_0 \subset \mathcal{U}_0$) of the leader-follower game (1)-(2)-(3) is a pair $(u^{0,\star}, U^\star) \in \mathcal{K}_0 \times \bar{\mathcal{U}}_1$ such that $U^\star(u^0) \in \mathcal{R}(u^0)$ for all $u^0 \in \mathcal{U}_0$ and*

$$J_0(u^{0,\star}, U^\star(u^{0,\star})) \leq J_0(u^0, U^\star(u^0)) \text{ for all } u^0 \in \mathcal{K}_0.$$

If it exists, a map $U^\star$ is called a best response map of the follower.

In general, $\mathcal{R}(u^0)$ can be empty for some $u^0 \in \mathcal{K}_0$. If this happens, the equilibrium will not exist. However, if the problem of the follower is convex enough, $\mathcal{R}(u^0)$ is reduced to a point $U^\star(u^0) \in \mathcal{U}_1$ that can be described by an FBSDE (Forward-Backward Stochastic Differential Equation). Thus, the existence of a Stackelberg equilibrium is reduced to the optimization of $u^0 \in \mathcal{K}_0 \mapsto J_0(u^0, U^\star(u^0))$ that we call the effective criterion of the leader. This can be done if $\mathcal{K}_0$ is compact or under additional structural assumption on the data as a form of control of solutions of FBSDEs. We will assume that the optimal response operator of the follower exists and possesses a minimal level of regularity.

Assumption 2 (Hölder Continuity of the Follower's Best Response) *There exists a Hölder continuous mapping $u^0 \in \mathcal{K}_0 \mapsto U^\star(u^0) \in \mathcal{U}_1$ so that $U^\star(u^0) \in \mathcal{R}(u^0)$ for all $u^0 \in \mathcal{K}_0$.*

Remark 2 *Though several of our universal approximation results (Theorem 8) can be applied only while assuming continuity of the follower's best response, our quantitative estimates fundamentally rely on Hölder continuity since we use properties of doubling metrics, building on the method of Kratsios et al. (2023b), and metric snowflakes; see Weaver (2018, page 66) for definitions.*

In Sections 4.2 and 5.1 below, we show Hölder dependence estimates of the optimal response if the optimization problem of the follower is strongly convex. We now provide the example of a deterministic game that shows that both $U^\star$ and $u^0 \mapsto J_0(u^0, U^\star(u^0))$ might not be continuous without the strong convexity assumption, even if the problem of the follower is convex. In such cases, our neural operator cannot approximate the optimal response. Thus, Assumption 2 is not merely a technical assumption.

Example 1 (The Effective Optimization Problem of the Leader can be Discontinuous) *Consider the deterministic single period game where the controls are $u^i \in [0, 1]$. Fix the loss functions $l_1(u^0, u^1) = u^0 u^1$ and $l_0(u^0, u^1) = -u^1$ so that the leader's problem is*

$$\inf_{u^0 \in [0,1]} \inf_{u^1 \in \mathcal{R}(u^0)} l_0(u^0, u^1) \tag{7}$$

and $\mathcal{R}(u^0) \stackrel{\text{def.}}{=} \{u^1 \in [0, 1] : l_1(u^0, u^1) \leq \inf l_1(u^0, \cdot)\}$. Note that l_1 is convex in u^1 but it lacks the strong convexity in u^1 needed in Proposition 13. Clearly $\mathcal{R}(u^0) = \{0\}$ if $u^0 > 0$ and $\mathcal{R}(0) = [0, 1]$. Thus,

$$\inf_{u^1 \in \mathcal{R}(u^0)} l_0(u^0, u^1) = -I_{\{u^0 = 0\}}$$

which is discontinuous in u^0.

3.2 Neural Operators

Fix an activation function $\sigma : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ and, for any $N \in \mathbb{N}_+$, define its componentwise composition with any vector $x \in \mathbb{R}^N$ with trainable parameter $\boldsymbol{\alpha} \in \mathbb{R}^N$, by $\sigma_{\boldsymbol{\alpha}} \bullet x \stackrel{\text{def.}}{=} \left(\sigma_{\boldsymbol{\alpha}_i}(x_i)\right)_{i=1}^{d}$. In our paper, we consider neural operators relying on either an untrainable activation function satisfying the condition of Kidger and Lyons (2020) or a trainable variant $\sigma \in C(\mathbb{R}^2)$ of the super-expressive activation function of Zhang et al. (2022); defined shortly. In both cases, our activation functions are gateable, in sense of Dziugaite and Roy (2025), meaning that they can modulate being a non-linearity and an identity or skip-type connection. We either consider the following activation function:

Example 2 (Standard *Trainable* Activation Functions) *We consider activation function $\sigma \in C(\mathbb{R}^2)$ with the property that: there is some $t_0 \in \mathbb{R}$ such that $\sigma_0(t_0) \stackrel{\text{def.}}{=} \sigma(0, t_0)$ is non-affine, differentiable at t_0, and $\sigma_0(t_0)' \neq 0$. Additionally, there is a $\alpha_0 \in \mathbb{R} \setminus \{0\}$ such that: $\sigma_{\alpha_0}(t) = t$ for all $t \in \mathbb{R}$.*

Observe that the $\sigma_\alpha(t) \stackrel{\text{def.}}{=} \max\{\alpha t, t\}$ satisfies this property than that $\max\{0t, t\} = \text{ReLU}(t)$. We also consider, and typically focus on, the following activation function.

Example 3 (Non-standard: Super-Expressive Activation with Neuron-Specific Skip-Connection) *We define the trainable variant of the activation function $\sigma : \mathbb{R}^2 \to \mathbb{R}$ mapping any $(\alpha, t) \in \mathbb{R}^2$ to*

$$\sigma_\alpha(t) \stackrel{\text{def.}}{=} \alpha t + (1 - \alpha) \begin{cases} |t \pmod{2}| & \text{if } t \geq 0, \\ \frac{t}{|t|+1} & \text{if } t < 0. \end{cases} \tag{8}$$

When $\alpha = 0$, then σ_0 coincides with the super-expressive activation function of Zhang et al. (2022). When $\alpha = 1$ then, $\sigma_1(t) = t$ for all $t \in \mathbb{R}$.

In both Example 2 and 3, the parameter α controls the transition between the identity map and a non-linear, non-affine activation.

We now define the deep learning backbone of our neural operator model, namely the multilayer perceptrons (MLP). Fix input and output dimensions n and m. An MLP with (trainable) activation function σ, *depth* $J \in \mathbb{N}_+$, and *width* $W \in \mathbb{N}_+$ is a map $f : \mathbb{R}^n \to \mathbb{R}^m$ with iterative representation,

$$
\begin{aligned}
f(x) &\overset{\text{def.}}{=} x^{(J)} + c, \\
x^{(j+1)} &\overset{\text{def.}}{=} A^{(j)} \sigma_{\alpha^{(j)}} \bullet (x^{(j)} + b^{(j)}), \\
x^{(0)} &\overset{\text{def.}}{=} x,
\end{aligned}
\tag{9}
$$

where for $j = 0, \ldots, J-1$: $A^{(j)} \in \mathbb{R}^{d_{j+1} \times d_j}$, $\alpha^{(j)}, b^{(j)} \in \mathbb{R}^{d_{j+1}}$, and $n = d_0, \ldots, m = d_J \leq W$. Denote the set of MLPs mapping $\mathbb{R}^n$ to $\mathbb{R}^m$ with depth at-most J and width at-most W by $\mathcal{NN}_{J,W:n,m}$.

3.3 Predictable Processes

Similarly to the Fourier neural operator, the neural operators of our approximation results will rely on an orthogonal basis[1] of $\mathcal{H}_T^2$. We begin by constructing a theortically and computationally convenient basis of $L^2(\mathcal{F}_t)$ based on the Wiener Chaos decomposition. A key feature of the following orthonormal basis is that no iterated stochastic integrals need to be explicitly computed, as is the case with general Wiener Chaos decompositions; see Appendix A.3 for additional details.

For any $i \in \mathbb{N}$, the Hermite polynomials $(h_i)_{i \in \mathbb{N}}$ are the eigenfunctions of the generator $\frac{d^2}{dx^2} - x \frac{d}{dx}$ of the Ornstein-Uhlenbeck process $dX_t = -X_t + \sqrt{2} dW_t$. For each $i \in \mathbb{N}_+$, the ith Hermite polynomial h_i is given recursively by Rodrigues' formula as

$$
h_i(x) = \frac{(-1)^i}{i!} e^{x^2/2} \frac{d^i}{dx^i} e^{-x^2/2}, \qquad h_0(x) = 1.
$$

The Hermite polynomials allow us to define a family of random variables $\{u_{i,j,k}^t : j \in \mathbb{N}, i, k \in \mathbb{N}, \frac{k+1}{2^i} \leq 1\} \subset L^2(\mathcal{F}_t)$ where each $u_{i,j,k}^t$ is defined by

$$
u_{i,j,k}^t \overset{\text{def.}}{=} \prod_{\tilde{j}=1}^{j} h_{\tilde{j}}\left(2^i W_{\frac{tk}{2^i}} - 2^{i+1} W_{\frac{t(1+2k)}{2^{i+1}}} + 2^i W_{\frac{t(k+1)}{2^i}} \right).
\tag{10}
$$

Using (10), we construct a predictable and dynamic version of the Wiener chaos decomposition. First recall the *Haar (wavelet) system* on the larger space $L^2([0,T])$ where $(\psi_{i,k})_{i,k \in \mathbb{N}; 0 \leq k, \frac{k+1}{2^i} \leq 1}$ and

$$
\psi_{i,k}(t) \overset{\text{def.}}{=} 2^i \left(I_{[T\frac{k}{2^i}, T\frac{1+2k}{2^{i+1}})}(t) - I_{[T\frac{1+2k}{2^{i+1}}, T\frac{k+1}{2^i})}(t), \right)
$$

---

1. Some authors emphasize that "orthonormal basis" is rather a complete orthonormal system since it is not a linear algebraic basis. However, we follow the common abuse of terminology standard in functional analysis.

where, for all $A \in \mathbb{R}$, I_A denotes the indicator function of a set A. We note that by Meyer (1990, Chapter 3), the set $\left(\psi^{i,j}\right)_{i,k \in \mathbb{N}; 0 \le k, \frac{k+1}{2^i} \le 1}$ is a complete orthonormal basis of $L^2([0,T])$.

The Haar wavelet system will allow us to activate/deactivate the random variables in Wiener Chaos, in (10), as a function of time. We focus on the collection of simple processes obtained as linear combinations of

$$\mathcal{S} \stackrel{\text{def.}}{=} \left\{ u_{i,j,k}^{s_1,s_2} : i,j,k,s_1,s_2 \in \mathbb{N},\ s_2 + 1 \le 2^{s_1}, \frac{k+1}{2^i} \le \frac{s_2}{2^{s_1}} \right\} \tag{11}$$
$$u_{i,j,k}^{s_1,s_2}(t,\omega) \stackrel{\text{def.}}{=} \psi_{s_1,s_2}(t) \cdot u_{i,j,k}^T(\omega).$$

It can be shown (see Lemma 19 in the proofs section) that $\mathcal{S}$ is an orthonormal basis of $\mathcal{H}_T^2$.

Furthermore, $\mathcal{S}$ has an elegant interpretation as a Haar wavelet expansion in time and the iterated Itô stochastic integral of a Haar wavelet expansion in space. The ability to explicitly compute the iterated Itô stochastic integrals of the Haar wavelet system in space makes this closed-form expansion particularly favourable, especially in higher dimensions where the computation of iterated stochastic integrals can be computationally intensive. Our results refer to compact classes of controls in $\mathcal{H}_T^2$ for which several explicit, and natural, examples are constructed in Appendix A.6.

3.3.1 THE ATTENTIONAL NEURAL OPERATOR MODEL

Neural operators (NOs) are natural, infinite dimensional extensions of classical (finite-dimensional) neural networks. We follow the general encoder-processor-decoder NO paradigm considered, e.g. in PCA-nets (Lanthaler, 2023), (Castro, 2023), or (Kratsios et al., 2023c). Our neural operator model, illustrated in Figure 1, maps inputs and outputs in the space structure $\mathcal{H}_T^2$, as opposed to standard neural operators which map functions on a Euclidean domain to functions in another Euclidean domain.
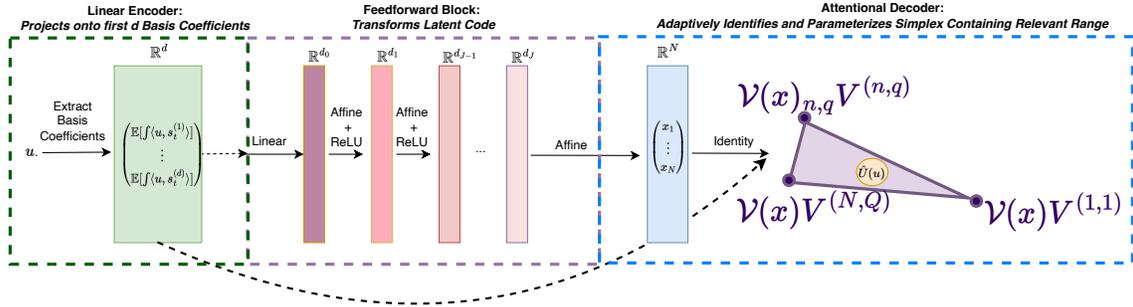

Figure 1: **Attentional Neural Operator Workflow:** Our *attentional* **n**eural **o**perator model maps controls $u.$ to square-integrable $\mathbb{F}$-adapted processes $\hat{U}(u.)$ in three phases. First, the (input) control is linearly projected onto the wavelet-like (in time) Wiener Chaos-like (in space) orthonormal basis of $\mathcal{H}_T^2$. Next, the basis coefficients are transformed by a feedforward neural network (MLP). Lastly, the basis coefficients are used to identify extremal points in a simplex in $\mathcal{H}_T^2$ and the outputs of the MLP are used to parameterize a prediction in its relative interior.

The neural operators in Figure 1 can be formalized as follows.

Definition 3 (Attentional Neural Operator) *Fix a trainable activation function $\sigma \in C(\mathbb{R}^2)$, an encoding dimension $d \in \mathbb{N}_+$, a queries dimension $Q \in \mathbb{N}_+$, a values dimension $N \in \mathbb{N}_+$, depth and width parameters $J, W \in \mathbb{N}_+$, and an orthonormal basis $\mathcal{S}$ of $\mathcal{H}_T^2$. The set $\mathcal{NO}_{N,Q,d,J,W:\mathcal{S}}^{\sigma}$ consists of all (non-linear) operators $U : \mathcal{H}_T^2 \to \mathcal{H}_T^2$ admitting the following representation: for each $u. \in \mathcal{H}_T^2$*

$$U(u.) \stackrel{\text{def.}}{=} \mathcal{D}\big(f \circ \mathcal{E}(u.), \mathcal{E}(u.)\big)$$

$$\mathcal{E}(u.) \stackrel{\text{def.}}{=} \left(\mathbb{E}\left[\int_0^T \langle u_t, s_t^{(i)}\rangle dt\right]\right)_{i=1}^d \qquad \text{(encoder)}$$

$$\mathcal{D}(w, x) \stackrel{\text{def.}}{=} \sum_{n=1}^N \text{softmax}(w)_n \sum_{q=1}^Q \mathcal{V}_{n,q}(x)\, V^{(n,q)}, \qquad \text{($\mathcal{H}_T^2$-attentional decoder)}$$

where $\{s^{(i)}\}_{i=1}^d$, $\{V^{(n,q)}\}_{n,q=1}^{N,Q} \subset \mathcal{S}$, $f \in \mathcal{NN}_{J,W:d,N}^{\sigma}$, $\mathcal{V} \in \mathcal{NN}_{J,W:d,N\times Q}$, and $\mathcal{D} : \mathbb{R}^N \times \mathbb{R}^d \to \mathcal{H}_T^2$.

The number of (non-zero trainable) parameters defining the neural operators U is at most

$$\underbrace{JW^2}_{MLP\ (f)} + \underbrace{NQ.}_{MLP\ (\mathcal{D})}$$

The map $\mathcal{E}$ is called an encoder, $\mathcal{D}$ is called an attention-based decoder, $(\mathcal{V}, \{V^{(n,q)}\}_{n,q=1}^{N,Q})$ are called values, and U is called a attentional neural operator.

We henceforth take $\mathcal{S}$ to be (11) and denote its elements $\{s^{(i)}\}_{i=1}^\infty$. Thus, we write $\mathcal{NO}_{N,Q,d,J,W}^{\sigma}$ in place of $\mathcal{NO}_{N,Q,d,J,W:\mathcal{S}}^{\sigma}$.

3.3.2 WHY THIS ATTENTION-BASED ARCHITECTURE?

Standard neural operator architectures—e.g., appropriately modified DeepONets (Lu et al., 2021), Fourier Neural Operators (Li et al., 2020a), PCA-Net–style encoder-decoder models (Korolev, 2022; Galimberti et al., 2022), or the attention-based models of Acciaio et al. (2023)—can in principle be adapted for our setting. The main modification, particularly for the latter two, is to replace their function bases with an orthonormal basis of $\mathcal{H}_T^2$ consisting of square-integrable $\mathbb{F}$-predictable processes. A practical basis should avoid relying solely on iterated Itô integrals (e.g., the fixed-time bases in Galimberti et al., 2022), which are computationally costly. Instead, our cheap orthonormal basis (cf. Lemma 19) provides closed-form access to these integrals (see (10)).

We choose to build upon the attention-based model of Acciaio et al. (2023) because it is not constrained to inputs and outputs lying in a vector space but can adapt to most metric spaces, unlike the other neural operator architectures mentioned above. This architecture is naturally compatible with extensions to Stackelberg games using randomized (measure-valued or "mixed") strategies (Conitzer, 2016; Von Stengel and Zamir, 2004), which, in general, do not lie in a Banach space. Because every Sobolev space is a Banach space, this excludes Graph Neural Operators (GNOs) (Li et al., 2020b) and related graph variants (Mousavi et al., 2025), which can efficiently approximate operators only between Sobolev spaces on irregular Euclidean domains.

While this paper focuses on "strong guarantees," i.e., approximation in the $\mathcal{H}_T^2$ sense depending on the path of the driving Brownian motion, the choice of architecture also allows for future modifications to take values in the space of probability measures. In that setting, weaker approximation guarantees "only in distribution" would likely be possible by combining our Theorem 8 (which allows infinite-dimensional inputs) with Acciaio et al. (2023, Theorem 3.8), which currently applies only to finite-dimensional inputs.

3.3.3 LINK BETWEEN OUR DECODER, ATTENTION, AND TRANSFORMERS

Before moving on, we discuss the relationship between the decoder of our neural operator and the attention layer in Bahdanau et al. (2015) in the standard transformer networks of Vaswani et al. (2017). For simplicity, we focus on a single attention head rather than the multi-head attention mechanism.

The standard attention mechanism maps N d-dimensional vectors $x_1, \ldots, x_N$, seen as an $N \times d$ matrix $x = (x_n)_{n=1}^d$ to another $N \times \tilde{d}$ matrix attention(x) for some $\tilde{d} \in \mathbb{N}$. This attention mechanism operates in two phases. First, it extracts *contextual weights* w^x in the N-simplex via

$$w^x \overset{\text{def.}}{=} \text{softmax}\left(\left(\langle x_n Q, x_j K \rangle / \sqrt{d} \right)_{j=1}^N \right)_{n=1}^N \tag{12}$$

where the key and query matrices K and Q are $\tilde{d} \times d$ matrices. Using these weights, one defines a set of N *values* $v_1, \ldots, v_N$, depending on the given input x, by

$$v_n^x \overset{\text{def.}}{=} V x_n \tag{13}$$

for $n = 1, \ldots, N$. The attention mechanism then uses the weights w^x to tweak the contextual importance of each value $v_1^x, \ldots, v_N^x$ when generating a weighted prediction by

$$\text{attention}(x) \overset{\text{def.}}{=} \sum_{n=1}^N w_n^x v_n^x. \tag{14}$$

In this way, the standard attention mechanism parameterizes the interior of the convex hull of the values $v_1^x, \ldots, v_N^x$ using the softmax weights of any input.

If we streamline the contextual weight extraction step in (12), by simply an input weight $w \in \mathbb{R}^N$ which is mapped to contextual importance weights w by only using the softmax function itself. One can relax the dependence on the contextual values $v_1^x, \ldots, v_N^x$ in (13) to be vectors in the target space, i.e. $\mathbb{R}^{N \times \tilde{d}}$ for classical transformers and $\mathcal{H}_T^2$ for our neural operator, depending non-linearly on the given input. Here, we parameterize this non-linear dependence using a values neural network $\mathcal{V}$ depending on the encoding instead of a values matrix V; doing so, one arrives at the following construction, which is precisely our $\mathcal{H}_T^2$-*attention*

$$\mathcal{D}(w) \overset{\text{def.}}{=} \sum_{n=1}^N \underbrace{\text{softmax}(w)_n}_{\text{Contextual Weights (12)}} \underbrace{\sum_{q=1}^Q \mathcal{V}_{n,q}(u.) V^{(n,q)}}_{\text{Contextual Values (13)}}.$$

Thus, one can interpret our $\mathcal{H}_T^2$-attention as an infinite-dimensional analogue of the standard attention mechanism of Bahdanau et al. (2015). Note that, since $\mathcal{D}$ receives inputs from an

MLP and since MLPs can approximately implement continuous functions then there is no need to rely on a vector of inner-products such as $(\langle Qx_n, Kx_j/\sqrt{d}\rangle)_{j=1}^N$, in (12), since that can be approximately implemented by the MLP in principle; by the universal approximation theorem.

Instead, we use the inner products to obtain low-dimensional approximate representations of any given input (control) in our encoding layer (encoder). Thus, one can view our neural operator as an infinite-dimensional take on the transformer network model.

4. Main Results

Our first result shows that that neural operators are rich enough to approximately minimize the response functional of the follower, to arbitrary precision on any given compact set of actions which the follower can take.

Theorem 4 (ε-Optimal Response Operators) *Let σ be either as in Example 2 or 3. Under Assumptions 1 and 2, for each compact $\mathcal{K}_0 \subseteq \mathcal{U}_0$, and each $\varepsilon > 0$ there is an encoding dimension $d \in \mathbb{N}_+$ and a neural operator $\hat{U} \in \mathcal{NO} : \mathcal{U}_0 \to \mathcal{U}_1$ satisfying*

$$\sup_{u^0 \in \mathcal{K}_0} \|U^\star(u^0) - \hat{U}(u^0)\|_{\mathcal{H}_T^2} \le \varepsilon. \tag{15}$$

A natural question is whether, once the best-response map $U^\star$ has been uniformly approximated over the compact action space $\mathcal{K}_0$ of the follower by a neural operator $\hat{U}$ (as in Theorem 4), our proxy $J_0(\cdot, \hat{U}(\cdot))$ faithfully approximates the follower's true objective $J_0(\cdot, U^\star(\cdot))$, and whether optimizing this proxy leads to an approximate Stackelberg equilibrium. The following result confirms that this is indeed the case, even when the follower's actions are further restricted to a finite-dimensional approximation $p_d(\mathcal{K}_0)$, where $p_d : \mathcal{H}_T^2 \to \mathrm{span}(\{s_i\}_{i=1}^d)$ denotes the orthogonal projection[2] for some hyperparameter $d \in \mathbb{N}+$. To keep the notation concise, we will often abbreviate $p_d(u^0)$ as

$$\hat{u}_d^0 \overset{\text{def.}}{=} p_d(u^0).$$

That is, if a neural operator $\hat{U}$ has been identified which approximates the best-response map $U^\star$, then an approximate Stackelberg equilibrium can be obtained by solving

$$\min_{u^0 \in \mathcal{K}_0} J_0(\hat{u}_d^0, \hat{U}(\hat{u}_d^0)). \tag{16}$$

As the following theorem demonstrates, this is justified since $J_0(\hat{u}_d^0, \hat{U}(\hat{u}_d^0))$ is approximately $J_0(u_0, U^\star(u_0))$.

Theorem 5 (Convergence of Approximated Objectives for Leader and Follower) *The following hold in the setting of Theorem 4.*

(i) *For every $\delta > 0$ there exist $\epsilon > 0$, and $d \in \mathbb{N}_+$ such that if $\hat{U}$ satisfies (15), then*

$$\sup_{u^0 \in \mathcal{K}_0} \left| J_0(u^0, U^\star(u^0)) - J_0(\hat{u}_d^0, \hat{U}(\hat{u}_d^0)) \right| < \delta. \tag{17}$$

---

2. That is, $p_d\left(\sum_{i=1}^\infty \beta_i s_i\right) = \sum_{i=1}^d \beta_i s_i$ for all $u = \sum_{i=1}^\infty \beta_i s_i \in \mathcal{H}_T^2$.

