

A Bachelor's Thesis at the FernUniversität in Hagen

A Functional Analysis Perspective on Neural Operators

submitted by
Artem Shramko

First examiner: Univ.-Prof. Dr. Mugnolo
Second examiner: Dr. Lu

Contents

1	Introduction	5
1.1	Motivation	5
1.2	Literature Review	5
1.3	Outline	6
2	Foundations	9
2.1	Problem Definition	9
2.2	Neural Operator	10
2.2.1	Motivation	11
2.2.2	Induced Nemytskii Operators Σ and $\mathcal{W}$	13
2.2.3	The Kernel Integral Operator	15
2.2.4	Single Neural Operator Layer	18
2.2.5	Lifting and Projection	19
2.2.6	Properties of the Neural Operator	20
2.2.7	Neural Operator Instantiations	21
3	Discretization Invariance	23
3.1	Discretization Invariance: Motivation	23
3.2	Discretization Invariance: Definitions	24
3.3	Discretization Invariance Theorem	25
3.3.1	Analysis of the Discretization Invariance Theorem	26
3.3.2	Refined Discretization Invariance Proof	29
3.4	Explicit convergence rates for discretized operators	34
4	Universal Approximation Theorems	41
4.1	Motivating Problem	41
4.1.1	Fixed-Grid based Approach to Universal Approximation	41
4.1.2	Analysis of the Fixed Grid UAT	46
4.2	Generalised <i>Encode-Approximate-Decode</i> Framework	47
4.2.1	On the role of the Approximation Property of Banach Spaces	47
4.2.2	Generalised UAT for Banach Spaces with AP	48
4.2.3	Discussion on the Error-Decomposition and Applied Architectures	51
4.3	Continuous Non-Locality Guarantees Discretization Invariance	53
4.3.1	Universal Approximation via Integral Kernel Operators	53
4.3.2	Relaxing Compactness to Boundedness for Sobolev Spaces	56
4.3.3	Integration as the Mechanism for Discretization Invariant Approximation	59
4.4	Synthesis: the Structure of Universal Approximation	61
4.4.1	Necessary Components for Universal Approximation	61
4.4.2	The Proof Template	62
4.4.3	The Tradeoff: Minimality vs. Quantitative Control	64

5	Conclusion	67
A	Neural Operator Properties	73
A.1	Neural Operators on $C(\bar{D})$ with sup-norm	73
A.2	Neural Operators on Sobolev spaces $W^{k,p}$ with $kp > d$	75
B	Discretization Invariance Proof	79

Chapter 1

Introduction

1.1 Motivation

Many dynamic systems in Science and Engineering are governed by Partial Differential Equations. Since closed-form solutions are often unavailable, relying on numerical methods, such as finite elements, finite differences or spectral methods, becomes prohibitively expensive as the required resolution increases. Given the rapid progress of Deep Learning in such fields as Computer Vision and Natural Language Processing within the last decade, a natural question arises: can similar methods be applied to learn the solution operator to a PDE directly from the observed data, instead of relying on the repeated numerical integration? This is the premise of operator learning. Unlike standard neural networks, which map finite dimensional vectors between fixed grids on Euclidean spaces, the goal is to learn mappings between infinite-dimensional Banach spaces. This requires architectures that can accept inputs sampled at arbitrary locations and generalize across discretizations, a property known as *Discretization Invariance*, which is central to this thesis.

However, a fundamental tension arises: neural networks are inherently finite dimensional objects, but the operators we wish to approximate act between infinite dimensional function spaces. Resolving this tension is precisely the domain of Functional Analysis, as it provides means of justifying the reduction of an infinite dimensional problem via finite dimensional approximations. We refer to the subclass of neural networks that learn mappings between function spaces as *Neural Operators*. The theoretical foundations of Neural Operators, from Universal Approximation guarantees to Discretization Invariance, rely on classical functional-analytic tools such as compactness, the approximation property of Banach spaces, and Sobolev Embedding Theorem. This thesis presents a detailed study of these theoretical foundations in the context of Neural Operators, with a particular focus on Discretization Invariance and the proof strategies for the Universal Approximation guarantees.

1.2 Literature Review

The first Universal Approximation results for neural networks, i.e. their ability to approximate continuous real valued functions defined on compact subsets of Euclidean domains, date back to Cybenko (1989) [1] and Hornik (1989) [2]. They are based on the classical Stone-Weierstrass Theorem and highlight the key role of *non-linearity* for the Universal Approximation guarantees. Chen and Chen (1995) [3] pioneered the field of operator learning by extending the Universal Approximation guaranties for neural networks to infinite dimensional Banach spaces, and introduced key components such as *non-locality*, as well as an approach on how to reduce the problem to finite dimensional approximation, referred to as *encode-approximate-decode* proof strategy, which is used extensively in the subsequent

literature.

Lu et. al. (2021) [4] extended their work by adding depth to the neural network architecture known as *DeepONet*, which has achieved success in practical applications and brought attention to the field. Subsequent works extended the architecture by applying Monte Carlo approximation of kernel methods (Nelsen et. al. (2021) [5]) or combining the ideas from the Principal Component Analysis (*PCA-Net*, Bhattacharya et al. (2021) [6]), among others. The main limitation of these approaches was the reliance on the fixed discretization grid, which requires the neural network parameters to be retrained upon the change of grid resolution.

A discretization invariant architecture known as the Fourier Neural Operator (FNO) was introduced by Li (2020) [7] and remains one of the most widely applied Neural Operator architectures to this day. Inspired by the Green operator, it relies on the Fast Fourier Transform algorithm to decompose the input function and achieves Discretization Invariance by implementing a continuous Kernel Integral Operator in its layers. Kovachki et al. (2021) [8] provided theoretical foundations and Universal Approximation guarantees for FNO on Hilbert spaces and later extended their results to a generalized class of Kernel Integral based Neural Operators on Banach spaces, including the Discretization Invariance guarantees [9].

While the majority of the previous literature was focused on the practical Neural Operator architectures, Lanthaler et al. (2025) [10] investigated the minimal requirements for the Universal Approximation guarantees to hold for Neural Operators. They illustrate based on a FNO with only one Fourier mode, referred to as Averaging Neural Operator (ANO), that *non-locality* and *non-linearity* are a sufficient criteria and show that many of the common Neural Operator architectures can be reduced to ANO as their special case, thus providing generalized Universal Approximation guarantees.

The focus of the recent literature on Neural Operators has shifted towards finding explicit convergence rates and investigating the discrepancies between theoretical guarantees and the observed empirical performance of Neural Operator architectures. Despite their state-of-the-art performance on many benchmarks, Fourier Neural Operators display a preference for learning low-frequency components and struggle on PDEs with strong oscillations, a phenomenon referred to as *Spectral bias* [11, 12].

Separately, the theoretical Discretization Invariance results in [9] promise resolution-independent evaluation, but the proofs are non-constructive and provide no convergence rate. Lanthaler, Stuart and Trautner (2024) [13] close this gap for the Fourier Neural Operator on periodic domains, establishing explicit algebraic rates via aliasing analysis, while Brugiapaglia, Franco and Nelsen [14] survey the parallel question of sample complexity, i.e. how many training examples suffice to learn an operator that achieves a certain target accuracy. To the best of our knowledge, explicit discretization convergence rates for general Kernel Integral Neural Operators on bounded domains $D \subset \mathbb{R}^d$ with Lipschitz boundary, with inputs in Sobolev spaces $W^{k,p}(D)$ satisfying $kp > d$, have not been established, which we address in Chapter 3 of this thesis.

1.3 Outline

We begin Chapter 2 by defining the general problem that Neural Operators are intended to solve in section 2.1 and providing formal definitions of Neural Operators in section 2.2. Specifically, we establish the core components of Neural Operators, provide their functional-analytic definitions and study their properties, filling in the gaps that we identified in the literature.

In Chapter 3, we analyze the Discretization Invariance property of Neural Operators, which guarantees that a trained model generalizes across different discretization resolu-

tions without retraining. We revisit the proof of Kovachki et al. [9] and clarify several implicit assumptions in [section 3.3](#). Building on the Sobolev Embedding Theorem and the Morrey inequality [15, Cor. E.14, p. 302], we provide an alternative proof that relaxes the compactness requirement to boundedness on Sobolev spaces in [subsection 3.3.2](#) and establishes explicit algebraic convergence rates for discretized operators in [section 3.4](#). To the best of our knowledge, these results are original to this thesis.

In [Chapter 4](#), we study Universal Approximation guarantees for Neural Operators. We begin with the theorem of Chen and Chen (1995) [3] in [subsection 4.1.1](#), which establishes sufficiency and implicitly introduces the key requirements of non-linearity, non-locality, and compactness, as well as the encode-approximate-decode proof strategy ([subsection 4.1.2](#)). We then present the generalization by Kovachki et al. [9], who extend the proof pattern to general Banach spaces via the Approximation Property ([section 4.2](#)) and establish Universal Approximation for Kernel Integral based Neural Operators ([section 4.3](#)). Analogously to our Discretization Invariance result, we further relax the compactness requirement to boundedness on Sobolev spaces via the Morrey inequality under additional assumptions on the target operator regularity ([subsection 4.2.2](#)). In the concluding synthesis ([section 4.4](#)), we formalize the necessary components and the shared proof template, and contrast the Kernel Integral approach of Kovachki et al. with the minimal Averaging Neural Operator of Lanthaler et al. (2025) [10], discussing the tradeoff between architectural richness and the possibility of obtaining explicit convergence rates.

Chapter 2

Foundations

2.1 Problem Definition

We consider the problem of learning a mapping between two infinite-dimensional Banach spaces. Let $D \subset \mathbb{R}^d$ be a bounded Lipschitz domain. We define the input space $\mathcal{A}$ as a Banach space of real-valued coefficient functions (representing physical parameters such as conductivity or permeability), and the output space $\mathcal{U}$ as the Banach space of solution functions, both defined over domain D . Our objective is to learn the non-linear operator $\mathcal{G}^\dagger : \mathcal{A} \rightarrow \mathcal{U}$, which maps a given coefficient $a \in \mathcal{A}$ to its corresponding PDE solution $u \in \mathcal{U}$.

The ground truth operator $\mathcal{G}^\dagger$ arises from the inversion of a parametric Partial Differential Equation. Specifically, we consider a system

$$(L_a u)(x) = f(x), \quad \forall x \in D$$

where $f \in \mathcal{U}^*$ is a fixed forcing term, and for each $a \in \mathcal{A}$, $L_a : \mathcal{U} \rightarrow \mathcal{U}^*$ is the associated differential operator. We formalize this as follows:

Definition 2.1.1 (Target Operator; [9, Sec. 2.1]). Let $D \subset \mathbb{R}^d$ be a bounded Lipschitz domain and let $\mathcal{A}, \mathcal{U}$ be admissible function spaces over D , i.e. each is one of $C(\bar{D})$, $L^p(D)$ or $W^{k,p}(D)$ with $p \in [1, \infty)$, equipped with its canonical norm. Let $f \in \mathcal{U}^*$ be a fixed forcing term and $L_a : \mathcal{U} \rightarrow \mathcal{U}^*$ the differential operator associated to the parameter $a \in \mathcal{A}$. The *target operator* $\mathcal{G}^\dagger : \mathcal{A} \rightarrow \mathcal{U}$ is defined as

$$\mathcal{G}^\dagger(a) := L_a^{-1}f. \tag{2.1}$$

Remark 2.1.2. We draw a distinction between the solution operator $L_a^{-1} : f \mapsto u$ and the operator $\mathcal{G}^\dagger : a \mapsto u$ that we want to learn. Given a *fixed parameter function* $a \in \mathcal{A}$, the *linear* operator L_a^{-1} maps the forcing function $f \in \mathcal{U}^*$ to the solution function $u \in \mathcal{U}$, whereas $\mathcal{G}^\dagger : \mathcal{A} \rightarrow \mathcal{U}$ maps any parameter function $a \in \mathcal{A}$ to the solution $u \in \mathcal{U}$, given a *fixed forcing function* f . Hence, the operator $\mathcal{G}^\dagger$ is generally *non-linear*.

The PDE setting above motivates the general operator learning problem. The theoretical results of this thesis depend only on *continuity* of $\mathcal{G}^\dagger$ and *regularity* of the domain D , not on the PDE structure itself. We formalize the learning task as follows.

Definition 2.1.3 (Operator Learning Problem; [9, Sec. 2.2]). Let $D \subset \mathbb{R}^d$ and $D' \subset \mathbb{R}^{d'}$ be bounded Lipschitz domains, and let $\mathcal{A}, \mathcal{U}$ be admissible function spaces over D and D' respectively, i.e. each is one of $C(\bar{D})$, $L^p(D)$ or $W^{k,p}(D)$ with $p \in [1, \infty)$, equipped with its canonical norm. Let $\mathcal{G}^\dagger : \mathcal{A} \rightarrow \mathcal{U}$ be continuous. The operator learning problem is to

construct a parametric approximation $\mathcal{G}_\theta : \mathcal{A} \rightarrow \mathcal{U}$, $\theta \in \mathbb{R}^p$, such that, given a compact set $K \subset \mathcal{A}$,

$$\sup_{a \in K} \|\mathcal{G}^\dagger(a) - \mathcal{G}_\theta(a)\|_{\mathcal{U}} \quad (2.2)$$

is minimized. Unlike standard deep learning approaches that map finite-dimensional vectors (discretized grids), $\mathcal{G}_\theta$ is defined directly on the Banach spaces.

Remark 2.1.4 (Neural Operator Notation). We retain the parameter $\theta \in \mathbb{R}^p$ in [Definition 2.1.3](#), following [\[9\]](#), because the learning problem is defined as minimization over θ . The focus of this thesis, however, is the existence of a Neural Operator approximating the true operator $\mathcal{G}^\dagger$ rather than its construction. In every such statement θ occurs only under an existential quantifier and thus carries no information about the operator. We therefore write $\mathcal{G}$ for a generic Neural Operator under consideration, restoring the subscript only where the value of θ is significant.

Remark 2.1.5 (Banach Spaces). The reviewed literature on Neural Operators commonly refers to the general "*Banach spaces of functions*" when discussing the spaces $\mathcal{A}$ and $\mathcal{U}$ [\[10, 9\]](#). To accommodate for different topologies used by Chen and Chen [\[3\]](#), Kovachki et al. [\[9\]](#) and Lanthaler et al. [\[10\]](#) in their Universal Approximation Theorems discussed in [Chapter 4](#), we define the spaces explicitly as in [Definition 2.1.1](#) and [Definition 2.1.3](#). Throughout, $D \subset \mathbb{R}^d$ denotes a bounded Lipschitz domain. We write $C(\bar{D}; \mathbb{R}^m)$ for the Banach space of continuous functions with the supremum norm, $L^p(D; \mathbb{R}^m)$ with $1 \leq p < \infty$ for the Lebesgue space of the equivalence classes of p -integrable functions with $\|\cdot\|_{L^p}$ norm, and $W^{k,p}(D; \mathbb{R}^m)$ for the Sobolev space of L^p functions whose weak derivatives up to k -th order lie in L^p , combined with the $\|\cdot\|_{W^{k,p}}$ norm. When mentioning Banach spaces, we refer to the spaces above with their respective norms, unless a different topology is specified explicitly.

2.2 Neural Operator

In this section, we discuss the general formulation of the Neural Operator as presented in Kovachki et al. (2023) [\[9\]](#). While empirical successes of specific architectures, such as the DeepONet [\[4\]](#), the Fourier Neural Operator (FNO) [\[7\]](#) and Graph Neural Operator (GNO) [\[16\]](#), have been previously established, Kovachki et al. [\[9\]](#) unify these architectures under a rigorous mathematical framework.

To understand the mechanics of this framework, it is crucial to establish *why* the Neural Operator takes its specific form.

Remark 2.2.1 (On the definitions of Neural Operators in the literature). The Neural Operator architecture appears under different but structurally equivalent descriptions across the examined literature ([section 1.2](#)), all of which decompose the operator into three components: an *encoding* or *lifting* stage, a *learning* stage, and a *decoding* or *projection* stage.

Kovachki et al. [\[9\]](#) present the architecture as a composition of a pointwise Lifting operator $\mathcal{P}$, iterative layers consisting of a sum of Kernel Integral operators $\mathcal{K}_t$ and linear Skip-connection operator $\mathcal{W}_t$ composed with pointwise nonlinearities σ , and a pointwise Projection operator $\mathcal{Q}$:

$$\mathcal{G}_\theta := \mathcal{Q} \circ \sigma_T(W_{T-1} + \mathcal{K}_{T-1} + b_{T-1}) \circ \cdots \circ \sigma_1(W_0 + \mathcal{K}_0 + b_0) \circ \mathcal{P}. \quad (2.3)$$

Lanthaler et al. [\[10\]](#) identify a hidden encoder-decoder structure inherent to the Neural Operator architectures. In their Averaging Neural Operator (ANO), the composition of a

non-local averaging operation R followed by a pointwise nonlinearity σ and integration L acts as a nonlinear encoder that maps the input function to a constant vector:

$$v := (L \circ R)(u) = \int \sigma(R(u(x), x)) dx \in \mathbb{R}^{d_c}, \quad (2.4)$$

which is then decoded by a pointwise map $\mathcal{Q} : \mathbb{R}^{d_c} \rightarrow Y$ to produce the output function. The *learning* and *encoding* stages are thus merged into a single non-local operation.

The recent survey by Brugiapaglia, Franco, and Nelsen (2026) [14] frames the construction as an instance of the *encode–approximate–decode* strategy that originates from Chen and Chen (1995) [3].

All three descriptions capture the same structural idea: an infinite-dimensional input function is mapped into a finite-dimensional representation, transformed, and projected back to the output space. The approaches differ in how explicitly they specify the intermediate operations. Kovachki et al. define the *learning* stage as a parameterized Kernel Integral operator, which makes the analytical properties of the architecture (continuity, Discretization Invariance, approximation rates) directly accessible and reflects the empirical architectures (FNO, GNO) most closely. In this thesis we adopt the Kovachki et al. formulation, as the explicit Kernel Integral structure, first proposed by Li et al. (2020) [16] and later generalized by Kovachki et al. (2023) [9], is essential for the Discretization Invariance analysis in Chapter 3 and the convergence rate estimates derived there.

The remainder of this section introduces the components of the Neural Operator architecture following Kovachki et al. [9].

Remark 2.2.2 (Operator-theoretic formulation of the Neural Operator). Kovachki et al. [9, sec. 3] define the Neural Operator components as maps on Euclidean spaces, effectively specifying the discretized architecture, and interpret these pointwise maps as inducing Nemytskii operators on Lebesgue and Sobolev spaces. Lanthaler et al. [10, Sec. 1.2] instead define the architecture on general Banach spaces of real-valued functions, without establishing the function space properties of the resulting operators. In the subsequent sections of Chapter 2, we make these properties explicit, providing operator-theoretic definitions of the Kernel Integral based Neural Operator and its components, instantiated on the Lebesgue spaces $L^p(D)$ of (equivalence classes of) p -integrable functions with $p \in [1, \infty)$. The equivalent definitions and proofs of selected operator properties, relevant to the Discretization Invariance and Universal Approximation results, are provided for the $C(\bar{D})$ space with sup-norm and $W^{k,p}(D)$ space in Appendix A.

2.2.1 Motivation

Kovachki et al. (2023) introduce the Neural Operator directly as a composed, deep neural architecture operating on latent feature spaces in $\mathbb{R}^d$ [9]. While this presentation efficiently sets up the subsequent Universal Approximation Theorem, which we discuss in detail in Chapter 4, it bypasses the classical functional-analytic roots that naturally motivate this specific design. To rigorously understand the Neural Operator architecture, we first reconstruct the underlying logic by analyzing a direct, single-step mapping between function spaces before expanding into the deep latent formulation.

Remark 2.2.3 (On well-posedness of the approximating operator). Our objective is to learn the true operator $\mathcal{G}^\dagger : \mathcal{A} \rightarrow \mathcal{U}$ (Definition 2.1.1). Because both $\mathcal{A}$ and $\mathcal{U}$ are infinite-dimensional function Banach spaces (as described in Remark 2.1.5), we cannot observe them directly. Empirically, we work with a finite L number of pointwise evaluations of the input function $a \in \mathcal{A}$ on a discrete subset D_L of the bounded domain $D \subset \mathbb{R}^d$. D_L is assumed to approximate D arbitrarily well, as we show in Chapter 3. Consequently,

the learned empirical operator $\hat{\mathcal{G}}_L$ takes the signature $\hat{\mathcal{G}}_L : \mathbb{R}^{L \times d} \rightarrow \mathcal{U}$. For this mapping to be mathematically well-posed, i.e. accepting a finite-dimensional discrete input while outputting a continuous function in $\mathcal{U}$, the core operations of $\hat{\mathcal{G}}_L$ must act globally and continuously over the domain D , effectively interpolating the discrete observations $\{x_i, a(x_i)\}_{i=1}^L$.

Remark 2.2.4 (Green’s function motivation). To motivate the construction of this mapping, we first consider a simplified, single-step transformation directly from an input function $a \in \mathcal{A}$ to an output function $u \in \mathcal{U}$. For this construction, we follow the argumentation of Li et al. (2020) [16, sec. 3] which illustrates the intuition behind the Kernel Integral operator being inspired by the Green’s function [15, Def. 3.18, p. 97]. In functional analysis, the solution to a linear PDE is often expressed via a Green’s function, which takes the form of a global integral operator. Inspired by this, the natural choice for our continuous mapping is a parameterized Kernel Integral Operator $\mathcal{K}$. If we were to approximate the map $a \rightarrow u$ directly, it would take the form:

$$(\mathcal{K}a)(x) = \int_D \kappa(x, y) a(y) d\mu(y), \quad \forall x \in D \quad (2.5)$$

where μ is a Borel measure on $\mathbb{R}^d$ (typically chosen as the Lebesgue measure [9]) and $\kappa \in C(\overline{D \times D}; \mathbb{R}^{d_u \times d_a})$ is a continuous integration kernel. We provide the formal definition of the Kernel Integral Operator and study its properties in subsection 2.2.3.

From a functional-analytic perspective, the operator $\mathcal{K}$ is inherently compact. We establish this fact rigorously in Proposition 2.2.17. This compactness has architectural consequences, as discussed in the following remark.

Remark 2.2.5 (Motivation behind the Skip Connection operator). Relying solely on a compact operator introduces a critical limitation. Compact integral operators map weakly convergent sequences to strongly convergent ones, fundamentally acting as low-pass filters that smooth their inputs. If the model relied exclusively on $\mathcal{K}$, it would suffer from over-smoothing and fail to approximate target operators that contain non-compact components, such as Differential Operators or the Identity map.

To resolve this, Li et al. (2023) [16] augment the global integration operator $\mathcal{K}$ with a local, pointwise linear transformation. Let W be a learnable weight matrix, which induces a pointwise operator $\mathcal{W}$ defined as $(\mathcal{W}a)(x) = W(a(x))$ (formal definition in subsection 2.2.2). Critically, the operator $\mathcal{W}$ is non-compact. By defining the mapping as $\mathcal{W} + \mathcal{K}$, the architecture breaks strict compactness. The compact operator $\mathcal{K}$ efficiently captures non-local, low-frequency dependencies, while the non-compact pointwise operator $\mathcal{W}$ acts as a skip connection that preserves and propagates local, high-frequency information [9].

Remark 2.2.6 (On construction of the hidden network layer). While this single-step construction, where we sum a global integral and a local linear map, is mathematically sound, a single application is insufficiently expressive to approximate highly non-linear operators. To achieve the capacity of a deep neural network, the architecture does not map directly from $\mathcal{A}$ to $\mathcal{U}$. Instead, the input $a(x)$ is first *lifted* to a higher-dimensional latent representation $v(x) \in \mathbb{R}^{d_v}$. The combined operator, followed by a continuous non-linear activation function σ , is then applied iteratively to generate a sequence of hidden continuous functions $v_t \rightarrow v_{t+1}$:

$$v_{t+1}(x) = \sigma\left((\mathcal{W}v_t)(x) + (\mathcal{K}v_t)(x)\right) \quad (2.6)$$

After T iterations, the final latent representation $v_T(x)$ is projected back down to the target space $\mathcal{U}$. This iterative formulation defines the core layers of the Neural Operator, retaining the functional-analytic guarantees of the single-step map while leveraging the expressive power of deep latent spaces. The Kernel Integral Layer is further discussed in subsection 2.2.4.