(ii) *There is a $\hat{u}_d^0 \in p_d(\mathcal{K}_0)$ such that the pair $(\hat{u}_d^0, \hat{U}(\hat{u}_d^0))$ is an ε-Stackelberg equilibrium in the sense that the pair $(\hat{u}_d^0, \hat{U}) \in \mathcal{K}_0 \times \bar{\mathcal{U}}_1$ satisfies for all $(u^0, u^1) \in \mathcal{K}_0 \times \mathcal{U}_1$ the inequalities*

$$J_1(u^0, \hat{U}(u^0)) \leq J_1(u^0, u^1) + \epsilon,$$
$$J_0(\hat{u}_d^0, \hat{U}(\hat{u}_d^0)) \leq J_0(u_0, \hat{U}(u_0)) + \epsilon.$$

As with classical uniform approximation results in deep learning, we worked on a compact set in the space of square-integrable $\mathbb{F}$-predictable processes. Though this is standard in the approximation theory literature, it is not so in game theory. However, reducing the problem to compact $\mathcal{K}_0$ is asymptotically coherent in the sense that if we take compact sets $\mathcal{K}_n \subset \mathcal{U}_0$ so that $\cup_n \mathcal{K}_n = \mathcal{U}_0$. Then, for any $\delta_n \downarrow 0$ and $(\hat{u}_d^n, \hat{U}^n)$ satisfying (17) with δ_n, we have that

$$\inf_{u^0 \in \mathcal{U}_0} J_0(u^0, U^\star(u^0)) = \lim_{n \to \infty} J_0(\hat{u}_d^n, \hat{U}^n(\hat{u}_d^n)).$$

Thus, by taking the compact set $\mathcal{K}_n$ and the NO larger, we approximate the optimum of the effective value of the leader $\inf_{u^0 \in \mathcal{U}_0} J_0(u^0, U^\star(u^0))$. Note that if $\mathcal{U}_0$ is not compact or if the problem of the leader does not have additional properties such as convexity or coercivity of $u^0 \mapsto J_0(u^0, U^\star(u^0))$, see e.g. Dal Maso (1993, Theorem 7.12), one cannot easily claim the existence of the optimizer of $\inf_{u^0 \in \mathcal{U}_0} J_0(u^0, U^\star(u^0))$ or the convergence of the family $(\hat{u}_d^n, \hat{U}^n)$. However, these additional structural assumptions are not needed for the purposes of computationally approximating the optimal value of the leader.

More insight can be gained into the inner workings of these results by inspecting the steps in their derivations, at least at a high level. The next section shows the overarching structure of an argument which one can use to derive any such result.

4.1 Rates for Perturbations of Closed-Loop Solutions to Linearized Games

This section shows that, contrary to the general approximation rates for neural operators (see e.g. Lanthaler et al., 2022a; Galimberti et al., 2022), which may be highly inefficient due to the infinite-dimensional nature of the involved spaces, there are stylized conditions which allow for efficient approximation rates of the optimal response map by our attentional neural operator.

Building on the strategy of Shi et al. (2023), we focus on the case where the leader begins by considering a proxy/ansatz/pre-trained control $\bar{u}$. Suppose that $\bar{u}_0 \in \mathcal{H}_T^2$ is a control for the *leader*. Suppose that the leader is comfortable deviating from this strategy to an open-loop control, but only marginally through "small residual perturbations". If those deviations are not too large, efficient approximation rates are possible. More precisely, consider the following situation. Note, for any $d_{\text{low}} \in \mathbb{N}_+$, we denote the d_{low}-dimensional Euclidean unit ball by $B_{d_{\text{low}}} \stackrel{\text{def.}}{=} \{x \in \mathbb{R}^{d_{\text{low}}} : \|x\|_2 \leq 1\}$.

Assumption 3 (Perturbations of an Ansatz) *Let $C \geq 0$, $r > 1$, d_{low}, $\bar{u} \in \mathcal{H}_T^2$, and suppose that Assumption 2 holds. Let $\tilde{\mathcal{K}}_0$ consist of all $u \in \mathcal{H}_T^2$ for which:*

(i) $|\langle u - \bar{u}, s_i \rangle| \leq C \, e^{-ri}$,

(ii) $|\langle U^\star(u - \bar{u}), s_i \rangle| \leq C \, e^{-ri}$.

Let $\pi : B_{d_{\text{low}}} \to \tilde{\mathcal{K}}_0$ be 1-Lipschitz, and suppose that $\mathcal{K}_0 \subseteq \pi(B_{d_{\text{low}}})$.

Remark 6 *Note that, for every $C \geq 0$ and $r > 1$, $\bar{u} \in \mathcal{H}_T^2$ the set $\mathcal{K}_0$ is non-empty since $\bar{u} \in \mathcal{K}_0$.*

In both of the following examples, we consider the sets of "residual controls". Let $C > 0$ and $r > 0$ and define $B_{C,r}(u)$ to be the "exponentially ellipsoidal" set of *open-loop* "residual controls" relative to a control $\bar{u} \in \mathcal{H}_T^2$ to consist of all $u \in \mathcal{H}_T^2$ such that $v \overset{\text{def.}}{=} u - \bar{u} = \sum_{i=1}^{\infty} \beta_i\, s_i \in \mathcal{H}_T^2$ satisfies

$$\max\{|\langle v, s_i\rangle|, |\langle U^*(v), s_i\rangle|\} \leq C\, e^{-ri} \ (\forall i \in \mathbb{N}_+). \tag{18}$$

Illustrated by Figure 2, our primary example of a compact set satisfying the perturbation of the Ansatz assumption above is given by first solving a linear-quadratic proxy of the general (non-linear) Stackelberg game. Then, once an optimal *feedback* control for the leader is derived, for the linear-quadratic proxy of our (non-linear) Stackelberg game, we build $\mathcal{K}_0$ by adding residual open-loop controls in $B_{C,r}$. Intuitively, these residual loops provide added freedom to the closed-loop control optimizing the linear-quadratic proxy required when playing the Stackelberg game.
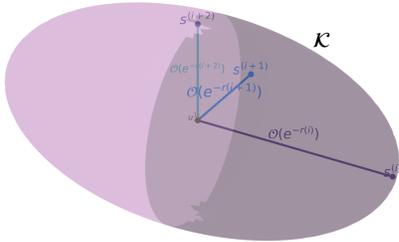
Explanation of Figure: The set $\tilde{\mathcal{K}}_0$ in Example 4 is a (compact) ellipsoidal region in $\mathcal{H}_T^2$ consisting of open-loop controls u which are a small perturbation of a base strategy/control $u^{1:\star}$. The base strategy $u^{1:\star}$ is the solution to a linearization, see (19)-(20), of the dynamic Stackelberg game. The perturbations u of $u^{1:\star}$ in $\tilde{\mathcal{K}}_0$ are built by adding a small residual term in each of the basic directions $s^{(i)} \in \mathcal{S}$ where we allow for possibly large perturbations of the linearized strategy $u^{1:\star}$ for the "low-frequency directions" (i.e. for small values of i) and much smaller perturbations of the linearized strategy for "high-frequency directions" (i.e. for small i). The key subtlety is that the size of the perturbations must decay exponentially in i.

Figure 2: The ellipsoidal compact set of Example 4.

Example 4 (Perturbations of Feedback Control for Linearized Problem—Pt. I)
Consider a finite subset $\{(x_n, v_n^0, v_n^1)\}_{n=1}^N \subset \mathbb{R}^d$ and let A, B_1, B_2, C, D_1, D_2 be matrices minimizing the following MSE problem over all matrices of compatible dimension

$$\sum_{n=1}^{N} \big| f(x_n, v_n^0, v_n^1) - \underbrace{(Ax_n + B_1 v_n^0 + B_2 v^1)}_{\textit{lin. approx. drift}} \big|^2 + \big| \sigma(x_n, v_n^0, v_n^1) - \underbrace{Cx_n + D_1 v_n^0 + D_2 v_n^1}_{\textit{lin. approx. diff.}} \big|^2.$$

Consider the (controlled) "linearized" state-space process $X_{\cdot}^{lin} \overset{\text{def.}}{=} (X_t^{lin})_{t \geq 0}$ given by

$$dX_t^{lin} = (AX_t^{lin} + B_1 u_t^0 + B_2 u_t^1)dt + (CX_t^{lin} + D_1 u_t^0 + D_2 u_t^1)dW_t. \tag{19}$$

Further, assume that $Q_1, Q_2, R_1, R_2, G_1, G_2$ are matrices minimizing the following MSE problem over all matrices of compatible dimension

$$\sum_{n=1}^{N} \sum_{i=1}^{2} \left| L_i(x_n, v_n^0, v_n^1) - \underbrace{\left((Q_i x_n)^\top x_n + (R_i v_n^i)^\top v_n^i \right)}_{\text{lin. approx. running cost}} \right|^2 + \left| \underbrace{g_i(x_n) - (G_i x_n)^\top x_n}_{\text{lin. approx. terminal}} \right|^2.$$

Under Yong (2002, Assumptions DI and H1) the computations on Yong (2002, pages 1034 and 1035), together with Yong (2002, Proposition 2.2 and Theorem 2.3), imply that there is an optimal control $u^{0:\star}$ for the leader minimizing the approximate objective function

$$\mathbb{E}\left[\int_0^T L_0(X_t, u_t^0, u_t^1)\, dt + (G_i X_T)^\top X_T \right] \tag{20}$$

across all controls in $\mathcal{H}_T^2$. Moreover, as shown in (Yong, 2002, Equation (5.12)), $u^{1:\star}$ is given by

$$u_t^{1:\star} = -R_1^{-1} B_1^\top p_t,$$

where $(p_t, q_t)_{t \geq 0}$ solve the BSDE

$$dp_t = p(A^\top p_+ Q_0 X_t)dt + q_t\, dW_t$$
$$p_T = G_0 X_T.$$

Let $\mathcal{K}_0$ be the set of controls for the leader $u \in \mathcal{H}_T^2$ for which there exists some open-loop "residual perturbations" $v \in B_{C,r}(u^{1:\star})$ such that

$$u = u^{1:\star} + v.$$

By construction, $\tilde{\mathcal{K}}_0$ satisfies conditions (i) and (ii) in Assumption 3. Now, fix $d_{\text{low}} \in \mathbb{N}_+$ and let $\pi : B_{d_{\text{low}}} \to \tilde{\mathcal{K}}_0$ be any 1-Lipschitz map. Then, $\mathcal{K}_0 \stackrel{\text{def.}}{=} \pi(B_{d_{\text{low}}})$ satisfies Assumption 3.

The set of open-loop perturbations of the closed-loop optimal control for the linearized game, just described in Example 4, can be efficiently represented using few dimensions, and likewise for its image under the optimal response map $U^\star$. Nevertheless, the set need not be coverable by few controls; that is, it may still have high metric entropy (see e.g. Lorentz, 1966, for an exposée). We, therefore, adopt a technique used in statistical learning/empirical process theory to construct classes of VC-dimension proportional to the metric entropy of high-dimensional balls (see e.g. van der Vaart and Wellner, 1996, Theorem 2.7.11). Namely, we postulate the existence of a latent unknown Lipschitz parameterization of a latent low-dimensional "manifold" of the set of open-loop controls $\mathcal{K}_0$ constructed in Example 4.

Example 5 (Perturbations of Feedback Control for Linearized Problem—Pt. II)
Fix a latent dimension $d \in \mathbb{N}_+$, a latent parameter space $B_d \stackrel{\text{def.}}{=} \{x \in \mathbb{R}^d : |x| \leq 1\}$, and a latent encoding given by a 1-Lipschitz map $\pi : \mathbb{R}^d \to \mathcal{K}_0$ with the property that $\pi(0) = u^{1:\star}$; i.e. parameterizing a latent low-dimensional structure which perturbs the optimal control for the linear-quadratic approximation of the general Stackelberg games. Let $K_{d,\pi} \stackrel{\text{def.}}{=} \pi(B_d)$. By

construction $K_{d,\pi} \subseteq \mathcal{K}_0$; thus, $K_{d,\pi}$ also satisfies Assumption 3. Finally, by van der Vaart and Wellner (1996, Theorem 2.7.11)[3] and Lorentz et al. (1996, Proposition 15.1.3), we have that: for each $\varepsilon > 0$ there exists at-most $N \le 3^d \left(\sqrt{d}2/\varepsilon\right)^d$ controls $\{u^{(n)}\}_{n=1}^N$ forming an ε-cover of $K_{d,\pi}$; that is,

$$\sup_{u. \in K_{d,\pi}} \min_{n=1,\ldots,N} \mathbb{E}\left[\int_0^T \left|v_t - u_t^{(n)}\right|^2 dt\right] \le \varepsilon. \tag{21}$$

In other words, if the latent dimension d is "small", then $K_{d,\pi}$ is small in metric entropy. In turn, the parameter N in our transformer can be taken to also be proportionally small (see Theorem 7).

When the conditions of Assumption 3 are met, we are able to guarantee that the approximating attention neural operator $\hat{U}$ to the best-response map $U^\star$, given by Theorem 4, is determined by relatively few parameters. This is the content of our last main result.

Theorem 7 (Improved Approximation Rates for the Best-Response Map) *Consider the setting of Theorem 4 and additionally suppose that $\mathcal{K}_0$ satisfies Assumption 3. For every $\varepsilon > 0$, the conclusion of Theorem 4 holds and there exists a neural operator $\hat{U} \in \mathcal{NO} : \mathcal{U}_0 \to \mathcal{U}_1$ satisfying the uniform estimate in (15) whose depth, width, decoding dimension, encoding dimension, and attentional complexity are bound above by the estimates in Table 1.*

No. Param	Depth	Width	Decoding Dim. (Q)	Encoding Dim. (d)	Att. Complexity (N)
$\mathcal{O}\left(\varepsilon^{-d_{\text{low}}} \ln(\varepsilon^{-1})^2\right)$	$\mathcal{O}\left(\ln(\varepsilon^{-1})\right)$	$\mathcal{O}(\varepsilon^{-d_{\text{low}}})$	$\mathcal{O}\left(\ln(\varepsilon^{-1/r})\right)$	$\mathcal{O}\left(\ln(\varepsilon^{-1/r})\right)$	$\mathcal{O}(\varepsilon^{-d_{\text{low}}})$

Table 1: Parametric complexity of attentional neural operator approximation of the best response map $U^\star$ on any compact set $\mathcal{K}_0$ satisfying Assumption 3 with constants $r > 0$ and $d_{\text{low}} \in \mathbb{N}_+$, provided that the neural operator is designed with the trainable activation function σ of Example 3.

4.2 Overview of Proof for Theorems 4 and 5

The derivation of Theorem 4 is undertaken in two steps. First, one must show that, under mild conditions, the best response operator set $\mathcal{R}$ is single-valued and there is a continuous selection therein, i.e., the best response operator $U^\star$ is a well-defined continuous non-linear operator. This is key since our neural operator architectures are continuous, and classes of continuous functions cannot uniformly approximate discontinuous functions by the Uniform Limit theorem, see e.g. Munkres (2000, Theorem 21.6).

Since we now know that the best response operator is well-defined and continuous, we need to show that our attentional neural operator class has the power to approximate it or, more generally, functions of the same regularity. This is guaranteed by the following Universal Approximation theorem, guaranteeing that our class of neural operators can approximate, uniformly on compacta, any continuous non-linear operator with square-integrable $\mathbb{F}$.-adapted processes as inputs and outputs.

---

3. We have used the upper-bound of the ε-bracketing numbers on the $\varepsilon/2$-covering numbers (see e.g. van der Vaart and Wellner, 1996, page 84)

Theorem 8 (Universal Approximation for Operators between $\mathbb{F}$-Adapted 2-Integrable Processes) *Let σ be either as in Example 2 or 3. For every (non-empty) compact subset $\mathcal{K}_0 \subset \mathcal{H}_T^2$, each continuous "target" function $f : \mathcal{K}_0 \to \mathcal{H}_T^2$, and every "approximation error" $\varepsilon > 0$ there exists an attentional neural operator $\hat{F} : \mathcal{H}_T^2 \to \mathcal{H}_T^2$ satisfying*

$$\max_{u \in \mathcal{K}_0} \|f(u) - \hat{F}(u)\|_{\mathcal{H}_T^2} \le \varepsilon.$$

In order to apply Theorem 8 and draw our our main conclusion, we need that the non-linear operator $U^\star$ be continuous. This is guaranteed by the following main technical regularity estimate.

Lemma 9 (Hölder Regularity of the Leader's Utility on Optimal Response) *Under Assumptions 1 and 2, there exist constants $C, \alpha > 0$, depending only on K, T, and the Hölder norm in Assumption 2, such that for each $u^0, \tilde{u}^0 \in \mathcal{U}_0 \cap \mathcal{H}_T^2$,*

$$\left| J_0(u^0, U^\star(u^0)) - J_0(\widetilde{u}^0, U^\star(\widetilde{u}^0)) \right| \le \tilde{\omega}\left(\|u^0 - \widetilde{u}^0\|_{\mathcal{H}_T^2} \right)$$

where $\tilde{\omega}(t) \stackrel{\text{def.}}{=} C \max\{|t|^\alpha, |t|\}$ for each $t \ge 0$.

Armed with Lemma 9, we may apply Theorem 8 to the target function $U^\star$ to deduce that that the best response map can be approximated. The main step in using these results to deduce Theorem 5 lies in the continuous dependence of the leader's response on the best response.

It remains to consider when favourable approximation rates can be achieved.

4.2.1 How to Unlock Favourable Approximation Rates Even in Infinite Dimensions?

Generally, deep learning suffers from the curse of dimensionality even in finite dimensions; see, e.g., Lanthaler and Stuart (2023). In the infinite-dimensional setting, this challenge is compounded by two distinct sources of error, whereas only one arises in the finite-dimensional case:

(1) *Dimension-reduction error:* the error introduced by explicitly (or implicitly) approximating an infinite-dimensional problem with a finite-dimensional surrogate;

(2) *Approximation error on Finite-Dimensional Surrogate:* the error incurred when approximating within that finite-dimensional surrogate problem.

The second source of error can be addressed by employing a non-standard choice of activation function, cf. Example 3. This construction is based on the "super-expressive" activation introduced in Zhang et al. (2008), which exploits the bit-extraction mechanism of Bartlett et al. (2019) in a way that standard activations—such as ReLU—simply cannot (cf. Shen et al., 2022, Theorem 2.4).

The first source of error, which arises only in infinite dimensions, can be mitigated by restricting attention not to arbitrary compact sets of inputs, but to those with a favourable structure, namely, compact sets that admit accurate approximation by very low-dimensional surrogates. These "well-structured" compact sets of inputs are defined as follows and they abstract Assumption 3.

Definition 10 ((r, f)-Exponentially Ellipsoidal) *Let $r > 0$, $f : \mathcal{H}_T^2 \to \mathcal{H}_T^2$, and fix an orthonormal basis $\{u_i\}_{i=0}^{\infty}$ of $\mathcal{H}_T^2$. A subset $\mathcal{K} \subseteq \mathcal{H}_T^2$ is of (r, f)-exponentially width if there exists a constant $C > 0$ such that the following holds for each $u \in \mathcal{K}$:*

(i) $u = \sum_{i=1}^{\infty} \beta_i u_i$ and $|\beta_i| \leq C\, e^{-ri}$,

(ii) $f(u) = \sum_{i=1}^{\infty} c_i u_i$ and $|c_i| \leq C\, e^{-ri}$.

Exponentially ellipsoidal compact sets can be well approximated by low-dimensional projections, yet they may still possess large metric entropy, which poses difficulties for standard activation functions (cf. Example 2). This challenge can be alleviated by focusing on low-dimensional subsets, understood as 1-Lipschitz images of low-dimensional Euclidean unit balls, in line with the 1-Lipschitz widths of Petrova and Wojtaszczyk (2023) and the stable optimal manifold widths of Cohen et al. (2022).

Definition 11 ((r, f, d_{low})-Exponential Manifold) *Fix $r > 0$, an "intrinsic dimension" $d_{\text{low}} \in \mathbb{N}_+$, and a mapping $f : \mathcal{H}_T^2 \to \mathcal{H}_T^2$. Fix an orthonormal basis $\{u_i\}_{i=0}^{\infty}$ of $\mathcal{H}_T^2$, and let $\widetilde{\mathcal{K}}$ be an (r, f)-exponentially ellipsoidal subset of $\mathcal{H}_T^2$. A compact subset $\mathcal{K} \subseteq \widetilde{\mathcal{K}}$ is said to be an (r, f, d)-Exponential Manifold if*

$$\mathcal{K} \subseteq \pi(B_{d_{\text{low}}})$$

for some 1-Lipschitz latent parameterization $\pi : B_{d_{\text{low}}} \to \mathcal{H}_T^2$.
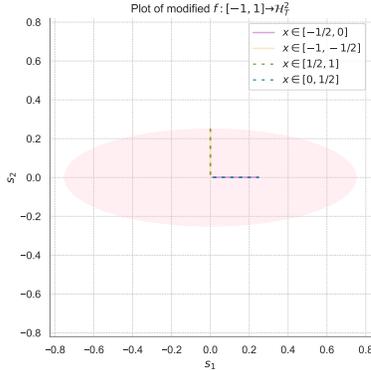
Explanation of Figure: We illustrate Example 8 where a 1-dimensional set embedded as a star-shaped subset of a $d^\star = 2$ dimensional subspace of $\mathcal{H}_T^2$ via π. The dashed line depicts the $(r, f, 1)$-exponential manifold $\mathcal{K}_0$ as traced out by the map π, and the light pink ellipse in the background depicts the projection of the (r, f)-exponential ellipsoid $\tilde{\mathcal{K}}_0$ in $\mathcal{H}_T^2$ onto the span of s_1, s_2.

Figure 3: The (r, f, d_{low})-ellipsoidal-manifold of Example 8 with $d_{\text{low}} = 1$.

Before moving on, we present three increasingly complex examples of exponential manifolds. The first is a simple embedding of a linear subspace when the target function is the identity.

Example 6 (Embeddings of Linear Subspaces) *In the notation of Assumption 3, let $\{s_i\}_{i=0}^{\infty}$ be an enumeration of the basis $\mathcal{S}$ in (11), and let $f : \mathcal{H}_T^2 \to \mathcal{H}_T^2$. Fix $d_{low} \in \mathbb{N}_+$, and define the map $\pi : \mathbb{R}^{d_{\text{low}}} \to \mathcal{H}_T^2$ for any $\beta \in \mathbb{R}^{d_{\text{low}}}$ by*

$$\pi(\beta) = \sum_{i=0}^{d_{\text{low}}-1} \beta_n\, s_i.$$

Then, the set $\mathcal{K}_0 \stackrel{\text{def.}}{=} \pi(B_{d_{\text{low}}})$ is an (r, f, d_{low})-exponential manifold for any $r > 0$.

Example 6 extends naturally to general maps f by intersecting with the pre-image under f of controls in $\mathcal{H}_T^2$ whose basis coefficients (relative to $\mathcal{S}$) decay exponentially.

Example 7 (Embeddings of Linear Subspaces (v2)) *In the context of Example 6, fix $r > 0$ and let $f : \mathcal{H}_T^2 \to \mathcal{H}_T^2$ be an 1-Lipschitz operator. Define*

$$\mathcal{K}_{r,f} \stackrel{\text{def.}}{=} \left\{ v \in \mathcal{H}_T^2 : v = \sum_{i=0}^{d_{\text{low}}-1} |\langle v, s_i \rangle| s_i, \ \text{where for each } i \in \mathbb{N} \, |\langle v, s_i \rangle| \, s_i \le e^{-ri} \right\}. \quad (22)$$

Then, the set $\mathcal{K}_0 \stackrel{\text{def.}}{=} \pi(B_{d_{\text{low}}}) \cap f^{-1}[\mathcal{K}_{r,f}]$ is an (r, f, d_{low})-exponential manifold.

The set traced out by π need not be linear, and can even span several dimensions even if it is intrinsically 1-dimensional, in the sense that the domain of π is $B_1 = [-1, 1]$.

Example 8 (A One-Dimensional Exponential Manifold Sitting in Arbitrarily-Many Dimensions) *In the context of Example 7, define the 1-Lipschitz map $\varphi : \mathbb{R} \to \left[0, \frac{1}{2} \right]$ given for each $x \in \mathbb{R}$ by*

$$\varphi(x) \stackrel{\text{def.}}{=} \left| \left\lfloor x + \tfrac{1}{2} \right\rfloor - x \right|.$$

Fix a dimension $d^\star \in \mathbb{N}_+$, and define the 1-Lipschitz map $\pi : B_1 = [-1, 1] \to \mathcal{H}_T^2$ for each $x \in B_1$ by

$$\pi(x) \stackrel{\text{def.}}{=} \sum_{i=0}^{d^\star} \left(\frac{\varphi(d^\star x)}{d^\star} \, I_{\frac{i}{d^\star} \le |x| \le \frac{i+1}{d^\star}} \right) s_i. \quad (23)$$

Then, set $\pi(B_1) \cap f^{-1} \left[\mathcal{K}_{r,f} \right]$ is an $(r, f, 1)$-exponential manifold; i.e. $d_{\text{low}} = 1$.

Using the trainable activation σ from Example 3 and letting $\mathcal{K}_0$ be an (r, f, d_{low})-exponential manifold for some $r, f > 0$ and small $d_{\text{low}} \in \mathbb{N}_+$, Theorem 8 still holds, with approximation rates as in Table 1. Combining these elements, we state our abstract main approximation result.

Proposition 12 (Approximation of Lipschitz Operators with Attentional Neural Operators) *Let $\mathcal{K}_0 \subseteq \mathcal{H}_T^2$ be compact, $F : \mathcal{H}_T^2 \to \mathcal{H}_T^2$ be an L-Lipschitz (non-linear) operator, and consider respective "dimension reduction" and "approximation" errors $\varepsilon_D, \varepsilon_A > 0$. There exists an attentional neural operator $\hat{F} : \mathcal{H}_T^2 \to \mathcal{H}_T^2$ satisfying*

$$\sup_{u \in \mathcal{K}_0} \|F(u) - \hat{F}(u)\|_{\mathcal{H}_T^2} \le \varepsilon_D + \varepsilon_A.$$

Furthermore, the complexity of the neural operator $\hat{F}$ is recorded in Table 2.