Remark 2.2.7 (Expansion to higher dimensions). As mentioned above, in order to improve the learning capacity of the empirical model, Kovachki et al. (2023) [9] expand the dimensionality of hidden representations $v_t(x)$ of inputs $\{a(x_i)\}_{i=1}^L$ by introducing a pointwise lifting operator $P : \mathbb{R}^{d_a} \rightarrow \mathbb{R}^{d_v}$ with $d_v > d_a$. Consequently, the output of the final T -th layer needs to be mapped back to the output dimension d_u , which defines a pointwise projection map $Q : \mathbb{R}^{d_v} \rightarrow \mathbb{R}^{d_u}$. The *lifting* and *projection* functions, as well as the Nemytskii operators $\mathcal{Q}$ and $\mathcal{P}$ that they induce, are formally defined in [subsection 2.2.5](#).

Combining the lifting, iterative layers, and projection results in the full Neural Operator architecture $\mathcal{G}_\theta : \mathcal{A} \rightarrow \mathcal{U}$ expressed as

$$\mathcal{G}_\theta := \mathcal{Q} \circ \sigma_T(\mathcal{W}_{T-1} + \mathcal{K}_{T-1} + b_{T-1}) \circ \dots \circ \sigma_1(\mathcal{W}_0 + \mathcal{K}_0 + b_0) \circ \mathcal{P} \quad (2.7)$$

The formal definition of the operator is discussed in [subsection 2.2.6](#).

However, Kovachki et al. (2023) deviate from this formulation in their proof of the Universal Approximation Theorem [9, Thm. 11] by dropping the non-compact linear map $\mathcal{W}$ entirely. This is not a mathematical necessity imposed by compactness constraints, but rather a simplification: the Kernel Integral Operators $\mathcal{K}$ alone suffice for Universal Approximation, as we discuss in [section 4.3](#).

2.2.2 Induced Nemytskii Operators Σ and $\mathcal{W}$

The maps σ, W, Q, P defined in [subsection 2.2.6](#) are all pointwise functions, or "local maps" as Kovachki et al. refer to them in [9, sec. 3]. However, since $\hat{\mathcal{G}}$ is defined to be a map between function spaces, a rigorous justification that those functions induce operators on Banach spaces is required. The authors briefly mention that "*P, Q, and σ_t can be thought of as defining Nemitsky (Nemytskii) operators ... when each of their components are assumed to be Borel measurable*" [9, sec. 3, p. 10]. However, these assumptions are not verified in the reviewed literature that offers definitions of Kernel Integral based Neural Operators [9, 7, 16]. Hence, we devote this section to verifying that the requirements for functions σ and W to be Nemytskii operators are fulfilled and address the lifting P and projection Q operators in [subsection 2.2.5](#).

Definition 2.2.8 (Nemytskii Operator, [17, Sec. 10.3.4]). Let $\Omega \subset \mathbb{R}^d$ be a measurable set. Let $f : \Omega \times \mathbb{R}^m \rightarrow \mathbb{R}^k$ be a function satisfying the Carathéodory conditions:

1. $f(x, \cdot) : \mathbb{R}^m \rightarrow \mathbb{R}^k$ is *continuous* for almost every $x \in \Omega$.
2. $f(\cdot, y) : \Omega \rightarrow \mathbb{R}^k$ is *measurable* for every $y \in \mathbb{R}^m$.

For any measurable function $u : \Omega \rightarrow \mathbb{R}^m$, the *induced Nemytskii operator* N_f is defined pointwise by:

$$[N_f(u)](x) = f(x, u(x)), \quad \text{for a.e. } x \in \Omega. \quad (2.8)$$

The operator N_f maps $L^p(\Omega; \mathbb{R}^m)$ into $L^q(\Omega; \mathbb{R}^k)$ (where $1 \leq p, q < \infty$) if and only if there exists a constant $C > 0$ and a function $g \in L^q(\Omega; \mathbb{R}^k)$ such that:

$$|f(x, y)| \leq g(x) + C|y|^{p/q} \quad \text{for a.e. } x \in \Omega \text{ and all } y \in \mathbb{R}^m. \quad (2.9)$$

We now provide formal definitions for the Nemytskii operators induced by pointwise functions in (2.6).

Definition 2.2.9 (Induced skip connection operator $\mathcal{W}$). Let $W \in \mathbb{R}^{d_v \times d_v}$ be a weight matrix. The *induced skip connection operator* $\mathcal{W} : L^p(D; \mathbb{R}^{d_v}) \rightarrow L^p(D; \mathbb{R}^{d_v})$ with $p \in [1, \infty)$ is the Nemytskii operator defined pointwise by

$$[\mathcal{W}v](x) = Wv(x), \quad \text{for a.e. } x \in D. \quad (2.10)$$

The following proposition verifies that $\mathcal{W}$ is a well-defined Nemytskii operator:

Proposition 2.2.10. *Let $D \subset \mathbb{R}^d$ be a measurable set and let $W : \mathbb{R}^{d_v} \rightarrow \mathbb{R}^{d_v}$ be a linear transformation represented by a finite-dimensional weight matrix. The map $\mathcal{W}$ defined by $[\mathcal{W}v](x) = Wv(x)$ is a well-defined Nemytskii operator on Lebesgue spaces mapping $L^p(D; \mathbb{R}^{d_v})$ to $L^p(D; \mathbb{R}^{d_v})$ for $1 \leq p < \infty$.*

Proof. We define the underlying pointwise function $f_W : D \times \mathbb{R}^{d_v} \rightarrow \mathbb{R}^{d_v}$ as $f_W(x, y) = Wy$. We first verify the Carathéodory conditions. For any fixed $y \in \mathbb{R}^{d_v}$, $f_W(\cdot, y)$ is constant with respect to x , and thus trivially Lebesgue measurable on D . For almost every $x \in D$, the map $f_W(x, \cdot)$ is simply matrix multiplication, which is a continuous operation on $\mathbb{R}^{d_v}$.

Next, we verify the growth condition. Since W is a finite-dimensional linear operator, it is bounded by its operator norm $\|W\|$, induced by the Euclidean norm $|\cdot|$ on $\mathbb{R}^{d_v}$. Hence, for all $y \in \mathbb{R}^{d_v}$ and almost every $x \in D$, we have:

$$|f_W(x, y)| = |Wy| \leq \|W\||y| \quad (2.11)$$

This satisfies the Nemytskii growth condition $|f(x, y)| \leq g(x) + C|y|$ with constant $C = \|W\| > 0$, and $g(x) = 0 \in L^p(D)$. Thus, $\mathcal{W}$ is a well-defined bounded Nemytskii operator on $L^p(D; \mathbb{R}^{d_v})$. Explicitly, for any $v \in L^p(D; \mathbb{R}^{d_v})$

$$\|\mathcal{W}v\|_{L^p}^p = \int_D |Wv(x)|^p dx \leq \|W\|^p \int_D |v(x)|^p dx = \|W\|^p \|v\|_{L^p}^p < \infty \quad (2.12)$$

it follows, that $\|\mathcal{W}v\|_{L^p} \leq \|W\| \|v\|_{L^p}$ and $\mathcal{W} \in \mathcal{L}(L^p(D; \mathbb{R}^{d_v}))$. □

Remark 2.2.11 (Growth constraint on activation function σ). While proving that any finite dimensional matrix W induces a well defined Nemytskii operator is trivial, [Proposition 2.2.13](#) shows that Kovachki et al. (2023) [9] omitted an important requirement on the activation function $\sigma : \mathbb{R}^{d_v} \rightarrow \mathbb{R}^{d_v}$, namely that it must grow at most linearly. While this requirement is satisfied for the most commonly used activation functions, such as ReLU, GeLU, or Tanh, a non-linear continuous function such as $\sigma(x) = e^x$ would fail to induce a valid Nemytskii operator mapping L^p to L^p .

Definition 2.2.12 (Induced activation operator Σ). Let $\sigma : \mathbb{R}^{d_v} \rightarrow \mathbb{R}^{d_v}$ be a continuous non-linear activation function with at most linear growth. The *induced activation operator* $\Sigma : L^p(D; \mathbb{R}^{d_v}) \rightarrow L^p(D; \mathbb{R}^{d_v})$ with $p \in [1, \infty)$ is the Nemytskii operator defined pointwise by

$$[\Sigma v](x) = \sigma(v(x)), \quad \text{for a.e. } x \in D. \quad (2.13)$$

The following proposition verifies that Σ is a well-defined Nemytskii operator given that it's induced by σ with at most linear growth.

Proposition 2.2.13. *Let $\sigma : \mathbb{R}^{d_v} \rightarrow \mathbb{R}^{d_v}$ be a continuous non-linear activation function with at most linear growth. The map Σ defined by $[\Sigma v](x) = \sigma(v(x))$ is a well-defined non-linear bounded Nemytskii operator on $L^p(D; \mathbb{R}^{d_v})$ with $1 \leq p < \infty$.*

Proof. Let the pointwise function be $f_\sigma(x, y) = \sigma(y)$ with $\sigma \in C(\mathbb{R}^{d_v})$. Similar to the linear case, f_σ is independent of the spatial variable $x \in D$, making it trivially measurable for any fixed $y \in \mathbb{R}^{d_v}$. By definition, σ is assumed to be continuous on $\mathbb{R}^{d_v}$, satisfying the second Carathéodory condition ([Definition 2.2.8](#)).

By the assumption of at most linear growth, there exist constants $C > 0$ and $C' \geq 0$ such that $|\sigma(y)| \leq C' + C|y|$ for all $y \in \mathbb{R}^{d_v}$. Because the spatial domain D is bounded,

it has finite Lebesgue measure, ensuring that the constant function $g(x) = C'$ belongs to $L^p(D; \mathbb{R}^{d_v})$. Thus, the inequality becomes:

$$|f_\sigma(x, y)| \leq g(x) + C|y| \quad (2.14)$$

This satisfies the required growth condition $L^p(D) \rightarrow L^p(D)$, confirming that Σ is a well-defined Nemytskii operator.

By raising the pointwise growth bound to the power $p \geq 1$ and due to the elementary inequality $(a + b)^p \leq 2^{p-1}(a^p + b^p)$ for $a, b \geq 0$ and $p \geq 1$, we show that the operator Σ is a bounded operator on $L^p(D; \mathbb{R}^{d_v})$. For any $v \in L^p(D; \mathbb{R}^{d_v})$ with $p \in [1, \infty)$ it holds

$$\|\Sigma v\|_{L^p}^p = \int_D |\sigma(v(x))|^p dx \quad (2.15)$$

$$\leq \int_D (C' + C|v(x)|)^p dx \quad (2.16)$$

$$\leq 2^{p-1}((C')^p |D| + C^p \|v\|_{L^p}^p) < \infty \quad (2.17)$$

Thus Σ is a well defined bounded Nemytskii operator on $L^p(D; \mathbb{R}^{d_v})$, satisfying the affine growth bound. $\square$

2.2.3 The Kernel Integral Operator

As established in [Remark 2.2.4](#), the core component allowing the Neural Operator to process infinite-dimensional continuous inputs is the integral operator. In classical PDEs, integral operators with specifically constructed kernels (such as Green's functions) are used to map forcing functions directly to solution functions.

Kovachki et al. (2023) [9] parameterize this concept for statistical learning. The authors define the Kernel Integral Operator using a neural network parameterized kernel $\kappa_\phi(x, y)$ [9]. While Li et al. (2021) [7] and Kovachki et al. (2023) [9] define the kernel integral operator at the pointwise level, the conditions under which it constitutes a well-defined bounded operator on Lebesgue spaces $L^p(D, \mathbb{R}^{d_v})$ are not explicitly verified.

Definition 2.2.14 (Kernel Integral Operator, [9, sec. 3, eq. 7]). Let $D \subset \mathbb{R}^d$ be a bounded measurable domain. Let $\mathcal{V} := L^p(D; \mathbb{R}^{d_v})$ denote the Lebesgue space with $1 \leq p < \infty$. Given a parameterized, continuous kernel function $\kappa_\phi \in C(\bar{D} \times \bar{D}; \mathbb{R}^{d_v \times d_v})$, the *Kernel Integral Operator* $\mathcal{K}_\phi : \mathcal{V} \rightarrow \mathcal{V}$ is defined as:

$$[\mathcal{K}_\phi v](x) = \int_D \kappa_\phi(x, y) v(y) d\mu(y), \quad \text{for a.e. } x \in D \quad (2.18)$$

where μ is the Lebesgue measure.

Lemma 2.2.15 (Boundedness). *Let $D \subset \mathbb{R}^d$ be a bounded domain, and let $\mathcal{K}$ be a Kernel Integral Operator with a continuous kernel $\kappa \in C(\bar{D} \times \bar{D})$. For any $p \in [1, \infty]$, the operator $\mathcal{K}$ defines a continuous map from $L^p(D)$ to $L^\infty(D)$. Additionally, $\mathcal{K}$ is a bounded linear operator on $L^p(D)$.*

Proof. Let $f \in L^p(D)$. The operator $\mathcal{K}$ is defined as:

$$(\mathcal{K}f)(x) := \int_D \kappa(x, y) f(y) dy, \quad \forall f \in L^p(D) \quad (2.19)$$

Because $D \subset \mathbb{R}^d$ is bounded, its closure $\bar{D}$ is compact by the Heine-Borel theorem. Hence, the product space $\bar{D} \times \bar{D}$ is also compact and given that κ is continuous, the image $\kappa(\bar{D} \times \bar{D})$ is compact [18, Prop. 1.8.20] and therefore bounded. That implies that

there exists $M < \infty$ such that $\|\kappa\|_\infty \leq M$ for κ restricted to $\bar{D} \times \bar{D}$. Furthermore, by Heine-Cantor theorem κ is uniformly continuous on $\bar{D} \times \bar{D}$ [18, Prop. 1.8.20].

By applying the triangle inequality for integrals, we find the bound for the pointwise absolute value:

$$|(\mathcal{K}f)(x)| \leq \int_D |\kappa(x, y)| |f(y)| dy \quad (2.20)$$

$$\leq \|\kappa\|_\infty \int_D |f(y)| dy \quad (2.21)$$

$$= \|\kappa\|_\infty \|f\|_{L^1} < \infty \quad (2.22)$$

Let q be the Hölder conjugate of p , where $\frac{1}{p} + \frac{1}{q} = 1$, with $q = \infty$ if $p = 1$. Applying Hölder's inequality to $\|f\|_{L^1}$:

$$\|f\|_{L^1} = \int_D |f(y)| \cdot 1 dy \leq \|f\|_{L^p} \|1\|_{L^q} \quad (2.23)$$

Note that $\|1\|_{L^q} = |D|^{1/q}$ for $q < \infty$, and $\|1\|_{L^\infty} = 1$. Thus, we can define a finite constant $C_q = \|1\|_{L^q}$ with $C_q < \infty$ because the bounded domain $D \subset \mathbb{R}^d$ has a finite Lebesgue measure. Substituting this bound into (2.22), we obtain a uniform bound on $\mathcal{K}f$ independent of x :

$$\sup_{x \in D} |(\mathcal{K}f)(x)| \leq \|\kappa\|_\infty \cdot C_q \cdot \|f\|_{L^p} \quad (2.24)$$

To show continuity of $\mathcal{K}f$, fix $\varepsilon > 0$. For $f = 0$ the map $\mathcal{K}0 = 0$ is continuous. Assume $\|f\|_{L^1} > 0$ and set $\varepsilon' = \varepsilon / \|f\|_{L^1}$. By uniform continuity of κ , there exists $\delta > 0$ such that for all $x, z \in D$ with $\|x - z\| < \delta$ it holds $\sup_{y \in D} (|\kappa(x, y) - \kappa(z, y)|) < \varepsilon'$. By linearity of the integral:

$$|(\mathcal{K}f)(x) - (\mathcal{K}f)(z)| \leq \int_D |\kappa(x, y) - \kappa(z, y)| |f(y)| dy \quad (2.25)$$

$$\leq \sup_{y \in D} (|\kappa(x, y) - \kappa(z, y)|) \int_D |f(y)| dy \quad (2.26)$$

$$\leq \varepsilon' \cdot \|f\|_{L^1} = \varepsilon \quad (2.27)$$

Hence the essential supremum $\|\mathcal{K}f\|_{L^\infty}$ coincides with supremum of $\mathcal{K}f$ due to continuity of $\mathcal{K}f$. This implies

$$\|\mathcal{K}f\|_{L^\infty} = \sup_{x \in D} |(\mathcal{K}f)(x)| \leq MC_q \|f\|_{L^p} \quad (2.28)$$

It follows that $\mathcal{K}f \in L^\infty(D)$ and $\mathcal{K} : L^p(D) \rightarrow L^\infty(D)$ is bounded.

The linearity of $\mathcal{K}$ follows directly from the linearity of the Lebesgue integral, as the integrals are finite for all $x \in D$.

To prove boundedness of $\mathcal{K}$ on $L^p(D)$, observe that since D has finite measure, $L^\infty(D) \subset L^p(D)$. We establish an explicit bound as follows:

$$\|\mathcal{K}f\|_{L^p} = \left(\int_D |(\mathcal{K}f)(x)|^p dx \right)^{1/p} \quad (2.29)$$

$$\leq \left(\int_D \|\mathcal{K}f\|_{L^\infty}^p dx \right)^{1/p} \quad (2.30)$$

$$= |D|^{1/p} \|\mathcal{K}f\|_{L^\infty} \quad (2.31)$$

Combining the inequalities (2.24) and (2.31):

$$\|\mathcal{K}f\|_{L^p} \leq |D|^{1/p} \cdot M \cdot C_q \cdot \|f\|_{L^p} < \infty \quad (2.32)$$

For $p = \infty$ the claim follows immediately from (2.24). Thus, $\mathcal{K}$ is a well-defined and bounded linear operator from $L^p(D)$ to $L^p(D)$. $\square$

Lemma 2.2.16 (Equicontinuity). *Let $D \subset \mathbb{R}^d$ be a bounded domain, $\mathcal{K}$ be a kernel integral operator with a continuous kernel $\kappa \in C(\bar{D} \times \bar{D})$, and $B \subset L^p(D)$ be a bounded set for $p \in [1, \infty)$. Then the image $\mathcal{K}(B)$ is uniformly equicontinuous on D .*

Proof. To prove that $\mathcal{K}(B)$ is uniformly equicontinuous, we must show that $\forall \varepsilon > 0, \exists \delta > 0 : \forall f \in B$ and $\forall x, z \in D$ with $|x - z| < \delta$ it follows that $|(\mathcal{K}f)(x) - (\mathcal{K}f)(z)| < \varepsilon$.

Because $D \subset \mathbb{R}^d$ is bounded, its closure $\bar{D}$ is compact by the Heine-Borel theorem and hence the product space $\bar{D} \times \bar{D}$ is also compact. Since $\kappa \in C(\bar{D} \times \bar{D})$ is a continuous function on a compact domain, it is uniformly continuous by the Heine-Cantor theorem [18, Prop. 1.8.20].

The uniform continuity of κ implies that for any $\varepsilon' > 0$, there exists $\delta > 0$ such that for all $(x, y), (z, y) \in \bar{D} \times \bar{D}$:

$$|x - z| < \delta \implies |\kappa(x, y) - \kappa(z, y)| < \varepsilon' \quad (2.33)$$

Since B is a bounded set in $L^p(D)$, there exists a constant $C < \infty$ such that

$$\sup_{f \in B} \|f\|_{L^p} \leq C$$

Let q be the Hölder conjugate to p with $\frac{1}{p} + \frac{1}{q} = 1$ and $q = \infty$ if $p = 1$. Define $C_q = \|1\|_{L^q(D)}$. Note, that $C_q = |D|^{1/q}$ for $q < \infty$ and $C_\infty = 1$.

Fix $\varepsilon > 0$ and choose ε' such that $\varepsilon' = \frac{\varepsilon}{C_q C}$. From uniform continuity of κ it follows that we can find $\delta > 0$ such that (2.33) holds. For any $f \in B$ and any $x, z \in D$ with $|x - z| < \delta$, we establish the bound:

$$|(\mathcal{K}f)(x) - (\mathcal{K}f)(z)| = \left| \int_D (\kappa(x, y) - \kappa(z, y)) f(y) dy \right| \quad (2.34)$$

$$\leq \int_D |\kappa(x, y) - \kappa(z, y)| |f(y)| dy \quad (2.35)$$

By applying the Hölder's inequality with $p \in (1, \infty)$:

$$|(\mathcal{K}f)(x) - (\mathcal{K}f)(z)| \leq \|\kappa(x, \cdot) - \kappa(z, \cdot)\|_{L^q} \|f\|_{L^p} \quad (2.36)$$

$$\leq \left(\int_D (\varepsilon')^q dy \right)^{1/q} \|f\|_{L^p} \quad (2.37)$$

$$< \varepsilon' \cdot C_q \cdot C \quad (2.38)$$

$$= \left(\frac{\varepsilon}{C_q C} \right) \cdot C_q \cdot C = \varepsilon \quad (2.39)$$

For the case $p = 1$ the (2.38) becomes

$$|(\mathcal{K}f)(x) - (\mathcal{K}f)(z)| \leq \|\kappa(x, \cdot) - \kappa(z, \cdot)\|_{L^\infty} \|f\|_{L^1} \quad (2.40)$$

$$< \varepsilon' \cdot \|f\|_{L^1} \quad (2.41)$$

$$= \left(\frac{\varepsilon}{C} \right) \cdot C = \varepsilon \quad (2.42)$$

Thus, $\mathcal{K}(B)$ is uniformly equicontinuous. $\square$

Proposition 2.2.17 (Compactness). *Let $D \subset \mathbb{R}^d$ be a bounded domain, and let $\mathcal{K}$ be a Kernel Integral Operator with a continuous kernel $\kappa \in C(\bar{D} \times \bar{D})$. Then $\mathcal{K}$ is a compact operator on $L^p(D)$ for any $p \in [1, \infty)$.*

Proof. Let $B \subset L^p(D)$ be a bounded set. Because $D \subset \mathbb{R}^d$ is bounded, its closure $\bar{D}$ is compact by the Heine-Borel Theorem.

By Lemma 2.2.15, $\mathcal{K}$ is a bounded linear operator on $L^p(D)$, meaning $\mathcal{K}(B) \subset C(\bar{D})$ and $\mathcal{K}(B)$ is uniformly bounded. Furthermore, the set $\mathcal{K}(B)$ is uniformly equicontinuous by Lemma 2.2.16. By the Arzelà-Ascoli theorem [18, Thm. 5.4.2, p. 196], $\mathcal{K}(B)$ is precompact in $C(\bar{D})$ with respect to the supremum norm.

Because D is bounded, the space $C(\bar{D})$ embeds continuously into $L^p(D)$. Continuous mappings preserve precompactness, meaning the image of $\mathcal{K}(B)$ under this embedding remains precompact in $L^p(D)$.

Hence, $\mathcal{K}$ maps bounded sets in $L^p(D)$ to precompact sets in $L^p(D)$, and is therefore compact according to [18, Def. 7.5.1]. $\square$

2.2.4 Single Neural Operator Layer

With the globally acting Kernel Integral Operator formally defined (Definition 2.2.14), we can now assemble the complete mapping for a single hidden layer of the Neural Operator.