The error $\varepsilon_D > 0$ in Table 2 expresses the "dimension reduction" error resulting from the encoding ($\mathcal{E}$) and decoding ($\mathcal{D}$) maps used in the definition of our attentional neural operator, in Definition 3. That is, ε_D expresses the error made in encoding infinite dimensional objects, namely $\mathbb{F}$-adapted processes, into finite-dimensional objects, namely vectors in some Euclidean spaces. Once the neural operator has implicitly transformed the

	(Case 1)	(Case 2)		(Case 3)
Activation Function (σ)	Example 2	Example 3		Example 3
Shape of Compact ($\mathcal{K}$)	General	(r, f)-**Exponentially Ellipsoidal**		(r, f, d)-**Exponential Manifold**
No. Param	Finite	$\mathcal{O}(Nd^2)$		$\mathcal{O}(Nd^2)$
Depth	Finite	$\mathcal{O}(d)$		$\mathcal{O}(d)$
Width	Finite	$\mathcal{O}(N)$		$\mathcal{O}(N)$
Decoding Dim. (Q)	Finite	$\mathcal{O}(\ln(\varepsilon_D^{-1/r}))$		$\mathcal{O}(\ln(\varepsilon_D^{-1/r}))$
Encoding Dim. (d)	Finite	$\mathcal{O}(\ln(\varepsilon_D^{-1/r}))$		$\mathcal{O}(\ln(\varepsilon_D^{-1/r}))$
Att. Complexity (N)	Finite	$r\left(1 + (C/\varepsilon_A)^2\right)^{1/2})^{C_1 \ln(\epsilon_D^{-1/r})^2}$		$\mathcal{O}\left(\varepsilon_A^{-d_{\text{low}}}\right)$

Table 2: **Complexity of the neural operator in the context of Proposition 12:** We consider three broad cases; with Case 3 being the one used to obtain the rates in Theorem 7. *Case 1:* $\mathcal{K}$ is an arbitrary non-empty compact subset of $\mathcal{H}_T^2$ and σ is the standard non-trainable activation functions of Example (2). *Case 2:* $\mathcal{K}$ is an (r, f)-exponentially ellipsoidal (cf. Definition 10) in controls in $\mathcal{H}_T^2$ and σ is the super-expressive activation function with neuron-specific skip-connections in (3); $c, \tilde{c} > 0$ are absolute constants and the constant $C > 0$ is as in Definition 10. *Case 3:* $\mathcal{K}$ is an (r, f, d)-exponential manifold (cf. Definition 11) in controls in $\mathcal{H}_T^2$ and σ is the super-expressive activation function with neuron-specific skip-connections in (3); $c, \tilde{c} > 0$ are absolute constants and the constant $C > 0$ is as in Definition 10.

approximation problem as an approximation problem between finite-dimensional spaces, it is approximated by the MLP (f) between the encoding and decoding layers of the attentional neural operator. The error $\varepsilon_A > 0$ in Table 2 expresses the error incurred in this finite-dimensional approximation step.

Having outlined our main proof, we note that our two-step argument—first establishing universal approximation for our neural operator class, then proving sufficient continuity of the target operator—also applies to several problems in differential game theory and control.

5. Discussion

5.1 Examples of Stackelberg Games Satisfying Assumption 2

An easily verifiable sufficient condition guarantees that Assumption 2 holds. The condition requires that the Hamiltonian of the follower satisfies a basic level of strong convexity; where the Hamiltonian of the follower is given by

$$H_1(x, u^0, u^1, p_1, q_1) \stackrel{\text{def.}}{=} p_1^\top f(x, u^0, u^1) + Tr(q_1^\top \sigma(x, u^0, u^1)) + L_1(x, u^0, u^1).$$

Proposition 13 (Hamiltonian Strong Convexity Implies Hölder Continuity) *Assume Assumption 1 and that there exists $\kappa > 0$ such that the functions*

$$(x, u^1) \in \mathbb{R}^d \mapsto H_1(x, u^0, u^1, p_1, q_1) - \frac{\kappa}{2}|u^1|^2$$

$$x \in \mathbb{R}^d \mapsto g_1(x)$$

are convex for all values of other variables. Assume also that $\nabla_x L_1$ and $\nabla_x g_1$ are Lipschitz continuous in their variables. Then, $\mathcal{R}(u^0)$ is single-valued and its unique element $\{U_t^\star(u^0)\} = \mathcal{R}(u^0)$ has the following continuous dependence

$$\frac{\kappa}{2}\mathbb{E}\Big[\int_0^T |U_t^\star(u^0) - U_t^\star(\tilde{u}^0)|^2 dt\Big] \le J_1(u^0, U_t^\star(\tilde{u}^0)) - J_1(u^0, U_t^\star(u^0)) \tag{24}$$

and Assumption 2 holds.

Remark 14 (Deriving Alternative Sufficient Conditions) *Another case where one can check the Assumption 2 would be to directly rely on Peng and Wu (1999, Section 3) and check their Assumption H3.1 uniformly in u^0 and rely on the existence of solution to a fully coupled FBSDE. Note that the case in Proposition 13 deviates from Peng and Wu (1999, Section 3). Indeed, our proof is self-contained and only requires the existence of a solution for a BSDE (Backward Stochastic Differential Equation) (44)–rather than an FBSDE–for a given $u^0, u^1, \tilde{u}^1$ which is a trivial problem. Although it is out of the scope of this work, one can then check that the optimally controlled state together with the solution to (44) leads to a solution to the same FBSDE.*

The following is a broad class of functions satisfying the assumption of Proposition 13.

Example 9 *Let A, A^σ be matrices of dimensions $d \times d$, and B, B^σ be matrices of dimensions $d \times d^1$. Let $(C, C^\sigma, C^L) : \mathbb{R}^{d_0} \to \mathbb{R} \times \mathbb{R} \times \mathbb{R}_+$, and $(D, D^\sigma) : \mathbb{R}^{d_0} \to (\mathbb{R}^d)^2$ be Lipschitz functions. Assume that $L_1^1 : \mathbb{R}^{d \times d_1} \to \mathbb{R}$ is convex in x and strongly convex in u^1, and that $L_1^2 : \mathbb{R}^{d_0} \to \mathbb{R}_+$ and $g : \mathbb{R}^d \to \mathbb{R}$ have Lipschitz gradients (in x). Define*

$$f(x, u^0, u^1) \stackrel{\text{def.}}{=} C(u^0)\left(A x + B u^1\right) + D(u^0)$$

$$\sigma(x, u^0, u^1) \stackrel{\text{def.}}{=} C^\sigma(u^0)\left(A^\sigma x + B^\sigma u^1\right) + D^\sigma(u^0)$$

$$L_1(x, u^0, u^1) \stackrel{\text{def.}}{=} C^L(u^0) L_1^1(x, u^1) + L_1^2(u^0).$$

Then the Hamiltonian associated with these f, σ, and L_1 and the map g satisfies the assumption of Proposition 13.

In linear-quadratic Stackelberg games, e.g., Yong (2002), the coefficients $f(x, u^0, u^1)$ and $\sigma(x, u^0, u^1)$ are linear in (x, u^0, u^1), and the costs $L_i(x, u^0, u^1)$ and $g_i(x)$ $i = 0, 1$ are quadratic forms of (x, u^0, u^1). With an additional assumption that the weight matrix of the quadratic form for u^1 in L_1 is positive definite, the assumption of Proposition 13 is satisfied.

5.2 A Reinsurance Example

In game theory, Stackelberg games are canonical models where there is a temporal order between the decisions of the agents. Within this vast literature, we present an example of a stochastic Stackelberg dynamic game between an insurer and a reinsurer (Chen and Shen, 2018). The reinsurer is the leader of the game and decides on an optimal reinsurance premium u^0 to charge, while the insurer, the follower of the game, chooses an optimal proportional reinsurance u^1 to purchase.

The insurer's surplus process X^1 and the reinsuer's surplus process X^0 follow

$$
\begin{aligned}
dX_t^1 &= \big[\rho_1 X_t^1 + c - (1 - u_t^1)u_t^0\big]dt - u_t^1 dC_t, \quad X_0^1 = x_{01} > 0, \\
dX_t^0 &= \big[\rho_0 X_t^0 + (1 - u_t^1)u_t^0\big]dt - (1 - u_t^1)dC_t, \quad X_0^0 = x_{00} > 0.
\end{aligned}
$$

Here ρ_0 and ρ_1 denote the instantaneous interest rates at which the reinsurer's and insurer's surplus are debited/credited, respectively. The claim process C_t is governed by $dC_t = a\,dt - \sigma\,dW_t$ with $a > 0$ and $\sigma > 0$ denoting the mean and volatility of the insurance risk model. The insurance premium rate is given by $c = (1 + \theta)a$, where $\theta > 0$ is the insurer's relative security loading. In practice, it is restricted that $u^0 \in [c, g]$ and $u^1 \in [0, 1]$, where g is an upper bound for the reinsurance premium charge.

With c and C_t substituted into the dynamics of X^0 and X^1, the dynamics of the vector-valued surplus process $X_t \overset{\text{def.}}{=} \big[X_t^0, X_t^1\big]^\top$ follows

$$
\begin{aligned}
dX_t &= \big[AX_t + B(u_t^0, u_t^1)\big]dt + D(u_t^1)dW_t, \quad t > 0, \\
X_0 &= x_0 \overset{\text{def.}}{=} [x_{00}, x_{01}]^\top,
\end{aligned}
\tag{25}
$$

where A, B, and D are defined by

$$
A = \begin{bmatrix} \rho_0 & 0 \\ 0 & \rho_1 \end{bmatrix},
$$

$$
B(u_t^0, u_t^1) = \big[(1 - u_t^1)(u_t^0 - u_t^1), \theta a - (1 - u_t^1)(u_t^0 - a)\big]^\top,
$$

$$
D(u_t^1) = \big[(1 - u_t^1)\sigma, u_t^1\sigma\big]^\top.
$$

Given the utility functions $g_0(\cdot)$, $g_1(\cdot) : \mathbb{R} \to \mathbb{R}$, the insurer and the reinsurer choose a pair (u^0, u^1) to maximize their expected utilities

$$
J_0(u^0, u^1) = \mathbb{E}\big[g_0(X_T^0)\big] \quad \text{and} \quad J_1(u^0, u^1) = \mathbb{E}\big[g_1(X_T^1)\big],
\tag{26}
$$

subject to (25). The above objectives J_0 and J_1 are special cases of the general model (2) and (3) with $L_0(\cdot) \equiv 0$ and $L_1(\cdot) \equiv 0$.

Denote the strategy spaces of the reinsurer (leader) and the insurer (follower), respectively, as

$$
\begin{aligned}
\mathcal{U}_0 &= \big\{u : [0, T] \times \Omega \to \mathbb{R} \,\big|\, u(\cdot) \text{ is } \{\mathcal{F}\}_t\text{-adapted}, u \in [c, g] \text{ on } [0, T] \times \Omega\big\}, \\
\mathcal{U}_1 &= \big\{u : [0, T] \times \Omega \to \mathbb{R} \,\big|\, u(\cdot) \text{ is } \{\mathcal{F}\}_t\text{-adapted}, u \in [0, 1] \text{ on } [0, T] \times \Omega\big\}.
\end{aligned}
$$

A pair of strategies (u^0, u^1) is said to be admissible if $(u^0, u^1) \in \mathcal{U}_0 \times \mathcal{U}_1$ and the state (25) under (u^0, u^1) has a unique strong solution satisfying $\big\{\mathbb{E}\big(\sup_{t \in [0,T]} |X(t)|^2\big)\big\}^{1/2} < \infty$.

A pair $(u^{0,\star}, U^{\star}(u^{0,\star}))$ of admissible strategies constitutes a Stackelberg equilibrium for the reinsurer and insurer if

$$J_1(x_0; u^0, U^{\star}(u^0)) \geq J_1(x_0; u^0, u^1), \quad \forall\, u^1 \in \mathcal{U}_1,$$
$$J_0(x_0; u^{0,\star}, U^{\star}(u^{0,\star})) \geq J_0(x_0; u^0, U^{\star}(u^0)), \quad \forall\, u^0 \in \mathcal{U}_0,$$

where $U^{\star}(\cdot) : \mathcal{U}_0 \to \mathcal{U}_1$ denotes the insurer's best response map to the reinsurer's strategy. It is shown in Chen and Shen (2018, Proposition 5.2) that when the reinsurer and insurer have exponential utility functions $g_0(x) = (-1/\gamma_0) \exp(-\gamma_0 x)$ and $g_1(x) = (-1/\gamma_1) \exp(-\gamma_1 x)$ with $\gamma_0 > 0$ and $\gamma_1 > 0$, respectively, then the reinsurer-insurer Stackelberg game formulated by (25) and (26) admits an explicit equilibrium. For general utility functions, approximate equilibria can be computed using our results.

5.3 Why the Space $\mathcal{H}_T^2$?

It is well known that in control theory, the space $\mathcal{H}_T^2$ is the appropriate space to obtain continuous dependence estimates, see the seminal works (Pardoux and Peng, 1990; El Karoui et al., 1997). Alternative frameworks, such as Wasserstein spaces over path spaces fail to preserve this causal structure in time (Backhoff-Veraguas et al., 2020). While the adapted Wasserstein distance on continuous controls respects causality, it suffers from two drawbacks: (i) we control the processes themselves, not their laws, and (ii) comparing laws across different probability spaces is unnecessary here and would not lead to continuous dependence estimates. Moreover, computing the adapted Wasserstein distance in continuous time is infeasible due to non-convergence of discrete approximations and the prohibitive cost of dynamic programming (Backhoff et al., 2017). Although path-signature spaces from rough path theory (Fermanian et al., 2023) capture causality, most control and dynamic game results are formulated in $\mathcal{H}_T^2$, making them less practical for our purposes.

5.4 Why this Neural Operator Build?

Several neural operator architectures exist, so it is natural to ask why we choose an attention-based design over classical alternatives. We exclude DeepONets (Lu et al., 2021) and Fourier Neural Operators (Li et al., 2020a) because they are defined on function spaces over Euclidean domains and thus are unsuitable for the natural space $\mathcal{H}_T^2$ of open-loop games.

Alternatively, encoder-decoder networks with linear encoders, such as PCA-net (Chan et al., 2015), benefit from universal approximation theorems (Galimberti et al., 2022; Lanthaler, 2023). However, we avoid this direction because linear encoding and decoding are provably less efficient than nonlinear counterparts (Pinkus, 2012; Cohen et al., 2022; Petrova and Wojtaszczyk, 2023). Attention mechanisms provide nonlinear encoding, unlike PCA-net's linear layers.

Finally, our attention mechanism is closely related to the geometric attention in Kratsios et al. (2023a); Acciaio et al. (2023), suggesting our architecture can be adapted for problems requiring relaxed controls—where controlling the law, rather than the process itself, is all which can be hoped for. Notably, these works differ in key ways: the former uses projection-based attention instead of the more common softmax mechanism.

5.4.1 The Role of Super-Expressive Activation Function (8)

Additionally, the approximation guarantees in the latter do not utilize super-expressive activation functions (Yarotsky, 2021; Jiao et al., 2023; Zhang et al., 2022) needed for dimension-free approximation rates of Lipschitz functions; the same is true for the approximation guarantees for PCA-Net (Lanthaler, 2023) and more general encoder-decoder networks (Galimberti et al., 2022).

The technical impact of incorporating such activation functions lies in addressing the two main sources of error in operator learning: (i) reduction of the infinite-dimensional problem to a finite-dimensional surrogate, and (ii) approximation error from solving the surrogate only approximately. Our activation choice effectively eliminates the latter, which, to the best of our knowledge, is new. This is crucial since the finite-dimensional error is often the main bottleneck (Lanthaler and Stuart, 2023). As a result, reasonable approximation rates can be achieved when the input set is "well-shaped" (exponentially ellipsoidal in the sense of Definition A.8), whereas such rates are generally unattainable with standard activations (Galimberti et al., 2022, Corollary 1).

5.5 Solving Stackelberg Games Guiding Training by a Theory-Informed Heuristic

We close our manuscript by briefly discussing how our theory can inform practical implementations of our model. Specifically, the necessary conditions established in our analysis can be used as optimization heuristics. This strategy follows a broader paradigm in machine learning, where intractable objectives are replaced by tractable necessary conditions that are easier to optimize. Well-known examples include maximum-principle–based heuristics (Li et al., 2018), the evidence lower bound (ELBO) in variational inference (Blei et al., 2017), or KKT-based necessary conditions as criteria for bi-level optimization problems (Outrata et al., 2013) such as Stackelberg games; all of which demonstrate how principled relaxations of exact criteria can yield effective and scalable training methods.

In our setting, since the action space of the leader $\mathcal{K}_0$ is compact, for any $\eta > 0$ one can select a finite η-net $\{u^{(i)}\}_{i=1}^N$ on $\mathcal{K}_0$ that captures every possible action of the leader up to an error of at most η; that is, for each $u^0 \in \mathcal{K}_0$, there exists some $i \in \{1, \ldots, N\}$ such that

$$\|u^{(i)} - u^0\|_{\mathcal{H}_T^2} \leq \eta.$$

Next, one may fix a compact set $\hat{\mathcal{NO}}$ of neural operators, for example by restricting depth, width, and the magnitudes of parameters such as weights and biases.

A natural practical approach is then to consider the optimization problem

$$\min_{\hat{U} \in \hat{\mathcal{NO}}} \sum_{i=1}^N J_1\big(\hat{u}_d^{(i)}, \hat{U}(\hat{u}_d^{(i)})\big), \tag{27}$$

where $\hat{u}_d^{(i)}$ denotes a suitable finite-dimensional projection of $u^{(i)}$. The finite-dimensional objective in (27) is guided by the necessary conditions in Theorem 5, and is expected to approximate the true objective of the leader provided the neural operator captures the best response map $U^\star$ well, as suggested by the approximability guarantee in Theorem 4.

In this sense, (27) provides a practical implementation heuristic, directly informed by our theoretical results. By Theorem 5, it yields an approximate Stackelberg equilibrium once the leader's optimal action is identified by solving (16).

6. Conclusion

We show that, the best response map for a broad class of dynamic Stackelberg games with random effects, can be approximated by neural operators (Theorem 4). This implies that Stackelberg games may be solved (approximate computation of equilibria in Theorem 5), and we can explicitly describe how they are played (approximate representation of best response map) without making highly stylized conditions needed in classical results needed to derive analytic expressions for the best response and the equilibirum itself (see e.g. Example 4). Thus, our approach allows one to specify realistic dynamics and action sets, and obtain an approximate solution.

We further show that if the space of actions for the leader consists of perturbations of the optimal solution for a linearized version of the Stackelberg game, then the approximation rates can be further improved beyond our worst-case upper-bounds (Theorem 7).

Future Research

Our universal approximation theorem for non-linear operators between spaces of $\mathbb{F}$-adapted square-integrable processes—namely, the quantitative version of Theorem 8 in the appendix (cf. Proposition 12)—provides explicit approximation guarantees. As with most approximation theorems in infinite-dimensional Hilbert spaces, the approximation rates themselves are particularly un-insightful. Nevertheless, we find that if one uses a trainable variant of the "super-expressive" activation function of Zhang et al. (2022) (see (8)) when defining our neural operator and if the compact subset on which the non-linear operator is sufficiently close to being finite-dimensional (see Definition 10) then exponential approximation rates can be achieved (see Table 2). The main challenge is then to verify that the non-linear operator being approximated and the relevant compact set of adapted open-loop controls satisfy the required compatibility conditions. Constructing examples of Stackelberg games with these properties, i.e. games whose best-response operator is efficiently approximable, is a highly non-trivial but interesting project and is the objective of our future research.

We include the relevant quantitative universal approximation theorem and conditions for exponential approximation rates in the appendix of this paper so that these results may be used in other operator learning problem in stochastic analysis and its applications; e.g. to economics and finance.

Acknowledgments

NEURAL OPERATORS CAN PLAY DYNAMIC STACKELBERG GAMES

Appendix A. Proofs

A.1 Continuous Dependence of J_i

For the next result, it is convenient to recall that the $\mathcal{H}_T^\infty$ norm of a process $X.$ in $\mathcal{H}_T^2$ is given by

$$\|X.\|_{\mathcal{H}_T^\infty} \overset{\text{def.}}{=} \mathbb{E}\left[\sup_{0 \leq t \leq T} |X_t|\right].$$

Lemma 15 *Under Assumption 1, we have the following uniform continuity guarantees*

$$\left\|\left|X_\cdot^{u^0,u^1} - X_\cdot^{u^0,\tilde{u}^1}\right|^2\right\|_{\mathcal{H}_T^\infty} \leq C \left\|u^1 - \tilde{u}^1\right\|_{\mathcal{H}_T^2}^2, \tag{28}$$

$$\left\|\left|X_\cdot^{u^0,u^1} - X_\cdot^{\tilde{u}^0,u^1}\right|^2\right\|_{\mathcal{H}_T^\infty} \leq C \left\|u^0 - \tilde{u}^0\right\|_{\mathcal{H}_T^2}^2, \tag{29}$$

where $C \geq 0$ is a constant that depends on T and the Lipschitz constant K in Assumption 1; moreover,

$$\left\|X_\cdot^{u^0,u^1} - X_\cdot^{u^0,\tilde{u}^1}\right\|_{\mathcal{H}_T^\infty} \leq C \left\|u^1 - \tilde{u}^1\right\|_{\mathcal{H}_T^2}, \tag{30}$$

$$\left\|X_\cdot^{u^0,u^1} - X_\cdot^{\tilde{u}^0,u^1}\right\|_{\mathcal{H}_T^\infty} \leq C \left\|u^0 - \tilde{u}^0\right\|_{\mathcal{H}_T^2}. \tag{31}$$

Proof We prove (28), noting that the derivation of (29) is carried out in a nearly identical similar manner. Let X and $\widetilde{X}$ be the strong solution of (1) under (u^0, u^1) and $(u^0, \tilde{u}^1)$, respectively, i.e., $X = X^{u^0,u^1}$, $\widetilde{X} = X^{u^0,\tilde{u}^1}$. Denote $h(s) = h(X_s, u_s^0, u_s^1)$ and $\tilde{h}(s) = h(\widetilde{X}_s, u_s^0, \tilde{u}_s^1)$, where h is as defined in Assumption 1.

By the Jensen's inequality, we have

$$\begin{aligned}
|X_r - \widetilde{X}_r|^2 &= \left|\int_0^r f(s) - \tilde{f}(s)ds + \int_0^r \sigma(s) - \tilde{\sigma}(s)dW_s\right|^2 \\
&\leq 2\left|\int_0^r f(s) - \tilde{f}(s)ds\right|^2 + 2\left|\int_0^r \sigma(s) - \tilde{\sigma}(s)dW_s\right|^2 \\
&\leq 2\int_0^r |f(s) - \tilde{f}(s)|^2 ds + 2\left|\int_0^r \sigma(s) - \tilde{\sigma}(s)dW_s\right|^2.
\end{aligned}$$

We deduce that

$$\sup_{0 \leq r \leq t} |X_r - \widetilde{X}_r|^2 \leq 2\int_0^t |f(s) - \tilde{f}(s)|^2 ds + 2\sup_{0 \leq r \leq t}\left|\int_0^r \sigma(s) - \tilde{\sigma}(s)dW_s\right|^2. \tag{32}$$

By the Burkholder-Davis-Gundy's inequality (Cohen and Elliott, 2015, Theorem 11.5.5), Tonelli's theorem (Cohen and Elliott, 2015, Theorem 1.4.6), and Assumption 1, we have

$$\begin{aligned}
\mathbb{E}\sup_{0 \leq r \leq t}\left|\int_0^r \sigma(s) - \tilde{\sigma}(s)dW_s\right|^2 &\leq C \cdot \mathbb{E}\int_0^t |\sigma(s) - \tilde{\sigma}(s)|^2 ds \\
&\leq C \cdot \int_0^t \mathbb{E}\sup_{0 \leq r \leq s} |X_r - \widetilde{X}_r|^2 + |u_s^1 - \tilde{u}_s^1|^2 \, ds. \tag{33}
\end{aligned}$$

28

By the Tonelli's theorem and Assumption 1, we have

$$
\mathbb{E} \int_0^t |f(s) - \widetilde{f}(s)|^2 \, ds \leq C \cdot \int_0^t \left|X_s - \widetilde{X}_s\right|^2 + |u_s^1 - \widetilde{u}_s^1|^2 \, ds
$$

$$
\leq C \cdot \int_0^t \mathbb{E} \sup_{0 \leq r \leq s} |X_r - \widetilde{X}_r|^2 + |u_s^1 - \widetilde{u}_s^1|^2 \, ds. \tag{34}
$$

Combining (32), (33), and (34), we obtain

$$
\mathbb{E} \sup_{0 \leq r \leq t} |X_r - \widetilde{X}_r|^2 \leq C \cdot \int_0^t \mathbb{E} \sup_{0 \leq r \leq s} |X_s - \widetilde{X}_s|^2 ds + C \cdot \mathbb{E} \int_0^t |u_s^1 - \widetilde{u}_s^1|^2 \, ds.
$$

From the above inequality and the Grönwall's inequality, we obtain

$$
\mathbb{E}\left[\sup_{0 \leq t \leq T} \left|X_t^{u^0, u^1} - X_t^{u^0, \widetilde{u}^1}\right|^2 \right] \leq C \cdot \mathbb{E} \int_0^T |u_t^1 - \widetilde{u}_t^1|^2 dt. \tag{35}
$$

Arguing nearly identically, we may also obtain

$$
\mathbb{E}\left[\sup_{0 \leq t \leq T} \left|X_t^{u^0, u^1} - X_t^{\widetilde{u}^0, u^1}\right|^2 \right] \leq C \cdot \mathbb{E} \int_0^T |u_t^0 - \widetilde{u}_t^0|^2 dt. \tag{36}
$$

Note that the right-hand side of (35) (resp. (36)) is simply a scalar multiple (by a factor of $C \geq 0$) of the squared norm between u^1 and $\widetilde{u}^1$ (resp. u^0 and $\widetilde{u}^0$). By definition of the $\|\cdot\|_{\mathcal{H}_T^\infty}$ norm applied to the "squared difference processes" $\left|X_\cdot^{u^0, u^1} - X_\cdot^{u^0, \widetilde{u}^1}\right|^2$ and $\left|X_\cdot^{u^0, u^1} - X_\cdot^{\widetilde{u}^0, u^1}\right|^2$, (35) and (36) can be re-expressed as (28) and (29). (30) and (31) follows from (28) and (29) by the Cauchy-Schwarz inequality; thus concluding our proof. ∎

Lemma 16 *The costs J_i, $i = 0, 1$ are Lipschitz in u^0 and u^1 such that for each $u^i, \tilde{u}^i \in \mathcal{U}_i \cap \mathcal{H}_T^2$, $i = 0, 1$,*

$$
\left|J_i(u^0, u^1) - J_i(\tilde{u}^0, \tilde{u}^1)\right| \leq C \left(\left\|u^0 - \widetilde{u}^0\right\|_{\mathcal{H}_T^2} + \left\|u^1 - \widetilde{u}^1\right\|_{\mathcal{H}_T^2}\right), \tag{37}
$$

where C is a constant depending on T and the Lipschitz constant K in Assumption 1.