Unlike standard neural networks that learn maps between Euclidean vectors (e.g., $\mathbb{R}^{64} \rightarrow \mathbb{R}^{64}$), a Neural Operator layer learns a continuous mapping between Banach spaces of functions. Kovachki et al. [9] define the update from the t -th hidden representation $v_t \in \mathcal{V}$ to the next representation $v_{t+1} \in \mathcal{V}$ with $\mathcal{V} := L^p(D; \mathbb{R}^{d_v})$, $p \in [1, \infty)$ as follows:

Definition 2.2.18 (Neural Operator Layer, [9, sec. 3, eq. 6]). Let $\mathcal{V} := L^p(D; \mathbb{R}^{d_v})$ with $p \in [1, \infty)$. Given a kernel $\kappa_\phi \in C(\bar{D} \times \bar{D}; \mathbb{R}^{d_v \times d_v})$, a weight matrix $W \in \mathbb{R}^{d_v \times d_v}$, and a continuous activation function $\sigma : \mathbb{R}^{d_v} \rightarrow \mathbb{R}^{d_v}$ with at most linear growth, the *Neural Operator layer* defines the update from the t -th hidden representation $v_t \in \mathcal{V}$ to the next representation $v_{t+1} \in \mathcal{V}$ by

$$v_{t+1}(x) = \sigma \left(W v_t(x) + \int_D \kappa_\phi(x, y) v_t(y) d\mu(y) \right), \quad \text{for a.e. } x \in D \quad (2.43)$$

where μ is the Lebesgue measure.

Notice that the argument of the activation function in (2.43) is a pointwise summation of two distinct operators: the global integration operator $\mathcal{K}_\phi$, and a pointwise linear map $W : \mathbb{R}^{d_v} \rightarrow \mathbb{R}^{d_v}$ acting strictly locally as a skip connection.

Remark 2.2.19 (Neural operator layer definition in operator notation). To treat equation (2.43) rigorously, we translate the pointwise algebraic formulation into a composition of the formal operators on $\mathcal{V}$ established in previous sections:

- $\mathcal{K}_\phi$ is the Kernel Integral Operator (Definition 2.2.14).
- $\mathcal{W}$ is the continuous Nemytskii operator induced by the finite-dimensional weight matrix W (Definition 2.2.9).
- Σ is the continuous Nemytskii operator induced by the activation function σ (Definition 2.2.12).

The Neural Operator Layer from (2.43) can then be written concisely as a non-linear operator $\mathcal{H} : \mathcal{V} \rightarrow \mathcal{V}$:

$$\mathcal{H}(v_t) = \Sigma \left(\mathcal{W} v_t + \mathcal{K}_\phi v_t \right) \quad (2.44)$$

We formally establish the well-posedness of this layer.

Proposition 2.2.20 (Boundedness and Continuity). *Let $D \subset \mathbb{R}^d$ be a bounded domain. Assume the kernel κ_ϕ is continuous on $\bar{D} \times \bar{D}$, the weight matrix W is finite-dimensional, and the activation $\sigma : \mathbb{R}^{d_v} \rightarrow \mathbb{R}^{d_v}$ is continuous with at most linear growth. Then the Neural Operator layer $\mathcal{H} := \Sigma \circ (\mathcal{W} + \mathcal{K}_\phi)$ is a bounded and continuous operator on Lebesgue space $L^p(D; \mathbb{R}^{d_v})$ with $p \in [1, \infty)$.*

Proof. Because $\mathcal{W}$ and $\mathcal{K}_\phi$ are both bounded linear operators on $L^p(D; \mathbb{R}^{d_v})$ with $p \in [1, \infty)$, as established in Proposition 2.2.10 and Lemma 2.2.15, their sum $\mathcal{L} := \mathcal{W} + \mathcal{K}_\phi$ is a bounded linear operator, as follows from the triangle inequality. By Proposition 2.2.13, the activation function σ induces a continuous and bounded non-linear Nemytskii operator $\Sigma : L^p \rightarrow L^p$. As a composition of continuous operators, $\mathcal{H}$ is continuous. From the boundedness of linear operator $\mathcal{L}$ it follows $\forall v \in L^p(D; \mathbb{R}^{d_v})$:

$$\|\mathcal{L}v\|_{L^p} \leq \|\mathcal{L}\|_{L^p} \|v\|_{L^p} \quad (2.45)$$

From Proposition 2.2.13 due to the affine bound of Σ it holds for all $v \in L^p(D; \mathbb{R}^{d_v})$:

$$\|\mathcal{H}v\|_{L^p} = \|\Sigma(\mathcal{L}v)\|_{L^p} \leq C(1 + \|\mathcal{L}v\|_{L^p}) \leq C(1 + \|\mathcal{L}\|_{L^p} \|v\|_{L^p}) < \infty \quad (2.46)$$

Hence, the full layer mapping $\mathcal{H} = \Sigma \circ \mathcal{L}$ is a well-defined bounded and continuous operator on $L^p(D; \mathbb{R}^{d_v})$, $p \in [1, \infty)$. Due to non-linearity of the induced operator Σ the operator $\mathcal{H}$ is non-linear. $\square$

2.2.5 Lifting and Projection

As motivated in Remark 2.2.7, the lifting and projection operators serve as the entry and exit points of the Neural Operator architecture, mapping between the input and output function spaces $\mathcal{A}$ and $\mathcal{U}$ and the higher-dimensional latent space $\mathcal{V}$ in which the Kernel Integral layers operate. While Kovachki et al. (2023) [9] treat these operators primarily as a mean of changing dimensions, Lanthaler et al. (2025) [10] reveal that they carry a deeper significance for the Universal Approximation guarantees. We present the formal definitions following Kovachki et al., and then discuss the contrasting perspectives.

Definition 2.2.21 (Lifting Operator, [9, sec. 3]). Let $\mathcal{A} := L^p(D; \mathbb{R}^{d_a})$ and $\mathcal{V} := L^p(D; \mathbb{R}^{d_v})$ with $p \in [1, \infty)$ and $d_v \geq d_a$. Given a pointwise function $P : \mathbb{R}^{d_a} \rightarrow \mathbb{R}^{d_v}$, the *lifting operator* $\mathcal{P} : \mathcal{A} \rightarrow \mathcal{V}$ is the Nemytskii operator defined by

$$[\mathcal{P}(a)](x) = P(a(x)), \quad \text{for a.e. } x \in D, a \in \mathcal{A}. \quad (2.47)$$

Definition 2.2.22 (Projection Operator, [9, sec. 3]). Let $\mathcal{V} := L^p(D; \mathbb{R}^{d_v})$ and $\mathcal{U} := L^p(D; \mathbb{R}^{d_u})$ with $d_v \geq d_u$. Given a pointwise function $Q : \mathbb{R}^{d_v} \rightarrow \mathbb{R}^{d_u}$, the *projection operator* $\mathcal{Q} : \mathcal{V} \rightarrow \mathcal{U}$ is the Nemytskii operator defined by

$$[\mathcal{Q}(v)](x) = Q(v(x)), \quad \text{for a.e. } x \in D, v \in \mathcal{V}. \quad (2.48)$$

As established in subsection 2.2.2, because P and Q are finite-dimensional linear maps, the induced operators $\mathcal{P}$ and $\mathcal{Q}$ are bounded and continuous. The proofs follow analogous to Proposition 2.2.10.

Remark 2.2.23 (Contrasting roles of lifting and projection). In Kovachki et al. (2023) [9], the lifting operator $\mathcal{P}$ embeds the d_a -dimensional input into a d_v -dimensional latent space, expanding the integration kernel to a $d_v \times d_v$ matrix field and allowing the network to apply multiple distinct continuous transformations simultaneously. The projection operator $\mathcal{Q}$ reverses this expansion after the final layer. However, in the proof of the Universal

Approximation Theorem [9, Thm. 11], the authors set $d_v = d_a = d_u$, effectively replacing $\mathcal{P}$ and $\mathcal{Q}$ by the identity. The entire Approximation-theoretic argumentation is based on the Kernel Integral layers $\mathcal{K}_\phi$ (Definition 2.2.14).

Lanthaler et al. (2025) [10] assign a fundamentally different role to these operators. In their Averaging Neural Operator (ANO), the Kernel Integral Operator (Definition 2.2.14) is simplified by reducing the kernel to the identity $\kappa(x, y) = I$, so that $\mathcal{K}$ acts as pure spatial averaging, $[\mathcal{K}v](x) = \int_D v(y) d\mu(y)$. In this configuration, the Kernel Integral layer is only responsible for adding *non-locality* to the architecture. All other work is redistributed to the lifting and projection operators. The lifting layer $\mathcal{R} : u \mapsto R(u(x), x)$ incorporates explicit positional encoding by depending on the input $x \in D$, applying a *nonlinear* transformation to the input function in contrast to the lifting operator $\mathcal{P}$ in Kovachki et al. definition. The subsequent averaging layer then integrates out this spatial dependence, resulting in a position-independent encoding vector $v \in \mathbb{R}^{d_c}$, acting as a finite-dimensional representation of the input function a . The composition $\mathcal{L} \circ \mathcal{R}$ acts as a *nonlinear encoder*, since it maps the entire input function to a constant vector, but, importantly, relies on the positional encoding in $\mathcal{R}$ to prevent translation equivariance, because otherwise the non-local averaging operator $\mathcal{K}$ would absorb the positional information in $x \in D$. The decoder $\mathcal{Q} : v \mapsto Q(v, x)$ then reconstructs the output function from this constant encoding vector with explicit spatial dependence.

This decomposition results in a clean theoretical separation: *nonlinearity* is provided by the encoder, *non-locality* by the averaging layer, and spatial reconstruction by the decoder. Each element contributes exactly one necessary component for the Universal Approximation, as discussed in section 4.4. In contrast, Kovachki et al.’s formulation describes the architecture closer to how it is implemented in practice, where the Kernel Integral layers are responsible for providing all three above mentioned components. However, the two perspectives are complementary rather than contradicting each other, and we examine the consequences for the respective Universal Approximation Theorems in Chapter 4.

2.2.6 Properties of the Neural Operator

Having formally established the functional-analytic properties of the lifting map, the Kernel Integration Operator, and the local pointwise transformations, we can now formally define the complete Neural Operator. The multi-layer Neural Operator is constructed as a direct composition of these layer-wise mappings.

Definition 2.2.24 (The Neural Operator [9, sec. 3]). Let $D \subset \mathbb{R}^d$ be a bounded domain. Let the input space $\mathcal{A} = L^p(D; \mathbb{R}^{d_a})$, the latent feature space $\mathcal{V} = L^p(D; \mathbb{R}^{d_v})$, and the output target space $\mathcal{U} = L^p(D; \mathbb{R}^{d_u})$ be Lebesgue spaces with $1 \leq p < \infty$. Given $T \in \mathbb{N}$ hidden layers, the *Neural Operator* $\mathcal{G} : \mathcal{A} \rightarrow \mathcal{U}$ is defined as the composition:

$$\mathcal{G} = \mathcal{Q} \circ \mathcal{H}_T \circ \mathcal{H}_{T-1} \circ \cdots \circ \mathcal{H}_1 \circ \mathcal{P} \quad (2.49)$$

where $\mathcal{P} : \mathcal{A} \rightarrow \mathcal{V}$ is the lifting Nemytskii operator (Definition 2.2.21), $\mathcal{H}_t : \mathcal{V} \rightarrow \mathcal{V}$ is the t -th Kernel Integral based hidden layer $\mathcal{H}_t := \Sigma_t \circ (\mathcal{W}_t + \mathcal{K}_{\phi,t})$ (Definition 2.2.18), $\mathcal{Q} : \mathcal{V} \rightarrow \mathcal{U}$ is the projection Nemytskii operator (Definition 2.2.22).

With the full operator defined, we state its core properties as the proofs are omitted in the original literature.

Proposition 2.2.25 (Properties of the Neural Operator). *Let $D \subset \mathbb{R}^d$ be a bounded domain. Let the input space $\mathcal{A} = L^p(D; \mathbb{R}^{d_a})$, the latent feature space $\mathcal{V} = L^p(D; \mathbb{R}^{d_v})$, and the output target space $\mathcal{U} = L^p(D; \mathbb{R}^{d_u})$ be Lebesgue spaces with $1 \leq p < \infty$.*

Assume the kernels $\kappa_{\phi,t} \in C(\bar{D} \times \bar{D}; \mathbb{R}^{d_v \times d_v})$ are continuous, the weight matrices P, Q , and W_t are finite-dimensional, and the activations σ_t are continuous with at most linear

growth. Then the Neural Operator $\mathcal{G} : \mathcal{A} \rightarrow \mathcal{U}$ is a well-defined, bounded, and continuous non-linear operator. Furthermore, in general, $\mathcal{G}$ is not a compact operator.

Proof. The proof of boundedness and continuity follows from the properties of function composition. In previous sections, we established that $\mathcal{P}$ and $\mathcal{Q}$ are bounded continuous Nemytskii operators (Proposition 2.2.10), and that each hidden layer $\mathcal{H}_t$ is a bounded continuous operator on $\mathcal{V}$ (Proposition 2.2.20). Because the composition of bounded and continuous operators between Banach spaces remains bounded and continuous, the full mapping $\mathcal{G}$ is bounded and continuous from $\mathcal{A}$ to $\mathcal{U}$.

To assess compactness, we examine the hidden layer $\mathcal{H}_t = \Sigma_t \circ (\mathcal{W}_t + \mathcal{K}_{\phi,t})$. While the integration operator with continuous kernel $\mathcal{K}_{\phi,t}$ is compact in $\mathcal{V} = L^p(D; \mathbb{R}^{d_v})$, $p \in [1, \infty)$ (Proposition 2.2.17), the pointwise linear operator $\mathcal{W}_t$ acts as the identity over the infinite-dimensional spatial domain and is strictly non-compact. The sum of a compact and a non-compact operator is non-compact. Because the activation Σ_t is continuous and non-compact, the entire hidden layer $\mathcal{H}_t$ is non-compact. Consequently, the composed operator $\mathcal{G}$ does not map bounded sets in $\mathcal{A}$ to precompact sets in $\mathcal{U}$, implying that the full Neural Operator is not compact. $\square$

Remark 2.2.26. In their proofs of the Discretization Invariance [9, Thm. 8] and the Universal Approximation Theorems, Kovachki et al. (2023) [9, Thm. 11] work with a simplified architecture in which the skip connections $\mathcal{W}_t$ are omitted and the lifting and projection dimensions are set to $d_v = d_a = d_u$. We provide the formal definition of the simplified operator as we will refer to it subsequently in Chapter 3 and Chapter 4.

Definition 2.2.27 (Neural Operator (theoretical form), [9, Thm. 11]). Let $D \subset \mathbb{R}^d$ be a bounded domain. Let the input space $\mathcal{A} = L^p(D; \mathbb{R}^{d_a})$ and the output target space $\mathcal{U} = L^p(D; \mathbb{R}^{d_u})$ be Lebesgue spaces with $1 \leq p < \infty$.

The *theoretical Neural Operator* $\mathcal{G} : \mathcal{A} \rightarrow \mathcal{U}$ is defined as

$$\mathcal{G} = \Sigma_T \circ \mathcal{K}_{\phi,T} \circ \cdots \circ \Sigma_1 \circ \mathcal{K}_{\phi,1} \quad (2.50)$$

for $T \in \mathbb{N}$, where each layer consists only of a Kernel Integral Operator $\mathcal{K}_{\phi,t}$ with continuous kernel $\kappa_{\phi,t} \in C(\bar{D} \times \bar{D}; \mathbb{R}^{d_v \times d_v})$, followed by a Nemytskii operator Σ_t induced by a non-linear activation function with at most linear growth σ_t .

Remark 2.2.28. As a composition of compact Kernel Integral Operators (Proposition 2.2.17) and bounded continuous Nemytskii operators (Proposition 2.2.13), the theoretical Neural Operator from Definition 2.2.27 is a compact nonlinear operator on $L^p(D; \mathbb{R}^{d_u})$ with $p \in [1, \infty)$ and $D \subset \mathbb{R}^{d_a}$ [18, Prop. 7.5.7, p. 269].

Remark 2.2.29 (Definitions on $C(\bar{D})$ with sup-norm and $W^{k,p}(D)$ spaces). The definitions of the Neural Operator in its theoretical form on $C(\bar{D})$ with sup-norm and $W^{k,p}(D)$ spaces, as well as the well-definiteness and boundedness proofs for its components, namely the induced Nemytskii Activation Operator Σ and the Kernel Integral Operator $\mathcal{K}_\phi$, are provided in Appendix A.

2.2.7 Neural Operator Instantiations

The Neural Operator framework (Definition 2.2.24) is deliberately general, so that the kernel κ_ϕ in the integral operator (Definition 2.2.14) can be parameterized in different ways, resulting in architectures with distinct computational and theoretical trade-offs. We present three instantiations that form a natural hierarchy and that motivate the theoretical developments on Discretization Invariance and Universal Approximation guarantees that we further discuss in Chapter 3 and Chapter 4.

Definition 2.2.30 (Graph Neural Operator (GNO), [16, eq. 8]). The *Graph Neural Operator* parameterizes the kernel κ_ϕ directly as a neural network $\kappa_\phi : \mathbb{R}^{2(d+d_a)} \rightarrow \mathbb{R}^{d_v \times d_v}$ that takes as input both spatial coordinates and function values:

$$[\mathcal{K}_\phi v](x) = \int_{B(x,r)} \kappa_\phi(x, y, a(x), a(y)) v(y) d\mu(y), \quad \forall x \in D \quad (2.51)$$

where integration is truncated to a ball $B(x, r)$ of radius $r > 0$. The integral is approximated via Monte Carlo summation on a graph whose nodes are the discretization points and whose edges connect points within radius r [16, sec. 3].

The GNO is the most general instantiation as its kernel is not constrained and can depend on the input function a , which comes at computational cost $\mathcal{O}(J^2)$ in the number of discretization points J [9, sec. 4.5]. Historically it's the first kernel integral based neural operator architecture [16] and it directly reflects the Green's function motivation discussed in Remark 2.2.4.

Definition 2.2.31 (Fourier Neural Operator (FNO), [9, eq. 26]; [7, def. 3]). The *Fourier Neural Operator* restricts the kernel to be translation-invariant, $\kappa_\phi(x, y) = \kappa_\phi(x - y)$, and parameterizes it directly in Fourier space. The kernel integral layer becomes:

$$[\mathcal{K}_\phi v](x) = \mathcal{F}^{-1} \left(R_\phi \cdot (\mathcal{F}v) \right)(x), \quad \forall x \in D \quad (2.52)$$

where $\mathcal{F}$ denotes the Fourier transform, $R_\phi \in \mathbb{C}^{k_{\max} \times d_v \times d_v}$ is a learnable complex-valued tensor of truncated Fourier modes, and $k_{\max}$ controls the frequency cutoff [9, sec. 4.4].

Because the kernel is translation-invariant, the integral reduces to a convolution, which can be computed efficiently in Fourier space at $\mathcal{O}(J \log J)$ via the Fast Fourier Transform. This restriction is natural from the perspective of fundamental solutions to PDEs with constant coefficients, where the Green's function depends only on $x - y$ [7, sec. 4]. The FNO is among the most widely adopted Neural Operator architectures in practice [14].

Definition 2.2.32 (Averaging Neural Operator (ANO), [10, sec. 2.2]). The *Averaging Neural Operator* sets the kernel to the identity $\kappa(x, y) = |\Omega|^{-1} I_{d_c \times d_c}$, reducing the integral operator to spatial averaging:

$$[\mathcal{K}v](x) = \frac{1}{|\Omega|} \int_{\Omega} v(y) dy \in \mathbb{R}^{d_c} \quad (2.53)$$

The ANO architecture consists of a single hidden layer $\Psi = \mathcal{Q} \circ \mathcal{L} \circ \mathcal{R}$, where $\mathcal{R} : u \mapsto R(u(x), x)$ and $\mathcal{Q} : v \mapsto Q(v, x)$ are single-hidden-layer neural networks with width d_c , acting as nonlinear encoder and decoder respectively [10, sec. 2.2].

The ANO is the minimal architecture for which Universal Approximation can be proved. By reducing the Kernel Integral layer to simple averaging, Lanthaler et al. (2025) [10] isolate the theoretical role of each component: the encoder $\mathcal{L} \circ \mathcal{R}$ provides nonlinearity, the averaging provides non-locality, and the decoder $\mathcal{Q}$ reconstructs spatial dependence (Remark 2.2.23). Since the ANO is a special case of the general Kernel Integral framework, its Universal Approximation property extends directly to the GNO and FNO as corollaries [10, Cor. 1][10, Cor. 3]. We analyze and compare the Universal Approximation theorems for the Kovachki and Lanthaler frameworks in section 4.4.

Chapter 3

Discretization Invariance

Discretization Invariance for the abstract Kernel Integral (Definition 2.2.14) based Neural Operators is one of the key contributions of Kovachki et al. (2023)[9]. It provides the mechanism by which Neural Operators, defined as maps between infinite-dimensional function spaces, can be trained and evaluated using finite pointwise observations. The core idea is to introduce non-locality in the neural network by formulating its hidden layers as integral operators with learnable kernels. Because integration is defined independently of any particular grid, the learned parameters remain valid across different discretization resolutions. In practice, the integral is approximated via quadrature over the observation points, and the Discretization Invariance property guarantees that this approximation converges as the grid is refined.

We begin this section by establishing the motivation behind the Discretization Invariance property of Neural Operators in section 3.1 and presenting core definitions in section 3.2. In section 3.3 we analyze the non-constructive Discretization Invariance theorem for general Neural Operator architectures [9, Thm. 8] and propose an alternative proof approach for operators on Sobolev spaces in subsection 3.3.2, which allows us to establish explicit convergence rates for general kernel integral based Neural Operators in section 3.4.

3.1 Discretization Invariance: Motivation

A central challenge in learning operators is reconciling the infinite-dimensional nature of the function spaces $\mathcal{A}$ and $\mathcal{U}$ with the finite nature of computational resources. The Neural Operator overcomes this by operating on *discretizations* of the input functions while learning a representation that is independent of the discretization resolution.

Let $D \subset \mathbb{R}^d$ be an open and bounded domain. We assume the input functions $a \in \mathcal{A}$ are observed on a finite grid $D_L = \{x_1, \dots, x_L\} \subset D$ of size L . Consequently, the computational input is a vector of evaluations $a|_{D_L} \in \mathbb{R}^{L \times d_{in}}$. The goal is to construct a parametric operator $\hat{\mathcal{G}}_\theta$ that approximates the true operator $\mathcal{G}^\dagger: \mathcal{A} \rightarrow \mathcal{U}$ (Definition 2.1.1) regardless of the grid size L used during training or inference.

The resolution independence of the learned operator is achieved through the kernel integration layer (Definition 2.2.14). In the continuous limit, the layer computes an integral operator. In practice, this integral is approximated via a quadrature rule (e.g., a Riemann sum) over the grid points. For an input function a and a learned kernel κ_ϕ , the approximation takes the form

$$\int_D \kappa_\phi(x, y) a(y) dy \approx \frac{|D|}{L} \sum_{j=1}^L \kappa_\phi(x, x_j) a(x_j) \quad (3.1)$$

Critically, the kernel parameters ϕ define the continuous function $\kappa_\phi(\cdot, \cdot)$ and are shared across all discretizations. This formulation has two implications for the resolution L :

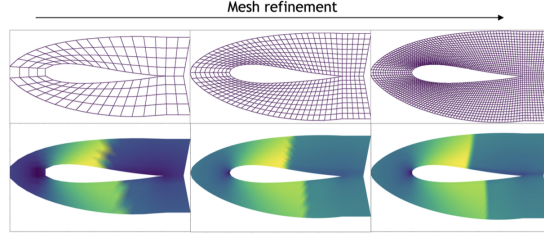


Figure 3.1: Mesh refinement. *Source: Kovachki et al. (2023) [9, fig. 1]*

1. *Resolution Independence*: Parameters ϕ learned on a coarse training grid (e.g., $L = 64$) define a valid continuous operator. This same operator can be evaluated on a finer grid (e.g., $L = 256$) during inference without retraining.
2. *Approximation Quality*: The grid size L controls the quadrature error of the integral approximation. As $L \rightarrow \infty$, the discrete sum converges to the exact integral, improving the fidelity of the model's prediction relative to the learned continuous operator, as depicted in Figure 3.1.