Proof For each $u^i, \tilde{u}^i \in \mathcal{U}_i \cap \mathcal{H}_T^2$, $i = 0, 1$, it follows from Assumption 1 that

$$
\left|J_i(u^0, u^1) - J_i(\widetilde{u}^0, \widetilde{u}^1)\right|
$$

$$
\leq \mathbb{E}\left[\int_0^T \left|L_i(X_t^{u^0, u^1}, u_t^0, u_t^1) - L_i(X_t^{\widetilde{u}^0, \widetilde{u}^1}, \widetilde{u}_t^0, \widetilde{u}_t^1)\right| dt + \left|g_i(X_T^{u^0, u^1}) - g_i(X_T^{\widetilde{u}^0, \widetilde{u}^1})\right| \right]
$$

$$
\leq \mathbb{E}\left[\int_0^T K(|X_t^{u^0, u^1} - X_t^{\widetilde{u}^0, \widetilde{u}^1}| + |u_t^0 - \widetilde{u}_t^0| + |u_t^1 - \widetilde{u}_t^1|) \, dt + K|X_T^{u^0, u^1} - X_T^{\widetilde{u}^0, \widetilde{u}^1}| \right]
$$

$$
\leq K(1 + T)\left\|X^{u^0, u^1} - X^{\widetilde{u}^0, \widetilde{u}^1}\right\|_{\mathcal{H}_T^\infty} + K\left(\left\|u^0 - \widetilde{u}^0\right\|_{\mathcal{H}_T^2} + \left\|u^0 - \widetilde{u}^0\right\|_{\mathcal{H}_T^2}\right),
$$

29

where the last inequality is due to the Cauchy-Schwarz inequality. By the triangle inequality and Lemma 15, we have

$$
\begin{aligned}
\left\| X_{\cdot}^{u^0,u^1} - X_{\cdot}^{\tilde{u}^0,\tilde{u}^1} \right\|_{\mathcal{H}_T^\infty} &\leq 2\left(\left\| X_{\cdot}^{u^0,u^1} - X_{\cdot}^{\tilde{u}^0,u^1} \right\|_{\mathcal{H}_T^\infty} + \left\| X_{\cdot}^{\tilde{u}^0,u^1} - X_{\cdot}^{\tilde{u}^0,\tilde{u}^1} \right\|_{\mathcal{H}_T^\infty} \right) \\
&\leq C\left(\left\| u^0 - \tilde{u}^0 \right\|_{\mathcal{H}_T^2} + \left\| u^1 - \tilde{u}^1 \right\|_{\mathcal{H}_T^2} \right).
\end{aligned}
$$

It then follows that

$$
\left| J_1(u^0,u^1) - J_1(\tilde{u}^0,\tilde{u}^1) \right| \leq C\left(\left\| u^0 - \tilde{u}^0 \right\|_{\mathcal{H}_T^2} + \left\| u^1 - \tilde{u}^1 \right\|_{\mathcal{H}_T^2} \right).
$$

∎

Lemma 17 *Assume Assumption 1 and suppose that for all $u^0 \in \mathcal{K}_0$, the set $\mathcal{R}(u^0)$ is not empty. Let $U^\star(u^0) \in \mathcal{R}(u^0)$ be a selected element. Then the map $u^0 \in \mathcal{K}_0 \mapsto J_1(u^0, U^\star(u^0))$ is continuous; particularly, for each $u^0, \tilde{u}^0 \in \mathcal{U}_0 \cap \mathcal{H}_T^2$, it satisfies for the follower that*

$$
\left| J_1(u^0, U^\star(u^0)) - J_1(\tilde{u}^0, U^\star(\tilde{u}^0)) \right| \leq C\left\| u^0 - \tilde{u}^0 \right\|_{\mathcal{H}_T^2}, \tag{38}
$$

where the constant C depends on T and the Lipschitz constant K in Assumption 1.

Remark 18 *With the assumption of Lemma 17, $u^0 \in \mathcal{K}_0 \mapsto J_1(u^0, U^\star(u^0))$ is continuous but as shown in the counterexample in Example 1, $u^0 \in \mathcal{K}_0 \mapsto J_0(u^0, U^\star(u^0))$ might fail to be continuous.*

Proof By (37), we have for any $u \in \mathcal{U}_0 \cap \mathcal{H}_T^2$,

$$
\begin{aligned}
J_1(u^0, u) &\leq J_1(\tilde{u}^0, u) + \left| J_1(\tilde{u}^0, u) - J_1(u^0, u) \right| \\
&\leq J_1(\tilde{u}^0, u) + C\|u^0 - \tilde{u}^0\|_{\mathcal{H}_T^2},
\end{aligned}
$$

and furthermore

$$
\begin{aligned}
J_1(u^0, U^\star(u^0)) = \inf_u J_1(u^0, u) &\leq \inf_u J_1(\tilde{u}^0, u) + C\|u^0 - \tilde{u}^0\|_{\mathcal{H}_T^2} \\
&= J_1(\tilde{u}^0, U^{1,\star}(\tilde{u}^0)) + C\|u^0 - \tilde{u}^0\|_{\mathcal{H}_T^2}. \tag{39}
\end{aligned}
$$

Exchanging the roles of $(u^0, U^{1,\star}(u^0))$ and $(\tilde{u}^0, U^{1,\star}(\tilde{u}^0))$ in (39), we have

$$
J_1(\tilde{u}^0, U^\star(\tilde{u}^0)) \leq J_1(u^0, U^\star(u^0)) + C\|u^0 - \tilde{u}^0\|_{\mathcal{H}_T^2}. \tag{40}
$$

Combining (39) and (40), we obtain (38). ∎

Proof of Lemma 9 By (39) and Assumption 2, we have

$$
\begin{aligned}
\left| J_0(u^0, U^\star(u^0)) - J_0(\tilde{u}^0, U^\star(\tilde{u}^0)) \right| &\leq C\|u^0 - \tilde{u}^0\|_{\mathcal{H}_T^2} + C\left\| U^\star(u^0) - U^\star(\tilde{u}^0) \right\|_{\mathcal{H}_T^2} \\
&\leq C\|u^0 - \tilde{u}^0\|_{\mathcal{H}_T^2} + (C\widetilde{C})\|u^0 - \tilde{u}^0\|_{\mathcal{H}_T^2}^\alpha, \tag{41}
\end{aligned}
$$

for some constant $C \geq 0$ depending only on T, K, and for $\tilde{C}, \alpha \geq 0$ the Hölder continuity parameters of the best response function given by Assumption 2. Define $C' \overset{\text{def.}}{=} C \max\{1, \tilde{C}\}/2$. Note that $a + b \leq 2 \max\{a, b\}$ for every $a, b \geq 0$. Then (41) implies that

$$\left| J_0(u^0, U^\star(u^0)) - J_0(\tilde{u}^0, U^\star(\tilde{u}^0)) \right| \leq C' \cdot \max \left\{ \|u^0 - \tilde{u}^0\|_{\mathcal{H}_T^2}, \|u^0 - \tilde{u}^0\|_{\mathcal{H}_T^2}^\alpha \right\}. \tag{42}$$

Define the modulus of continuity $\tilde{\omega}(t) \overset{\text{def.}}{=} C' \max\{|t|^\alpha, |t|\}$. Since $t \leq t^\alpha$ when $t \in [0, 1)$ and $t \geq t^\alpha$ when $t \in [1, \infty)$ then, (42) cleans up as

$$\left| J_0(u^0, U^\star(u^0)) - J_0(\tilde{u}^0, U^\star(\tilde{u}^0)) \right| \leq \tilde{\omega}\left(\|u^0 - \tilde{u}^0\|_{\mathcal{H}_T^2} \right). \tag{43}$$

Relabelling C as C' concludes our proof. ∎

A.2 Proof of the Sufficient Conditions for Assumption 2 in Proposition 13

Proof of Proposition 13 We fix $u^0 \in \mathcal{U}_0$ and choose $u^1, \tilde{u}^1 \in \mathcal{U}_1$, $\lambda \in [0, 1]$. Denote X_t^λ the state controlled by the pair $(u^0, u^\lambda) = (u^0, \lambda u^1 + (1 - \lambda)\tilde{u}^1)$. Thus,

$$J_1(u^0, u^\lambda) - \lambda J_1(u^0, u^1) - (1 - \lambda)J_1(u^0, \tilde{u}^1)$$

$$= \mathbb{E}\left[\int_0^T L_1(X_t^\lambda, u_t^0, u_t^\lambda) - \lambda L_1(X_t^1, u_t^0, u_t^1) - (1 - \lambda)L_1(X_t^0, u_t^0, \tilde{u}_t^1)dt \right]$$

$$+ \mathbb{E}\left[g_1(X_T^\lambda) - \lambda g_1(X_T^1) - (1 - \lambda)g_1(X_T^0) \right]$$

$$= \mathbb{E}\left[\int_0^T L_1(X_t^\lambda, u_t^0, u_t^\lambda) - \lambda L_1(X_t^1, u_t^0, u_t^1) - (1 - \lambda)L_1(X_t^0, u_t^0, \tilde{u}_t^1)dt \right]$$

$$- \mathbb{E}\left[\lambda(g_1(X_T^1) - g_1(X_T^\lambda)) + (1 - \lambda)(g_1(X_T^0) - g_1(X_T^\lambda)) \right]$$

$$\leq - \lambda \mathbb{E}\left[\int_0^T L_1(X_t^1, u_t^0, u_t^1) - L_1(X_t^\lambda, u_t^0, u_t^\lambda)dt + \nabla_x g_1(X_T^\lambda)^\top (X_T^1 - X_T^\lambda) \right]$$

$$- (1 - \lambda)\mathbb{E}\left[\int L_1(X_t^0, u_t^0, \tilde{u}_t^1) - L_1(X_t^\lambda, u_t^0, u_t^\lambda)dt + \nabla g_1(X_T^\lambda)^\top (X_T^0 - X_T^\lambda) \right]$$

where we used the the convexity of g_1 to obtain the last line. We now provide an upper bound for the last two terms. For a fixed λ, by the definition of X^λ, u^λ and our assumptions on H_1, the function $(y, z) \mapsto \nabla_x H_1(X_t^\lambda, u_t^0, u_t^\lambda, y, z)$ is uniformly Lipschitz continuous and $\nabla g_1(X_T^\lambda)$ is square integrable. Thus, there exists a unique solution $(Y_t^\lambda, Z_t^\lambda)$ to the BSDE

$$dY_t^\lambda = -\nabla_x H_1(X_t^\lambda, u_t^0, u_t^\lambda, Y_t^\lambda, Z_t^\lambda)dt + Z_t^\lambda dW_t, \tag{44}$$

$$Y_T^\lambda = \nabla g_1(X_T^\lambda), \tag{45}$$

so that by Itô's formula we have

$$\mathbb{E}\left[\nabla g_1(X_T^\lambda)^\top (X_T^0 - X_T^\lambda) \right] = \mathbb{E}\left[Y_T^{\lambda\top}(X_T^0 - X_T^\lambda) \right]$$

$$= \mathbb{E}\left[\int_0^T Y_t^{\lambda\top}(f(X_t^0, u_t^0, \tilde{u}_t^1) - f(X_t^\lambda, u_t^0, u_t^\lambda))dt \right]$$

$$+ \mathbb{E}\Big[Tr\left(Z_t^{\lambda\top}(\sigma(X_t^0, u_t^0, \tilde{u}_t^1) - \sigma(X_t^\lambda, u_t^0, u_t^\lambda))\right) dt\Big]$$

$$- \mathbb{E}\Big[\nabla_x H_1(X_t^\lambda, u_t^0, u_t^\lambda, Y_t^\lambda, Z_t^\lambda)(X_t^0 - X_t^\lambda)dt\Big]$$

$$= \mathbb{E}\Big[\int_0^T H_1(X_t^0, u_t^0, \tilde{u}_t^1, Y_t^\lambda, Z_t^\lambda) - H_1(X_t^\lambda, u_t^0, u_t^\lambda, Y_t^\lambda, Z_t^\lambda)dt\Big]$$

$$- \mathbb{E}\Big[\nabla_x H_1(X_t^\lambda, u_t^0, u_t^\lambda, Y_t^\lambda, Z_t^\lambda)(X_t^0 - X_t^\lambda)dt\Big]$$

$$- \mathbb{E}\Big[\int_0^T L_1(t, X_t^0, u_t^0, \tilde{u}_t^1) - L_1(t, X_t^\lambda, u_t^0, u_t^\lambda)dt\Big].$$

Thus, we obtain

$$\mathbb{E}\Big[\nabla g_1(X_T^\lambda)^\top(X_T^0 - X_T^\lambda)\Big] + \mathbb{E}\Big[\int_0^T L_1(X_t^0, u_t^0, \tilde{u}_t^1) - L_1(X_t^\lambda, u_t^0, u_t^\lambda)dt\Big]$$

$$= \mathbb{E}\Big[\int_0^T H_1(X_t^0, u_t^0, \tilde{u}_t^1, Y_t^\lambda, Z_t^\lambda) - H_1(X_t^\lambda, u_t^0, u_t^\lambda, Y_t^\lambda, Z_t^\lambda)dt\Big]$$

$$- \mathbb{E}\Big[\nabla_x H_1(X_t^\lambda, u_t^0, u_t^\lambda, Y_t^\lambda, Z_t^\lambda)(X_t^0 - X_t^\lambda)dt\Big]$$

and

$$\mathbb{E}\Big[\nabla g_1(X_T^\lambda)^\top(X_T^1 - X_T^\lambda)\Big] + \mathbb{E}\Big[\int_0^T L_1(X_t^1, u_t^0, u_t^1) - L_1(X_t^\lambda, u_t^0, u_t^\lambda)dt\Big]$$

$$= \mathbb{E}\Big[\int_0^T H_1(X_t^1, u_t^0, u_t^1, Y_t^\lambda, Z_t^\lambda) - H_1(X_t^\lambda, u_t^0, u_t^\lambda, Y_t^\lambda, Z_t^\lambda)dt\Big]$$

$$- \mathbb{E}\Big[\nabla_x H_1(X_t^\lambda, u_t^0, u_t^\lambda, Y_t^\lambda, Z_t^\lambda)(X_t^1 - X_t^\lambda)dt\Big].$$

Combining these equalities, we obtain

$$- \lambda\mathbb{E}\Big[\int_0^T L_1(X_t^1, u_t^0, u_t^1) - L_1(X_t^\lambda, u_t^0, u_t^\lambda)dt + \nabla_x g_1(X_T^\lambda)^\top(X_T^1 - X_T^\lambda)\Big]$$

$$- (1-\lambda)\mathbb{E}\Big[\int L_1(X_t^0, u_t^0, \tilde{u}_t^1) - L_1(X_t^\lambda, u_t^0, u_t^\lambda)dt + \nabla_x g_1(X_T^\lambda)^\top(X_T^0 - X_T^\lambda)\Big]$$

$$= \mathbb{E}\Big[\int_0^T H_1(\lambda X_t^0 + (1-\lambda)X_t^1, u_t^0, u_t^\lambda, Y_t^\lambda, Z_t^\lambda) - \lambda H_1(X_t^1, u_t^0, u_t^1, Y_t^\lambda, Z_t^\lambda)dt\Big]$$

$$- \mathbb{E}\Big[\int_0^T (1-\lambda)H_1(X_t^0, u_t^0, \tilde{u}_t^1, Y_t^\lambda, Z_t^\lambda)dt\Big]$$

$$+ \mathbb{E}\Big[\int_0^T H_1(X_t^\lambda, u_t^0, u_t^\lambda, Y_t^\lambda, Z_t^\lambda) - H_1(\lambda X_t^0 + (1-\lambda)X_t^1, u_t^0, u_t^\lambda, Y_t^\lambda, Z_t^\lambda)dt\Big]$$

$$+ \mathbb{E}\Big[\int_0^T \nabla_x H_1(X_t^\lambda, u_t^0, u_t^\lambda, Y_t^\lambda, Z_t^\lambda)(\lambda X_t^1 + (1-\lambda)X_t^0 - X_t^\lambda)dt\Big].$$

By the convexity of H_1 in x and strong convexity in u^1 we have that

$$\mathbb{E}\Big[\int_0^T H_1(\lambda X_t^0 + (1-\lambda)X_t^1, u_t^0, u_t^\lambda, Y_t^\lambda, Z_t^\lambda) - \lambda H_1(X_t^1, u_t^0, u_t^1, Y_t^\lambda, Z_t^\lambda)\Big]$$

$$- \mathbb{E}\Big[\int_0^T (1-\lambda) H_1(X_t^0, u_t^0, \tilde{u}_t^1, Y_t^\lambda, Z_t^\lambda) dt \Big] \leq -\frac{\kappa\lambda(1-\lambda)}{2} \mathbb{E}\Big[\int_0^T |u_t^1 - \tilde{u}_t^1|^2 dt \Big]$$

and

$$H_1(x, u^0, u^1, y, z) - H_1(\tilde{x}, u^0, u^1, y, z) + \nabla_x H_1(x, u^0, u^1, y, z)(\tilde{x} - x) \leq 0.$$

Thus,

$$J_1(u^0, u^\lambda) - \lambda J_1(u^0, u^1) - (1-\lambda) J_1(u^0, \tilde{u}^1) \leq -\frac{\kappa\lambda(1-\lambda)}{2} \mathbb{E}\Big[\int_0^T |u_t^1 - \tilde{u}_t^1|^2 dt \Big]$$

which is the strong convexity in u^1.

Due to this strong convexity, for all u^0, there exists a unique minimizer for the optimal response of the follower that we denote $U^\star(u^0)$. To prove (24), we use the first order optimality condition for $U^\star(u^0)$ which reads

$$J_1(u^0, u^1) - J_1(u^0, U^\star(u^0)) \geq \frac{\kappa}{2} \mathbb{E}\Big[\int_0^T |u_t^1 - U_t^\star(u^0)|^2 dt \Big]$$

which is (24) for $u^1 = U^\star(\tilde{u}^0)$.

To conclude the proof of the Proposition it remains to prove the $1/2$-Hölder continuity of the best response function which allows us to verify Assumption 2. By (24), we have

$$\begin{aligned}
&\big\| U^\star(u^0) - U^\star(\tilde{u}^0) \big\|_{\mathcal{H}_T^2} \\
&\leq \frac{2}{\kappa^{1/2}} \Big\{ \big| J_1(\tilde{u}^0, U^\star(\tilde{u}^0)) - J_1(u^0, U^\star(u^0)) \big|^{1/2} + \big| J_1(u^0, U^\star(\tilde{u}^0)) - J_1(\tilde{u}^0, U^\star(\tilde{u}^0)) \big|^{1/2} \Big\}.
\end{aligned} \tag{46}$$

By (37) and (38), we have

$$\big| J_1(u^0, U^\star(\tilde{u}^0)) - J_1(\tilde{u}^0, U^\star(\tilde{u}^0)) \big| \leq C \cdot \| u^0 - \tilde{u}^0 \|_{\mathcal{H}_T^2}, \tag{47}$$

$$\big| J_1(\tilde{u}^0, U^\star(\tilde{u}^0)) - J_1(u^0, U^\star(u^0)) \big| < C \cdot \| u^0 - \tilde{u}^0 \|_{\mathcal{H}_T^2}. \tag{48}$$

Then the desired result follows from (46), (47) and (48). ∎

A.3 Additional Background on the Wiener Chaos

For any time $0 \leq t \leq T$, since the σ-algebra $\mathcal{F}_t$ is generated by $\{W_s\}_{0 \leq s \leq t}$, any $u \in L^2(\mathcal{F}_t)$ admits the following *Wiener chaos expansion*

$$u = \mathbb{E}[u] + \sum_{i=0}^\infty \int_0^T \int_0^{s_n} \cdots \int_0^{s_2} f_i(s_n, \ldots, s_1) \, dW_{s_1} \ldots dW_{s_n} \tag{49}$$

where, for each $i \in \mathbb{N}_+$, the deterministic functions f_i belongs to $L^2(\mathcal{S}_{i,T})$, with $\mathcal{S}_{i,T} \stackrel{\text{def.}}{=} \{(s_j)_{j=1}^i \in [0,T]^i : 0 < s_1 < \cdots < s_i < T\}$.

We consider an alternative description of the Wiener chaos expansion of any random variable $u \in L^2(\mathcal{F}_t)$ for a given $0 \leq t \leq T$, which both extends more easily to multiple dimensions and does not require lengthy computation of multiple iterated stochastic

integrals. For any $i \in \mathbb{N}$, the Hermite polynomials $(h_i)_{i \in \mathbb{N}}$ are the eigenfunctions of the generator $\frac{d^2}{dx^2} - x\frac{d}{dx}$ of the Ornstein-Uhlenbeck process $dX_t = -X_t + \sqrt{2}dW_t$. For each $i \in \mathbb{N}_+$, the ith Hermite polynomial h_i is given recursively by Rodrigues' formula as

$$h_i(x) = \frac{(-1)^i}{i!} e^{x^2/2} \frac{d^i}{dx^i} e^{-x^2/2} \qquad h_0(x) = 1.$$

The multi-dimensional version of the ith iterated integral in (49) is given by

$$\sum_{|\alpha|=i} \beta_{i,\alpha_j} \prod_{j=1}^{J_i} h_{\alpha_j} \left(\int_0^T \psi_{i,k}(s)\, dB_s \right) \tag{50}$$

where $\alpha \stackrel{\text{def.}}{=} (\alpha_1, \dots, \alpha_{J_i})$ is a multi-index consisting of positive integers with $i = |\alpha| \stackrel{\text{def.}}{=} \sum_{j=1}^{J_i} \alpha_i$, $\beta_{i,\alpha_j} \in \mathbb{R}$, and where $(\psi_{i,k})_{i,k \in \mathbb{Z}; 0 \le k, \frac{k+1}{2^n} \le t}$ is an orthonormal basis of $L^2([0,t])$. Here, we will consider the *Haar (wavelet) system* given by

$$\psi_{i,k}(s) \stackrel{\text{def.}}{=} 2^i \left(I_{[t\frac{k}{2^i}, t\frac{1+2k}{2^{i+1}})}(s) - I_{[t\frac{1+2k}{2^{i+1}}, t\frac{k+1}{2^i})}(s) \right).$$

Since $L^2([0,t]) \subset L^2([0,T])$ we can replace the t in indicator functions with T and we can consider the *Haar (wavelet) system* on the larger space $L^2([0,T])$ where $(\psi_{i,k})_{i,k \in \mathbb{N}; 0 \le k, \frac{k+1}{2^i} \le 1}$ and

$$\psi_{i,k}(s) \stackrel{\text{def.}}{=} 2^i \left(I_{[T\frac{k}{2^i}, T\frac{1+2k}{2^{i+1}})}(s) - I_{[T\frac{1+2k}{2^{i+1}}, T\frac{k+1}{2^i})}(s) \right).$$

By elementary functional analytic considerations, for each $0 \le t \le T$, $L^2(\mathcal{F}_t)$ is closure of the span on the of orthogonal $L^2(\mathcal{F}_t)$ random variables

$$\prod_{\tilde{j}=1}^{j} h_{\alpha_{\tilde{j}}} \left(\int_0^t \psi_{i,k}(s)\, dB_s \right) \tag{51}$$

where $j \in \mathbb{N}$, and $i, k \in \mathbb{N}; 0 \le k, \frac{k+1}{2^i} \le \frac{t}{T}$, for some $I \in \mathbb{N}$, $\beta_0, \beta_{1,\alpha_1}, \dots, \beta_{I,\alpha_{J_I}} \in \mathbb{R}$, and $(f_i)_{i \in \mathbb{N}}$ is an orthonormal basis of $L^2([0,t])$. Since $(\psi_{i,k})_{i,k \in \mathbb{N}; 0 \le k, \frac{k+1}{2^i} \le \frac{t}{T}}$ is piecewise constant, then the stochastic integrals in (51) simplify to

$$\int_0^t \psi_{i,k}(s)\, dB_s = 2^i W_{\frac{tk}{2^i}} - 2^{i+1} W_{\frac{t(1+2k)}{2^{i+1}}} + 2^i W_{\frac{t(k+1)}{2^i}}$$

where $T\frac{k+1}{2^i} \le t$. Thus, the random variables $\{u_{i,j,k} : j \in \mathbb{N}, i, k \in \mathbb{N}, \frac{k+1}{2^i} \le \frac{t}{T}\} \subset L^2(\mathcal{F}_t)$ where

$$u_{i,j,k}^{(t)} \stackrel{\text{def.}}{=} \prod_{\tilde{j}=1}^{j} h_{\alpha_{\tilde{j}}} \left(2^i W_{\frac{tk}{2^i}} - 2^{i+1} W_{\frac{t(1+2k)}{2^{i+1}}} + 2^i W_{\frac{t(k+1)}{2^i}} \right) \tag{52}$$

form an *orthonormal* basis of $L^2(\mathcal{F}_t)$. Conveniently, each of the $u_{i,j,k}$ can be computed without any explicit stochastic integration. We refer to Nualart (2006) for more details on Wiener Chaos.