Remark 3.1.1 (Relation to prior work). The convergence of the discrete approximation to the continuous operator is established by Kovachki et al. [9, Thm. 8], whose proof relies on a precompactness assumption and does not offer explicit convergence rates. Architecture-specific rates have been obtained for particular instantiations: Lanthaler et al. (2024) [13] establishes $\mathcal{O}(N^{-s})$ for the aliasing error of Fourier Neural Operators, while Li et al. [16] derive $\mathcal{O}(m^{-1/2})$ via Monte Carlo quadrature for Graph Kernel Networks. However, to the best of our knowledge, no explicit rate has been established at the level of the abstract Kernel Integral Operator. In the subsequent sections, we address this gap by providing an alternative proof under Sobolev regularity assumptions that establishes explicit convergence rates.

3.2 Discretization Invariance: Definitions

Definition 3.2.1 (Discrete Refinement; [9, Def. 1]). Given the domain $D \subset \mathbb{R}^d$, define a sequence $D_1 \subset \dots \subset D_L \subset D$ with $|D_L| = L$ and $L \in \mathbb{N}$ as a *discrete refinement* of D such that $\forall \varepsilon > 0 \exists L = L(\varepsilon) \in \mathbb{N}$ such that

$$D \subseteq \bigcup_{x \in D_L} \{y \in \mathbb{R}^d : \|y - x\|_2 < \varepsilon\} \quad (3.2)$$

Given a discrete refinement $(D_L)_{L=1}^\infty$, any member D_L is called a *discretization* of D .

Remark 3.2.2 (Discretization as ε -net). Definition 3.2.1 implies that for any $\varepsilon > 0$ there exists a finite ε -net of centroids $D_L := \{x_i\}_{i=1}^L$ such that D can be covered by finitely many open balls $B_\varepsilon(x_i)$, $i \in \{1, \dots, L\}$. From the definition it follows that D is *totally bounded*. In particular, $S = \bigcup_{L=1}^\infty D_L$ is dense in D .

Given the parameter function $a \in \mathcal{A}$, $a: D \subset \mathbb{R}^d \rightarrow \mathbb{R}^m$, pointwise evaluation on the discretization D_L results in the data set $\{(x_l, a(x_l))\}_{l=1}^L \subset \mathbb{R}^{Ld} \times \mathbb{R}^{Lm}$ [9, sec. 2.3].

Definition 3.2.3 (Discretization Invariant Operator; cf. [9, Def. 4]). Let $\Theta \subseteq \mathbb{R}^p$ be a finite-dimensional parameter space and $\mathcal{G}: \mathcal{A} \times \Theta \rightarrow \mathcal{U}$ a map representing a parametric class of operators with parameters $\theta \in \Theta$. Given a discrete refinement $(D_n)_{n=1}^\infty$ of the domain $D \subset \mathbb{R}^d$ and a sequence of operators $\{\hat{\mathcal{G}}_L\}_{L=1}^\infty$ with $\hat{\mathcal{G}}_L: \mathbb{R}^{Ld} \times \mathbb{R}^{Lm} \times \Theta \rightarrow \mathcal{U}$, we

say $\mathcal{G}$ is *discretization-invariant* if for any $\theta \in \Theta$, any bounded set $K \subset \mathcal{A}$, and $\forall \varepsilon > 0$ $\exists N \in \mathbb{N}$ such that $\forall L \geq N$ it holds

$$\sup_{a \in K} \|\hat{\mathcal{G}}(a|_{D_L}, \theta) - \mathcal{G}(a, \theta)\|_{\mathcal{U}} < \varepsilon$$

Remark 3.2.4 (Relation to [9, Defs. 3–4]). [Definition 3.2.3](#) differs from the formulation of Kovachki et al. [9] in two respects.

First, the term “risk” is dropped. Kovachki et al. define a *discretized uniform risk* $R_K(\mathcal{G}, \hat{\mathcal{G}}, D_L) = \sup_{a \in K} \|\hat{\mathcal{G}}(D_L, a|_{D_L}) - \mathcal{G}(a)\|_{\mathcal{U}}$ [9, Def. 3] and formulate Discretization Invariance as $\lim_{L \rightarrow \infty} R_K = 0$ [9, Def. 4]. Standard risk $R(f) = \mathbb{E}[\ell(f(X), Y)]$ inherently requires a probability space and an expectation over a data-generating distribution. The quantity R_K is a uniform approximation error over the set K , measuring structural discretization error rather than generalization error. [Definition 3.2.3](#) uses this characterization directly.

Second, the compactness hypothesis on $K \subset \mathcal{A}$ is weakened to boundedness, which is strictly weaker in infinite-dimensional Banach spaces. [Definition 3.2.3](#) therefore admits a larger class of parameter families. The price is that establishing Discretization Invariance in this stronger sense requires additional regularity on $\mathcal{G}$: in the proof of [Theorem 3.3.10](#), the Morrey embedding $W^{k,p}(D) \hookrightarrow C^{0,\alpha}(\bar{D})$ supplies uniform Hölder continuity on K , which provides the equicontinuity that Kovachki et al. obtain from the precompactness assumption. The ε – N formulation replacing the limit is purely notational and anticipates the explicit rates derived in [Corollary 3.4.3](#).

Remark 3.2.5 (On the map $\hat{\mathcal{G}}_L$). The discrete operator $\hat{\mathcal{G}}_L$ maps the finite-dimensional input $\{(x_l, a(x_l))\}_{l=1}^L \in \mathbb{R}^{Ld} \times \mathbb{R}^{Lm}$ to a function in the infinite-dimensional space $\mathcal{U}$. This is well-defined because the kernel integral formulation evaluates $\kappa_\phi(x, \cdot)$ at all $x \in \bar{D}$, producing a continuous output from discrete input data. The space of observations $\{(x_l, a(x_l))\}_{l=1}^L \subset (\mathbb{R}^d \times \mathbb{R}^m)^L$ naturally embeds in $\mathbb{R}^{Ld} \times \mathbb{R}^{Lm}$ via the canonical coordinate identification, which is an isometric isomorphism.

3.3 Discretization Invariance Theorem

One of the key contributions of Kovachki et al. (2023) is the Discretization Invariance Theorem for Neural Operators [9, Thm. 8], which establishes that operators between Banach spaces can be learned from discrete sets of function evaluations. In this section we outline the proof and address its limitations. Specifically, we identify a discrepancy between the theorem statement and the proof: the proof relies on assumptions, such as Sobolev regularity for the embedding theorem as well as equicontinuity and pointwise boundedness to justify the existence of a precompact set, that are not explicitly included in the theorem statement. We identify these implicit assumptions and, in [subsection 3.3.2](#), propose an alternative proof that makes all assumptions explicit and extends the Discretization Invariance guarantee to a wider class of functions. A detailed step-by-step analysis of the original proof is presented in [Appendix B](#).

Remark 3.3.1. The theorem statement below reflects the formulation as provided in Kovachki et al. (2023) [9, Thm. 8]. In [Theorem 3.3.8](#) we propose an adjusted theorem statement based on our analysis of the theorem and its proof in [subsection 3.3.1](#).

Theorem 3.3.2 (Discretization Invariance; [9, Thm. 8]). *Let $D \subset \mathbb{R}^d$ and $D' \subset \mathbb{R}^{d'}$ be two domains for some $d, d' \in \mathbb{N}$. Let $\mathcal{A}$ and $\mathcal{U}$ be real-valued Banach function spaces on D and D' respectively. Suppose that $\mathcal{A}$ and $\mathcal{U}$ can be continuously embedded in $C(\bar{D})$ and $C(\bar{D}')$ respectively and that $\sigma_1, \sigma_2, \sigma_3 \in C(\mathbb{R})$. Then, for any $n \in \mathbb{N}$, the set of Neural Operators $NO_n(\sigma_1, \sigma_2, \sigma_3; D, D')$ whose elements are viewed as maps $\mathcal{A} \rightarrow \mathcal{U}$ is discretization-invariant. [9, Thm. 8]*

Proof outline; [9, Thm. 8, Appdx. E]. Without loss of generality, the domains are aligned as $D = D'$ and the continuous embedding is invoked to place the parameter space $\mathcal{A}$ and solution space $\mathcal{U}$ in $C(\bar{D})$, which is necessary to allow pointwise evaluations on the discretized grid.

The multi-layer Neural Operator (Definition 2.2.24) is then reduced to a single-layer operator consisting of a kernel integral with continuous kernel $\kappa \in C(\bar{D} \times \bar{D})$, continuous bias $b \in C(\bar{D})$, and continuous activation function σ :

$$\mathcal{G}(a)(x) = \sigma\left(\int_D \kappa(x, y) a(y) dy + b(x)\right) \quad (3.3)$$

as in Definition 2.2.27. The extension to the multi-layer case follows by composition and continuity of each layer.

Given a discretization $\{y_k\}_{k=1}^j = D_j$ of the domain D (Definition 3.2.1), define a family of pairwise disjoint partitions (P_j) such that each $P_j^{(k)}$ contains exactly one point y_k of the discretization and

$$\bigcup_{k=1}^j P_j^{(k)} = D, \quad P_j^{(k)} \cap P_j^{(l)} = \emptyset \quad \text{for } k \neq l \quad (3.4)$$

This allows the definition of the Riemann sum approximation of the single-layer Neural Operator for any $x \in D$ and $a \in C(\bar{D})$:

$$\hat{\mathcal{G}}_j(a)(x) = \sigma\left(\sum_{k=1}^j \kappa(x, y_k) a(y_k) |P_j^{(k)}| + b(x)\right) \quad (3.5)$$

Consider a compact set $K \subset \mathcal{A}$ with the bound $\|a\|_{\mathcal{A}} \leq M$ for all $a \in K$, and construct a finite ε -net of N representative functions $\{a_n\}_{n=1}^N \subset K$.

This concludes the setup. To show convergence of $\hat{\mathcal{G}}_j$ to $\mathcal{G}$, the proof constructs a three- ε argument for the pre-activation expressions and applies the modulus of continuity of the activation function σ .

First, the bounds on the pre-activation expression are obtained as:

$$\sup_{x \in \bar{D}} \left| \int_D \kappa(x, y) a(y) dy + b(x) \right| \leq M |D| \|\kappa\|_{C(\bar{D} \times \bar{D})} + \|b\|_{C(\bar{D})} \quad (3.6)$$

Equivalent bounds hold for the discretized Riemann sum operator (3.5), which restricts the domain of the activation function σ to a compact subset of $\mathbb{R}$. By the Heine-Cantor Theorem [18, Prop. 1.8.20, p. 32], σ is uniformly continuous on this compact set.

As the final step, a three- ε argument establishes that the pre-activation approximation error decomposes into three components: the continuous approximation error of a by a_n , the discrete approximation error of a by a_n , and the Riemann sum convergence for the Neural Operator evaluated on the finite set of representative functions $\{a_n\}_{n=1}^N \subset K$. This concludes the proof. $\square$

For a detailed analysis of the proof, including the omitted steps and explicit justification statements, we refer to Appendix B.

3.3.1 Analysis of the Discretization Invariance Theorem

On implicit assumptions in the proof setup

We now identify several implicit assumptions in the proof of Theorem 3.3.2 that are not reflected in the theorem statement.

Remark 3.3.3 (Role of continuity). We begin our analysis of the proof by addressing the *continuous embedding* assumption, stated under the "without loss of generality" claim. Given that the proof is based on approximating Neural Operator via Riemann sums, evaluated at the points of the discretized grid $D_j \subset D$, the continuity of $\mathcal{A}$ is necessary for allowing pointwise evaluations of parameter functions $a \in \mathcal{A}$. In the Lebesgue or Sobolev spaces the discretized grid $D_j = \{y_k\}_{k=1}^j$ has a zero mass as a countable union of points. Besides, pointwise evaluations of $a(y_k)$ may be undefined. Hence, continuity is central to the proof in order to enable well defined pointwise evaluations of the operator on the discretized domain.

Remark 3.3.4 (Missing justification for continuous embedding claim). The statement of the [Theorem 3.3.2](#) defines $\mathcal{A}, U$ as general *Banach function spaces*. The continuous embedding claim is non-trivial and requires additional assumptions based on the choice of the Banach space. For instance, for the $L^p(D)$ Lebesgue spaces such continuous embedding does not exist. Since according to their problem definition [9, sec. 2.2] Kovachki et al. intend to approximate operators $\mathcal{G} : \mathcal{A} \rightarrow \mathcal{U}$ that output solution functions to PDEs, it is justified to assume that $\mathcal{A}, U \subset W^{k,p}(D)$ are Sobolev spaces. For the Sobolev spaces the existence of the continuous embedding is guaranteed by the Sobolev Embedding Theorem. However, in order to invoke it, additional assumptions on the regularity of the spaces $\mathcal{A}, U$ as well as the boundary ∂D are required to be stated explicitly. To formalize, given $D \subset \mathbb{R}^d$ we define $\mathcal{A}, U \subset W^{k,p}(D)$ with $k > d/p$ and a bounded C^1 -boundary ∂D . Then, by the Sobolev Embedding Theorem [18, Thm. 5.3.28, p. 193] there exists a continuous embedding $W^{k,p} \hookrightarrow C(\bar{D})$ with

$$\|f\|_\infty \leq C\|f\|_{W^{k,p}}$$

for $C > 0$.

Remark 3.3.5 (Necessary assumptions for compactness of K). Kovachki et al. (2023) [9] rely on an additional strong assumption that needs addressing. They define a compact set $K \subset C(\bar{D})$ to construct the ε -net in the parameter function space $\mathcal{A}$. Since $C(\bar{D})$ is infinite-dimensional, the closed unit ball is not compact by the Riesz Theorem [18, Thm. 2.3.2, p. 80], so compactness of K is non-trivial and requires additional assumptions. Given that $\mathcal{A} \hookrightarrow C(\bar{D})$, we are required to invoke Ascoli-Arzelà Theorem to construct a *precompact* set in $C(\bar{D})$. The precompactness of K is sufficient for existence of ε -net due to total boundedness. However, by Ascoli-Arzelà, set K must contain only *equicontinuous* and *pointwise bounded* functions on a compact domain D [18, Thm. 5.4.2, p. 196]. Given that $\bar{D} \subset \mathbb{R}^d$ is closed, the compactness of $\bar{D}$ is guaranteed by the Heine-Borel Theorem under an additional assumption that the domain D is *bounded*, which is a standard requirement for PDEs with a boundary condition, as defined in the problem setting [section 2.1](#). The requirement on the equicontinuity and pointwise boundedness of functions in K , as well as boundedness of the domain D must be stated explicitly in the theorem definition, which further narrows the set of functions for which Discretization Invariance holds.

Having identified the missing assumptions in the theorem statement, we now address our attention to the key mechanisms that the proof is based on.

On the role of Compactness in the proof of the Discretization Invariance Theorem

The compactness of K is required in order to construct the ε -net on the parameter function space $\mathcal{A}$, which is later used in the three- ε argument of the proof to show convergence of the pre-activation expression. Essentially, the main objective of the proof is to establish convergence of the Riemann sums on the discretized grid D_j to the integral expression for any parameter function $a \in K$. Since the rate of Riemann convergence depends on the

particular function a , and K contains infinitely many functions, there is no guarantee that a single grid resolution suffices for all $a \in K$ simultaneously. Hence, the authors invoke the ε -net to reduce this to a finite problem: by fixing N representative functions $\{a_n\}_{n=1}^N$, they can choose a grid refinement $J \in \mathbb{N}$ such that for all $j \geq J$ and all a_n in the finite ε -net, the Riemann sums

$$\left| \int_D \kappa(x, y) a_n(y) dy - \sum_{k=1}^j \kappa(x, y_k) a_n(y_k) |P_j^{(k)}| \right| \quad (3.7)$$

converge. In order to show

$$\left| \int_D \kappa(x, y) a(y) dy - \sum_{k=1}^j \kappa(x, y_k) a(y_k) |P_j^{(k)}| \right| < \delta \quad (3.8)$$

the authors choose to expand this expression via the triangle inequality into components

$$\int_D \kappa(x, y) |a(y) - a_n(y)| dy \quad (3.9)$$

and

$$\sum_{k=1}^j \kappa(x, y_k) |a(y_k) - a_n(y_k)| |P_j^{(k)}| \quad (3.10)$$

as well as the Riemann convergence for the fixed set of representative functions $\{a_n\}_{n=1}^N$ as stated above. A detailed derivation is presented in [Appendix B](#).

Remark 3.3.6 (Alternative approach for constructing ε -net). It was in fact not necessary to assume the compactness of K and thus impose additional assumptions on the parameter space $\mathcal{A}$. Given that we have established $\mathcal{A}$ as a Sobolev space, by the Rellich-Kondrachov Theorem [18, Thm. 5.4.6] a bounded subset K has a precompact image under the embedding into $C(\bar{D})$, given that $\bar{D} \subset \mathbb{R}^d$ is bounded with C^1 boundary. However, this requirement is already satisfied given that we invoke the Sobolev Embedding Theorem earlier. As a consequence of Rellich-Kondrachov, the precompactness of the image of K in $C(\bar{D})$ guarantees the existence of a finite ε -net as required by the subsequent proof steps.

Remark 3.3.7 (Absence of explicit convergence rates). However, the ε -net based approach does not yield explicit convergence rates, as it relies on asymptotic convergence of Riemann sums for the expression

$$\left| \int_D \kappa(x, y) a_n(y) dy - \sum_{k=1}^j \kappa(x, y_k) a_n(y_k) |P_j^{(k)}| \right| < \delta \quad (3.11)$$

in the three- ε argument. Thus, we propose an alternative, constructive proof that allows us to establish explicit convergence rates on each interval of partition $P_j^{(k)}$ and forgoes the construction of the ε -net over K , which is presented in [subsection 3.3.2](#).

This concludes our analysis of the Theorem on Discretization Invariance for Neural Operators by Kovachki et al. [9, Thm. 8]. A detailed analysis, combined with expanded derivation of the statements of the proof is presented in [Appendix B](#).

Refined theorem statement

Finally, we establish the adjusted Discretization Invariance Theorem statement, which explicitly mentions required assumptions to reflect the presented proof by Kovachki et al. (2023) [9, Thm. 8]:

Theorem 3.3.8 (Discretization Invariance, single-layer; restatement of [9, Thm. 8]). *Let $D \subset \mathbb{R}^d$ be a bounded open domain with a bounded C^1 boundary ∂D . Let $(D_j)_{j=1}^\infty$ denote its discrete refinement. Assume the parameter space $\mathcal{A}$ and solution space $\mathcal{U}$ are subsets of the Sobolev space $W^{k,p}(D)$ with $k > d/p$.*

Let $K \subset \mathcal{A}$ be a uniformly bounded and equicontinuous family of parameter functions. Consider a single-layer Neural Operator $\mathcal{G} : \mathcal{A} \rightarrow \mathcal{U}$ defined by a continuous kernel $\kappa \in C(\bar{D} \times \bar{D})$ and a continuous activation function σ .

Then, for any $\varepsilon > 0$, there exists a sufficiently dense discrete grid D_j such that the empirical operator $\hat{\mathcal{G}}_j$, evaluated via Riemann sums on D_j , satisfies:

$$\sup_{a \in K} \|\mathcal{G}(a) - \hat{\mathcal{G}}_j(a|_{D_j})\|_{C(\bar{D})} < \varepsilon$$

Remark 3.3.9. [Theorem 3.3.8](#) is a restatement of the single-layer Discretization Invariance result of [9, Thm. 8]. The conclusion and the proof strategy (Riemann-sum approximation of the kernel integral) are theirs. The hypotheses stated above make explicit the regularity and compactness conditions on which the argument actually relies. In particular, uniform boundedness and equicontinuity of $K \subset \mathcal{A}$, the Sobolev embedding threshold $k > d/p$, and the C^1 regularity of ∂D are introduced here to close gaps in the original formulation. The alternative proof presented in [subsection 3.3.2](#) weakens several of these hypotheses, in particular replacing the equicontinuity and compactness requirements with boundedness in $W^{k,p}(D)$.

3.3.2 Refined Discretization Invariance Proof

In this section we present an alternative proof of the Discretization Invariance Theorem, in contrast to the proof of Kovachki et al (2023), described in [Theorem 3.3.2](#). The alternative proof utilizes the guarantees of the Sobolev Embedding Theorem [18, Thm. 5.3.28] under weaker assumptions on the function space $\mathcal{A}$, replacing equicontinuity and uniform boundedness assumptions of input set K with boundedness in $W^{k,p}(D)$, as this proof does not rely on finding a compact subset K of $\mathcal{A}$ that the proof presented in Kovachki et al. (2023) [9, Thm. 8, Appdx. E] is based on.

As discussed in [Remark 3.3.6](#), the assumptions of equicontinuity and uniform boundedness of the input set $K \subset \mathcal{A}$ can be omitted by invoking the Rellich-Kondrachov Theorem under the Sobolev Embedding Theorem guarantees. However, our alternative proof omits constructing the ε -net on K entirely, which allows us to define explicit rates of convergence of the discretized operator, as it does not rely on the asymptotic convergence of Riemann sums for finitely many representative functions, as argued in [Remark 3.3.7](#). We first present the alternative proof strategy for the Discretization Invariance Theorem in [Theorem 3.3.10](#), which allows us to replace the reliance on asymptotic Riemann sum convergence for finitely many ε -net representatives with explicit bounds on each partition element via Hölder continuity due to the Sobolev Embedding Theorem, and then establish explicit convergence rates in [Corollary 3.4.3](#) under additional assumptions of Lipschitz continuity of the integral kernel.

Theorem 3.3.10 (Discretization Invariance for a Single-Layer Neural Operator on Sobolev Spaces). *Let $D \subset \mathbb{R}^d$ be a bounded open domain with a Lipschitz continuous boundary. Let $k \in \mathbb{N}$ and $p \in [1, \infty)$ be regularity parameters satisfying $kp > d$. Assume the parameter space $\mathcal{A}$ is a subset of the Sobolev space $W^{k,p}(D)$, and solution space $\mathcal{U} \subset C(\bar{D})$ with $\|\cdot\|_\infty$ norm. Let K be a bounded set of equivalence classes of functions in $W^{k,p}(D)$.*

Consider a single-layer Neural Operator $\mathcal{G} : \mathcal{A} \rightarrow \mathcal{U}$ defined by a continuous kernel $\kappa \in C(\bar{D} \times \bar{D})$, a continuous bias $b \in C(\bar{D})$, and a continuous activation function $\sigma \in C(\mathbb{R})$. Let $(D_j)_{j=1}^\infty$ denote a discrete refinement of the domain D .

Then, for any $\varepsilon > 0$, there exists a sufficiently dense discrete grid D_m such that the empirical operator $\hat{\mathcal{G}}_m$, evaluated via Riemann sums on $D_m \in (D_j)_{j=1}^\infty$, satisfies:

$$\sup_{a \in K} \|\mathcal{G}(a) - \hat{\mathcal{G}}_m(a|_{D_m})\|_{C(\bar{D})} < \varepsilon \quad (3.12)$$

Remark 3.3.11 (Limitation). Theorem 3.3.10 is restricted to the single-layer case. The gap in [9, Thm. 8] identified in Remark 3.3.4 arises in the composition of multiple layers, where the original proof passes from layerwise convergence to convergence of the full operator by an unjustified continuity argument. The present result makes the single-layer mechanism rigorous but does not address that composition step, which remains open and is discussed as future work in Chapter 5.

Remark 3.3.12 (Attribution and relation to prior work). Theorem 3.3.10 and its proof are original to this thesis. The result is inspired by, and generalizes, the single-layer discretization-invariance result of [9, Thm. 8], which assumes the parameter family K to be precompact in $C(\bar{D})$. The present formulation requires only that K be bounded in $W^{k,p}(D)$ with $kp > d$. The equicontinuity that the Riemann-sum approximation argument needs is provided by the Morrey embedding $W^{k,p}(D) \hookrightarrow C^{0,\alpha}(\bar{D})$ ([15, Cor. E.14]), rather than assumed directly. This substitution expands the class of admissible parameter families and allows us to establish in Corollary 3.4.3 an explicit algebraic convergence rate that the precompactness hypothesis alone does not provide.