Next, we will derive our universal approximation guarantees for our transformer model.

A.4 Proof of Universal Approximation Theorem 8

Our main universal approximation theorem relies on the following orthonormal basis of simple processes in $\mathcal{H}_T^2$, defined by linear combinations of the elementary processes in (11).

Lemma 19 (Orthonormal Basis of $\mathcal{H}_T^2$) *The collection of simple processes*
$\mathcal{S} \overset{\text{def.}}{=} \{u_{i,j,k}^{s_1,s_2} : i,j,k,s_1,s_2 \in \mathbb{N}, \, s_2 + 1 \le 2^{s_1}, \frac{k+1}{2^i} \le \frac{s_2}{2^{s_1}}\}$ *where*

$$u_{i,j,k}^{s_1,s_2}(t,\omega) \overset{\text{def.}}{=} \psi_{s_1,s_2}(t) \cdot u_{i,j,k}^T(\omega)$$

is an orthonormal basis of $\mathcal{H}_T^2$.

Proof We denote the set of indices

$$\mathcal{I} \overset{\text{def.}}{=} \{(i,j,k,s_1,s_2) \in \mathbb{N}^5 : \frac{k+1}{2^i} \le \frac{s_2}{2^{s_1}} \le 1, \frac{s_2+1}{2^{s_1}} \le 1\},$$

$$\mathcal{I}(s_1,s_2) \overset{\text{def.}}{=} \{(i,j,k) \in \mathbb{N}^3 : (i,j,k,s_1,s_2) \in \mathcal{I}\}, \quad (s_1,s_2) \in \mathbb{N}^2,$$

and observe the following equality

$$\mathcal{S} = \left\{u_{i,j,k}^{s_1,s_2}\right\}_{(i,j,k,s_1,s_2)\in\mathcal{I}} \subset \mathcal{H}^2([0,T]).$$

It is easy to check that $\mathcal{S}$ is orthonormal in $\mathcal{H}^2([0,T])$ as the set $\{\psi_{s_1,s_2}\}_{(s_1,s_2)\in\mathbb{N}^2}$ is orthornormal in $L^2([0,T])$, and for all $(s_1,s_2) \in \mathbb{N}^2$, $\left(u_{i,j,k}^T\right)_{(i,j,k)\in\mathcal{I}(s_1,s_2)}$ is orthonormal in $L^2(\mathcal{F}_T)$. It remains to show that $\mathcal{H}^2([0,T])$ is the closure of the span of $\mathcal{S}$.

We denote $\hat{\mathcal{H}}_0$ to the set of simple processes $Z \in \mathcal{H}^2([0,T])$ satisfying

$$Z \overset{\text{def.}}{=} \sum_{l=1}^{m} \xi_l I_{[t_{2l-1},t_{2l})}, \tag{53}$$

where $m \in \mathbb{N}$, $(t_i)_{i=1}^{m+1}$ is a strictly increasing sequence of real numbers satisfying $t_1 = 0, t_{2m} < T$, and $\xi_l \in L^2(\mathcal{F}_{t_{2l-1}})$, for $1 \le l \le m$.

We notice that $\bar{\hat{\mathcal{H}}}_0 = \mathcal{H}^2([0,T])$. Indeed, we observe that every simple process $Z := \sum_{l=1}^{m} \xi_l I_{[t_l,t_{l+1})}$ with $t_1 = 0, t_{m+1} = T$ can be written as the $\mathcal{H}^2([0,T])$-limit of the sequence $(Z^n)_{n\in\mathbb{N}} \subset \hat{\mathcal{H}}_0$:

$$Z_t^n = \sum_{l=1}^{m} \xi_l I_{[t_l,t_{l+1}-\frac{1}{n})}. \tag{54}$$

The previous shows that $\bar{\hat{\mathcal{H}}}_0 = \bar{\mathcal{H}}_0 = \mathcal{H}^2([0,T])$.
Next, assume that $Z \in \hat{\mathcal{H}}_0$ is such that: for all $U \in \mathcal{S}$

$$\int_0^T \mathbb{E}\left[U_t Z_t\right] dt = 0. \tag{55}$$

We notice that for any $\tilde{s}_1 \in \mathbb{N}$ sufficiently large there exists $\tilde{s}_2 \in \mathbb{N}$ satisfying

$$1 + \tilde{s}_2 \le 2^{\tilde{s}_1},$$

$$t_1 < \frac{\tilde{s}_2 T}{2^{\tilde{s}_1}} \le t_2 < \frac{(1 + 2\tilde{s}_2)T}{2^{\tilde{s}_1 + 1}} < \frac{(1 + \tilde{s}_2)T}{2^{\tilde{s}_1}} < t_3. \tag{56}$$

Then, for any $(i, j, k) \in \mathcal{I}(\tilde{s}_1, \tilde{s}_2)$, we set

$$U \overset{\text{def.}}{=} u_{i,j,k}^{\tilde{s}_1, \tilde{s}_2} \in \mathcal{S}. \tag{57}$$

Plugging in the process (57) into (55), we obtain

$$\int_0^T \mathbb{E}\left[U_t Z_t\right] dt = 2^{\tilde{s}_1} \left(t_2 - \frac{\tilde{s}_2 T}{2^{\tilde{s}_1}}\right) \mathbb{E}\left[u_{i,j,k}^T \xi_1\right] = 0, \quad \forall (i, j, k) \in \mathcal{I}(\tilde{s}_1, \tilde{s}_2). \tag{58}$$

Finally, we will prove that the closure of the span of $\left(u_{i,j,k}^T\right)_{(i,j,k) \in \mathcal{I}(\tilde{s}_1, \tilde{s}_2)}$ contains $L^2\left(\mathcal{F}_{t_1}\right)$. Using (56), we observe that for all $(i, k) \in \mathbb{N}^2$, $1 + k \le 2^i$ the following inequality holds:

$$0 \le \frac{t_1(k + 1)}{T 2^i} < \frac{\tilde{s}_2}{2^{\tilde{s}_1}} \le 1.$$

Using that the set of the dyadic rationals is dense in $[0, 1]$, there exists a sequence $(i_n, k_n) \in \mathbb{N}^2$, $k_n + 1 \le 2^{i_n}$, satisfying:

$$\lim_{n \to \infty} \frac{(k_n + 1)}{2^{i_n}} = \frac{t_1(k + 1)}{T 2^i},$$
$$\frac{t_1(k + 1)}{T 2^i} \le \frac{(k_n + 1)}{2^{i_n}} \le \frac{\tilde{s}_2}{2^{\tilde{s}_1}}, \quad n \in \mathbb{N}.$$

Hence, for all $j \in \mathbb{N}$, $\left(u_{i_n,j,k_n}^T\right)_{n \in \mathbb{N}} \subset \mathcal{S}$. By continuity of the paths of the Brownian motion and the Hermite polynomials, we obtain that $(u_{i_n,j,k_n}^T)_{n \in \mathbb{N}}$ converges $\mathbb{P}$-a.s. to $u_{i,j,k}^{(t_1)}$. Moreover, we observe that $\left(\xi_1 u_{i_n,j,k_n}^T\right)_{n \in \mathbb{N}}$ is uniformly integrable. Indeed, for $0 < \epsilon_0$ small enough, applying the Hölder's inequality, there exist constants $p_1 := \frac{3}{2(1 + \epsilon_0)}$, and $p_2 := \frac{1}{1 - \frac{1}{p_1}}$ such that

$$\sup_{n \in \mathbb{N}} \mathbb{E}\left[\left|\xi_1 u_{i_n,j,k_n}^T\right|^{1 + \epsilon_0}\right]$$

$$\le \mathbb{E}\left[|\xi_1|^{(1 + \epsilon_0)p_1}\right] \sup_{n \in \mathbb{N}} \mathbb{E}\left[\prod_{\tilde{j}=1}^j \left|h_{\tilde{j}}\left(2^{i_n} W_{\frac{T k_n}{2^{i_n}}} - 2^{i_n + 1} W_{\frac{T(1 + 2k_n)}{2^{i_n + 1}}} + 2^{i_n} W_{\frac{T(k_n + 1)}{2^{i_n}}}\right)\right|^{p_2}\right]$$

$$< \infty.$$

We therefore deduce that

$$\mathbb{E}\left[\xi_1 u_{i,j,k}^{(t_1)}\right] = \lim_{n \to \infty} \mathbb{E}\left[\xi_1 u_{i_n,j,k_n}^T\right] = 0. \tag{59}$$

Using completeness of the basis $\{u_{i,j,k}^{(t_1)}\}_{(i,j,k) \in \mathbb{N}^3}$ in $L^2(\mathcal{F}_{t_1})$ we obtain that

$$\xi_1 = 0, \quad \mathbb{P}\text{-a.s.}$$

Finally, we repeat the same argument in (56) for every addend in (53), obtaining

$$Z = 0, \quad dt \otimes \mathbb{P} - a.e.$$

Hence,

$$\overline{\text{span}(\mathcal{S})}^{\perp} \cap \hat{\mathcal{H}}_0 = \{0\}, \tag{60}$$

where $\mathcal{H}_0$ denotes the set of simple processes in $\mathcal{H}^2([0,T])$. Finally, using the decomposition

$$\mathcal{H}^2([0,T]) = \overline{\text{span}(\mathcal{S})} \oplus \overline{\text{span}(\mathcal{S})}^{\perp},$$

and (60), we have that $\hat{\mathcal{H}}_0 \subset \overline{\text{span}(\mathcal{S})}$. The later implies that $\overline{\text{span}(\mathcal{S})} = \mathcal{H}^2([0,T])$. ∎

We now prove our main general abstract approximation result (Proposition 12).

Proof of Proposition 12 Fix respective "dimension reduction" and "approximation" errors $\varepsilon_D, \bar{\varepsilon}_A > 0$.

Step 1 - Finite Dimensional Encoding

Enumerate $\mathcal{S} = \{s_i\}_{i=1}^{\infty}$, where $\mathcal{S}$ is as in Lemma 19. For any $d \in \mathbb{N}$ (which we fix retroactively) define the 1-Lipschitz encoder $\mathcal{E}_d : \mathcal{H}_T^2 \to \mathbb{R}^d$ given, for each $u \in \mathcal{H}_T^2$, by

$$\mathcal{E}_d(u.) \overset{\text{def.}}{=} (\langle u, s_j \rangle_{\mathcal{H}_T^2})_{j=1}^d.$$

Consider its right-inverse $\iota_d : \mathbb{R}^d \to \mathcal{H}_T^2$ given, for each $x \in \mathbb{R}^d$, by

$$\iota_d(x) \overset{\text{def.}}{=} \sum_{i=1}^{d} x_i \, s_i.$$

Observe that ι_d is an isometric embedding; we will come back to this point shortly.

- **$\mathcal{K}_0$ is the exponentially ellipsoidal $\mathcal{K}$ in Definition 10:** By the exponential decay condition in Definition 10, we have that: for each $u \in \mathcal{K}$, with representation $u = \sum_{i=1}^{\infty} \beta_i \, s_i$ the following error estimate holds by orthonormality of the $(s_i)_{i=1}^{\infty}$

$$\begin{aligned}
\left\| u - \iota_d \circ \mathcal{E}_d(u) \right\|_{\mathcal{H}_T^2}^2 &= \left\| \sum_{i=1}^{\infty} \beta_i \, s_i - \iota_d \circ \mathcal{E}_d(u) \right\|_{\mathcal{H}_T^2}^2 \\
&= \sum_{i=1}^{\infty} |\beta_i|^2 \, I_{i \leq d} \, \|s_i\|_{\mathcal{H}_T^2}^2 \\
&= \sum_{i=d+1}^{\infty} |\beta_i|^2 \\
&\leq \frac{C \, e^{-rd}}{1 - e^{-r}} \overset{\text{def.}}{=} \tilde{C}_K \, e^{-rd}
\end{aligned} \tag{61}$$

where $\tilde{C}_K \overset{\text{def.}}{=} C/(1-e^{-r})$. Fix $\varepsilon_0 > 0$. We now can retroactively set $d \overset{\text{def.}}{=} \ln\left(\frac{2^r \tilde{C}_K^r}{(1-e^{-r})^r} \varepsilon_0^{-r} \right) = \ln(C_K \varepsilon_0^{-r}) \in \mathcal{O}(\ln(\varepsilon_0^{-1/r}))$ where $C_K \overset{\text{def.}}{=} \left(\frac{C}{3L(1-e^{-r})} \right)^{1/r} > 0$.

37

- **Exponential Sub-manifold:** If $\mathcal{K}_0$ satisfies Definition 11, then this case is implied by the previous case (i.e. that of exponentially ellipsoidal compacta),

- **General $\mathcal{K}_0$:** If $\mathcal{K}_0$ is general, then by the 1-bounded approximation property (e.g. see Szarek, 1987) of Hilbert spaces with orthonormal bases (which are simply Banach spaces with Schauder bases), for every $\epsilon_0 > 0$ there is some $d \in \mathbb{N}$ for which $\sup_{x \in \mathcal{K}_0} |x - \mathcal{E}_d(x)| \le \epsilon_0$.

In each case, we have that

$$\sup_{u \in \mathcal{K}_0} |u - \iota_d \circ \mathcal{E}_d(u)| \le \varepsilon_0. \tag{62}$$

Let $L \ge 0$ denote the optimal Lipschitz constant of F. We note that the map

$$f^{(1)} \stackrel{\text{def.}}{=} F \circ \iota_d : \mathbb{R}^d \to \mathcal{H}_T^2$$

is (L, α)-Hölder since F is (L, α)-Hölder and since is 1-Lipschitz. In particular, the Lipschitz constant of $f^{(1)}$ is independent of d.

Step 2 - Quantization of the Image and the Domain:

Since $\mathcal{K}_0$ is compact and since $\mathcal{E}_d$ is continuous, then $\mathcal{E}_d(\mathcal{K}_0)$ is compact and thus $\mathcal{E}_d(\mathcal{K}_0)$ is totally bounded. Therefore, for every $\varepsilon_1 > 0$ there exists a finite subset $\{x_n\}_{n=1}^{N_{\varepsilon_1}} \subseteq \mathcal{E}_d(\mathcal{K}_0)$, of minimal cardinality $N \stackrel{\text{def.}}{=} N_{\varepsilon_1}$ (a so-called minimal ε_1-net) such that:

$$\max_{x \in \mathcal{K}_0} \min_{n=1,\dots,N_{\varepsilon_1}} \|x - x_n\|_2 < \left(\frac{1}{L}\frac{\varepsilon_1}{3}\right)^{1/\alpha}. \tag{63}$$

Since, $f^{(1)}$ is an L-Lipschitz surjection of $\mathcal{K}_0$ onto $f(\mathcal{K}_0)$ then (63) implies that

$$\max_{x \in \mathcal{K}_0} \min_{n=1,\dots,N_{\varepsilon_1}} \|f^{(1)}(x) - f^{(1)}(x_n)\|_{\mathcal{H}_T^2} \le L \max_{x \in \mathcal{K}_0} \min_{n=1,\dots,N_{\varepsilon_1}} \|x - x_2\|_2^{\alpha}$$
$$< L\left(\left(\frac{\varepsilon_1}{3L}\right)^{1/\alpha}\right)^{\alpha} = \frac{\varepsilon_1}{3}. \tag{64}$$

By Lemma 19, $(s_i)_{i=1}^{\infty}$ is an orthonormal basis of the Hilbert space $\mathcal{H}_T^2$. Therefore, it realizes the 1-bounded approximation property. This means that since $F(\mathcal{K}_0)$ is compact then, for each $Q \in \mathbb{N}$

$$\max_{y \in F(\mathcal{K}_0)} \|y - P_Q(y)\|_{\mathcal{H}_T^2} \stackrel{N \mapsto \infty}{\to} 0 \text{ and } \|P_Q\|_{op} \le 1 \tag{65}$$

where $P_Q : \mathcal{H}_T^2 \to \mathcal{H}_T^2$ is the (rank Q) orthogonal projection operator of $\mathcal{H}_T^2$ onto $\text{span}(\{s_i\}_{i=1}^Q)$; that is, $P_Q(u) \mapsto \sum_{i=1}^Q \langle s_i, u \rangle_{\mathcal{H}_T^2}$; where $\|\cdot\|_{op}$ denotes the operator norm. Thus, for each $\varepsilon_2 > 0$ (to be fixed retroactively) there exists some $Q \stackrel{\text{def.}}{=} Q_{\varepsilon_2} \in \mathbb{N}$ for which

$$\max_{y \in F(\mathcal{K}_0)} \|y - P_Q(y)\|_{\mathcal{H}_T^2} \le \varepsilon_2.$$

In the special case where $F(\mathcal{K}_0)$ satisfies Definition (10) then, by a similar computation to Step 1 (61), we obtain the following bounds on $Q \stackrel{\text{def.}}{=} Q_{\varepsilon_2}$:

- **Exponentially Ellipsoidal $\mathcal{K}_0$:** $Q \in \mathcal{O}(\ln(\varepsilon_D^{-1/r}))$,

- **Exponential Sub-manifold:** As before, if $\mathcal{K}_0$ satisfies Definition 11, then this case is implied by the previous case,

- **General $\mathcal{K}_0$:** $Q \to \infty$ as $\varepsilon_2 \to 0$.

To summarize this step, the set $\{x_n\}_{n=1}^N$ discretized $\mathcal{E}_d(\mathcal{K}_0)$ and the set $\{y_n\}_{n=1}^N$ discretized the "finitely parameterized" image of $\mathcal{K}_0$ under F.

Step 3 - Simplicialization of Target Function:

Our next objective is to replace the target function F, with a function which maps between $\mathcal{E}_d(\mathcal{K}_0)$ to an N-simplex and which, informally speaking, is an approximate continuous selection to the nearest neighbour problem

$$x \mapsto \operatorname*{argmin}_{n=1,\ldots,N} |F(x) - y_n|.$$

First, we construct an "projection-like/extremal" version of this solution, as in Step 4 of the proof of Acciaio et al. (2023, Theorem 3.8). In the second step, we "mollify" that function to make it comparable with the softmax operation.

Fix $\varepsilon_3 > 0$. Let $\mathcal{P}_1(\{x_n\}_{n=1}^N, \mathcal{W}_1)$ denote the 1-Wasserstein space over $\{x_n\}_{n=1}^N$ with respect to the α-snowflaked Euclidean distance $\|\cdot\|_2^\alpha$ on the inherited finite set $\{x_n\}_{n=1}^N$; i.e. the metric $\mathbb{R}^d \times \mathbb{R}^d \ni (x, \tilde{x}) \mapsto \|x - \tilde{x}\|_2^\alpha$ restricted to the set $\{x_n\}_{n=1}^N$. By Bruè et al. (2021, Theorem 3.2), there exists a Lipschitz map (a so-called weak random projection) $\Pi : (\mathcal{E}_d(\mathcal{K}_0), \|\cdot\|_2) \to \mathcal{P}_1(\{x_n\}_{n=1}^N, \mathcal{W}_1)$ with the property that: for each $x \in \mathcal{E}_d(\mathcal{K}_0)$, if $x \in \{x_n\}_{n=1}^N$, then $\Pi(x) = \delta_x$; where δ_x is the pointmass on x. Furthermore, the Lipschitz constant L_Π of Π is at-most $c \log_2(C_{(\{x_n\}_{n=1}^N, \|\cdot\|_2^\alpha)})$ where $c > 0$ is an absolute constant and $C_{(\{x_n\}_{n=1}^N, \|\cdot\|_2^\alpha)} > 0$ is the doubling constant of the set $\{x_n\}_{n=1}^N$ with respect to the α-snowflake of the Euclidean distance restricted to $\{x_n\}_{n=1}^N$. By Robinson (2011, Lemma 9.3), since the inclusion of $\{x_n\}_{n=1}^N$ into $\mathbb{R}^d$ is an isometric embedding (with respect to the Euclidean distance on $\mathbb{R}^d$) then the doubling constant of $(\{x_n\}_{n=1}^N, \|\cdot\|_2)$ is no larger than that of $(\mathbb{R}^d, \|\cdot\|_2)$. By Robinson (2011, Lemma 9.3), the doubling constant of $\mathbb{R}^d$ in the Euclidean metric is 2^{d+1}. Thus, the first statement in Acciaio et al. (2023, Lemma 7.1) implies that doubling of $(\{x_n\}_{n=1}^N, \|\cdot\|_2^\alpha)$ is at-most equal to the doubling constant of $\mathbb{R}^d$ to the power of $\lceil \frac{1}{\alpha} \rceil$. Hence, the Lipschitz constant L_Π of Π is

$$L_\Pi \leq c \log_2(C_{(\{x_n\}_{n=1}^N, \|\cdot\|_2^\alpha)}) \leq c \left\lceil \frac{1}{\alpha} \right\rceil \log_2(C_{(\{x_n\}_{n=1}^N, \|\cdot\|_2)}) \leq c \left\lceil \frac{1}{\alpha} \right\rceil (d+1) \leq \tilde{c} \left\lceil \frac{1}{\alpha} \right\rceil d \overset{\text{def.}}{=} C_\Pi \tag{66}$$

where $\tilde{c} \overset{\text{def.}}{=} 2 \max\{1, c\} > 0$.

As shown in Acciaio et al. (2023, Equations 41-46), the map $\iota_N : \mathcal{P}(\{x_n\}_{n=1}^N, \mathcal{W}_1) \ni \mu = \sum_{n=1}^N w_n \delta_{x_n} \to (w_n)_{n=1}^N \in (\Delta_N, \|\cdot\|_2)$ is $\frac{2}{\varepsilon_1}$-Lipschitz since the minimal distance between any distinct pairs of points in $\{x_n\}_{n=1}^N$ is ε_1. Together with the right-hand side of (66), this shows that the α-Hölder constant $L_{f^{(2)}}$ of composite map $f^{(2)} : \iota_N \circ \Pi_N : \mathcal{E}_d(\mathcal{K}_0) \to \Delta_N$ is bounded-above by

$$L_{f^{(2)}} \leq L_\Pi \frac{2}{\varepsilon_1} \leq \tilde{c} \left\lceil \frac{1}{\alpha} \right\rceil N \frac{2}{\varepsilon_1} \overset{\text{def.}}{=} \tilde{L}_{f^{(2)}}. \tag{67}$$

39

For $n = 1, \ldots, N$, let $y_n \overset{\text{def.}}{=} P_{Q_{\varepsilon_3}}(f^{(1)}(x_n)) = \sum_{i=1}^{Q_{\varepsilon_3}} \langle f^{(1)}(u), s_i \rangle_{\mathcal{H}_T^2} s_i$. Define the Lipschitz map $\eta_N : \Delta_N \to \mathcal{H}_T^2$ by

$$\eta_N : w \mapsto \sum_{i=1}^{N} w_i\, y_n. \tag{68}$$

Note that $\eta_N : (\Delta_N, \| \cdot \|_2) \to (\mathcal{H}_T^2, \| \cdot \|_{\mathcal{H}_T^2})$ is $L_{\varepsilon_1,\varepsilon_2}^\eta$-Lipschitz with an optimal Lipschitz constant, which we denote by $\mathrm{Lip}(\eta_{\epsilon_N}) \geq 0$, bounded-above by the constant $L_{\varepsilon_1,\varepsilon_2}^\eta > 0$ defined by

$$\left| \eta_N(w) - \eta_N(v) \right| \leq \sum_{i=1}^{N} \|y_i\|_{\mathcal{H}_T^2} |w_i - v_i| \tag{69}$$

$$\leq (\mathrm{diam}(f(\mathcal{E}_d(\mathcal{K}_0))) + 2\epsilon_2) \sum_{i=1}^{N} |w_i - v_i| \leq (L \,\mathrm{diam}(\mathcal{K}_0)^\alpha + 2\epsilon_2) \sum_{i=1}^{N} |w_i - v_i|$$

$$\leq (L \,\mathrm{diam}(\mathcal{K}_0)^\alpha + 2\epsilon_2)\sqrt{N}\, \|w - v\|_2 \overset{\text{def.}}{=} L_{\varepsilon_1,\varepsilon_2}^\eta \|w - v\|_2 \tag{70}$$

where $w, v \in \Delta_N$ are arbitrary and we have used the inequality $\| \cdot \|_1 \leq \sqrt{N} \| \cdot \|_2$ (on $\mathbb{R}^N$).

Next, we show that $\eta_N \circ f^{(2)}$ *approximates* F *on* $\mathcal{E}_d(\mathcal{K}_0)$.

Since $\mathcal{H}_T^2$ is a QAS space–see Acciaio et al. (2023, Definition 3.4) with $p = 1$ and $C_\eta = 1$, as shown in Acciaio et al. (2023, Example 5.1)–Step 4 of the proof of Acciaio et al. (2023, Theorem 3.8) holds unaltered in our setting (with $\mathcal{X} = \mathcal{E}_d(\mathcal{K}_0)$, $(\mathcal{Y}, d_\mathcal{Y} = (\mathcal{H}_T^2, \| \cdot \|_{\mathcal{H}_T^2})$, the α-Hölder target function with respect to the α-Hölder seminorm $L > 0$ of $f^{(1)}$).

Set[5] $\varepsilon_1 \overset{\text{def.}}{=} \frac{\varepsilon_D}{3 \cdot 3 L C_\Pi} = \frac{\varepsilon_D}{9 L \tilde{c} d} > 0$ and $\varepsilon_2 \overset{\text{def.}}{=} \frac{\varepsilon_D}{9} > 0$. Arguing identically to the (Acciaio et al., 2023, Proof of Theorem 3.8 - Step 3), following Acciaio et al. (2023, Equation 51), we conclude that

$$\sup_{x \in \mathcal{E}_d(\mathcal{K}_0)} \|f^{(1)}(x) - \eta_N \circ f^{(2)}(x)\|_{\mathcal{H}_T^2} \leq \frac{\varepsilon_D}{3}, \tag{71}$$

where, for us, our improved estimate on C_Π was given in (66); as is summarized in (71).

We now modify the map $f^{(2)}$ *so that it takes values in the range of the softmax function; that is, in the relative interior of the N-simplex, i.e., in the set* $\mathrm{int}(\Delta_N) \overset{\text{def.}}{=} \{w \in (0,1)^N : \sum_{n=1}^{N} w_n = 1\}$. *We subsequently associate* $f^{(2)}$ *to a map taking values* $\mathbb{R}^{N-1}$. *Finally, this latter map will be approximated by a neural network in step* 3.

Fix $0 < \varepsilon_3 \leq 1$, to be determined retroactively. Consider the 1-Lipschitz homotopy $H : [0,1] \times \Delta_N \to \Delta_N$ given by $H(t, w) \overset{\text{def.}}{=} t(w - \bar{\Delta}_N) + \bar{\Delta}_N$, where[6] $\bar{\Delta}_N = (1/N, \ldots, 1/N) \in \Delta_N$ is the barycenter of the N-simplex. Observe that, for each $t \in [0,1)$ we have $H(t, \Delta_N) \subset \mathrm{int}(\Delta_N)$ and for each $0 \leq \varepsilon_3 \leq \max_{w \in \Delta_N} \|w - \bar{\Delta}_N\|_2 1 - \frac{1}{N}$ there exists a $t_{\varepsilon_3} \in [0,1)$ satisfying[7]

$$\max_{w \in \Delta_N} \|H(t_{\varepsilon_3}, w) - w\|_2 \leq \max_{w \in \Delta_N} \|H(t_{\varepsilon_3}, w) - w\|_1 \leq \varepsilon_3. \tag{72}$$

---

5. In the notation of Acciaio et al. (2023, Proof of Theorem 3.8 - Step 3), we have set $\bar{\varepsilon}_A \overset{\text{def.}}{=} \varepsilon_Q \overset{\text{def.}}{=} \varepsilon_D/3$.

6. I.e. $\bar{\Delta}_N \overset{\text{def.}}{=} (1/N, \ldots, 1/N)$ is the barycenter of the N-simplex Δ_N.

7. For the interested reader: we have just noted that the boundary of Δ_N is a $\mathcal{Z}$-set, in the sense of van Mill (2001, Section 5.1).

The right-hand inequality can be solved explicitly for the largest value of t_{ε_3} in $[0, 1)$; this is because $w^\star \in \Delta_N$ given by $w_1^\star = 1$ and $w_j^\star = 0$ for $j = 2, \ldots, N$ is a non-unique maximizer of $\max_{w \in \Delta_N} \|H(t_{\varepsilon_3}, w) - w\|_1$. We therefore compute

$$\|H(t_{\varepsilon_3}, w) - w\|_1 = \underbrace{|(t_{\varepsilon_3}(1 - 1/N) + 1/N) - 1|}_{1^{rst} \text{ component}}$$

$$+ (N - 1)\underbrace{|(t_{\varepsilon_3}(0 - 1/N) + 1/N) - 0|}_{j > 1^{rst} \text{ components}} \tag{73}$$

$$= |1 - t_{\varepsilon_3}| |(1 - 1/N)| + (N - 1)|1 - t_{\varepsilon_3}| |1/N|$$

$$= |1 - t_{\varepsilon_3}| (N - 1)/N + |1 - t_{\varepsilon_3}| (N - 1)/N$$

$$= (1 - t_{\varepsilon_3}) \frac{2(N - 1)}{N}. \tag{74}$$

If $\varepsilon_3 > 0$ is small enough, [8] then setting the right-hand side of (74) equal to ε_3 and solving for t_{ε_3} yields $t_{\varepsilon_3} = 1 - \frac{N\varepsilon_3}{2(N-1)}$. For general values of ε_3, we may set

$$t_{\varepsilon_3} = 1 - \min\left\{\frac{N\varepsilon_3}{2(N - 1)}, 1/2\right\}. \tag{75}$$

Consider the map

$$\rho \overset{\text{def.}}{=} \text{softmax}_N \circ W : \mathbb{R}^{N-1} \to \text{int}(\Delta_N) \tag{76}$$

where $W : \mathbb{R}^{N-1} \ni x \to (x_1, \ldots, x_{N-1}, 1) \in \mathbb{R}^N$. A right-inverse of the smooth function $R : \text{int}(\Delta_N) \to \mathbb{R}^{N-1}$ given for each $y \in \text{int}(\Delta_N)$ by

$$R(y) \overset{\text{def.}}{=} \left((\ln(y_i) - \ln(y_N) + 1)\right)_{i=1}^{N}. \tag{77}$$

Finally, we define the "mollified simplicial target function" $f^{(3)} \overset{\text{def.}}{=} R \circ H(t_{\varepsilon_3}, \cdot) \circ f^{(2)} : \mathbb{R}^d \to \mathbb{R}^{N-1}$.
We therefore, have the following uniform estimate between $\rho \circ f^{(3)}$ and $f^{(2)}$

$$\max_{x \in \mathcal{E}_d(\mathcal{K}_0)} \|f^{(2)}(x) - \rho \circ f^{(3)}(x)\|_{\mathcal{H}_T^2} = \max_{x \in \mathcal{E}_d(\mathcal{K})} \|f^{(2)}(x) - \rho \circ \left(R \circ H(t_{\varepsilon_3}, \cdot) \circ f^{(2)}(x)\right)\|_{\mathcal{H}_T^2}$$

$$\tag{78}$$

$$= \max_{x \in \mathcal{E}_d(\mathcal{K})} \|f^{(2)}(x) - H(t_{\varepsilon_3}, \cdot) \circ f^{(2)}(x)\|_{\mathcal{H}_T^2}$$

$$= \max_{w \in \Delta_N} \|w - H(t_{\varepsilon_3}, w)\|_{\mathcal{H}_T^2} \tag{79}$$

$$\leq \varepsilon_3, \tag{80}$$

where we have use the fact that $f^{(2)}$ takes values in the N-simplex in deducing (79), and (80) held by (72). Combining our estimate in (71) with those in (78)-(80) yields

$$\sup_{x \in \mathcal{E}_d(\mathcal{K}_0)} \|f^{(1)}(x) - \eta_N \circ \rho \circ f^{(3)}(x)\|_{\mathcal{H}_T^2} \leq \sup_{x \in \mathcal{E}_d(\mathcal{K}_0)} \|f^{(1)}(x) - \eta_N \circ f^{(2)}(x)\|_{\mathcal{H}_T^2} \tag{81}$$

---

8. Namely, one needs that $0 < \varepsilon_3 < \frac{2(N-1)}{N}$.

$$+ \sup_{x \in \mathcal{E}_d(\mathcal{K}_0)} \|\eta_N \circ f^{(2)}(x) - \eta_N \circ \rho \circ f^{(3)}(x)\|_{\mathcal{H}_T^2}$$

$$\leq \sup_{x \in \mathcal{E}_d(\mathcal{K}_0)} \|f^{(1)}(x) - \eta_N \circ f^{(2)}(x)\|_{\mathcal{H}_T^2}$$

$$+ \operatorname{Lip}(\eta_N) \sup_{x \in \mathcal{E}_d(\mathcal{K}_0)} \|f^{(2)}(x) - \rho \circ f^{(3)}(x)\|_{\mathcal{H}_T^2}$$

$$\leq \frac{\varepsilon_D}{3} + \operatorname{Lip}(\eta_N)\, \varepsilon_3$$

$$\leq \frac{\varepsilon_D}{3} + \left((L \operatorname{diam}(\mathcal{K}_0)^\alpha + 2\varepsilon_2)\sqrt{N} \right) \varepsilon_3, \qquad (82)$$

where $\operatorname{Lip}(\eta_N)$ denotes the optimal Lipschitz constant of the map η_N which we have bounded above by $L_{\varepsilon_1, \varepsilon_2}^\eta$, in (70). Combining the estimates in (64) and in (62) with those in (81)-(82) yields

$$\sup_{u \in \mathcal{K}_0} \|F(u) - \eta_N \circ \rho \circ f^{(3)} \circ \mathcal{E}_d(u)\|_{\mathcal{H}_T^2} \leq \sup_{u \in \mathcal{K}_0} \|F(u) - F \circ \iota_d \circ \mathcal{E}_d(u)\|_{\mathcal{H}_T^2} \qquad (83)$$

$$+ \sup_{u \in \mathcal{K}_0} \|F \circ \iota_d \circ \mathcal{E}_d(u) - \eta_N \circ \rho \circ f^{(3)} \circ \mathcal{E}_d(u)\|_{\mathcal{H}_T^2}$$

$$\leq L \sup_{u \in \mathcal{K}_0} \|u - \iota_d \circ \mathcal{E}_d(u)\|_{\mathcal{H}_T^2}$$

$$+ \sup_{u \in \mathcal{K}_0} \|F \circ \iota_d \circ \mathcal{E}_d(u) - \eta_N \circ \rho \circ f^{(3)} \circ \mathcal{E}_d(u)\|_{\mathcal{H}_T^2}$$

$$= L \sup_{u \in \mathcal{K}_0} \|u - \iota_d \circ \mathcal{E}_d(u)\|_{\mathcal{H}_T^2}$$

$$+ \sup_{u \in \mathcal{E}_d(\mathcal{K}_0)} \|f^{(1)}(x) - \eta_N \circ \rho \circ f^{(3)}(x)\|_{\mathcal{H}_T^2}$$

$$\leq L \varepsilon_0 + \sup_{u \in \mathcal{E}_d(\mathcal{K}_0)} \|f^{(1)}(x) - \eta_N \circ \rho \circ f^{(3)}(x)\|_{\mathcal{H}_T^2}$$

$$\leq L \varepsilon_0 + \frac{\varepsilon_D}{3} + \left((L \operatorname{diam}(\mathcal{K}_0)^\alpha + 2\varepsilon_2)\sqrt{N} \right) \varepsilon_3. \qquad (84)$$

Retroactively set

$$\varepsilon_0 \overset{\text{def.}}{=} \frac{\varepsilon_D}{3L} \text{ and } \varepsilon_3 \overset{\text{def.}}{=} \frac{\varepsilon_D}{\min\left\{1, 3(L \operatorname{diam}(\mathcal{K}_0)^\alpha)\sqrt{N}\right\}}, \qquad (85)$$

implying that $\varepsilon_0 \in \mathcal{O}(\varepsilon_D)$ and $\varepsilon_3 \in \mathcal{O}(\varepsilon_D/\sqrt{N})$. Consequentially,

$$\sup_{u \in \mathcal{K}_0} \|F(u) - \eta_N \circ \rho \circ f^{(3)} \circ \mathcal{E}_d(u)\|_{\mathcal{H}_T^2} \leq \varepsilon_D. \qquad (86)$$

Next, we will obtain an upper-bound for the best α-Hölder constant of $f^{(3)}$ to derive quantitative parameter estimates on our MLP, which will approximate $f^{(3)}$.

Step 4 - Computing the Regularity of the Surrogate Target Function To apply a *quantitative* universal approximation theorem, we need a handle on the regularity of the target function being approximated. We first observe that, on the set $\Delta_N^{\epsilon_2} \overset{\text{def.}}{=} H(t_{\epsilon_2}, \Delta_N)$ the map R, defined in (77), is $L_{f^{(3)}\epsilon_2, N}$-Lipschitz with constant given by

$$L_{f^{(3)}\varepsilon_2, N} = \sup_{w \in \Delta_N^{\varepsilon_3}} \|\nabla R(w)\|_2 \leq \sup_{w \in \Delta_N^{\varepsilon_3}} \|\nabla R(w)\|_1 \qquad (87)$$

$$= \sup_{w \in \Delta_N} \sum_{i=1}^{N} \frac{1}{|t_{\varepsilon_3}(w_i - 1/N) + 1/N|} \tag{88}$$

$$\leq \sum_{i=1}^{N} \frac{1}{\min_{w \in \Delta_N} \lambda |t_{\varepsilon_3}(w_i - 1/N) + 1/N|}$$

$$= \frac{N^2}{(1 - t_{\varepsilon_3})}$$

$$= N \max \left\{ 2N, \frac{2(N-1)}{\varepsilon_3} \right\}$$

$$\leq \frac{2 N^2}{\min\{\varepsilon_3^{-1}, 1/2\}} \stackrel{\text{def.}}{=} \tilde{L}_{f^{(3)} \varepsilon_3, N} \tag{89}$$

where ∇R denotes the Jacobian of R and where the inequality (87) holds by the Rademacher-Stephanov theorem, see e.g. Federer (1969, Theorems 3.1.6-3.1.9).

We thus conclude that, from the estimates in (67), (87)-(89), and the observation that H is 1-Lipschitz, $f^{(3)}$ is $L_{f^{(3)}}$-Lipschitz; where

$$L_{f^{(3)}} \leq L_{f^{(2)}} L_{f^{(3)} \epsilon_2, N} = \frac{\bar{c} N^3}{\varepsilon_1 \min\{\min\{1, \epsilon_3\}, 1/2\}}$$

$$= \frac{9 \bar{c} d N^2}{\varepsilon_D \min\{\varepsilon_D/3, 1/2\}} \stackrel{\text{def.}}{=} \tilde{L}, \tag{90}$$

where $\bar{c} \stackrel{\text{def.}}{=} 4\tilde{c} > 0$. We now construct our deep-learning approximation.

Step 5 - Neural Approximation of Surrogate Target Function $\hat{f}^{(3)}$:
We consider two cases; in the former, the activation function is smooth and in the latter, it is the trainable super-expressive activation (defined in (8)).

1. **Case 1 - σ as in Example 2:** By Kratsios and Papon (2022, Proposition 53), there is an MLP $\hat{f} : \mathbb{R}^d \to \mathbb{R}^N$ with activation function $\sigma_0 \in C(\mathbb{R})$ (in the notation of Example 2) satisfying

$$\sup_{x \in \mathcal{E}_d(\mathcal{K}_0)} \|\hat{f}(x) - f^{(3)}(x)\|_2 < \bar{\varepsilon}_A \tag{91}$$

   with depth and width given by:

   - **Width:** $d + N + 2$,
   - **Depth:** Finite, and if σ is non-affine and smooth then: $\mathcal{O}\Big(N((1-d/4)N)^{2d/\alpha} (2C)^{2d} \varepsilon^{-2d/\alpha} \Big)$,

   where we use the fact that the diameter of $\mathcal{K}_0$ is at-most $2C$.

2. If σ is the trainable super-expressive activation function in (8), then: for $i = 1, \ldots, N$ (Gao et al., 2022, Theorem 1) there exists an MLP $\hat{f}_i : \mathbb{R}^d \to \mathbb{R}$ with activation function σ_0 satisfying

$$\max_{i=1,\ldots,d} \sup_{x \in \mathcal{E}_d(\mathcal{K}_0)} \|\langle f^{(3)}(x), e_i \rangle - \hat{f}_i(x)\|_2 < \bar{\varepsilon}_A/N \tag{92}$$

43

where $\{e_i\}_{i=1}^N$ is the standard orthonormal basis of $\mathbb{R}^N$. Moreover, by Gao et al. (2022, Theorem 1), and the remark directly after, the width, depth, and number of non-zero parameters determining each network is exactly

- **Width**: 11,
- **Depth:** $36\,(2d+1)$,
- **No. Params:** $5437\,(d+1)\,(2d+1)$.

Since $\sigma_1(x) = x$, for all $x \in \mathbb{R}$ then, the parallelization construction in Petersen and Zech (2024, Section 5.1.3), and its direct proof in Petersen and Zech (2024, Lemma 5.3), goes through (with the identity ReLU-blocks in Petersen and Zech (2024, Section 5.1.3) replaced identity blocks using the σ_1 activation). Thus, there exists an MLP $\hat{f} : \mathbb{R}^d \to \mathbb{R}^N$ with activation function σ satisfying: for each $x \in \mathbb{R}^d$

$$\hat{f}(x) = \sum_{i=1}^N f_i(x)\, e_i.$$

Furthermore, the with, depth, and number of non-zero determining $\hat{f}$ are

- **Width**: $22N \in \mathcal{O}(N)$,
- **Depth:** $36(2d+1) \in \mathcal{O}(d)$,
- **No. Parameters:** at-most $10874N\,(d+1)\,(2d+1) \in \mathcal{O}(Nd^2)$.

Consequentially, (92) implies that

$$\sup_{x \in \mathcal{E}_d(\mathcal{K}_0)} \|f^{(3)}(x) - \hat{f}(x)\|_2 \leq \sum_{i=1}^N \sup_{x \in \mathcal{E}_d(\mathcal{K}_0)} \|\langle f^{(3)}(x), e_i\rangle - \hat{f}_i(x)\|_2 < N\frac{\bar{\bar{\varepsilon}}_A}{N} = \bar{\varepsilon}_A \quad (93)$$

where e_i is the ith standard basis vector in $\mathbb{R}^N$ with 1 in the ith coordinate and 0 otherwise. We are now ready to complete the proof by combining the estimates from the previous steps.

Step 6 - Putting It All Together:
Set $\hat{F} \stackrel{\text{def.}}{=} \eta_{N_{\varepsilon_1}} \circ \rho \circ \hat{f} \circ \mathcal{E}_d : \mathcal{H}_T^2 \to \mathcal{H}_T^2$. The estimates in (86) with those in (91) (resp. (93)) yield

$$
\begin{aligned}
\sup_{u \in \mathcal{K}_0} \|F(u) - \hat{F}(u)\|_{\mathcal{H}_T^2} &\leq \sup_{u \in \mathcal{K}_0} \|F(u) - \eta_{N_{\varepsilon_1}} \circ \rho \circ f^{(3)} \circ \mathcal{E}_d(u)\|_{\mathcal{H}_T^2} \\
&\quad + \sup_{u \in \mathcal{K}_0} \|\eta_{N_{\varepsilon_1}} \circ \rho \circ f^{(3)} \circ \mathcal{E}_d(u) - \hat{F}(u)\|_{\mathcal{H}_T^2} \\
&\leq \varepsilon_D + \sup_{u \in \mathcal{K}_0} \|\eta_{N_{\varepsilon_1}} \circ \rho \circ f^{(3)} \circ \mathcal{E}_d(u) - \hat{F}(u)\|_{\mathcal{H}_T^2} \\
&= \varepsilon_D + \sup_{x \in \mathcal{E}_d(\mathcal{K}_0)} \|\eta_{N_{\varepsilon_1}} \circ \rho_1 \circ \hat{f}(x) - \eta_{N_{\varepsilon_1}} \circ \rho_1 \circ \hat{f}(x)\|_{\mathcal{H}_T^2} \\
&= \varepsilon_D + \mathrm{Lip}(\eta_{N_{\varepsilon_1}} \circ \rho) \sup_{x \in \mathcal{E}_d(\mathcal{K}_0)} \|f^{(3)}(x) - \hat{f}(x)\|_2 \\
&= \varepsilon_D + \mathrm{Lip}(\eta_{N_{\varepsilon_1}} \circ \rho)\, \bar{\varepsilon}_A. \quad (94)
\end{aligned}
$$

44

Since W is an isometric embedding and the softmax function is at-most 1-Lipschitz, then ρ is at-most 1-Lipschitz and $\mathrm{Lip}(\eta_{N_{\varepsilon_1}} \circ \rho) = \mathrm{Lip}(\eta_{N_{\varepsilon_1}})$, which by (70) is at-most

$$\mathrm{Lip}(\eta_{N_{\varepsilon_1}}) \leq (L \, \mathrm{diam}(\mathcal{K}_0)^\alpha + 2\epsilon_D/9)\sqrt{N}.$$

Since this upper-bound on $\mathrm{Lip}(\eta_{N_{\varepsilon_1}} \circ \rho))$ depends only on ε_D and is independent of $\bar{\varepsilon}_A$. Consequentially for the right-hand side of (94) can be made arbitrarily small by choosing ε_D and $\bar{\varepsilon}_A$ large enough.

It remains to bound N explicitly. There are three cases which we consider here, each of which corresponds to the respective assumptions made on $\mathcal{K}_0$ and its relationship to the target (non-linear) operator f:

1. **Exponentially Ellipsoidal:** Suppose that $\mathcal{K}_0 \subset \mathcal{H}_T^2$ is such that: for each $\mathcal{K}_0 \ni x = \sum_{i=1}^\infty \beta_i s_i$ we have that $|\beta_i| \leq C \, r^i$. Recall that, e.g. as noted on Dumer et al. (2004, in Remark 1), the ϵ_A-covering number of $p_d(\varepsilon_A^{-1} \cdot \mathcal{K}_0)$ (where $\delta \mathcal{K}_0$ denotes the δ-thickening of $\mathcal{K}_0$ in $\mathbb{R}^d$). Therefore, for each such $x \in \mathcal{K}_0$ we have that

$$\sum_{i=1}^d \frac{|\beta_i|^2}{\theta_i^2} \leq 1$$

where the scaling constants $(\theta_i)_{i=1}^\infty$ (independent of d) are given by

$$\theta_i = \left(r \left(1 + \left(\frac{C}{\varepsilon_A} \right)^2 \right)^{1/2} \right)^i. \tag{95}$$

Therefore, by Dumer et al. (2004, Theorem 2) (cf. the proof on Duchi et al. (2011, Equation 40) for the precise expression of $o(1)$ in the statement of Dumer et al. (2004, Theorem 2)) and (95) we find that

$$\begin{aligned} N &\leq \exp\Big(\sum_{i=1}^d i \log \big(Cr \, (C^{-2} + \varepsilon_A^{-2})^{1/2} \big) \Big) \\ &= \exp\Big(\frac{d(d+1)}{2} \log \big(Cr \, (C^{-2} + \varepsilon_A^{-2})^{1/2} \big) \Big) \\ &= \big(r \, (1 + (C/\varepsilon_A)^2)^{1/2} \big)^{\frac{d(d+1)}{2}}. \end{aligned} \tag{96}$$

Since, in this case, $d \in \mathcal{O}(\ln(\epsilon_D^{-1/r}))$ then there exists some $C_1 > 0$ such that (96) reduces to

$$N \leq \big(r \, (1 + (C/\varepsilon_A)^2)^{1/2} \big)^{C_1 \ln(\epsilon_D^{-1/r})^2}. \tag{97}$$

2. **Exponential Manifold:** Suppose that $\mathcal{K}_0$ satisfies Definition 11. For every $\tilde{\varepsilon}_A > 0$ (to be fixed momentarily), Lorentz et al. (1996, Proposition 15.1.3) implies that $\tilde{\varepsilon}_A$-covering number $\tilde{N}$ of the Euclidean unit ball $B_{d_{\mathrm{low}}}$ is bounded above and below by

$$2^{-d_{\mathrm{low}}} \big(\sqrt{d_{\mathrm{low}}}/\tilde{\varepsilon}_A \big)^{d_{\mathrm{low}}} \leq \tilde{N} \leq 3^{d_{\mathrm{low}}} \big(\sqrt{d_{\mathrm{low}}}/\tilde{\varepsilon}_A \big)^{d_{\mathrm{low}}}. \tag{98}$$

45

Since the "latent encoding" map $\pi : \mathbb{R}^{d_{\text{low}}} \to \mathcal{H}_T^2$ was assumed to be 1-Lipschitz and maps onto $\mathcal{K}_0$ then, the image of every $\tilde{\varepsilon}_A$ of $B_{d_{\text{low}}}$ under π must be $1 \cdot \tilde{\varepsilon}_A$ covering of $\mathcal{K}_0$. Set $\tilde{\varepsilon}_A = \varepsilon_A$. Then, (98) implies that

$$N \le \left(3\sqrt{d_{\text{low}}}/\varepsilon_A\right)^{d_{\text{low}}}. \tag{99}$$

Whence $N \in \mathcal{O}\left(\varepsilon_A^{-d_{\text{low}}} d_{\text{low}}^{d_{\text{low}}/2}\right) = \mathcal{O}\left(\varepsilon_A^{-d_{\text{low}}}\right)$.