Remark 3.3.13 (Proof outline: Discretization invariance via Hölder continuity). Given the discretizations (D_j) of domain D and Lebesgue measurable partitions (P_j) , on each cell of the partition $P_j^{(i)}$, we approximate the value of $\int_{P_j^{(i)}} \kappa(x, y) a(y) dy$ for a given $x \in D$ by evaluating the kernel κ and the parameter function a at the representative point of that interval $y_i \in D_j$, resulting in the approximation $\kappa(x, y_i) a(y_i) |P_j^{(i)}|$. We utilize the uniform continuity of the continuous integral kernel κ defined on the compact domain $\bar{D} \times \bar{D}$, as well as the Hölder continuity bounds for $a \in C^{0,\alpha}$, guaranteed by the Sobolev Embedding Theorem, to bound the approximation error on each interval $P_j^{(i)}$. We further construct Riemann sums to approximate the Neural Operator on the whole domain D .

Remark 3.3.14 (Explicit convergence rates requirements). Under an additional assumption of Lipschitz continuity of the activation function σ (Definition 2.2.12) and integral kernel κ (Definition 2.2.14), we further establish explicit convergence rates for the discretized operator, which to our knowledge are not present in the existing literature (Remark 3.3.12). The results are presented in Corollary 3.4.3.

Proof of Theorem 3.3.10. Given a bounded open domain $D \subset \mathbb{R}^d$ with a Lipschitz continuous boundary and the Sobolev space $\mathcal{A} \subset W^{k,p}(D)$ with $kp > d$, by Sobolev Embedding Theorem [15, Cor. E.14, p. 302] and Morrey's Inequality [15, Thm E.13, p. 302] we obtain continuous embedding $\mathcal{A} \hookrightarrow C^{0,\alpha}(\bar{D})$ with exponent $\alpha = \left\lceil \frac{d}{p} \right\rceil + 1 - \frac{d}{p} \in (0, 1]$ (for $d/p \notin \mathbb{Z}$). Each equivalence class $a \in \mathcal{A}$ thus admits a continuous representative in $C^{0,\alpha}(\bar{D})$, on which pointwise evaluation is well-defined. Let $K \subset \mathcal{A}$ be a bounded set of equivalence classes of functions in $W^{k,p}$, each of which has a continuous representative in $C^{0,\alpha}(\bar{D})$.

Define a family of pairwise disjoint, Lebesgue-measurable partitions (P_j) of D such that each $P_j^{(i)}$ contains exactly one point y_i of the discretization D_j of D :

$$\bigcup_{i=1}^j P_j^{(i)} = D \quad (3.13)$$

with $P_j^{(i)} \cap P_j^{(l)} = \emptyset$ for $l \neq i \forall l, i \in \{1, \dots, j\}$. Further we define the diameter of $P_j^{(i)}$ as $h_j^{(i)} := \text{diam}(P_j^{(i)})$ and $h_j = \max_i h_j^{(i)}$ as the maximum diameter of a disjoint element of the partition P_j . By definition of the discrete refinement (D_j) , $h_j \rightarrow 0$ as $j \rightarrow \infty$.

Given a discrete refinement $(D_j)_{j=1}^\infty$ of D , define a discrete approximation $\hat{\mathcal{G}}_j : \mathbb{R}^{jd} \times \mathbb{R}^j \rightarrow C(\bar{D})$ of an operator $\mathcal{G} \in NO$ with $\kappa \in C(\bar{D} \times \bar{D}, \mathbb{R}), b \in C(\bar{D}, \mathbb{R}), \sigma \in C(\mathbb{R})$ as a Riemann sum

$$\hat{\mathcal{G}}_j(y_1, \dots, y_j, w_1, \dots, w_j) = \sigma \left(\sum_{i=1}^j \kappa(\cdot, y_i) w_i |P_j^{(i)}| + b(\cdot) \right) \quad (3.14)$$

where $w_i := a(y_i)$ are the pointwise evaluations of the parameter function $a \in K$ on the element of the discretized domain $y_i \in D_j$.

Recall the single layer Neural Operator $\mathcal{G}$, defined as

$$\mathcal{G}(a)(x) = \sigma \left(\int_D \kappa(x, y) a(y) dy + b(x) \right) \quad (3.15)$$

Denote the expression within the integral as $f(x, y) = \kappa(x, y) a(y)$ with $a \in K$ and $\kappa(x, y) \in C(\bar{D} \times \bar{D}, \mathbb{R})$.

By the Heine-Borel Theorem, the domain $\bar{D} \times \bar{D}$ is compact as it is a Cartesian product of closed and bounded subsets of the finite dimensional space $\mathbb{R}^d$. The integral kernel function $\kappa \in C(\bar{D} \times \bar{D})$ is uniformly continuous, as it is a continuous function defined on a compact domain [18, Prop. 1.8.20, p. 32]. Additionally, by [18, Prop. 1.8.20, p. 32], the image of the integral kernel function κ is compact and hence κ is bounded, i.e.

$$C_\kappa := \sup_{x \in \bar{D}} \sup_{y \in \bar{D}} |\kappa(x, y)| = \|\kappa\|_\infty < \infty, \quad (3.16)$$

and there exists a modulus of continuity ω_κ such that for all $x \in \bar{D}$ and all $y, y' \in \bar{D}$

$$|\kappa(x, y) - \kappa(x, y')| \leq \omega_\kappa(\|y - y'\|) \quad (3.17)$$

The modulus of continuity ω_κ with respect to the second variable holds uniformly for all $x \in \bar{D}$ due to uniform continuity of κ . Without loss of generality, take ω_κ non-decreasing by replacing it with $t \mapsto \sup_{\|y - y'\| \leq t} |\kappa(x, y) - \kappa(x, y')|$, which dominates it and is non-decreasing by construction.

Given that the equivalence class of functions $a \in K$ is bounded with $M > 0$:

$$\|a\|_\infty \leq \|a\|_{C^{0,\alpha}} \leq C \|a\|_{W^{k,p}} \leq CM \quad (3.18)$$

for a $C > 0$. Denote $C_M = CM$. Furthermore, by the Sobolev Embedding Theorem [15, Cor. E.14, p. 302], $a \in K$ is Hölder continuous. By the definition of Hölder continuity [19, Ch. 5.1, p. 254], for all $y, y' \in \bar{D}$:

$$|a(y) - a(y')| \leq \|a\|_{C^{0,\alpha}} \|y - y'\|^\alpha \leq C_M \|y - y'\|^\alpha \quad (3.19)$$

Since the constants only depend on $\|a\|_{W^{k,p}}$, the above bounds hold uniformly for all $a \in K$.

Recall, that on each partition element $P_j^{(i)}$ we use exactly one representative point y_i from the discretized grid D_j . We will now construct a bound for the approximation of the integral operator. Fix an arbitrary evaluation point $x \in \bar{D}$. For this fixed x , the integrand becomes a scalar-valued function $f(x, y) = \kappa(x, y) a(y)$. As illustrated below, the obtained

bounds are independent of the choice of $a \in K$.

$$|\kappa(x, y)a(y) - \kappa(x, y_i)a(y_i)| \leq |\kappa(x, y)a(y) - \kappa(x, y)a(y_i)| + |\kappa(x, y)a(y_i) - \kappa(x, y_i)a(y_i)| \quad (3.20)$$

$$\leq |\kappa(x, y)| |a(y) - a(y_i)| + |\kappa(x, y) - \kappa(x, y_i)| |a(y_i)| \quad (3.21)$$

$$\leq \|\kappa\|_\infty |a(y) - a(y_i)| + |\kappa(x, y) - \kappa(x, y_i)| \|a\|_\infty \quad (3.22)$$

$$\leq C_\kappa C_M \|y - y_i\|^\alpha + C_M \omega_\kappa(\|y - y_i\|) \quad (3.23)$$

Denote $\tilde{C} = C_\kappa C_M$ with C_κ from (3.16) and C_M obtained in (3.18). The equation (3.23) shows that convergence rate depends on $h_j^{(i)}$ via the Hölder exponent α and the kernel modulus ω_κ (3.17). Since $y, y_i \in P_j^{(i)}$, we have

$$\sup_{y \in P_j^{(i)}} \|y - y_i\| \leq h_j^{(i)} \quad (3.24)$$

for each element $P_j^{(i)}$ of the partition P_j .

Since $t \mapsto t^\alpha$ is non-decreasing on $[0, \infty)$ for $\alpha > 0$ and due to monotonicity of the modulus of continuity ω_κ , the bound $\|y - y_i\| \leq h_j^{(i)}$ from (3.24) makes the bound from (3.23) independent of choice of y . Hence, for all $y \in P_j^{(i)}$,

$$|\kappa(x, y)a(y) - \kappa(x, y_i)a(y_i)| \leq \tilde{C} (h_j^{(i)})^\alpha + C_M \omega_\kappa(h_j^{(i)}). \quad (3.25)$$

Given a partition element $P_j^{(i)}$ and by applying the triangle inequality for integrals, we bound the approximation error on $P_j^{(i)}$ via Riemann sums, using (3.25), as:

$$\left| \int_{P_j^{(i)}} f(x, y) dy - f(x, y_i) |P_j^{(i)}| \right| = \left| \int_{P_j^{(i)}} (f(x, y) - f(x, y_i)) dy \right| \quad (3.26)$$

$$\leq \int_{P_j^{(i)}} |f(x, y) - f(x, y_i)| dy \quad (3.27)$$

$$= \int_{P_j^{(i)}} |\kappa(x, y)a(y) - \kappa(x, y_i)a(y_i)| dy \quad (3.28)$$

$$\leq \int_{P_j^{(i)}} \left(\tilde{C} (h_j^{(i)})^\alpha + C_M \omega_\kappa(h_j^{(i)}) \right) dy \quad (3.29)$$

$$= \left(\tilde{C} (h_j^{(i)})^\alpha + C_M \omega_\kappa(h_j^{(i)}) \right) |P_j^{(i)}| \quad (3.30)$$

Thus, by splitting the integral over the partition and applying the triangle inequality, followed by the result from (3.30), we can define the approximation error of the Integral

Kernel Operator via Riemann sums on the whole domain D as follows:

$$\left| \int_D f(x, y) dy - \sum_{i=1}^j f(x, y_i) |P_j^{(i)}| \right| = \left| \sum_{i=1}^j \int_{P_j^{(i)}} f(x, y) dy - \sum_{i=1}^j f(x, y_i) |P_j^{(i)}| \right| \quad (3.31)$$

$$= \left| \sum_{i=1}^j \int_{P_j^{(i)}} (f(x, y) - f(x, y_i)) dy \right| \quad (3.32)$$

$$\leq \sum_{i=1}^j \int_{P_j^{(i)}} |f(x, y) - f(x, y_i)| dy \quad (3.33)$$

$$\leq \sum_{i=1}^j \left(\tilde{C}(h_j^{(i)})^\alpha + C_M \omega_\kappa(h_j^{(i)}) \right) |P_j^{(i)}| \quad (3.34)$$

$$\leq \sum_{i=1}^j \left(\tilde{C}(h_j)^\alpha + C_M \omega_\kappa(h_j) \right) |P_j^{(i)}| \quad (3.35)$$

$$\leq \left(\tilde{C} h_j^\alpha + C_M \omega_\kappa(h_j) \right) |D| \quad (3.36)$$

where $h_j = \max_i \text{diam}(P_j^{(i)})$. The bound is independent of the choice of $x \in \bar{D}$. Hence, as the mesh becomes more refined, i.e. $h_j \rightarrow 0$, the approximation error goes to zero asymptotically. Its convergence rate depends on the modulus of continuity of the integral kernel function ω_κ (3.17).

To complete the proof, we need to show that

$$\|\mathcal{G}(a) - \hat{\mathcal{G}}_j(a|_{D_j})\|_{C(\bar{D})} < \varepsilon \quad (3.37)$$

As established in (3.18) and (3.16), $a \in K$ is bounded by C_M and κ by C_κ . Besides, $\|a\|_\infty \leq \|a\|_{C^{0,\alpha}} \leq C_M$ follows from (3.18). By applying the triangle inequality, we obtain the upper bound of the integration operator:

$$\left| \int_D \kappa(x, y) a(y) dy \right| \leq \int_D |\kappa(x, y) a(y)| dy \quad (3.38)$$

$$\leq \int_D \|a\|_\infty \|\kappa\|_\infty dy \quad (3.39)$$

$$\leq C_M C_\kappa |D| \quad (3.40)$$

Define the pre-activation function $z \in C(\bar{D})$ as

$$z(x) = \int_D \kappa(x, y) a(y) dy + b(x) \quad (3.41)$$

The bias term $b(x) \in C(\bar{D})$, as a continuous function on a compact domain, is bounded by $\|b\|_\infty < \infty$ due to the Extreme Value Theorem [18, Thm. 1.8.25]. Combined with (3.40) we obtain the bound for z :

$$\sup_{x \in \bar{D}} |z(x)| \leq C_M C_\kappa |D| + \|b\|_\infty \quad (3.42)$$

Similarly, we define the discretized pre-activation function $\hat{z} \in C(\bar{D})$ as

$$\hat{z}(x) = \sum_{i=1}^j \kappa(x, y_i) a(y_i) |P_j^{(i)}| + b(x) \quad (3.43)$$

and establish the bound via triangle inequality

$$\sup_{x \in \tilde{D}} |\hat{z}(x)| \leq \sum_{i=1}^j \|\kappa\|_{\infty} \|a\|_{\infty} |P_j^{(i)}| + \|b\|_{\infty} \quad (3.44)$$

$$\leq C_M C_{\kappa} |D| + \|b\|_{\infty} \quad (3.45)$$

Denote the bound as $R := C_M C_{\kappa} |D| + \|b\|_{\infty}$.

The domain of the continuous activation function $\sigma \in C(\mathbb{R})$, restricted to the outputs of $z(x)$ and $\hat{z}(x)$, becomes a compact set $[-R, R] \subset \mathbb{R}$. By the Heine-Cantor Theorem [18, Prop. 1.8.20, p. 32], σ is uniformly continuous on $[-R, R]$, which implies, that there exists a modulus of continuity ω_{σ} , with $\omega_{\sigma}(0) = 0$. Without loss of generality, take ω_{σ} non-decreasing by replacing it with $s \mapsto \sup_{|u-v| \leq s} |\sigma(u) - \sigma(v)|$, the supremum over $u, v \in [-R, R]$, which is non-decreasing by construction. It follows:

$$\left| \sigma \left(\int_D f(x, y) dy + b(x) \right) - \sigma \left(\sum_{i=1}^j f(x, y_i) |P_j^{(i)}| + b(x) \right) \right| \quad (3.46)$$

$$\leq \omega_{\sigma} \left(\left| \int_D f(x, y) dy + b(x) - \sum_{i=1}^j f(x, y_i) |P_j^{(i)}| - b(x) \right| \right) \quad (3.47)$$

$$\leq \omega_{\sigma} \left(\tilde{C} h_j^{\alpha} |D| + C_M \omega_{\kappa}(h_j) |D| \right) \quad (3.48)$$

Given $\varepsilon > 0$, choose $\delta > 0$ such that $\omega_{\sigma}(\delta) < \varepsilon$, which is possible since $\omega_{\sigma}(0) = 0$ and ω_{σ} is continuous. By (3.36), because $\lim_{h_j \rightarrow 0} h_j^{\alpha} = 0$ and $\lim_{h_j \rightarrow 0} \omega_{\kappa}(h_j) = 0$, there exists $N \in \mathbb{N}$ such that for all $j \geq N$:

$$\left(\tilde{C} h_j^{\alpha} + C_M \omega_{\kappa}(h_j) \right) |D| \leq \delta \quad (3.49)$$

Then, by monotonicity of ω_{σ} and (3.48):

$$\omega_{\sigma} \left(\left(\tilde{C} h_j^{\alpha} + C_M \omega_{\kappa}(h_j) \right) |D| \right) \leq \omega_{\sigma}(\delta) < \varepsilon \quad (3.50)$$

Since N depends only on C_M , C_{κ} , $|D|$, and δ and not on the particular $a \in K$, the bound holds uniformly over K , confirming (3.37). This concludes the proof. $\square$

Remark 3.3.15 (Moduli of continuity and explicit convergence rates). From (3.36) and (3.50) it follows that we can derive explicit convergence rates for the Neural Operator approximations by imposing additional requirements on the regularity of the integral kernel $\kappa(x, y)$ (Definition 2.2.14) and the activation function σ (Definition 2.2.12), as we discuss in the subsequent Corollary 3.4.3.

3.4 Explicit convergence rates for discretized operators

Having established the convergence of the discretized operator in (3.50) of theorem 3.3.10 proof, we can derive explicit convergence rates by replacing the modulus of continuity functions for activation function ω_{σ} as well as the integral kernel function ω_{κ} at the cost of imposing additional assumptions.

Remark 3.4.1 (On regularity of the activation function σ). In the definition of the simplified Neural Operator used in the Discretization Invariance and Universal Approximation

Theorem proofs, Kovachki et al. (2023) [9, sec. 8.1] restrict the set of activation functions used in the Neural Operator to a set of continuous functions on $\mathbb{R}$ that map the input set to a compact set in $C^m(K)$ with K being a compact subset of $\mathbb{R}^d$. Furthermore, they require any such function σ to be Borel measurable and have at most linear growth, i.e. $\sup_{x \in \mathbb{R}} \frac{|\sigma(x)|}{1+|x|} < \infty$. Kovachki et al. (2023) indicate that any Lipschitz continuous activation function, such as commonly used ReLU, GeLU, Tanh or Sigmoid, would satisfy this requirement [9, Section 8.1, p. 49]. Commonly used activation functions with their respective Lipschitz constants are presented in Table 3.1.

As a consequence, we can replace the modulus of continuity ω_σ with Lipschitz continuity constant in eq. 3.50, obtaining:

$$\|\mathcal{G}(a) - \hat{\mathcal{G}}_j(a|_{D_j})\|_{C(\bar{D})} \leq L_\sigma |D| \left(\tilde{C} h_j^\alpha + C_M \omega_\kappa(h_j) \right) \quad (3.51)$$

Remark 3.4.2 (On regularity of the integral kernel κ). The assumption of Lipschitz continuity of κ requires justification, since the true solution operator (Green’s function) of a PDE need not be globally continuous. However, the Discretization Invariance result concerns the *learned* Neural Operator, not the true solution operator itself. In practice, the two dominant empirical architectures parameterize κ in a way that guarantees Lipschitz continuity: in Graph Neural Operators (GNOs) (Definition 2.2.30), the kernel is learned via a Multilayer Perceptron κ_ϕ , which is a composition of affine maps and Lipschitz continuous activation functions and is therefore Lipschitz. In Fourier Neural Operators (FNOs) (Definition 2.2.31), the kernel is parameterized in the frequency domain and restricted to a finite set of Fourier modes, which results in a smooth and hence Lipschitz continuous function. Therefore, for both classes of architectures the assumption of Lipschitz continuity of κ is justified.

Activation Function	Expression $\sigma(x)$	Lipschitz Constant L_σ
ReLU	$\max(0, x)$	1
Tanh	$\tanh(x)$	1
Sigmoid	$(1 + e^{-x})^{-1}$	0.25
GELU	$x \cdot \Phi(x)$	≈ 1.129

Table 3.1: Standard activation functions utilized in Neural Operator architectures and their corresponding global Lipschitz constants.

Having provided the justification for our assumptions on the Lipschitz continuity of the Nemytskii operator induced by the Activation function σ (Definition 2.2.12) and the Kernel Integral Operator with kernel κ (Definition 2.2.14) in the Neural Operator (Definition 2.2.24), we proceed to establishing the explicit convergence rates for the discretized Neural Operator.

Corollary 3.4.3 (Convergence Rates for the Lipschitz Continuous Discretized Neural Operator). *Let $D \subset \mathbb{R}^d$ be a bounded open domain with a Lipschitz continuous boundary. Let $k \in \mathbb{N}$ and $p \in [1, \infty)$ be regularity parameters satisfying $kp > d$, with $\alpha = \left\lceil \frac{d}{p} \right\rceil + 1 - \frac{d}{p} \in (0, 1]$ (for $d/p \notin \mathbb{Z}$) being the Hölder continuity exponent of the embedding $W^{k,p}(D) \hookrightarrow C^{0,\alpha}(\bar{D})$ [15, Cor. E.13, E. 14]. Let the parameter space $\mathcal{A}$ be a subset of the Sobolev space $W^{k,p}(D)$ and and solution space $\mathcal{U} \subset C(\bar{D})$. Let K be a bounded set of equivalence classes of functions in $\mathcal{A}$.*

*Consider a single-layer Neural Operator $\mathcal{G} : \mathcal{A} \rightarrow \mathcal{U}$ defined by a **Lipschitz continuous** kernel $\kappa \in C(\bar{D} \times \bar{D})$ with Lipschitz constant L_κ and supremum bound $C_\kappa = \|\kappa\|_\infty$, and*

a **Lipschitz continuous** activation function σ with Lipschitz constant L_σ . Let $(D_j)_{j=1}^\infty$ denote a discrete refinement of the domain D , and let $C_M = C M$ where $C > 0$ is the Sobolev embedding constant and $M = \sup_{a \in K} \|a\|_{W^{k,p}}$.

Then, for any discretization $D_j \in (D_j)_{j=1}^\infty$ with mesh width $h_j = \max_i \text{diam}(P_j^{(i)})$, the empirical operator $\hat{\mathcal{G}}_j$, evaluated via Riemann sums on D_j , satisfies:

$$\sup_{a \in K} \|\mathcal{G}(a) - \hat{\mathcal{G}}_j(a|_{D_j})\|_{C(\bar{D})} \leq L_\sigma C_M |D| (C_\kappa h_j^\alpha + L_\kappa h_j) \quad (3.52)$$

Proof. The proof follows from [Theorem 3.3.10](#) by replacing the moduli of continuity in inequalities (3.23) and (3.36) with the respective Lipschitz constants.

Given that the integral kernel κ is Lipschitz continuous with constant L_κ , the Lipschitz bound $|\kappa(x, y) - \kappa(x, y_i)| \leq L_\kappa \|y - y_i\|$ replaces the modulus ω_κ in (3.23). By the monotonicity argument preceding (3.25) (using $\|y - y_i\| \leq h_j^{(i)}$), we obtain, for all $y \in P_j^{(i)}$,

$$|\kappa(x, y)a(y) - \kappa(x, y_i)a(y_i)| \leq \|\kappa\|_\infty |a(y) - a(y_i)| + |\kappa(x, y) - \kappa(x, y_i)| \|a\|_\infty \quad (3.53)$$

$$\leq C_\kappa C_M \|y - y_i\|^\alpha + L_\kappa C_M \|y - y_i\| \quad (3.54)$$

$$\leq C_M (C_\kappa (h_j^{(i)})^\alpha + L_\kappa h_j^{(i)}) \quad (3.55)$$

where C_M , C_κ , and $h_j^{(i)}$ are as defined in the proof of [Theorem 3.3.10](#).

Define $\phi(t) := C_\kappa t^\alpha + L_\kappa t$ for $t \geq 0$. Since $C_\kappa, L_\kappa \geq 0$ and $\alpha > 0$, ϕ is a monotonic non-decreasing function on $[0, \infty)$. Since $h_j^{(i)} := \text{diam}(P_j^{(i)}) \leq h_j$ for every i , we have $\phi(h_j^{(i)}) \leq \phi(h_j)$. Hence, by applying the result from (3.55) to (3.36) and by monotonicity of ϕ with $C_M \geq 0$ and $|P_j^{(i)}| \geq 0$ we obtain the bounds on the integral approximation via Riemann sums:

$$\left| \int_D f(y) dy - \sum_{i=1}^j f(y_i) |P_j^{(i)}| \right| \leq \sum_{i=1}^j C_M (C_\kappa (h_j^{(i)})^\alpha + L_\kappa h_j^{(i)}) |P_j^{(i)}| \quad (3.56)$$

$$= \sum_{i=1}^j C_M \phi(h_j^{(i)}) |P_j^{(i)}| \quad (3.57)$$

$$\leq C_M \phi(h_j) \sum_{i=1}^j |P_j^{(i)}| \quad (3.58)$$

$$= C_M (C_\kappa h_j^\alpha + L_\kappa h_j) \sum_{i=1}^j |P_j^{(i)}| \quad (3.59)$$

$$= C_M (C_\kappa h_j^\alpha + L_\kappa h_j) |D| \quad (3.60)$$

where the last equality uses $\sum_{i=1}^j |P_j^{(i)}| = |D|$ since $\{P_j^{(i)}\}$ partitions D .

For the following step we adopt the notation for the pre-activation expressions z and $\hat{z}$ from (3.41) and (3.43). Since the activation function σ is Lipschitz continuous with constant L_σ , for every $x \in \bar{D}$ it holds

$$|\sigma(z(x)) - \sigma(\hat{z}(x))| \leq L_\sigma |z(x) - \hat{z}(x)| \quad (3.61)$$

$$= L_\sigma \left| \int_D f(y) dy - \sum_{i=1}^j f(y_i) |P_j^{(i)}| \right|, \quad (3.62)$$

where the bias term $b(x)$ cancels. The right-hand side is independent of x by (3.60), so taking the supremum over $x \in \bar{D}$,

$$\|\mathcal{G}(a) - \hat{\mathcal{G}}_j(a|_{D_j})\|_{C(\bar{D})} = \|\sigma(z) - \sigma(\hat{z})\|_{C(\bar{D})} \quad (3.63)$$

$$= \sup_{x \in \bar{D}} |\sigma(z(x)) - \sigma(\hat{z}(x))| \quad (3.64)$$

$$\leq L_\sigma C_M |D| (C_\kappa h_j^\alpha + L_\kappa h_j). \quad (3.65)$$

Observe that the final bound depends only on the global constants and the grid spacing h_j , and, hence, holds uniformly over $a \in K$. By taking the supremum over $a \in K$ we obtain:

$$\sup_{a \in K} \|\mathcal{G}(a) - \hat{\mathcal{G}}_j(a|_{D_j})\|_{C(\bar{D})} \leq L_\sigma C_M |D| (C_\kappa h_j^\alpha + L_\kappa h_j) \quad (3.66)$$

which completes the proof. $\square$

Remark 3.4.4 (Approximation error decomposition). The bound in [Corollary 3.4.3](#) decomposes the approximation error into two components:

$$\text{Total error} \leq C_1 h_j^\alpha + C_2 h_j \quad (3.67)$$

where $C_1, C_2 > 0$ are constant for a given bounded domain D with $C_1 := L_\sigma C_M C_\kappa |D|$, $C_2 := L_\sigma C_M L_\kappa |D|$ and $\alpha \in (0, 1]$ is the Hölder exponent of the embedding $W^{k,p}(D) \hookrightarrow C^{0,\alpha}(\bar{D})$. The first component $C_1 h_j^\alpha$ arises from the pointwise approximation of the input function a between grid points and is governed by α , reflecting the regularity of the input function class. The second component $C_2 h_j$ arises from the Lipschitz continuity of the learned kernel κ and is controlled by the choice of architecture.

Since $\alpha \leq 1$, the input term decays no faster than the kernel term, and the convergence rate equals α . For $\alpha < 1$ the input term h_j^α dominates for sufficiently fine grids, and the rate is limited by the regularity of the input function class. At $\alpha = 1$ both terms decay at the same order h_j , and the rate is 1. In either case the rate is limited by the input regularity α and not by the kernel, so imposing regularity on κ beyond Lipschitz continuity does not improve the discretization rate.

Remark 3.4.5 (Lower bound on the number of grid partitions N). The explicit convergence bound for the discretized operator in [Corollary 3.4.3](#) is expressed in terms of the maximum partition cell diameter $h_j := \max_k \text{diam}(P_j^{(i)})$. This quantity is intrinsic to the bound. In practice, however, the relevant quantity is the number of sample points N , chosen by a practitioner with regards to the associated computational costs. Hence, expressing the bound in terms of N has a practical application. Establishing a *convergence rate* in N would require an additional assumption of quasi-uniformity of the partition. Such a rate is obtained by Lanthaler et al. (2024) [13] for the FNO ([Definition 2.2.31](#)), where a uniform grid is intrinsic to the Fast-Fourier-Transform-based architecture. Without imposing any uniformity assumptions on the discretization in term of [Definition 3.2.1](#), we can still establish a *lower bound* on the number of sample points N required for a given maximum partition cell diameter h_j , which is the purpose of [Proposition 3.4.6](#).

Proposition 3.4.6 (Mesh cardinality lower bound). *Let $D \subset \mathbb{R}^d$ be a bounded domain, and let $(P_j^{(i)})_{i=1}^N$ be a partition of D into N measurable cells with $\bigcup_{i=1}^N P_j^{(i)} = D$ and $P_j^{(i)} \cap P_j^{(l)} = \emptyset$ for $i \neq l$. Let $h_j = \max_{1 \leq i \leq N} \text{diam}(P_j^{(i)})$ denote the maximum mesh size. Then the lower bound on the amount of partitions N is given by*

$$N \geq \frac{|D|}{\nu_d} h_j^{-d}, \quad (3.68)$$

where $|D|$ is the Lebesgue measure of D and $\nu_d = \frac{\pi^{d/2}}{\Gamma(d/2+1)}$ is the volume of a unit ball in $\mathbb{R}^d$.

Proof. Fix a partition element $P_j^{(i)}$ and any point $y_0 \in P_j^{(i)}$. Since $\text{diam}(P_j^{(i)}) \leq h_j$, every $y \in P_j^{(i)}$ satisfies $\|y - y_0\| \leq h_j$, so $P_j^{(i)} \subseteq \overline{B}(y_0, h_j)$. By monotonicity of Lebesgue measure,

$$|P_j^{(i)}| \leq \nu_d h_j^d, \quad (3.69)$$

where $\nu_d = \frac{\pi^{d/2}}{\Gamma(d/2+1)}$ is the volume of the unit ball in $\mathbb{R}^d$. As the cells form a disjoint partition of D , their measures sum to $|D|$:

$$|D| = \sum_{i=1}^N |P_j^{(i)}| \leq \sum_{i=1}^N \nu_d h_j^d = N \nu_d h_j^d. \quad (3.70)$$

Dividing by $\nu_d h_j^d > 0$ gives $N \geq |D| \nu_d^{-1} h_j^{-d}$. $\square$

Remark 3.4.7 (Practical application). The error decomposition from (3.67) combined with Proposition 3.4.6 gives an explicit lower estimate on the sample resolution required for a target accuracy. Fix $\varepsilon > 0$ and let $\alpha \in (0, 1]$ be the regularity exponent of the input class in spatial dimension d . By Corollary 3.4.3, the discretization error is bounded by $C_1 h_j^\alpha + C_2 h_j$. Since $\alpha \leq 1$, a sufficient condition for the error to be bounded by ε is

$$C_1 h_j^\alpha \leq \frac{\varepsilon}{2} \quad \text{and} \quad C_2 h_j \leq \frac{\varepsilon}{2}, \quad (3.71)$$

the first of which is the more restrictive constraint for sufficiently small ε and requires

$$h_j \leq \left(\frac{\varepsilon}{2C_1} \right)^{1/\alpha}, \quad (3.72)$$

expressed as a lower bound on h_j^{-d} :

$$\left(\frac{\varepsilon}{2C_1} \right)^{-1/\alpha} \leq h_j^{-1} \quad (3.73)$$

$$\left(\frac{2C_1}{\varepsilon} \right)^{d/\alpha} \leq h_j^{-d} \quad (3.74)$$

By Proposition 3.4.6, any partition with maximal cell diameter h_j on the bounded domain D satisfies $N \geq \frac{|D|}{\nu_d} h_j^{-d}$. Substituting the required mesh size (3.74) into this lower bound, the number of sample points must satisfy

$$N \geq \frac{|D|}{\nu_d} h_j^{-d} \geq \frac{|D|}{\nu_d} \left(\frac{2C_1}{\varepsilon} \right)^{d/\alpha} \quad (3.75)$$

hence, the required number of sample points estimate follows directly from the regularity of the input function class, and the estimate holds for any partition fine enough to achieve ε without an assumption of grid uniformity.

Remark 3.4.8 (Curse of dimensionality). The estimate $N \gtrsim \varepsilon^{-d/\alpha}$ (absorbing constants) from (3.75) of Remark 3.4.7 demonstrates the curse of dimensionality. The exponent d/α grows linearly in the spatial dimension d , so the number of sample points sufficient to guarantee a discretization error below ε grows exponentially in d for any fixed error $\varepsilon < 1$ and regularity $\alpha \in (0, 1]$. Equivalently, reducing the target error by half would require scaling the sample size by at least $2^{d/\alpha}$.

The obtained estimate $N \gtrsim \varepsilon^{-d/\alpha}$ describes the lower bound on the sample count that the discretization error bound (3.66) obtained in Corollary 3.4.3 requires, not a rate of convergence of the discretization error in N . It holds for any discrete refinement satisfying Definition 3.2.1, with no assumption of grid uniformity or cell regularity.

The dependence on α highlights the impact of input function class regularity on the sample requirements. Rougher input classes (smaller α) increase the exponent d/α and hence require more sample points N for a given error target ε , while $\alpha = 1$, the cap imposed by the Hölder embedding, gives the mildest exponent d .

Chapter 4

Universal Approximation Theorems

4.1 Motivating Problem

The Universal Approximation Theorem for finite dimensional spaces, as proven by Cybenko (1989) [1], shows that a single layer neural network composed of finite sums of affine transformations followed by a nonlinear activation function acts as universal approximator for continuous functions defined on a unit hypercube $I_n \subset \mathbb{R}$. The proof by Cybenko demonstrates that a set of such finite sums $\hat{f}(x) = \sum_{i=1}^n c_i \sigma(w_i \cdot x + b_i)$, where σ is a nonlinear activation function such as sigmoid, is dense in $C(I_n)$ with respect to the supremum norm, in the same spirit as classical density results like Stone-Weierstrass approximation [18, Thm. 2.4.1, p. 61]. Despite the fact that the proof is non-constructive, as it relies on the Hahn-Banach Theorem [18, Thm. 4.2.4, p. 124] and the Riesz Representation Theorem [18, Thm. 4.3.8, p. 134] to construct an argument by contradiction, it sets the foundation for the first key component of the subsequent Universal Approximation Theorems for infinite dimensional spaces, the *non-linearity* of activation function σ [3].

Our objective differs from Cybenko's in a structural way. Instead of approximating functions $f : I_n \rightarrow \mathbb{R}^n$ that map points between finite dimensional spaces, we want to approximate operators $\mathcal{G} : \mathcal{A} \rightarrow \mathcal{U}$ between infinite dimensional Banach spaces of functions (Remark 2.1.5). Two obstacles arise immediately. First, neural networks are by construction maps between finite dimensional vector spaces, which prevents us from ingesting elements of the Banach space $\mathcal{A}$ directly. Second, the compactness arguments that Cybenko-like density results are based on are not freely available in the infinite dimensional spaces, as per Riesz Theorem [18, Thm. 2.3.2, p. 60] the unit ball is not compact in $\mathcal{A}$, and without some form of compactness no finite-dimensional reduction can be uniform. One approach to work around this limitation is to factor the element $a \in \mathcal{A}$ into a finite dimensional representation in $\mathbb{R}^n$, approximate the finite dimensional map via a neural network, and decode the output back into $\mathcal{U}$. This is the approach taken by Chen & Chen (1995) [3, Thm. 5] and, as we will demonstrate, the prerequisite is precisely the *compactness* of the input set, as was previously identified, and which now has to be assumed on the input space $K \subset \mathcal{A}$.

4.1.1 Fixed-Grid based Approach to Universal Approximation

As discussed above, we begin our analysis by presenting the Universal Approximation Theorem (UAT) statement by Chen and Chen (1995):

Theorem 4.1.1 (Fixed-Grid UAT by Chen & Chen, 1995 [3, Thm. 5]). *Let $\mathcal{A} = C(D)$ and $\mathcal{U} = C(D')$, where $D \subset \mathbb{R}^d$ and $D' \subset \mathbb{R}^{d'}$ are compact. Let $K \subset \mathcal{A}$ be compact, and let $\mathcal{G} : K \rightarrow \mathcal{U}$ be a continuous (nonlinear) operator. Let $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ be any nonpolynomial continuous function. Then for every $\varepsilon > 0$ there exist positive integers $N, M, m \in \mathbb{N}$,*

sampling points $x_1, \dots, x_m \in D$, and real parameters $c_i^k, \xi_{ij}^k, \theta_i^k, \zeta_k \in \mathbb{R}$ and $\omega_k \in \mathbb{R}^{d'}$, for $k = 1, \dots, N$ and $i = 1, \dots, M$ and $j = 1, \dots, m$, such that

$$\left| \mathcal{G}(a)(y) - \sum_{k=1}^N \sum_{i=1}^M c_i^k \sigma \left(\sum_{j=1}^m \xi_{ij}^k a(x_j) + \theta_i^k \right) \sigma(\omega_k \cdot y + \zeta_k) \right| < \varepsilon \quad (4.1)$$

holds for all $a \in K$ and all $y \in D'$.

Remark 4.1.2 (On notation in Chen and Chen (1995) [3]). We restate the result of Chen and Chen (1995) [3] in the notation adopted throughout this thesis. The translation is summarized in the following table.

Chen & Chen (1995)	This thesis
$u \in V \subset C(K_1)$	$a \in K \subset \mathcal{A}$
$G(u) \in C(K_2)$	$\mathcal{G}(a) \in \mathcal{U}$
$K_1 \subset \mathbb{R}^n$	$D \subset \mathbb{R}^d$
$K_2 \subset \mathbb{R}^n$	$D' \subset \mathbb{R}^{d'}$
g (Tauber-Wiener function)	σ (nonpolynomial continuous activation)

In particular, the symbol u in the original paper denotes an *input* function (parameter), whereas in our setting $\mathcal{U}$ denotes the *output* (solution) space. To avoid this collision, we consistently use a for the input parameter and reserve $\mathcal{U}$ for the solution space, in line with the convention established in [section 2.1](#). Likewise, the original paper uses K_1, K_2 for the spatial domains on which input and output functions are defined, while we reserve K for a compact subset of the parameter space $\mathcal{A}$ and write D, D' for the spatial domains.

Proof outline (following Chen and Chen (1995) [3, Thm. 5]). Two structural observations drive the construction of the approximating function:

First, the output. For a fixed $a \in \mathcal{A}$, the function $y \mapsto \mathcal{G}(a)(y)$ is a continuous map on a compact set $D' \subset \mathbb{R}^{d'}$. The classical UAT by Cybenko states that this function can be approximated to an arbitrary accuracy by a finite sum

$$\sum_{k=1}^N c_k \sigma(\omega_k \cdot y + \zeta_k) \quad (4.2)$$

[1, Thm. 2]. However, given that the output function $\mathcal{G}(a)$ varies dependent on the input function $a \in K$, classical Cybenko is not enough, since we need to find unique set of weights ω_k, ζ_k that works *uniformly* for the whole family of functions $\mathcal{G}(K)$. Otherwise, the weights ω_k, ζ_k would need to be retrained for each input function a , which defeats the purpose. Hence, compactness of $K \subset \mathcal{A}$ is required, as it guarantees via Ascoli-Arzelà Theorem [18, Thm. 5.4.2, p. 196] the equicontinuity of the input set of functions K , which in turn allows a uniform choice of parameters ω_k, ζ_k .

Second, the dependence on the input function $a \in K$. Once the basis for the output $\{\sigma(\omega_k \cdot y + \zeta_k)\}$ is fixed uniformly, the only free variable left in (4.2) is the coefficient c_k . Given that the input function a varies across the set K , the coefficient c_k must depend on the choice of a . The non-trivial result is that this dependence is continuous, i.e. the mapping $a \mapsto c_k$ defines a continuous functional on K . This observation reduces the operator approximation problem to approximating a finite set of continuous functionals on a compact set K , which can be solved by sampling evaluation of a at a fixed grid of points $\{x_i\}_{i=1}^m \subset D$ and applying Cybenko for the second time on the resulting finite dimensional vector $\{a(x_i)\}_{i=1}^m \subset \mathbb{R}^m$.

The two observations combined provide the architecture of the proof, where we apply the "outer" Cybenko UAT [1] in y to fix the basis, followed by the "inner" Cybenko UAT [1] in the sampled evaluations of a to find the coefficients. The proof formalises each step. $\square$

Remark 4.1.3. Chen and Chen (1995) divided their proof of the UAT into 5 separate theorems, which combined deliver the approximation claim. Before presenting the formal proof, we state the relevant results [3, Thm. 3, Thm. 4] from Chen and Chen (1995) that are required for the formal proof as lemmas and provide proof ideas.

Remark 4.1.4 (Universal Approximation via non-polynomial continuous functions). Theorems 1-3 of Chen and Chen [3] generalize the Universal Approximation results for continuous nonlinear functions established by Cybenko [1] and Hornik [2]. Their key contribution is to show that any *Tauber-Wiener* function qualifies as an activation function in the hidden layer, where a continuous function $\sigma \in S'(\mathbb{R})$ is Tauber-Wiener if and only if it is not a polynomial. Throughout this thesis we refer to σ simply as a nonpolynomial continuous function, omitting the Tauber-Wiener terminology to avoid introducing definitions from the Tempered Distribution Theory that are not otherwise used.

Lemma 4.1.5 (Cybenko approximation with uniform weights; [3, Thm. 3]). *Let $\Omega \subset \mathbb{R}^p$ be compact and let $V \subset C(\Omega)$ be a compact subset. Let $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ be any nonpolynomial continuous function. Then for every $\varepsilon > 0$ there exist a positive integer $N \in \mathbb{N}$, parameters $\omega_k \in \mathbb{R}^p$ and $\zeta_k \in \mathbb{R}$ for $k = 1, \dots, N$, independent of $f \in V$, and continuous linear functionals $c_k : V \rightarrow \mathbb{R}$ such that*

$$\sup_{f \in V} \sup_{z \in \Omega} \left| f(z) - \sum_{k=1}^N c_k(f) \sigma(\omega_k \cdot z + \zeta_k) \right| < \varepsilon. \quad (4.3)$$

Remark 4.1.6 (The learned basis for equicontinuous set). Lemma 4.1.5 implies that given a compact set of equicontinuous functions $V \subset C(\Omega)$ acting on a compact domain $\Omega \subset \mathbb{R}^p$, and an error tolerance $\varepsilon > 0$, there exists a *fixed* set of non-linear continuous functions $\{\beta_k\}_{k=1}^N$ with learned $\beta_k(z) = \sigma(\omega_k \cdot z + \zeta_k)$, such that the continuous transformation described by *any* function $f \in V$ on Ω can be approximated as a linear combination of the $\{\beta_k\}_{k=1}^N$, where the weights $\alpha_k = c_k(f)$ only depend on the choice of function $f \in V$ and *not* the input $z \in \Omega$. Hence the set $\{\beta_k\}_{k=1}^N$ can be thought of as forming a basis for the $V \subset C(\Omega)$ analogous to a uniform Fourier truncation.

Proof outline. (following [3, Theorem 3]) . For a single $f \in C(\Omega)$, the density of sigmoidal sums is guaranteed by the Cybenko UAT [1]. The challenge is to find a uniform set of weights ω_k, ζ_k that are shared across the whole compact set $V \subset C(\Omega)$.

Assume WLOG that $\Omega = [0, 1]^n$ and extend each $f \in V$ to an even periodic function f^* on $[-1, 1]^n$. The compact set of extensions V^* remains compact by [3, Lemma 3]. Since the extension is even, the sin coefficients in the Fourier expansion vanish by symmetry, and each f^* admits a pure cosine expansion with real-valued Fourier coefficients $d_m(f^*)$.

By the Ascoli-Arzelà Theorem [18, Thm. 5.4.2, p. 196], compactness of V^* implies equicontinuity, which guarantees the existence of a single modulus of continuity $\omega(\delta)$ bounding $|f^*(t) - f^*(t + \delta)|$ for all $f^* \in V^*$ and all t . Since the Fourier coefficients satisfy $|d_m(f^*)| \leq C \cdot \omega(1/|m|)$ for all $f^* \in V^*$ with the same constant C , a single truncation radius R controls the approximation error uniformly over V^* . More precisely, Chen & Chen use a smoothed truncation via Bochner-Riesz means on a multivariate domain, we refer to [3, Lemma 4] for details. The result is that

$$\left\| f^*(x) - \sum_{|m| \leq R} d_m(f^*) \cos(\pi m \cdot x) \right\| < \varepsilon/2$$

holds for all $f^* \in V^*$ simultaneously.

It remains to approximate the cosine basis via the sigmoidal sum. Each $\cos(\pi m \cdot x)$ is a continuous function of the single variable $u = \pi m \cdot x \in \mathbb{R}$. Since σ is nonpolynomial

and hence Tauber-Wiener by [3, Thm. 1], there exist $M \in \mathbb{N}$ and constants s_j, ξ_j, η_j such that $|\cos(u) - \sum_{j=1}^M s_j \sigma(\xi_j u + \eta_j)| < \varepsilon/(2L)$ on the relevant bounded interval, where $L = \sum_{|m| \leq R} |d_m|$ absorbs the Fourier coefficients. Substituting into the truncated Fourier sum yields a double sum over pairs (m, j) , which can be rearranged into a single sum $\sum_{k=1}^N c_k(f) \sigma(\omega_k \cdot z + \zeta_k)$ with $N = |\{|m| \leq R\}| \cdot M$. The weights $\omega_k = \xi_j \pi m$ and biases $\zeta_k = \eta_j$ depend only on the Fourier frequencies and the TW approximation to cosine, not on f . The output coefficients $c_k(f) = d_m(f^*) \cdot s_j$ are defined as products of a linear continuous functional of f with a fixed constant, hence they remain linear continuous functionals of f . Combined, this results in the expression of (4.3). $\square$

Lemma 4.1.7 (Approximation of continuous functionals on a compact subset of $\mathcal{A}$; [3, Thm. 4]). *Let $K \subset \mathcal{A} = C(D)$ be compact, with $D \subset \mathbb{R}^d$ compact, and let $F : K \rightarrow \mathbb{R}$ be a continuous functional. Let $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ be any nonpolynomial continuous function. Then for every $\varepsilon > 0$ there exist positive integers $M, m \in \mathbb{N}$, sampling points $x_1, \dots, x_m \in D$, and real parameters $c_i, \theta_i, \xi_{ij} \in \mathbb{R}$ for $i = 1, \dots, M$ and $j = 1, \dots, m$ such that*

$$\sup_{a \in K} \left| F(a) - \sum_{i=1}^M c_i \sigma \left(\sum_{j=1}^m \xi_{ij} a(x_j) + \theta_i \right) \right| < \varepsilon. \quad (4.4)$$

Proof idea. (following [3, Thm. 4]). The construction of the approximating function can be divided into two parts. First, given compactness of $K \subset \mathcal{A}$, by the Ascoli-Arzelà Theorem [18, Thm. 5.4.2, p. 196] the set K is equicontinuous, i.e. $\forall \eta > 0 \exists \delta > 0 : \forall a \in K, \forall x, y \in D$ it holds that $|x - y| < \delta$ implies $|a(x) - a(y)| < \eta$. By compactness of $D \subset \mathbb{R}^d$, we construct an δ -net $\{x_j\}_{j=1}^m$ and a corresponding partition of unity $\{T_j\}_{j=1}^m$ [18, Lemma 5.3.14], such that the discretized function

$$a_\eta(x) := \sum_{j=1}^m a(x_j) T_j(x) \quad (4.5)$$

satisfies $\|a_\eta - a\|_{C(D)} < \eta$ uniformly for all $a \in K$. As a result, any $a \in K$ can be approximated with accuracy η by the finite dimensional vector $(a(x_1), \dots, a(x_m)) \in \mathbb{R}^m$ uniformly across the whole set of functions K . We call this map the *encoder* and denote it as $\mathcal{E} : K \rightarrow \mathbb{R}^m$. Given the desired accuracy η , it samples the point grid $\{x_j\}_{j=1}^m \subset D$ to approximate the identity on K as

$$\mathcal{E}(a) = (a(x_1), \dots, a(x_m)) \in \mathbb{R}^m$$

We further denote the discretization of K for a given accuracy η as $K^\eta := \{a_\eta : a \in K\}$.