3. **General Case:** In the case of general $\mathcal{K}_0$, Acciaio et al. (2023, Lemma 7.1) and the upper-bound of 2^{d+1} on the doubling constant of $\mathcal{E}_d(\mathcal{K}_0)$, just prior to Equation (67), imply that

$$N \le \left(2^{(d+1)}\right)^{\log_2(\text{diam}(\mathcal{K}_0)) - \frac{1}{\alpha}\log_2(\varepsilon_D/\tilde{L}) + \frac{1}{\alpha}\log_2(\tilde{c}d)}$$

for some absolute constant $\tilde{c} > 0$.

Setting $\varepsilon_A \stackrel{\text{def.}}{=} \bar{\varepsilon}_A/\left(c(L\,\text{diam}(\mathcal{K}_0)^\alpha + \varepsilon_D)\sqrt{N}\right) \in \mathcal{O}\left(\frac{\bar{\varepsilon}_A}{\epsilon_D\sqrt{N}}\right)$ yields the conclusion.

Step 7 - Elucidating the Model

Let $V \in \mathbb{R}^{N \times Q}$ be such that, for $n = 1, \ldots, N$ and $i = 1, \ldots, Q$, $V_{n,q} \stackrel{\text{def.}}{=} \langle F \circ \iota_d(x_n), s_i \rangle_{\mathcal{H}_T^2}$. Then, (68) implies that: for each $w \in \Delta_N$

$$\eta(w) = \sum_{n=1}^{N} \left(\sum_{i=1}^{Q} \langle F \circ \iota_d(x_n), s_i \rangle_{\mathcal{H}_T^2} \right) = \sum_{n=1}^{N} \sum_{i=1}^{Q} w_n\, V_{n,q}\,. \tag{100}$$

For either σ is smooth or σ as in (92), consider the MLP with σ activation function $\mathcal{V} : \mathbb{R}^d \to \mathbb{R}^{N \times Q}$ given for each $x \in \mathbb{R}^d$ by

$$\mathcal{V}(x) \stackrel{\text{def.}}{=} \mathbf{0}^{(1)}\sigma \bullet \left(\mathbf{0}^{(2)}x + \mathbf{0}^{(3)}\right) + V$$

where $\mathbf{0}^{(1)}$ is the $ND \times 1$ zero matrix, $\mathbf{0}^{(2)}$ is the $1 \times d$ zero matrix, and $\mathbf{0}^{(3)} = (0) \in \mathbb{R}$, and where we have identified $\mathbb{R}^{N \times D}$ with $\mathbb{R}^{ND}$. In either case, observe that the number of non-zero parameters defining $\mathcal{V}$ are at-most ND.

For each $n = 1, \ldots, N$ and $i = 1, \ldots, Q$, we let $V^{(n,i)} \stackrel{\text{def.}}{=} s_i$. By construction: for each $u \in \mathcal{H}_T^2$ and every $w \in \Delta_N$ we have that

$$\mathcal{D}(w, u) \stackrel{\text{def.}}{=} \sum_{n=1}^{N} w_n\, [\mathcal{V} \circ \mathcal{E}_d(u)]_n\, V^{(n,q)} = \eta_N(w). \tag{101}$$

Since the map W in the definition of ρ (see the line just below (76)) was affine, then our approximation $\hat{F} = \eta_{N_{\varepsilon_1}} \circ \rho \circ \hat{f} \circ \mathcal{E}_d : \mathcal{H}_T^2 \to \mathcal{H}_T^2$ is of the form in Definition 3.

The number of non-zero parameters defining the model $\hat{F}$ are, therefore, at-most

$$\underbrace{\text{Depth}(\hat{f})\,\text{Width}(\hat{f})^2}_{\text{No. Param. } \hat{f}} + \underbrace{NQ}_{\text{No. Param. } \mathcal{V}} \tag{102}$$

where the depth and width of $\hat{f}$ were computed in step 4. In particular, if σ is the trainable super-expressive activation function in (92) then the quantity in (102) is $\mathcal{O}\left(N(N^2d^4 + Q)\right)$.

■

Proof of Theorem 8 Fix a non-empty compact subset $\mathcal{K}_0 \subset \mathcal{H}_T^2$, a continuous function $f : \mathcal{K}_0 \to \mathcal{H}_T^2$, and a $\varepsilon > 0$.

Let $\mathcal{K}_0 \subset \mathcal{H}_T^2$ be non-empty and compact. We would like to use Miculescu (2002/03, Theorem 1) to reduce the problem of approximating f to ε precision, to the problem of approximating an $\varepsilon/2$ Lipschitz approximation of our target function f to $\varepsilon/2$ precision. Thus, we will be able to employ our technical approximation theorem for Lipschitz maps from $\mathcal{H}_T^2$ to itself, to deduce our conclusion. On a technical note, we do not argue on the domain $\mathcal{H}_T^2$ but rather on the compact subspace $\mathcal{K}_0$, since all continuous functions are both bounded and uniformly continuous thereon; which then directly allows us to (Miculescu, 2002/03, Theorem 1) which only allows us to uniformly approximate bounded continuous functions on compacta.

Step 1 - Verification of Lipschitz Extension Property

In order to apply Miculescu (2002/03, Theorem 1) we will need to show that the pair $(\mathcal{K}_0, \| \cdot \|_{\mathcal{H}_T^2})$ and $\mathcal{H}_T^2$ have the so-called "Lipschitz extension property" (as named in Miculescu, 2002/03, Theorem 1). This means that the Lipschitz function from $(B, \| \cdot \|_{\mathcal{H}_T^2})$ to $\mathcal{H}_T^2$, for any subset B of $\mathcal{K}_0$, can be extended to a Lipschitz function of all of $\mathcal{K}_0$ with roughly the same Lipschitz constant.

Let $B \subseteq \mathcal{K}_0$. Note that, as $\mathcal{K}_0 \subset \mathcal{H}_T^2$ then, B is a subset of $\mathcal{H}_T^2$. Since $\mathcal{H}_T^2$ is a separable Hilbert space then the extension theorem of Benyamini and Lindenstrauss (2000, Theorem 1.12): for every $L \geq 0$ and each L-Lipschitz (non-linear operator) $g : (B, \| \cdot \|_{\mathcal{H}_T^2}) \to \mathcal{H}_T^2$ there exists an L-Lipschitz extension $\tilde{G} : \mathcal{H}_T^2 \to \mathcal{H}_T^2$; i.e. $\tilde{G}$ is L-Lipschitz such that

$$\tilde{G}|_B = g. \tag{103}$$

Since the restriction operator $\iota_{\mathcal{K}_0} : \mathcal{H}_T^2 \ni \tilde{g} \to \tilde{g}|_K \in \mathcal{K}_0$ is 1-Lipschitz then the composite map $G \stackrel{\text{def.}}{=} \iota_{\mathcal{K}_0} \circ \tilde{G} = \tilde{G}|_{\mathcal{K}_0} : (\mathcal{K}_0, \| \cdot \|_{\mathcal{H}_T^2}) \to \mathcal{H}_T^2$ is L-Lipschitz. Furthermore, (103) and the inclusion of B in $\mathcal{K}_0$ imply that

$$G|_B = (\tilde{G}|_B)|_{\mathcal{K}_0} = g. \tag{104}$$

Thus, G is an L-Lipschitz extension of g to $(\mathcal{K}_0, \| \cdot \|_{\mathcal{H}_T^2})$. Thus, the pair $(\mathcal{K}_0, \| \cdot \|_{\mathcal{H}_T^2})$ and $\mathcal{H}_T^2$ has the Lipschitz extension property; thus Miculescu (2002/03, Theorem 1) implies that the space of Lipschitz functions from $(\mathcal{K}_0, \| \cdot \|_{\mathcal{H}_T^2})$ to $\mathcal{H}_T^2$ is *dense* in the space of uniformly continuous and bounded functions from $(\mathcal{K}_0, \| \cdot \|_{\mathcal{H}_T^2})$ to $\mathcal{H}_T^2$ with respect to the uniform norm.

Step 2 - $\varepsilon/2$-Approximation of f by Lipschitz Maps

Since $\mathcal{K}_0$ is compact and f is continuous on $\mathcal{K}_0$ then f is uniformly continuous and bounded thereon. By Miculescu (2002/03, Theorem 1), we deduce that there exists a Lipschitz function $\tilde{f}_\varepsilon : (\mathcal{K}_0, \| \cdot \|_{\mathcal{H}_T^2}) \to \mathcal{H}_T^2$ satisfying

$$\max_{u \in \mathcal{K}_0} \|f(u) - \tilde{f}_\varepsilon(u)\|_{\mathcal{H}_T^2} \leq \varepsilon/2. \tag{105}$$

Again applying Benyamini and Lindenstrauss (2000, Theorem 1.12), we deduce that $\tilde{f}_\varepsilon$ admits a Lipschitz extension $f_\varepsilon : \mathcal{H}_T^2 \to \mathcal{H}_T^2$, with the same Lipschitz constant. Since f_ε is

a Lipschitz extension of $\tilde{f}_\varepsilon$, beyond $\mathcal{K}_0$, then (105) implies that

$$\max_{u \in \mathcal{K}_0} \|f(u) - f_\varepsilon(u)\|_{\mathcal{H}_T^2} = \max_{u \in \mathcal{K}_0} \|f(u) - \tilde{f}_\varepsilon(u)\|_{\mathcal{H}_T^2} \le \varepsilon/2. \tag{106}$$

Step 3 - $\varepsilon/2$-Approximation of f_ε by Attentional Neural Operator
Since $\mathcal{K}_0$ is a compact subset of $\mathcal{H}_T^2$ and $f_\varepsilon : \mathcal{H}_T^2 \to \mathcal{H}_T^2$ is Lipschitz then Proposition 12
applies. Whence, there exists an attentional neural operator $\hat{F} : \mathcal{H}_T^2 \to \mathcal{H}_T^2$ satisfying

$$\max_{u \in \mathcal{K}_0} \|f_\varepsilon(u) - \hat{F}(u)\|_{\mathcal{H}_T^2} \le \epsilon/2. \tag{107}$$

Combining (106) and (107) yields

$$\max_{u \in \mathcal{K}_0} \|f(u) - \hat{F}(u)\|_{\mathcal{H}_T^2} \le \max_{u \in \mathcal{K}_0} \|f(u) - f_\varepsilon(u)\|_{\mathcal{H}_T^2} + \max_{u \in \mathcal{K}_0} \|f_\varepsilon(u) - \hat{F}(u)\|_{\mathcal{H}_T^2} \le \varepsilon/2 + \varepsilon/2 = \varepsilon,$$

which concludes our proof. ∎

A.5 Proof of Main Stackelberg Equilibria Results

Theorems 4 and 5 are two parts of a larger whole. As such, their derivation is most naturally
merged into a single proof; we now do.
Proof (Joint Proof of Theorems 4, 5, and 7) Let $\hat{U} : \mathcal{H}_T^2 \to \mathcal{H}_T^2$ be a map, to be fixed
retroactively.
For any $d \in \mathbb{N}_+$, let $p_d : \mathcal{H}_T^2 \to \text{span}\{s_i\}_{i=1}^d$ be the orthogonal projection; i.e. $p_d\big(\sum_{i=1}^\infty \beta_i s_i\big) = \sum_{i=1}^d \beta_i s_i$ for all $u = \sum_{i=1}^\infty \beta_i s_i \in \mathcal{H}_T^2$. We will retroactively adjust d. Fix $u^0 \in \mathcal{K}_0$, denote
$\hat{u}_d^0 \stackrel{\text{def.}}{=} p_d(u^0)$ and compute

$$\left| J_0(u^0, U^\star(u^0)) - J_0\big(\hat{u}_d^0, \hat{U}(\hat{u}_d^0)\big) \right| \le \underbrace{\left| J_0(u^0, U^\star(u^0)) - J_0\big(\hat{u}_d^0, U^\star(\hat{u}_d^0)\big) \right|}_{(I)} \tag{108}$$
$$+ \underbrace{\left| J_0\big(\hat{u}_d^0, U^\star(\hat{u}_d^0)\big) - J_0\big(\hat{u}_d^0, \hat{U}(\hat{u}_d^0)\big) \right|}_{(II)}.$$

Step 1 - Bounding Term Term (I) By Lemma 9 , we can bound Term (I) from above
by

$$(I) = \left| J_0(u^0, U^\star(u^0)) - J_0\big(\hat{u}_d^0, U^\star(\hat{u}_d^0)\big) \right| \le \widetilde{\omega}\big(\|u^0 - \hat{u}_d^0\|_{\mathcal{H}_T^2}\big) \tag{109}$$

where $\widetilde{\omega}(t) = C \max\{|t|, |t|^{1/2}\}$ for each $t \in \mathbb{R}$ and for some constant $C \ge 0$ depending only
on T. Note that, $t \mapsto \widetilde{\omega}(t)$ is continuous, monotonically increasing on $[0, \infty)$ and subjective,
thus it is a homomorphism of $[0, \infty)$ to itself with continuous inverse given by

$$\bar{\omega}(t) \stackrel{\text{def.}}{=} \begin{cases} (t/C)^2 & \text{if } 0 \le t \le 1, \\ (t/C) & \text{if } 1 \le t. \end{cases}$$

We emphasize that, $\bar{\omega}$ has range $[0, \infty)$.

Since $\mathcal{H}_T^2$ has the 1-bounded approximation property implemented by the (finite-rank) projection operators $(p_d)_{d\in\mathbb{N}_+}$ then choosing d large enough, we may ensure that $\sup_{u^0\in\mathcal{K}_0}\|u^0 - \hat{u}_d^0\|_{\mathcal{H}_T^2} < \bar{\omega}(\varepsilon/2)$. The right-hand side of (109) can be bounded-above as follows

$$\text{(I)} \leq \widetilde{\omega}\big(\|u^0 - \hat{u}_d^0\|_{\mathcal{H}_T^2}\big) \leq \widetilde{\omega}\left(\sup_{u^0\in\mathcal{K}_0}\|u^0 - \hat{u}_d^0\|_{\mathcal{H}_T^2}\right) \leq \widetilde{\omega}\left(\bar{\omega}\left(\frac{\varepsilon}{2}\right)\right) = \frac{\varepsilon}{2}. \tag{110}$$

It remains to bound Term (II).

Step 2 - Bounding Term Term (II) By Lemma 16, we find that

$$\text{(II)} = \big|J_0(\hat{u}_d^0, U^\star(\hat{u}_d^0)) - J_0(\hat{u}_d^0, \hat{U}(\hat{u}_d^0))\big| \leq C \cdot \big\|U^\star(\hat{u}_d^0) - \hat{U}(\hat{u}_d^0)\big\|_{\mathcal{H}_T^2}. \tag{111}$$

By our universal approximation theorem, in Proposition 12, we have that there exists an attentional neural operator $\hat{U} : \mathcal{H}_T^2 \to \mathcal{H}_T^2$, as in Definition 3

$$\sup_{v\in p_d(\mathcal{K}_0)} \big\|U^\star(v) - \hat{U}(v)\big\|_{\mathcal{H}_T^2} \leq \frac{\varepsilon}{2C} \tag{112}$$

where we have set $\varepsilon_D \stackrel{\text{def.}}{=} \varepsilon_A \stackrel{\text{def.}}{=} \varepsilon/4C$. Thus, the estimate in (112) implies that the right-hand side of (111) can be bounded above as follows

$$\text{(II)} \leq C\big\|\hat{U}(\hat{u}_d^0) - U^\star(\hat{u}_d^0)\big\|_{\mathcal{H}_T^2} \leq C \sup_{v\in p_d(\mathcal{K}_0)} \big\|U^\star(v) - \hat{U}(v)\big\|_{\mathcal{H}_T^2} \leq C\frac{\varepsilon}{2C} = \frac{\varepsilon}{2}. \tag{113}$$

Upon combining the estimates in (110) and in (113) we obtain the following upper-bound for the right-hand side of (108)

$$\big|J_0(u^0, U^\star(u^0)) - J_0\big(\hat{u}_d^0, \hat{U}(\hat{u}_d^0)\big)\big| \leq \text{(I)} + \text{(II)} \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon. \tag{114}$$

Step 3 Suppose additionally that $u^0 \in \mathcal{K}_0$ is such that $(u^0, U^\star(u^0))$ is a $(0-)$Stackelberg equilibrium. If $(u^0, U^\star(u^0))$ is a Stackelberg equilibrium, see Definition 1, then this pair is optimal for J_1 and the left-hand side of (114) becomes

$$0 \leq \big|J_0(u^0, U^\star(u^0)) - J_0\big(\hat{u}_d^0, \hat{U}(\hat{u}_d^0)\big)\big| = J_0\big(\hat{u}_d^0, \hat{U}(\hat{u}_d^0)\big) - J_0(u^0, U^\star(u^0)). \tag{115}$$

Consequentially, (115) can be rearranged yielding

$$J_0\big(\hat{u}_d^0, \hat{U}(\hat{u}_d^0)\big) \leq J_0(u^0, U^\star(u^0)) + \varepsilon. \tag{116}$$

Since the left-hand side of (116) is optimal, then taking infima overall v in $\mathcal{H}_T^2$ does not reduce it further. Thus, (116) becomes

$$\inf_{u^0\in\mathcal{K}_0} J_0(u^0, U^\star(u^0)) = J_0(u^0, U^\star(u^0)). \tag{117}$$

Combining (116) and (117) yields

$$J_0\big(\hat{u}_d^0, \hat{U}(\hat{u}_d^0)\big) \leq J_0(u^0, U^\star(u^0)) + \varepsilon = \inf_{u^0\in\mathcal{K}_0} J_0(u^0, U^\star(u^0)) + \varepsilon.$$

This concludes our proofs of Theorems 4 and 5.

To obtain the conclusion of Theorem 7, we first note that the non-linear operator $U^\star$ is $1/2$-Hölder continuous. Since Example 5 showed that K is an exponential manifold (in the sense of Definition 11) then the complexity estimates for the attentional neural operator $\hat{U}$ selected in (112) must be as in Table 2. Consider the special case where $\varepsilon_D = \varepsilon_A$ completes the proof of Theorem 7. ∎

A.6 Examples of Compact Sets of Square-Integrable Controls

We will often be considering compact subsets of $\mathcal{H}_T^2$ whereon we establish the existence of Stackelberg equilibria and the possibility of uniform approximation. There are several other examples of a compact subset of $\mathcal{H}_T^2$ routinely encountered in the literature, for example, sets of processes which are Malliavin differentiable with uniformly bounded Malliavin Derivative, see Baños et al. (2018, Corollary C.3.), or perturbations of closed-loop controls which are efficiently approximable (see Section 4.1). Two illustrative, but broad classes, of examples are now constructed; building on compactness results in classical function spaces.

Example 10 (Compactness Via Regularity of the Malliavin Derivative) *For given* $C \geq 0$, *and* $\alpha \in (0,1)$, *define* $\mathcal{K}_{\alpha,C} \subset \mathcal{H}_T^2$ *as the set of* $H \in \mathcal{H}_T^2$ *so that*

$$\sup_{0 \leq s \leq t \leq T} \mathbb{E}[|D_s H_t|^2 + |H_t|^2] \leq C, \quad \sup_{0 \leq s < t \leq T} \frac{\mathbb{E}[|H_t - H_s|^2]}{|t-s|^\alpha} \leq C, \ and \ \sup_{r, 0 \leq s < t \leq T} \frac{\mathbb{E}[|D_t H_r - D_s H_r|^2]}{|t-s|^\alpha} \leq C.$$

For $H \in \mathcal{K}_{\alpha,C}$, *thanks to Nualart (2006, Proposition 1.3.8) we can compute the Malliavin derivative* $D_t \int_0^T H_r dW_r = H_t + \int_t^T D_t H_r dW_r$ *for* $0 \leq s < t \leq T$ *so that we have the following estimate*

$$\frac{\mathbb{E}[|D_t \int_0^T H_r dW_r - D_s \int_0^T H_r dW_r|^2]}{|t-s|^\alpha}$$

$$\leq 2 \frac{\mathbb{E}[|H_t - H_s|^2] + \mathbb{E}[|\int_s^t D_s H_r dW_r|^2] + \mathbb{E}[|\int_t^T D_t H_r - D_s H_r dW_r|^2]}{|t-s|^\alpha}$$

$$\leq 2 \frac{\mathbb{E}[|H_t - H_s|^2]}{|t-s|^\alpha} + 2 \frac{\int_s^t \mathbb{E}[|D_s H_r|^2] dr}{|t-s|^\alpha} + 2 \int_t^T \frac{\mathbb{E}[|D_t H_r - D_s H_r|^2]}{|t-s|^\alpha} dr \leq 2C(1 + T^{1-\alpha} + T).$$

Thanks to Baños et al. (2018, Corollary C3), this estimate implies that the set of random variables $\{\int_0^T H_s dW_s : H \in \mathcal{K}_{\alpha,C}\}$ *is relatively compact in the set of square integrable random variables. This implies by Itô's isometry that* $\mathcal{K}_{\alpha,C}$ *is relatively compact in* $\mathcal{H}_T^2$.

In a very special case of Example 10, one may require the predictable process H. to be a (non-random) smooth function. This next example shows precisely this, and it elucidates the link between classical families of smooth functions and compact sets of open-loop controls.

Example 11 (Compactness Via Continuously Differentiable Martingale Controls) *Fix* $T > 0$. *Let* $W^{1,2}([0,1])$ *denote the Sobolev space on the unit interval* $[0,1]$; *and denote*

its norm by $\|\cdot\|_{W^{1,2}}$. By the Rellich-Kondrashov Theorem, see e.g. Evans (2010, Theorem 5.1), the set of function $\varsigma : [0,1] \to \mathbb{R}$ satisfying

$$\|\varsigma\|_{W^{1,2}} \leq 1 \tag{118}$$

is compact in $L^2([0,1])$. By the Itô's isometry the set of $\mathcal{F}_T$-measurable random variables $\int_0^T \varsigma(t)\, dW_t$ is therefore compact in $L^2(\mathcal{F}_T)$. Since conditional expectations are non-expansive (1-Lipschitz) then the set $\mathcal{K} \subset \mathcal{H}_T^2$ of processes/open-loop controls $u. \overset{\text{def.}}{=} (u_t)_{0 \leq t \leq T}$ of the form

$$u_t = \mathbb{E}\left[\int_0^T \varsigma(s)\, dW_s \Big| \mathcal{F}_t \right] = \int_0^t \varsigma(s)\, dW_s$$

where ς satisfies (118), is compact in $\mathcal{H}_T^2$, and the last inequality held by the Martingale property of the Itô (stochastic) integral.

Example 12 (Deterministic Hölder Continuous Controls) *Fix $T > 0$, $0 < \alpha \leq 1$, and consider the set $\mathcal{K}_\alpha$ of deterministic controls $u. = (u_t)_{t \geq 0}$ in $\mathcal{H}_T^2$ where $t \mapsto u_t$ is an α-Hölder function mapping $[0,T]$ to $[-1,1]^d$. In this case, for each $u., v. \in \mathcal{K}_\alpha$ we have*

$$\mathbb{E}\left[\int_0^T |u_t - v_t|^2 \right]^{1/2} = \left(\int_0^T |u_t - v_t|^2 \right)^{1/2} \leq \max_{0 \leq t \leq T} |u_t - v_t|.$$

Therefore, the map $C([0,1], \mathbb{R}^d) \to \mathcal{H}_T^2$ is a 1-Lipschitz embedding when the domain is equipped with the uniform norm. By the Arzelá–Ascoli Theorem, we have that any set of uniformly bounded α-Hölder functions is relatively compact in $C([0,1], \mathbb{R}^d)$; thus, $\mathcal{K}_\alpha$ is relatively compact in $\mathcal{H}_T^2$.

One can easily extend the construction in Example 11 to non-Martingale controls iterated integrals using the isometries between the space of symmetric functions in $L^2([0,1]^q)$, for any $q \in \mathbb{N}_+$, and the qth Wiener Chaos (see e.g. Nualart, 2006, Theorem 1.1.1); however, do not do so for simplicity of presentation.

Example 13 (Conditioned Lipschitz Perturbations of Random Variables at Terminal Time) *Fix $X \in L^2(\mathcal{F}_T)$, and fix $T > 0$. Consider the set $\mathcal{X}$ of 1-Lipschitz function $f : \mathbb{R}^d \to [-1,1]^d$ which are supported on the hypercube $[-1,1]^d$; i.e. $f(x) = 0$ if $x \notin [-1,1]^d$. By the Arzela-Ascoli Theorem, $\mathcal{X}$ is relatively compact in $C(\mathbb{R}^d, \mathbb{R}^d)$. Since the map sending any $f \in C(\mathbb{R}^d, \mathbb{R}^d)$ to $f(X) \in L^2(\mathcal{F}_T)$ is 1-Lipschitz then the set of random variables $\{f(X) \in L^2(\mathcal{F}_T) : f \in \mathcal{X}\}$ is compact in $L^2(\mathcal{F}_T)$. As in Example (11), since conditional expectations are 1-Lipschitz then the set of controls $u. \overset{\text{def.}}{=} (u_t)_{t \geq 0} \in \mathcal{K} \subset \mathcal{H}_T^2$ of the form*

$$u_t = \mathbb{E}\big[f(X) | \mathcal{F}_t \big]$$

where $f \in \mathcal{X}$, is relatively compact in $\mathcal{H}_T^2$. As a concrete example, one may take f to belong to the set of 1-Lipschitz ReLU Neural Networks with output restricted to belong to $[-1,1]^d$; see e.g. Hong and Kratsios (2024, Theorem 1.1).