Second, given that F is a continuous functional on a compact set K , by Heine-Cantor Theorem [18, 1.8.20, p. 32] it is uniformly continuous on K . However, F may not be defined on K^η . Chen and Chen address this by defining a continuous extension $F^* : K^* \rightarrow \mathbb{R}$ on the compact set $K^* = K \cup K^\eta$ via Tietze extension. We refer for details to [3, Lemma 7 and Thm. 4 proof].

By the Heine-Cantor Theorem, the uniform continuity of F^* implies that we can find η small enough such that $|F^*(a) - F^*(a_\eta)| < \varepsilon/2$ for all $a \in K$. The restriction $F^*|_{K^\eta}$ depends only on the discretized vector $(a(x_1), \dots, a(x_m)) \in \mathbb{R}^m$, hence the map $v \mapsto F^*(a_\eta^v)$ with $a_\eta^v := \sum_{j=1}^m v_j T_j(x)$ is a continuous function on the compact image $\mathcal{E}(K) \subset \mathbb{R}^m$. By the classical Cybenko UAT theorem [1, Thm. 2] applied to this function, there exist a function $\hat{F}^*$ with $M \in \mathbb{N}$ and parameters c_i, θ_i, ξ_{ij} such that

$$|F^*(a_\eta) - \hat{F}^*(a_\eta)| = \left| F^*(a_\eta) - \sum_{i=1}^M c_i \sigma \left(\sum_{j=1}^m \xi_{ij} a(x_j) + \theta_i \right) \right| < \frac{\varepsilon}{2} \quad (4.6)$$

Given that $F^*|_K = F$ and by triangle inequality we obtain the final approximation

$$|F(a) - \hat{F}^*(a_\eta)| = |F^*(a) - \hat{F}^*(a_\eta)| \quad (4.7)$$

$$\leq |F^*(a) - F^*(a_\eta)| + |F^*(a_\eta) - \hat{F}^*(a_\eta)| \quad (4.8)$$

$$< \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon \quad (4.9)$$

□

Having both components in place, we can now proceed to proving the [Theorem 4.1.1](#) as outlined earlier in this section.

Proof of the [Theorem 4.1.1](#) (following [3, Thm. 5]). Given a compact set $K \subset \mathcal{A}$ and continuous operator $\mathcal{G} : \mathcal{A} \rightarrow \mathcal{U}$, the image $V := \mathcal{G}(K) \subset \mathcal{U}$ is compact in $C(D')$. Since V is a compact subset of $C(D')$ and $D' \subset \mathbb{R}^{d'}$ is compact by assumption, we may apply [Lemma 4.1.5](#) with $\Omega = D'$ and the compact family V . Formally, $\exists N \in \mathbb{N}$, parameters $\omega_k \in \mathbb{R}'$, $\zeta_k \in \mathbb{R}$ independent of choice of $a \in K$, non-polynomial continuous activation function σ and continuous linear functionals $c_k : V \rightarrow \mathbb{R}$ such that

$$\left| \mathcal{G}(a)(y) - \sum_{k=1}^N c_k(\mathcal{G}(a)) \sigma(\omega_k \cdot y + \zeta_k) \right| < \frac{\varepsilon}{2} \quad (4.10)$$

for all $a \in K$ and all $y \in D'$.

The (4.10) can be interpreted as a truncated Fourier-type expansion of $\mathcal{G}(a)$ in the sigmoidal basis $\{\sigma(\omega_k \cdot y + \zeta_k)\}_{k=1}^N$ with coefficients $c_k(\mathcal{G}(a))$. Only the coefficients c_k vary with a , the basis is fixed and independent of the choice of the input function $a \in K$.

Next, for each fixed $k \in \{1, \dots, N\}$ define the functionals F_k as

$$F_k : K \rightarrow \mathbb{R}, \quad F_k(a) := c_k(\mathcal{G}(a)) \quad (4.11)$$

As a composition of a continuous operator $\mathcal{G} : K \rightarrow V$ and a continuous linear functional $c_k : V \rightarrow \mathbb{R}$, the functional F_k is continuous. Note, that F_k is generally *non-linear* in a due to non-linearity of the operator $\mathcal{G}$.

This allows us to reduce the problem of approximating operators between Banach spaces to the approximation of N continuous functionals $\{F_k\}_{k=1}^N$ on the compact set $K \subset C(D)$. We can now apply [Lemma 4.1.7](#) to each functional F_k with error bounds $\varepsilon/2NL$, where

$$L := \sum_{k=1}^N \sup_{y \in D'} |\sigma(\omega_k \cdot y + \zeta_k)| \quad (4.12)$$

The lemma provides for each $k \in \{1, \dots, N\}$ a $M_k \in \mathbb{N}$ with sampling points $\{x_i\}_{i=1}^{M_k}$ and real valued parameters $c_i^k, \xi_{ij}^k, \theta_i^k$, which can be unified to a single grid $\{x_i\}_{i=1}^m \subset D$ of size m by taking unions and padding unused weights ξ_{ij}^k with zeros, such that

$$\left| F_k(a) - \sum_{i=1}^M c_i^k \sigma \left(\sum_{j=1}^m \xi_{ij}^k \cdot a(x_j) + \theta_i^k \right) \right| < \frac{\varepsilon}{2NL} \quad (4.13)$$

for all $a \in K$ with $M := \max_k M_k$.

Substituting (4.13) into (4.10) and applying the triangle inequality we obtain the approximation bounds:

$$\left| \mathcal{G}(a)(y) - \sum_{k=1}^N F_k(a) \sigma(\omega_k \cdot y + \zeta_k) \right| = \quad (4.14)$$

$$= \left| \mathcal{G}(a)(y) - \sum_{k=1}^N \sum_{i=1}^M c_i^k \sigma \left(\sum_{j=1}^m \xi_{ij}^k \cdot a(x_j) + \theta_i^k \right) \sigma(\omega_k \cdot y + \zeta_k) \right| \quad (4.15)$$

$$\leq \frac{\varepsilon}{2} + \sum_{k=1}^N \left| F_k(a) - \hat{F}_k(a) \right| \cdot |\sigma(\omega_k \cdot y + \zeta_k)| \quad (4.16)$$

$$\leq \frac{\varepsilon}{2} + N \cdot \frac{\varepsilon}{2NL} \cdot L = \varepsilon \quad (4.17)$$

for all $a \in K$ and all $y \in D'$, where $\hat{F}_k(a)$ denotes the inner sigmoidal sum approximating $F_k(a)$. This concludes the proof. $\square$

4.1.2 Analysis of the Fixed Grid UAT

Remark 4.1.8 (On discretization in the UAT by Chen & Chen). The architecture of the approximating function in [Theorem 4.1.1](#) requires two distinct fixed discretization choices, both of which are determined at design time and neither of them admitting canonical refinement.

First, the input sampling grid $\{x_1, \dots, x_m\} \subset D$ depends on the equicontinuity modulus of the compact set of input functions $K \subset \mathcal{A}$. Refining the grid, e.g. increasing from m to $2m$ sampling points to improve accuracy, results in a different encoder $\mathcal{E}' : K \rightarrow \mathbb{R}^{2m}$, hence a different finite-dimensional target map, and consequently a different approximating neural network with no canonical relationship to the original. The entire architecture must be reconstructed and retrained.

Second, the output basis $\{\sigma(\omega_k \cdot y + \zeta_k)\}_{k=1}^N$ is determined by the equicontinuity modulus of the image $\mathcal{G}(K) \subset \mathcal{U}$. A different choice of N or the weights ω_k, ζ_k produces a different basis, and again no refinement structure relates the resulting approximators to one another.

Remark 4.1.9 (On compactness of the input set $K \subset \mathcal{A}$). The proof of [Theorem 4.1.1](#) is entirely based on two applications of compactness via the Ascoli-Arzelà Theorem: once for $K \subset \mathcal{A}$, to construct the encoder through equicontinuity of the input family, and once for $\mathcal{G}(K) \subset \mathcal{U}$, to construct the output basis through equicontinuity of the output family. Without compactness, neither the uniform δ -net on D nor the uniform choice of output weights ω_k, ζ_k can be guaranteed to work across the entire set of input functions $K \subset \mathcal{A}$ simultaneously.

This reliance on compactness, however, is not specific to the Chen and Chen (1995) [\[3\]](#) proof technique. It is a structural requirement of the *encode-approximate-decode* strategy itself. Any approach that reduces an infinite-dimensional operator approximation problem to a finite-dimensional one via pointwise sampling must control the variation of the set of input functions uniformly, and compactness (or an equivalent condition such as total boundedness) is what provides that control. As we will see in the following sections, subsequent refinements of the Universal Approximation Theorem relax other aspects of the Chen and Chen (1995) [\[3\]](#) construction, such as dependence on a fixed discretization grid, but the compactness assumption on the input set K propagates through all of them. The scope and necessity of this assumption will be examined in detail when we compare the approaches of Kovachki et al. [\[9\]](#) and Lanthaler et al. [\[10\]](#) in [section 4.3](#) and [subsection 4.4.2](#).

Remark 4.1.10 (On the *encode-approximate-decode* pattern). The *encode-approximate-decode* strategy introduced by Chen and Chen bridges the conceptual gap between infinite-dimensional function spaces and the inherently finite-dimensional nature of neural networks.

This three-stage decomposition, where we first encode an infinite-dimensional input into a finite-dimensional representation, apply a finite-dimensional universal approximator, and then decode back into the output function space, is a structural pattern that recurs in every subsequent Universal Approximation result for operators discussed in this thesis.

What varies between approaches is the realization of each stage. For instance, the encoder may be pointwise evaluation on a fixed grid (Chen and Chen, (1995) [3]), projection onto a finite Fourier basis ([9]), or a non-local averaging operator ([10]). The architectural properties demonstrated in subsequent sections, such as Discretization Invariance and generality beyond $C(D)$, arise from different choices at each stage. However, the *encode-approximate-decode* proof strategy remains invariant across the entire literature.

Both discretizations identified in Remark 4.1.8 are consequences of Chen and Chen’s [3] specific instantiation of the *encode-approximate-decode* pattern, namely pointwise evaluation as the encoder and finite sigmoidal expansion as the decoder. Improving the accuracy of the approximation under this instantiation requires not only retraining parameters but also restructuring the architecture itself.

The Chapter 3 of this thesis addresses this limitation directly. There we show how both fixed discretizations can be replaced by a non-local Kernel Integral Operator that acts on the input function $a \in \mathcal{A}$ through integration rather than pointwise evaluation. The integral formulation offers a natural refinement structure, since discretizations of the kernel integral converge to the continuous operator as the quadrature grid is refined, and most importantly, the approximating operator is defined independently of any particular grid. This is the property we call *Discretization Invariance* (Definition 3.2.1), and the explicit convergence rates derived in Chapter 3 quantify how the approximation error decreases as the grid is refined.

In the following sections, we generalize the *encode-approximate-decode* strategy used by Chen and Chen and present the discretization-invariant formulation of the Universal Approximation Theorem due to Kovachki et al. (2023) [9].

4.2 Generalised *Encode-Approximate-Decode* Framework

In this section, we present the generalized Universal Approximation Theorem for Banach spaces with Approximation Property (AP), an important theoretical result that guarantees the existence of the finite dimensional approximation for the operators between Banach spaces and provides a unifying theoretical framework for the approximator architectures observed in applied literature. The proof formalizes the *encode-approximate-decode* strategy and relies on two main components, the Approximation Property of Banach spaces and compactness.

4.2.1 On the role of the Approximation Property of Banach Spaces

We begin by introducing the concept of the Approximation Property and discussing its implications, as it appears to be the foundational component in operator learning that can be traced back to all practical approaches discussed in this thesis. Informally, it guarantees that elements of compact subsets can be approximated by images of finite rank operators.

Definition 4.2.1 (Approximation Property for Banach Spaces; [9, Def. 17, p. 72]). A Banach space X is said to have an Approximation Property (AP) if for any compact set $K \subset X$ and $\varepsilon > 0$ there exists a finite rank operator $T : X \rightarrow X$ such that

$$\|x - Tx\|_X < \varepsilon$$

for all $x \in K$.

An important result that aids in understanding how AP enables interpolation, which is essential to any UAT proof, is that every Banach space that admits a basis has the approximation property [20, Thm. 4.1.33]. In order to illustrate this idea, we will require a formal definition of the Schauder basis:

Definition 4.2.2 (Schauder Basis; [20, def. 4.1.1, p. 350]). Let $\Omega \subset \mathbb{R}^n$ and $X(\Omega)$ a Banach space. A Schauder basis is a sequence of elements $\{\beta_i\}_{i=1}^\infty \subset X$ such that for every $x \in X$ there exists a unique sequence of scalars $\{c_i\}_{i=1}^\infty$, with $c_i \in \mathbb{R} \forall i \in \mathbb{N}$, such that

$$x = \sum_{i=1}^{\infty} c_i \beta_i$$

Remark 4.2.3. Unlike the Hamel basis, the Schauder basis is countable and its elements must be ordered. The coefficients $\{c_i\}_{i=1}^\infty$ from Definition 4.2.2 can be represented as evaluations of a sequence of functionals $\{f_i\}_{i=1}^\infty \subset X^*$ on a Banach space X with Schauder basis $\mathcal{B}(X) := \{\beta_i\}_{i=1}^\infty$, and the expansion $x = \sum_{i=1}^\infty f_i(x) \beta_i$ converges uniformly on every compact subset of X [21, Thm 9.6, p. 115]. Hence, given a compact subset $K \subset X$ and an error tolerance $\varepsilon > 0$, we can find a subset $\{\beta_i\}_{i=1}^N \subset \mathcal{B}(X)$ and a sequence of functionals $\{f_i\}_{i=1}^N \subset X^*$ with $N < \infty$ such that

$$\|x - \sum_{i=1}^N f_i(x) \beta_i\|_X < \varepsilon, \quad \forall x \in K$$

Remark 4.2.4. Recall from the proof of the Theorem 4.1.1 that this approach was used by Chen and Chen in (4.10) to construct the approximation of the operator output as a truncated Fourier-type expansion of $\mathcal{G}(a) \in \mathcal{U}$ in the sigmoidal basis $\{\sigma(\omega_k \cdot y + \zeta_k)\}_{k=1}^N$ with coefficients $c_k(\mathcal{G}(a))$ acting as functionals on the output space $\mathcal{U} \subseteq C(\mathbb{R}^m)$.

We conclude this interlude by presenting practical results that guarantee the existence of approximation property in the spaces of our interest: C , L^p , and $W^{k,p}$. Since those results aim to aid us in proving the generalized UAT, we state them in this work as lemmas.

Lemma 4.2.5 ([20, Thm. 4.1.33]). *Every Banach space with a Basis has the approximation property.*

Remark 4.2.6. Kovachki et al. (2023) verify explicitly that the Approximation Property (def. 4.2.1) holds for $L^p(D)$, $W^{m,p}(D)$, and $C^m(\bar{D})$ on bounded Lipschitz domains [9, Lemma 26]. The argument proceeds by establishing that each of these spaces either has a Schauder basis (which implies the AP by 4.2.5) or is isomorphic to a space that does. In particular, it is well-known result that $L^p(0,1)$ admits a Haar system which is a Schauder basis [20, Example 4.1.27], and the extension to $L^p(D)$ follows from a measure-preserving isomorphism [9, Lemma 26]. The Sobolev cases $W^{m,p}(D)$ are reduced to L^p -type spaces via Stein's extension theorem [9, Lemma 26], and the C^m cases use a classical basis construction due to Ciesielski and Domsta (1972) [22, Thm. 5].

4.2.2 Generalised UAT for Banach Spaces with AP

Theorem 4.2.7 (Generalised Universal Approximation Theorem for Banach Spaces with AP; [9, Lemma 22]). *Let X, Y be Banach spaces with AP and let $\mathcal{G} : X \rightarrow Y$ be a continuous operator. For every compact set $K \subset X$ and $\varepsilon > 0$ there exists $J, J' \in \mathbb{N}$ and continuous*

linear maps $F_J : X \rightarrow \mathbb{R}^J$ and $G_{J'} : \mathbb{R}^{J'} \rightarrow Y$ as well as a continuous map $\varphi \in C(\mathbb{R}^J, \mathbb{R}^{J'})$ such that

$$\sup_{x \in K} \|\mathcal{G}(x) - (G_{J'} \circ \varphi \circ F_J)(x)\|_Y < \varepsilon \quad (4.18)$$

Furthermore there exist functionals $w_1, \dots, w_J \in X^*$ such that F_J has the form

$$F_J(x) = (w_1(x), \dots, w_J(x)), \quad \forall x \in X \quad (4.19)$$

and there exist $\beta_1, \dots, \beta_{J'} \in Y$ such that $G_{J'}$ has the form

$$G_{J'}(v) = \sum_{j=1}^{J'} v_j \beta_j, \quad \forall v \in \mathbb{R}^{J'} \quad (4.20)$$

If Y admits a basis, then $\{\beta_j\}_{j=1}^{J'}$ can be chosen such that there exists an extension $\{\beta_j\}_{j=1}^\infty \subset Y$ which is a basis for Y .

Before presenting the proof, we introduce an auxiliary result without proof that shows the conditions under which an infinite union of compact sets is compact.

Lemma 4.2.8 ([9, Lemma 21]). *Let X, Y be Banach spaces and $T : X \rightarrow Y$ a continuous operator. Let $K \subset X$ be a compact set and $\{F_n : X \rightarrow Y\}_{n=1}^\infty$ a sequence of continuous operators such that*

$$\lim_{n \rightarrow \infty} \sup_{x \in K} \|F(x) - F_n(x)\|_Y = 0$$

Then the set

$$W := \bigcup_{n=1}^\infty F_n(K) \cup F(K)$$

is compact in Y

The proof below follows the argument of [9, Kovachki et al. (2023), Lemma 22]. We present it in expanded form with the construction of the encoder-decoder factorization made explicit.

Proof of Theorem 4.2.7 (following [9, Lemma 22]). First, in order to approximate a continuous operator between two infinite dimensional Banach spaces $G : X \rightarrow Y$ via a finite dimensional function, we need to discretize the input and output Banach spaces X and Y .

We begin with the input Banach space X . Since X has Approximation Property (Definition 4.2.1), for all compact sets $K \subset X$ and $\forall \varepsilon > 0 \exists$ finite rank operator $I_J^X : X \rightarrow X$ with $J \in \mathbb{N}$ such that

$$\|x - I_J^X x\|_X < \varepsilon \quad \forall x \in K.$$

Since the identity operator I_J^X has finite rank, it can be interpreted as constructing an element $x \in K$ as a linear combination of a finite set of elements $\{\alpha_i\}_{i=1}^J$ of X :

$$I_J^X(x) = \sum_{i=1}^J c_i \alpha_i \quad (4.21)$$

Note, if X admits a basis, then $\{\alpha_i\}_{i=1}^J$ is its finite subset.

Since the set $\{\alpha_i\}_{i=1}^J$ is fixed for all $x \in K$, the weights $\{c_i\}_{i=1}^J \in \mathbb{R}^J$ must depend on the input $x \in K$. Hence, we can define them as $c_i = \omega_i(x)$ with $\omega_i \in X^*$, where $\{\omega_i\}_{i=1}^J \in X^*$ is a finite set of functionals. We can re-define (4.21) as:

$$I_J^X(x) = \sum_{i=1}^J \omega_i(x) \alpha_i \quad (4.22)$$

Given that the identity $I_J^X : X \rightarrow X$ has a finite rank, we factorize it as a chain of two operators $I_J^X = G_J^X \circ F_J^X$, with $F_J^X : X \rightarrow \mathbb{R}^J$ being the *encoder*, defined as

$$F_J^X(x) := (\omega_1(x), \dots, \omega_J(x)) \quad (4.23)$$

with $\{\omega_i\}_{i=1}^J \in X^*$, and a *decoder* $G_J^X : \mathbb{R}^J \rightarrow X$, which constructs any element $x \in K$ as a linear combination of the set of elements $\{\alpha_i\}_{i=1}^J \in X$:

$$G_J^X(v) := \sum_{i=1}^J v_i \alpha_i \quad (4.24)$$

where $(v_i)^J \in \mathbb{R}^J$ is a finite input vector of real valued coefficients.

Similarly, we construct the encoder and decoder operators for the output space Y . First we need to extend the compact set K to include the elements $I_J^X(K)$. We define the set

$$Z = \bigcup_{n=1}^{\infty} I_n^X(K) \cup K \quad (4.25)$$

which is compact by [Lemma 4.2.8](#). Note that the definitions of F_J^X and G_J^X as the compact set $K \subset X$ is chosen arbitrarily.

Now, we define the set

$$W := (\mathcal{G} \circ I_J^X)(K) \subset Y \quad (4.26)$$

Given that $\mathcal{G} \circ I_J^X$ is continuous as a product of continuous maps and K is compact, the set W is compact in Y . Note that W depends on the choice of J in I_J^X .

Similar to X , we construct the finite rank operator $I_{J'}^Y : Y \rightarrow Y$, with $F_{J'}^Y : Y \rightarrow \mathbb{R}^{J'}$ and $G_{J'}^Y : \mathbb{R}^{J'} \rightarrow Y$ for the Banach space Y , with

$$F_{J'}^Y(y) := (q_1(y), \dots, q_{J'}(y)), \quad \forall y \in Y \quad (4.27)$$

$$G_{J'}^Y(v) := \sum_{i=1}^{J'} v_i \beta_i, \quad \forall v \in \mathbb{R}^{J'} \quad (4.28)$$

with $\{\beta_i\}_{i=1}^{J'} \in Y$ and $\{q_i\}_{i=1}^{J'} \in Y^*$ such that $I_{J'}^Y = G_{J'}^Y \circ F_{J'}^Y$. If the Banach space Y admits a basis $\mathcal{B}(Y)$, the set $\{\beta_i\}_{i=1}^{J'} \subset \mathcal{B}(Y)$ can be chosen as a finite subset of the basis of Y .

Now that we have defined the discretization of Banach spaces X and Y , we can represent the non-linear transformation $\mathcal{G}$ with a continuous function $\varphi \in C(\mathbb{R}^J, \mathbb{R}^{J'})$, defined as:

$$\varphi := F_{J'}^Y \circ \mathcal{G} \circ G_J^X \quad (4.29)$$

Note, that

$$G_{J'}^Y \circ \varphi \circ F_J^X = I_{J'}^Y \circ \mathcal{G} \circ I_J^X \quad (4.30)$$

We have defined all the components required for the proof. The *encoder* operator F_J^X reduces the input $x \in X$ to a finite-dimensional representation, the *approximator* map $\varphi \in C(\mathbb{R}^J, \mathbb{R}^{J'})$ learns the non-linear transformation defined by the operator $\mathcal{G}$, and the *decoder* operator $G_{J'}^Y$ constructs a vector in Y based on the output weights of φ . It remains to show that the operator $G_{J'}^Y \circ \varphi \circ F_J^X : X \rightarrow Y$ converges to $\mathcal{G}$ for all $x \in K$.

Observe that the continuous operator $\mathcal{G}$ is uniformly continuous for each compact set $Z \in X$, which implies that $\forall \varepsilon > 0 \exists \delta > 0$ such that $\forall x_1, x_2 \in Z$ with $\|x_1 - x_2\|_X < \delta$: $\|\mathcal{G}(x_1) - \mathcal{G}(x_2)\|_Y < \varepsilon/2$. Due to the Approximation Property of Banach space X , we can find a $J > 0$ such that $\|x - I_J^X x\|_X < \delta$, which implies $\|\mathcal{G}(x) - \mathcal{G}(I_J^X x)\|_Y < \varepsilon/2$ for all $x \in Z$.

Now, for the output Banach space Y the Approximation Property guarantees that we can find J' such that $\|y - I_{J'}^Y y\|_Y < \varepsilon/2$ for all $y \in W$. Recall from (4.26) that the set W depends on our choice of J for I_J^X .