Remark 20 (Alternative Proofs of Compactness Directly Via Example 10) *Example 13 can also be obtained from Example 10 upon adding Malliavin differentiability requirements on X and additional regularity. Examples 11 and 12 can alternatively be obtained as straightforward consequences of Example 10. We opted for self-contained presentations for each example to illustrate various construction methods for compacta in $\mathcal{H}_T^2$.*

In practice, one often discretizes their space when implementing it on a digital machine. In these cases, the set of controls is finite and, therefore, compact.

Example 14 (Finite Sets of Controls) *Let $I \in \mathbb{N}$ and $\mathcal{K} \stackrel{\text{def.}}{=} \{u_i\}_{i=1}^I \subset \mathcal{H}_T^2$. Then, $\mathcal{K}$ is compact.*

References

Beatrice Acciaio, Anastasis Kratsios, and Gudmund Pammer. Designing universal causal deep learning models: The geometric (hyper) transformer. *Mathematical Finance*, 2023.

Guillermo Alonso Alvarez, Sergey Nadtochiy, and Kevin Webster. Optimal brokerage contracts in Almgren–Chriss model with multiple clients. *SIAM Journal on Financial Mathematics*, 14(3):855–878, 2023.

Bo An, Milind Tambe, Fernando Ordonez, Eric Shieh, and Christopher Kiekintveld. Refinement of strong Stackelberg equilibria in security games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 25, pages 587–593, 2011.

Anima Anandkumar, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Nikola Kovachki, Zongyi Li, Burigede Liu, and Andrew Stuart. Neural operator: Graph kernel network for partial differential equations. In *ICLR 2020 Workshop on Integration of Deep Neural Models and Differential Equations*, 2020.

Alexander Aurell, Rene Carmona, Gokce Dayanikli, and Mathieu Lauriere. Optimal incentives to mitigate epidemics: a Stackelberg mean field game approach. *SIAM Journal on Control and Optimization*, 60(2):S294–S322, 2022.

David R. Baños, Sindre Duedahl, Thilo Meyer-Brandis, and Frank Proske. Construction of Malliavin differentiable strong solutions of SDEs under an integrability condition on the drift without the Yamada-Watanabe principle. *Ann. Inst. Henri Poincaré Probab. Stat.*, 54(3):1464–1491, 2018. ISSN 0246-0203,1778-7017. doi: 10.1214/17-AIHP845. URL https://doi.org/10.1214/17-AIHP845.

Julio Backhoff, Mathias Beiglbock, Yiqing Lin, and Anastasiia Zalashko. Causal transport in discrete time and applications. *SIAM Journal on Optimization*, 27(4):2528–2562, 2017.

Julio Backhoff-Veraguas, Daniel Bartl, Mathias Beiglböck, and Manu Eder. All adapted topologies are equal. *Probability Theory and Related Fields*, 178(3):1125–1172, 2020.

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *3rd International Conference on Learning Representations, ICLR 2015*, 2015.

Peter L. Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks. *J. Mach. Learn. Res.*, 20:Paper No. 63, 17, 2019. ISSN 1532-4435,1533-7928.

Francesca Bartolucci, Emmanuel de Bezenac, Bogdan Raonic, Roberto Molinaro, Siddhartha Mishra, and Rima Alaifari. Representation equivalent neural operators: a framework for alias-free operator learning. *Advances in Neural Information Processing Systems*, 36, 2024.

J Benitez, Takashi Furuya, Florian Faucher, Anastasis Kratsios, Xavier Tricoche, and Maarten V de Hoop. Out-of-distributional risk bounds for neural operators with applications to the Helmholtz equation. *arXiv preprint arXiv:2301.11509*, 2023.

Sarah Bensalem, Nicolás Hernández Santibáñez, and Nabil Kazi-Tani. Prevention efforts, insurance demand and price incentives under coherent risk measures. *Insurance: Mathematics and Economics*, 93:369–386, 2020.

Alain Bensoussan, Shaokuan Chen, and Suresh P Sethi. The maximum principle for global solutions of stochastic Stackelberg differential games. *SIAM Journal on Control and Optimization*, 53(4):1956–1981, 2015.

Yoav Benyamini and Joram Lindenstrauss. *Geometric nonlinear functional analysis. Vol. 1*, volume 48 of *American Mathematical Society Colloquium Publications*. American Mathematical Society, Providence, RI, 2000. ISBN 0-8218-0835-4. doi: 10.1090/coll/048. URL https://doi.org/10.1090/coll/048.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

Elia Bruè, Simone Di Marino, and Federico Stra. Linear Lipschitz and C^1 extension operators through random projection. *J. Funct. Anal.*, 280(4):Paper No. 108868, 21, 2021. ISSN 0022-1236,1096-0783. doi: 10.1016/j.jfa.2020.108868. URL https://doi.org/10.1016/j.jfa.2020.108868.

Carlos Calderon-Macias. *Artificial neural systems for interpretation and inversion of seismic data*. The University of Texas at Austin, 1997.

Edoardo Calvello, Nikola B Kovachki, Matthew E Levine, and Andrew M Stuart. Continuum attention for neural operators. *arXiv preprint arXiv:2406.06486*, 2024.

Jingyi Cao, Dongchen Li, Virginia R Young, and Bin Zou. Stackelberg differential game for insurance under model ambiguity. *Insurance: Mathematics and Economics*, 106:128–145, 2022.

Shuhao Cao. Choose a transformer: Fourier or galerkin. *Advances in neural information processing systems*, 34:24924–24940, 2021.

René Carmona and Gökçe Dayanıklı. Mean field game model for an advertising competition in a duopoly. *International Game Theory Review*, 23(04):2150024, 2021.

Javier Castro. The Kolmogorov infinite dimensional equation in a Hilbert space via deep learning methods. *Journal of Mathematical Analysis and Applications*, 527(2):127413, 2023.

Tsung-Han Chan, Kui Jia, Shenghua Gao, Jiwen Lu, Zinan Zeng, and Yi Ma. Pcanet: A simple deep learning baseline for image classification? *IEEE transactions on image processing*, 24(12):5017–5032, 2015.

Lv Chen and Yang Shen. On a new paradigm of optimal reinsurance: A stochastic STACK-ELBERG differential game between an insurer and a reinsurer. *ASTIN Bulletin*, 48(2): 905–960, 2018. doi: 10.1017/asb.2018.3.

Albert Cohen, Ronald DeVore, Guergana Petrova, and Przemyslaw Wojtaszczyk. Optimal stable nonlinear approximation. *Foundations of Computational Mathematics*, 22(3):607–648, 2022.

Samuel N. Cohen and Robert J. Elliott. *Stochastic calculus and applications*. Probability and its Applications. Springer, Cham, second edition, 2015. ISBN 978-1-4939-2866-8; 978-1-4939-2867-5. doi: 10.1007/978-1-4939-2867-5. URL `https://doi.org/10.1007/978-1-4939-2867-5`.

Vincent Conitzer. On stackelberg mixed strategies. *Synthese*, 193(3):689–703, 2016.

Vincent Conitzer and Tuomas Sandholm. Computing the optimal strategy to commit to. In *Proceedings of the 7th ACM conference on Electronic commerce*, pages 82–90, 2006.

Gianni Dal Maso. *An introduction to Γ-convergence*, volume 8 of *Progress in Nonlinear Differential Equations and their Applications*. Birkhäuser Boston, Inc., Boston, MA, 1993. ISBN 0-8176-3679-X. doi: 10.1007/978-1-4612-0327-8. URL `https://doi.org/10.1007/978-1-4612-0327-8`.

Gokce Dayanikli and Mathieu Lauriere. A machine learning method for Stackelberg mean field games. *arXiv preprint arXiv:2302.10440*, 2023.

Maarten de Hoop, J. Antonio Lara B., Anastasis Kratsios, Matti Lassas, and Takashi Furuya. Mixture of experts soften the curse of dimensionality in operator learning. *arXiv preprint arXiv:2404.09101*, 2024.

Maarten V de Hoop, Matti Lassas, and Christopher A Wong. Deep learning architectures for nonlinear operator functions and nonlinear inverse problems. *Mathematical Statistics and Learning*, 4(1):1–86, 2022.

Tim De Ryck, Ameya D Jagtap, and Siddhartha Mishra. Error estimates for physics-informed neural networks approximating the Navier–Stokes equations. *IMA Journal of Numerical Analysis*, 44(1):83–119, 2024.

Ludwig Dierks and Sven Seuken. Cloud pricing: The spot market strikes back. *Management Science*, 68(1):105–122, 2022.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.*, 12:2121–2159, 2011. ISSN 1532-4435.

Ilya Dumer, Mark S. Pinsker, and Viacheslav V. Prelov. On coverings of ellipsoids in Euclidean spaces. *IEEE Trans. Inform. Theory*, 50(10):2348–2356, 2004. ISSN 0018-9448,1557-9654. doi: 10.1109/TIT.2004.834759. URL `https://doi.org/10.1109/TIT.2004.834759`.

Gintare Karolina Dziugaite and Daniel M Roy. The size of teachers as a measure of data complexity: Pac-bayes excess risk bounds and scaling laws. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025.

Nicole El Karoui, Shige Peng, and Marie Claire Quenez. Backward stochastic differential equations in finance. *Mathematical finance*, 7(1):1–71, 1997.

Romuald Elie, Thibaut Mastrolia, and Dylan Possamaï. A tale of a principal and many, many agents. *Mathematics of Operations Research*, 44(2):440–467, 2019.

Lawrence C. Evans. *Partial differential equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, second edition, 2010. ISBN 978-0-8218-4974-3. doi: 10.1090/gsm/019. URL `https://doi.org/10.1090/gsm/019`.

VS Fanaskov and Ivan V Oseledets. Spectral neural operators. In *Doklady Mathematics*, pages 1–7. Springer, 2024.

Herbert Federer. *Geometric measure theory*. Die Grundlehren der mathematischen Wissenschaften, Band 153. Springer-Verlag New York, Inc., New York, 1969.

Adeline Fermanian, Terry Lyons, James Morrill, and Cristopher Salvi. New directions in the applications of rough path theory. *IEEE BITS the Information Theory Magazine*, 3 (2):41–53, 2023.

Takashi Furuya and Anastasis Kratsios. Simultaneously solving families of FBSDEs with neural operators of logarithmic depth, constant width, and sub-linear rank. *arXiv preprint arXiv:2409.12335*, 2024.

Luca Galimberti, Anastasis Kratsios, and Giulia Livieri. Designing universal causal deep learning models: The case of infinite-dimensional dynamical systems from stochastic analysis. *arXiv preprint arXiv:2210.13300*, 2022.

Xiao-shan Gao, Shuang Liu, and Lijia Yu. Achieving optimal adversarial accuracy for adversarial deep learning using Stackelberg games. *Acta Mathematica Scientia*, 42(6): 2399–2418, 2022.

Matthias Gerstgrasser and David C Parkes. Oracles & followers: Stackelberg equilibria in deep multi-agent reinforcement learning. In *International Conference on Machine Learning*, pages 11213–11236. PMLR, 2023.

Mario Ghossoub and Michael B Zhu. Stackelberg equilibria with multiple policyholders. *Insurance: Mathematics and Economics*, 116:189–201, 2024.

Somdatta Goswami, Minglang Yin, Yue Yu, and George Em Karniadakis. A physics-informed variational deeponet for predicting crack path in quasi-brittle materials. *Computer Methods in Applied Mechanics and Engineering*, 391:114587, 2022.

Nika Haghtalab, Chara Podimata, and Kunhe Yang. Calibrated Stackelberg games: Learning optimal commitments against calibrated agents. *Advances in Neural Information Processing Systems*, 36, 2023.

Zhongkai Hao, Zhengyi Wang, Hang Su, Chengyang Ying, Yinpeng Dong, Songming Liu, Ze Cheng, Jian Song, and Jun Zhu. Gnot: A general neural operator transformer for operator learning. In *International Conference on Machine Learning*, pages 12556–12569. PMLR, 2023.

Keegan Harris, Steven Wu, and Maria Florina Balcan. Stackelberg games with side information. In *Multi-Agent Security Workshop @ NeurIPS'23*, 2023. URL https://openreview.net/forum?id=4RFv4ODWkp.

Xiuli He, Ashutosh Prasad, and Suresh P Sethi. Cooperative advertising and pricing in a dynamic stochastic supply chain: Feedback Stackelberg strategies. In *PICMET'08-2008 Portland International Conference on Management of Engineering & Technology*, pages 1634–1649. IEEE, 2008.

Camilo Hernández and Dylan Possamaï. Time-inconsistent contract theory. *Mathematical Finance*, 34(3):1022–1085, 2024.

Camilo Hernández, Nicolás Hernández Santibáñez, Emma Hubert, and Dylan Possamaï. Closed-loop equilibria for Stackelberg games: it's all about stochastic targets. *arXiv preprint arXiv:2406.19607*, 2024.

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.

Mingshang Hu, Shaolin Ji, and Xiaole Xue. A global stochastic maximum principle for fully coupled forward-backward stochastic systems. *SIAM Journal on Control and Optimization*, 56(6):4309–4335, 2018. doi: 10.1137/18M1179547.

Emma Hubert, Thibaut Mastrolia, Dylan Possamaï, and Xavier Warin. Incentives, lockdown, and testing: from Thucydides' analysis to the covid-19 pandemic. *Journal of mathematical biology*, 84(5):37, 2022.

Kazufumi Ito, Christoph Reisinger, and Yufei Zhang. A neural network-based policy iteration algorithm with global H^2-superlinear convergence for stochastic games on domains. *Found. Comput. Math.*, 21(2):331–374, 2021. ISSN 1615-3375,1615-3383. doi: 10.1007/s10208-020-09460-1. URL https://doi.org/10.1007/s10208-020-09460-1.

Yuling Jiao, Yanming Lai, Xiliang Lu, Fengru Wang, Jerry Zhijian Yang, and Yuanyuan Yang. Deep neural networks with ReLU-sine-exponential activations break curse of dimensionality in approximation on Hölder class. *SIAM Journal on Mathematical Analysis*, 55(4):3635–3649, 2023.

Debarun Kar, Thanh H Nguyen, Fei Fang, Matthew Brown, Arunesh Sinha, Milind Tambe, and Albert Xin Jiang. Trends and applications in Stackelberg security games. *Handbook of dynamic game theory*, pages 1–47, 2017.

Jussi Keppo, Nizar Touzi, and Ruiting Zuo. Dynamic contracting in asset management under the investor-partner-manager relationship. *Operations Research*, 72(3):903–915, 2024.

Patrick Kidger and Terry Lyons. Universal Approximation with Deep Narrow Networks. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2306–2327. PMLR, 09–12 Jul 2020. URL https://proceedings.mlr.press/v125/kidger20a.html.

Kai A Konrad and Wolfgang Leininger. The generalized Stackelberg equilibrium of the all-pay auction with complete information. *Review of Economic Design*, 11:165–174, 2007.

Yury Korolev. Two-layer neural networks with values in a banach space. *SIAM Journal on Mathematical Analysis*, 54(6):6358–6389, 2022.

Nikola Kovachki, Samuel Lanthaler, and Siddhartha Mishra. On universal approximation and error bounds for Fourier neural operators. *J. Mach. Learn. Res.*, 22:Paper No. [290], 76, 2021. ISSN 1532-4435,1533-7928.

Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Learning maps between function spaces with applications to PDEs. *Journal of Machine Learning Research*, 24 (89):1–97, 2023.

Anastasis Kratsios and Léonie Papon. Universal approximation theorems for differentiable geometric deep learning. *J. Mach. Learn. Res.*, 23:Paper No. [196], 73, 2022. ISSN 1532-4435,1533-7928.

Anastasis Kratsios, Chong Liu, Matti Lassas, Maarten V de Hoop, and Ivan Dokmanić. An approximation theory for metric space-valued functions. *Submitted*, 2023a.

Anastasis Kratsios, Chong Liu, Matti Lassas, Maarten V de Hoop, and Ivan Dokmanić. An approximation theory for metric space-valued functions with a view towards deep learning. *arXiv preprint arXiv:2304.12231*, 2023b.

Anastasis Kratsios, Chong Liu, Matti Lassas, Maarten V de Hoop, and Ivan Dokmanić. Universal geometric deep learning via geometric attention. *arXiv preprint arXiv:2304.12231*, 2023c.

Anastasis Kratsios, Ariel Neufeld, and Philipp Schmocker. Generative neural operators of log-complexity can simultaneously solve infinitely many convex programs. *arXiv preprint arXiv:2508.14995*, 2025.

Emma Kroell, Sebastian Jaimungal, and Silvana M Pesenti. Optimal robust reinsurance with multiple insurers. *arXiv preprint arXiv:2308.11828*, 2023.

Samuel Lanthaler. Operator learning with PCA-Net: upper and lower complexity bounds. *Journal of Machine Learning Research*, 24(318):1–67, 2023.

Samuel Lanthaler and Andrew M Stuart. The curse of dimensionality in operator learning. *arXiv preprint arXiv:2306.15924*, 2023.

Samuel Lanthaler, Siddhartha Mishra, and George E Karniadakis. Error estimates for deeponets: A deep learning framework in infinite dimensions. *Transactions of Mathematics and Its Applications*, 6(1):tnac001, 2022a.

Samuel Lanthaler, Roberto Molinaro, Patrik Hadorn, and Siddhartha Mishra. Nonlinear reconstruction for operator learning of PDEs with discontinuities. *arXiv preprint arXiv:2210.01074*, 2022b.

Jae Yong Lee, SungWoong CHO, and Hyung Ju Hwang. HyperdeepONet: learning operator with complex target function space using the limited resources via hypernetwork. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=OAw6V3ZAhSd`.

Guanguan Li, Qiqiang Li, Yi Liu, Huimin Liu, Wen Song, and Ran Ding. A cooperative Stackelberg game based energy management considering price discrimination and risk assessment. *International Journal of Electrical Power & Energy Systems*, 135:107461, 2022.

Qianxiao Li, Long Chen, Cheng Tai, and Weinan E. Maximum principle based algorithms for deep learning. *Journal of Machine Learning Research*, 18(165):1–29, 2018. URL `http://jmlr.org/papers/v18/17-653.html`.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*, 2020a.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Graph kernel network for partial differential equations. *arXiv preprint arXiv:2003.03485*, 2020b.

Zongyi Li, Daniel Zhengyu Huang, Burigede Liu, and Anima Anandkumar. Fourier neural operator with learned deformations for PDEs on general geometries. *Journal of Machine Learning Research*, 24(388):1–26, 2023.

G. G. Lorentz. Metric entropy and approximation. *Bull. Amer. Math. Soc.*, 72:903–937, 1966. ISSN 0002-9904. doi: 10.1090/S0002-9904-1966-11586-0. URL `https://doi.org/10.1090/S0002-9904-1966-11586-0`.

George G. Lorentz, Manfred v. Golitschek, and Yuly Makovoz. *Constructive approximation*, volume 304 of *Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1996. ISBN 3-540-57028-4. doi: 10.1007/978-3-642-60932-9. URL https://doi.org/10.1007/978-3-642-60932-9. Advanced problems.

Lu Lu, Pengzhan Jin, and George Em Karniadakis. Deeponet: Learning nonlinear operators for identifying differential equations based on the universal approximation theorem of operators. *arXiv preprint arXiv:1910.03193*, 2019.

Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via deeponet based on the universal approximation theorem of operators. *Nature machine intelligence*, 3(3):218–229, 2021.

Carlo Marcati and Christoph Schwab. Exponential convergence of deep operator networks for elliptic partial differential equations. *SIAM Journal on Numerical Analysis*, 61(3): 1513–1545, 2023.

Yves Meyer. *Ondelettes et opérateurs. I.* Actualités Mathématiques. [Current Mathematical Topics]. Hermann, Paris, 1990. ISBN 2-7056-6125-0. Ondelettes. [Wavelets].

Radu Miculescu. Approximations by Lipschitz functions generated by extensions. *Real Anal. Exchange*, 28(1):33–40, 2002/03. ISSN 0147-1937,1930-1219. doi: 10.14321/realanalexch. 28.1.0033. URL https://doi.org/10.14321/realanalexch.28.1.0033.

Roberto Molinaro, Yunan Yang, Björn Engquist, and Siddhartha Mishra. Neural inverse operators for solving PDE inverse problems. In *Proceedings of the 40th International Conference on Machine Learning*, pages 25105–25139, 2023.

Sepehr Mousavi, Shizheng Wen, Levi Lingsch, Maximilian Herde, Bogdan Raonić, and Siddhartha Mishra. Rigno: A graph-based framework for robust and accurate operator learning for pdes on arbitrary domains. *arXiv preprint arXiv:2501.19205*, 2025.

James R. Munkres. *Topology*. Prentice Hall, Inc., Upper Saddle River, NJ, second edition, 2000. ISBN 0-13-181629-2.

David Nualart. *The Malliavin calculus and related topics*. Probability and its Applications (New York). Springer-Verlag, Berlin, second edition, 2006. ISBN 978-3-540-28328-7; 3-540-28328-5.

Jiri Outrata, Michal Kocvara, and Jochem Zowe. *Nonsmooth Approach to Optimization Problems with Equilibrium Constraints: Theory, Applications and Numerical Results*, volume 28 of *Nonconvex Optimization and Its Applications*. Springer Science & Business Media, Boston, MA, 2013.

Etienne Pardoux and Shige Peng. Adapted solution of a backward stochastic differential equation. *Systems & control letters*, 14(1):55–61, 1990.

Shige Peng and Zhen Wu. Fully coupled forward-backward stochastic differential equations and applications to optimal control. *SIAM Journal on Control and Optimization*, 37(3): 825–843, 1999.

Philipp Petersen and Jakob Zech. Mathematical theory of deep learning. *arXiv preprint arXiv:2407.18384*, 2024.

Guergana Petrova and Przemysław Wojtaszczyk. Lipschitz widths. *Constructive Approximation*, 57(2):759–805, 2023.

Allan Pinkus. *N-widths in Approximation Theory*, volume 7. Springer Science & Business Media, 2012.

Bogdan Raonic, Roberto Molinaro, Tim De Ryck, Tobias Rohner, Francesca Bartolucci, Rima Alaifari, Siddhartha Mishra, and Emmanuel de Bézenac. Convolutional neural operators for robust and accurate learning of PDEs. *Advances in Neural Information Processing Systems*, 36, 2024.

Christoph Reisinger and Yufei Zhang. Rectified deep neural networks overcome the curse of dimensionality for nonsmooth value functions in zero-sum games of nonlinear stiff systems. *Analysis and Applications*, 18(06):951–999, 2020.

James C. Robinson. *Dimensions, embeddings, and attractors*, volume 186 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 2011. ISBN 978-0-521-89805-8.

Zuowei Shen, Haizhao Yang, and Shijun Zhang. Optimal approximation rate of ReLU networks in terms of width and depth. *J. Math. Pures Appl. (9)*, 157:101–135, 2022. ISSN 0021-7824. doi: 10.1016/j.matpur.2021.07.009. URL https://doi.org/10.1016/j.matpur.2021.07.009.

Xiaofei Shi, Daran Xu, and Zhanhao Zhang. Deep learning algorithms for hedging with frictions. *Digital Finance*, 5(1):113–147, 2023.

Stanisław J. Szarek. A Banach space without a basis which has the bounded approximation property. *Acta Math.*, 159(1-2):81–98, 1987. ISSN 0001-5962,1871-2509. doi: 10.1007/BF02392555. URL https://doi.org/10.1007/BF02392555.

Aad W. van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. ISBN 0-387-94640-3. doi: 10.1007/978-1-4757-2545-2. URL https://doi.org/10.1007/978-1-4757-2545-2. With applications to statistics.

Jan van Mill. *The infinite-dimensional topology of function spaces*, volume 64 of *North-Holland Mathematical Library*. North-Holland Publishing Co., Amsterdam, 2001. ISBN 0-444-50557-1.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Bernhard Von Stengel and Shmuel Zamir. Leadership with commitment to mixed strategies. Technical report, Technical Report LSE-CDAM-2004-01, CDAM Research Report, 2004.

Nik Weaver. *Lipschitz algebras*. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, second edition, 2018. ISBN 978-981-4740-63-0.

Dmitry Yarotsky. Elementary superexpressive activations. In *International conference on machine learning*, pages 11932–11940. PMLR, 2021.

Jiongmin Yong. A leader-follower stochastic linear quadratic differential game. *SIAM Journal on Control and Optimization*, 41(4):1015–1041, 2002.

Jiongmin Yong and Xun Yu Zhou. *Stochastic Controls: Hamiltonian Systems and HJB equations*. Springer New York, NY, 1 edition, 1999. ISBN 978-1-4612-1466-3.

Hui Zhang, Fengrui Zhang, Bing Gong, Xuan Zhang, and Yifan Zhu. The optimization of supply chain financing for bank green credit using Stackelberg game theory in digital economy under internet of things. *Journal of Organizational and End User Computing (JOEUC)*, 35(3):1–16, 2023.

Kai Zhang, Ivor W Tsang, and James T Kwok. Improved Nyström low-rank approximation and error analysis. In *Proceedings of the 25th international conference on Machine learning*, pages 1232–1239, 2008.

Shijun Zhang, Zuowei Shen, and Haizhao Yang. Deep network approximation: Achieving arbitrary accuracy with fixed number of neurons. *Journal of Machine Learning Research*, 23(276):1–60, 2022.

Zijie Zheng, Lingyang Song, Zhu Han, Geoffrey Ye Li, and H Vincent Poor. A Stackelberg game approach to proactive caching in large-scale mobile edge networks. *IEEE Transactions on Wireless Communications*, 17(8):5198–5211, 2018.