Applying the triangle inequality and using the identity $G_{J'}^Y \circ \varphi \circ F_J^X = I_{J'}^Y \circ \mathcal{G} \circ I_J^X$ from (4.30) results in:

$$\|\mathcal{G}(x) - (G_{J'}^Y \circ \varphi \circ F_J^X)(x)\|_Y = \|\mathcal{G}(x) - I_{J'}^Y(\mathcal{G}(I_J^X x))\|_Y \quad (4.31)$$

$$\leq \|\mathcal{G}(x) - \mathcal{G}(I_J^X x)\|_Y + \|\mathcal{G}(I_J^X x) - I_{J'}^Y(\mathcal{G}(I_J^X x))\|_Y \quad (4.32)$$

$$< \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon \quad (4.33)$$

for all $x \in K$. This concludes the proof. $\square$

4.2.3 Discussion on the Error-Decomposition and Applied Architectures

Remark 4.2.9 (Applied Instantiations of the *Encode-Approximate-Decode* Framework). Several prominent operator learning architectures, including the original Chen & Chen network [3], DeepONet [4], PCA-Net [6], and the Fourier Neural Operator [7], can each be interpreted as a specific realization of the composition $(G_{J'} \circ \varphi \circ F_J)$ from Theorem 4.2.7, (4.29). The main difference arises in their choice of basis, encoder functionals, and the approximator function.

Chen and Chen (1995) [3]. The original operator UAT (Theorem 4.1.1) works on $X = C(K_1)$, $Y = C(K_2)$ with compact $K_1, K_2 \subset \mathbb{R}^d$. The encoder F_J evaluates the input function at J fixed sensor points $\{s_1, \dots, s_J\} \subset K_1$, so $F_J(x) = (x(s_1), \dots, x(s_J))$ (4.11) with each $\omega_i = \delta_{s_i}$ being a Dirac evaluation functional. The decoder $G_{J'}$ reconstructs the output as a linear combination $\sum_{k=1}^{J'} v_k \sigma(\omega_k \cdot y + \zeta_k)$ in a sigmoidal basis of $C(K_2)$ (4.10), and the approximator φ is itself a neural network mapping sensor values to the expansion coefficients (4.15).

DeepONet (Lu et al., 2019). The Deep Operator Network [4] separates φ into a *branch net* (processing the discretized input) and a *trunk net* (producing basis functions for the output), which are connected via an inner product. The encoder remains pointwise evaluation at sensor locations, as in Chen & Chen. The main difference is that the trunk net learns the output basis $\{\beta_k\}_{k=1}^{J'}$ instead of fixing it a priori, which can be interpreted as learning a data-adapted decoder $G_{J'}$ end-to-end. The existence guarantee of Theorem 4.2.7 applies to this architecture since learned basis functions in Y are still elements of Y .

PCA-Net (Bhattacharya et al., 2021). The PCA-Net [6] restricts to Hilbert spaces $X = Y = L^2$, where the Schauder basis (Definition 4.2.2) is replaced by a PCA basis which is derived from the data. The encoder F_J projects onto the first J empirically derived eigenfunctions of the input covariance via inner products $\omega_i(x) = \langle x, \phi_i^X \rangle$, and the decoder $G_{J'}$ reconstructs in the leading J' eigenfunctions of the output covariance. The approximator φ is a feed forward network between the two latent spaces $\mathbb{R}^J$ and $\mathbb{R}^{J'}$. Since Hilbert spaces admit an orthonormal basis [18, Thm. 6.5.13, p. 230] (and hence the AP by Lemma 4.2.5), this architecture fits within the framework defined in Theorem 4.2.7.

Fourier Neural Operator (Li et al., 2020). The FNO (introduced in [7] and formalised in [8]) operates on $L^2(\mathbb{T}^d)$, where $\mathbb{T}^d := [0, 1]^d$ denotes the d -dimensional torus with periodic boundary conditions. It uses the Fourier basis $\{s \mapsto e^{2\pi i k \cdot s}\}_{k \in \mathbb{Z}^d}$ as the fixed orthonormal

basis for both encoding and decoding. The encoder F_J truncates the input function $a \in \mathcal{A}$ to the first J Fourier modes, the approximator φ applies learnable linear transformations in Fourier space (pointwise multiplication of mode coefficients), and the decoder $G_{J'}$ reconstructs via the inverse discrete Fourier transform. This is an instance of [Theorem 4.2.7](#) on $L^2(\mathbb{T}^d)$, where the Fourier system is the canonical orthonormal Schauder basis.

In all cases, the abstract result in [Theorem 4.2.7](#) guarantees existence of a sufficient finite-dimensional factorization. The applied architectures differ in three main components: (i) the *choice of basis* (fixed analytic [\[3\]](#), learned [\[4\]](#), or data-driven [\[6\]](#)), (ii) the *form of the encoder functionals* (pointwise evaluation [\[3\]](#), inner products [\[4\]](#), or integral transforms [\[8\]](#)), and (iii) the *parametrization of φ* (shallow network [\[3\]](#), deep network [\[6\]](#) [\[4\]](#), or Kernel Integral Operator [\[8\]](#)).

Remark 4.2.10 (On the Error Structure and the Role of the Output Bottleneck). The non-constructive proof of [Theorem 4.2.7](#) shows how the approximation error due to discretization can be decomposed into two components, both governed by the Approximation Property applied separately to X and Y . The first term in [\(4.31\)](#), $\|\mathcal{G}(x) - \mathcal{G}(I_J^X x)\|_Y$, is the *encoding error*: compactness of the input set K forces $\mathcal{G}$ to be uniformly continuous on K , which lets us translate a small input deviations $\|x - I_J^X x\|_X < \delta$ into a small output deviations uniformly over K . The second term, $\|\mathcal{G}(I_J^X x) - I_{J'}^Y(\mathcal{G}(I_J^X x))\|_Y$, is the *decoding error*: it measures how well the finite-rank reconstruction of the output function in Y captures the image of the already-discretized operator, controlled by the AP of Y applied to the compact image set $W := (\mathcal{G} \circ I_J^X)(K)$.

Note, that the finite-dimensional map $\varphi = F_{J'}^Y \circ \mathcal{G} \circ G_J^X$ does not introduce an explicit additional error in this existence result, since it is defined exactly. In practical architectures discussed in [Remark 4.2.9](#), φ is unknown and must be approximated by a parametric model $\hat{\varphi} \approx \varphi$ (e.g. a neural network), which introduces a third error term controlled by classical finite-dimensional universal approximation results [\[2, 1\]](#).

Furthermore, the decoding error is not always defined as an independent source of error in practical architectures, as the decoding step is often combined with the approximation of non-linear transformation φ by design. In all bottleneck architectures, such as Chen & Chen [\[3\]](#), DeepONet [\[4\]](#), PCA-Net [\[6\]](#), the output of $\hat{\varphi}$ is defined in the fixed span of finite basis $\{\beta_j\}_{j=1}^{J'}$ by construction, and the network is trained to minimize the error $\|(G_{J'} \circ \chi \circ F_J)(x) - \mathcal{G}(x)\|_Y$ from the data. This gives the appearance that the decoding error is absorbed into the approximation error of $\hat{\varphi}$.

However, the decoding error controlled by the AP of Y defines an effective lower bound on the approximation error in bottleneck architectures: if a non-trivial component of $\mathcal{G}(x)$ lies in the orthogonal complement of $\text{span}\{\beta_j\}_{j=1}^{J'}$, no choice of the approximation function $\hat{\varphi}$ can recover it. In PCA-Net [\[6\]](#), this lower bound is precisely the tail $\sum_{j>J'} \lambda_j^Y$ of the output covariance spectrum, and in Chen and Chen [\[3\]](#) and DeepONet [\[4\]](#) it is controlled by the expressiveness of the chosen output basis, defined as sigmoidal in the former, and as learned trunk functions in the latter. Hence, even assuming that the non-linear transformation φ can be approximated arbitrarily well, the approximation error for the operator $\mathcal{G}$ remains limited by the choice of basis $\{\beta_j\}_{j=1}^{J'}$ used to reconstruct the output function.

The Neural Operator framework of Kovachki et al. [\[9\]](#), presented in the next chapter, removes this bottleneck by replacing the fixed finite-rank factorization through $\mathbb{R}^{J'}$ with a kernel integral layer ([Definition 2.2.14](#)) that acts directly on function spaces. As a consequence, the error decomposition simplifies to two terms: the *input discretization error* and the *operator approximation error*. This structure is reflected in the explicit convergence rates that we derived in [Corollary 3.4.3](#), where the operator approximation error is decomposed into a Hölder regularity contribution from the discretization of the input space and a Lipschitz kernel approximation contribution from the Neural Operator ([Definition 2.2.24](#)) architecture, effectively removing the lower bound on the latter defined

by the output truncation.

4.3 Continuous Non-Locality Guarantees Discretization Invariance

In this section we show how the Kernel Integral Operator (Definition 2.2.14) provides a realization of the *encode-approximate-decode* framework (Remark 4.1.10) that achieves Discretization Invariance (Definition 3.2.3) through continuous non-locality. We first outline the Universal Approximation Theorem for Kernel Integral Operators due to Kovachki *et al.* [9] in Theorem 4.3.1, compressing the key ideas of their proof into the *encode-approximate-decode* structure established in the preceding section. In Corollary 4.3.5, we extend this result from compact to bounded subsets of Sobolev spaces, yielding explicit approximation guarantees under weaker hypotheses on the input set, consistent with the generalized Discretization Invariance rates established in Corollary 3.4.3. We conclude with a discussion in Remark 4.3.7 of how continuous non-locality via kernel integration enables Discretization Invariance, and show that the fixed-grid approach of Chen and Chen (Theorem 4.1.1) can be viewed as a special case of this generalized framework.

4.3.1 Universal Approximation via Integral Kernel Operators

Theorem 4.3.1 (Discretization-Invariant UAT; [9, Thm. 11]). *Let $D \subset \mathbb{R}^d$ and $D' \subset \mathbb{R}^{d'}$ be bounded Lipschitz domains. Let the input space $\mathcal{A}$ be one of $L^{p_1}(D)$ or $W^{k_1, p_1}(D)$ for some $1 \leq p_1 < \infty$, $k_1 \in \mathbb{N}$ or $C(\bar{D})$, and let the output space $\mathcal{U}$ be one of $L^{p_2}(D')$ or $W^{k_2, p_2}(D')$ for some $1 \leq p_2 < \infty$, $k_2 \in \mathbb{N}$ or $C(\bar{D}')$. Let $\mathcal{G}^\dagger : \mathcal{A} \rightarrow \mathcal{U}$ be a continuous operator and let the activation function σ in NO_T (Definition 2.2.27) be globally Lipschitz and non-polynomial. Then for any compact set $K \subset \mathcal{A}$ and $0 < \varepsilon \leq 1$, there exist $T \in \mathbb{N}$ and a Neural Operator $\mathcal{G} \in \text{NO}_T$ such that*

$$\sup_{a \in K} \|\mathcal{G}^\dagger(a) - \mathcal{G}(a)\|_{\mathcal{U}} \leq \varepsilon. \quad (4.34)$$

Remark 4.3.2. The proof of Theorem 11 in Kovachki et al. (2023) [9] is distributed across six core lemmas (Lemmas 21, 22, 26, 32, 34, 35) and several auxiliary results (Lemmas 18–20, 27–30) spanning approximately 30 pages of appendices. The auxiliary lemmas establish functional-analytic foundations, such as Riesz-type representation theorems for $W^{k,p}$ [9, Lemma 28] and C^m [9, Lemma 27] spaces, mollification arguments for approximating Radon measures by smooth kernels [9, Lemma 30], and verification of the Approximation Property for the relevant Banach spaces via Schauder basis theory and extension operators [9, Lemmas 18, 19, 20]. The core lemmas then show that each component of the *encode-approximate-decode* factorization from Theorem 4.2.7 can be realized within the Neural Operator architecture using integral kernel operators with neural-network-parametrized kernels.

In the following proof outline, we consolidate the logic of these results into a single argument, organized around the three components of the factorization. Instead of restating each lemma separately, we focus on *how* the Kernel Integral Operator (Definition 2.2.14) serves as the unifying architectural component that enables all three steps, namely encoding, transformation, and decoding, while guaranteeing the Discretization Invariance (Definition 3.2.3). We demonstrate the proof ideas on the example of $\mathcal{A} = L^p(D)$ and invoke the auxiliary functional-analytic results for $W^{k,p}(D)$ and $C(\bar{D})$ spaces with references to the corresponding lemmas by Kovachki et al. [9] when required.

Proof outline. The proof is based on the existence result from Theorem 4.2.7 and proceeds by parametrizing each component of the $\tilde{\mathcal{G}} \approx G_{J'} \circ \varphi \circ F_J$ *encode-approximate-decode*

architecture with Kernel Integral operators (Definition 2.2.14), which provide continuous non-locality that guarantees Discretization Invariance. The total error budget is divided in two between the asymptotic result from Theorem 4.2.7 and the approximation of the theoretical operator $\tilde{\mathcal{G}}$ with $\mathcal{G} \in NO_T$, which we focus on in this proof outline.

Step 1. Abstract factorization via Theorem 4.2.7. Since $L^{p_1}(D)$, $W^{k_1, p_1}(D)$ and $C(\bar{D})$ have Approximation Property (Definition 4.2.1) as discussed in Remark 4.2.6, by Theorem 4.2.7 there exist continuous maps $F_J : \mathcal{A} \rightarrow \mathbb{R}^J$, $G_{J'} : \mathbb{R}^{J'} \rightarrow \mathcal{U}$ and $\varphi : \mathbb{R}^J \rightarrow \mathbb{R}^{J'}$ such that

$$\sup_{a \in K} \|\mathcal{G}^\dagger(a) - (G_{J'} \circ \varphi \circ F_J)(a)\|_{\mathcal{U}} < \frac{\varepsilon}{2} \quad (4.35)$$

where the map F_J is realized as J continuous functionals $c_j \in \mathcal{A}^*$ (4.23) and $G_{J'}$ reconstructs elements of $\mathcal{U}$ based on fixed J' representing elements $\{\beta_j\}_{j=1}^{J'} \subset U$, as defined in (4.28). We denote $\tilde{\mathcal{G}} := G_{J'} \circ \varphi \circ F_J$.

Step 2. Encoding input via a Kernel Integral Operator. The encoder $F : \mathcal{A} \rightarrow \mathbb{R}^J$ is defined by $F(a) := (c_1(a), \dots, c_J(a))$ where each $c_j \in \mathcal{A}^*$. We first show that F can be approximated by an integral kernel operator on the example of $\mathcal{A} = L^p(D)$ and then discuss how this result is transferred to $W^{k, p}$ and $C(\bar{D})$ spaces.

Fix $\delta > 0$. For $\mathcal{A} = L^p(D)$, the Riesz Representation Theorem [18, Thm. 5.1.27] represents each functional as $c_j(a) = \int_D v_j a \, dx$ for a unique $v_j \in L^q(D)$ with $\frac{1}{p} + \frac{1}{q} = 1$. Since $C_0^\infty(D)$ is dense in $L^q(D)$ [18, Thm. 5.1.23], there exist $f_j \in C_0^\infty(\bar{D})$ with $\|f_j - v_j\|_{L^q} < \delta/(2M)$, where $M := \sup_{a \in K} \|a\|_{L^p}$. By Hölder's inequality:

$$\left| c_j(a) - \int_D f_j a \, dx \right| = \left| \int_D (v_j - f_j) a \, dx \right| \leq \|v_j - f_j\|_{L^q} \|a\|_{L^p} < \frac{\delta}{2}. \quad (4.36)$$

Since $\bar{D}$ is compact, $f_j|_{\bar{D}}$ is uniformly continuous, and the classical Universal Approximation Theorem¹ guarantees that there exist a neural network $\tilde{f}_j$ with non-polynomial activation such that $\sup_{x \in \bar{D}} |f_j(x) - \tilde{f}_j(x)| \leq \delta/(2M|D|^{1/q})$. Applying Hölder's inequality gives:

$$\left| \int_D (f_j - \tilde{f}_j) a \, dx \right| \leq \|f_j - \tilde{f}_j\|_{L^\infty} |D|^{1/q} \|a\|_{L^p} < \frac{\delta}{2}. \quad (4.37)$$

Combining both approximations from (4.36) and (4.37) and applying the triangle inequality:

$$\left| c_j(a) - \int_D \tilde{f}_j a \, dx \right| < \delta, \quad \forall a \in K. \quad (4.38)$$

Finally, define the integral kernel $\kappa : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^{J \times 1}$ by stacking:

$$\kappa(y, x) = \begin{pmatrix} \tilde{f}_1(x) \\ \vdots \\ \tilde{f}_J(x) \end{pmatrix} \quad (4.39)$$

Note, that the kernel κ does not depend on y , and the weights in the neural network associated with the first argument are set to zero. We can now define the neural network based encoder approximation F_δ as:

$$F_\delta(a)(y) := \int_D \kappa(y, x) a(x) \, dx = \begin{pmatrix} \int_D \tilde{f}_1(x) a(x) \, dx \\ \vdots \\ \int_D \tilde{f}_J(x) a(x) \, dx \end{pmatrix} \approx F(a) \quad (4.40)$$

¹Any of Cybenko (1989) [1], Hornik et al. (1989) [2], or Pinkus (1999)[23] applies here. Kovachki et al.[9] invoke Pinkus.

which is a spatially constant $\mathbb{R}^J$ -valued function that outputs the same vector for every $y \in \bar{D}$, hence it creates a consistent encoding of any function $a \in K$ on the discretized $\mathbb{R}^J$ grid. The y -independence is a necessary architecture choice as it allows to apply a standard UAT for approximating the continuous transformation φ from (4.35).

Step 3. Approximating continuous transformation φ with classical UAT. In the first step we defined our approximation function as a composition $\tilde{\mathcal{G}} := G_{J'} \circ \varphi \circ F_J$, where $\varphi : \mathbb{R}^J \rightarrow \mathbb{R}^{J'}$ describes the non-linear continuous transformation. Since φ is a continuous function between finite dimensional spaces, by the classic UAT [1] it can be approximated via a neural network with non-polynomial activation function. Kovachki et al. leverage this result by showing that a Neural Operator (Definition 2.2.24) is a generalization of the standard feed-forward neural network and hence the classic UAT applies [9, Lemma 35].

By choosing a kernel with no spatial dependence $\kappa^{(t)}(x, y) = W_t/|D|$ with $W_t \in \mathbb{R}^{d' \times d}$ and a constant bias $b_t(x) = b_t$, for a constant input function $v = w\mathbf{1}$ with $w \in \mathbb{R}^J$ the Kernel Integral operator (Definition 2.2.14) reduces to

$$\int_D \kappa^{(t)}(x, y)v(y)dy = \int_D \kappa^{(t)}(x, y)w\mathbf{1} dy = \frac{W_t w}{|D|} \cdot |D| \quad (4.41)$$

which is a standard neural network layer $w \mapsto \sigma(W_t w + b_t)$. Hence, for every feed forward neural network ϕ there exists a Neural Operator $\mathcal{G} \in NO_T$ such that

$$\phi(w) = \mathcal{G}(w\mathbf{1})(x), \quad \forall w \in \mathbb{R}^J, \forall x \in D \quad (4.42)$$

Recall that in *Step 2* we defined the encoder F_δ ((4.40)) as a constant map on $\bar{D}$, hence the continuous transformation function φ can be approximated via a Neural Operator $\mathcal{G} \in NO_T$ [9, Lemma 35]. We denote the resulting approximation function of φ with $\tilde{\varphi} : \mathbb{R}^J \rightarrow \mathbb{R}^{J'}$.

Step 4. Decoding output via a Kernel Integral Operator. The decoder function $G : \mathbb{R}^{J'} \rightarrow \mathcal{U}$ reconstructs an element of Banach space $\mathcal{U}$ as a linear combination of J' basis elements $\{\beta_j\}_{j=1}^{J'} \subset \mathcal{U}$, as guaranteed by the Approximation Property (Lemma 4.2.5). The challenge we address in this step is how to define the map G as a Kernel Integral Operator. The approach mirrors the idea described in *Step 2*, with a key difference that while in (4.39) the kernel $\kappa(y, x)$ was defined as a non-local function constant in $y \in \bar{D}$, the output kernel $\kappa^{out}(y, x)$ is designed such that it ignores the integral variable x and produces a spatially varying output in $\mathcal{U}$.

Formally, given the transformed vector of coefficients $w \in \mathbb{R}^{J'}$ with $w = \tilde{\varphi}(F_\delta(a))$ and a subset of functions $\{\beta_j\}_{j=1}^{J'} \subset \mathcal{U}$, we can approximate the function

$$G(w) = \sum_{j=1}^{J'} w_j \beta_j \quad (4.43)$$

by defining the kernel $\kappa^{out} : \bar{D}' \times \bar{D}' \rightarrow \mathbb{R}^{1 \times J'}$ as a row vector:

$$\kappa^{out}(y, x) = \frac{1}{|\bar{D}'|} [\tilde{\beta}_1(y), \dots, \tilde{\beta}_{J'}(y)] \quad (4.44)$$

where each $\tilde{\beta}_j$ is a neural network approximation of basis element β_j , as we will discuss below. Since κ^{out} is independent of the integral variable x , integrating against a constant function $v = w\mathbf{1}$ yields:

$$\int_{D'} \kappa^{out}(y, x)w\mathbf{1}(x)dx = \frac{1}{|D'|} \sum_{j=1}^{J'} w_j \tilde{\beta}_j(y) \int_{D'} 1dx \quad (4.45)$$

$$= \sum_{j=1}^{J'} w_j \tilde{\beta}_j(y) \quad (4.46)$$

which results in an element of $\mathcal{U}$ that approximates $G(w) = \sum_{j=1}^{J'} w_j \beta_j$.

The approximation of the basis elements $\{\beta_j\}_{j=1}^{J'} \subset \mathcal{U}$ differs from the approach used in *Step 2*, where we rely on the Riesz Representation Theorem and approximate the elements of the dual space $\mathcal{A}^*$, in a way that we are approximating the elements of the space $\mathcal{U}$ directly. By the density of smooth functions in $\mathcal{U}$, each basis element β_j is approximated by $\hat{\beta}_j \in C^\infty(\bar{D}')$, cf. Kovachki et al. (2023) [9, Lemma 34]. Note, for our example with $\mathcal{U} = L^p(D)$ this follows directly from the density of smooth functions with compact support C_c^∞ in L^p [18, Thm. 5. 1. 23] and the fact that $C_c^\infty(D') \subset C^\infty(\bar{D}')$. We then invoke Pinkus (1999)[23] to approximate each $\hat{\beta}_j \in C^\infty(\bar{D}')$ with a NN $\tilde{\beta}_j$ with non-polynomial activation function. Denoting the decoder approximation from (4.45) as $\tilde{G} : \mathbb{R}^{J'} \rightarrow \mathcal{U}$ we obtain

$$\|G(w) - \tilde{G}(w)\| = \left\| \sum_{j=1}^{J'} w_j \beta_j - \tilde{\beta}_j \right\| \quad (4.47)$$

$$\leq \|w\|_1 \cdot \max_j \|\beta_j - \tilde{\beta}_j\|_{\mathcal{U}} \quad (4.48)$$

$$\leq \|w\|_1 \cdot \max_j (\|\beta_j - \hat{\beta}_j\|_{\mathcal{U}} + \|\hat{\beta}_j - \tilde{\beta}_j\|_{\mathcal{U}}) \quad (4.49)$$

$$\leq \|w\|_1 \left(\frac{\delta}{2} + \frac{\delta}{2} \right) \quad (4.50)$$

We need to bound $w = \tilde{\varphi}(F_\delta(a))$ uniformly on K because it depends on the choice of $a \in K$. Since K is compact and $\tilde{\varphi} \circ F_\delta$ is continuous, $C := \sup_{a \in K} \|\tilde{\varphi}(F_\delta(a))\|_1 < \infty$, and choosing $\delta < \varepsilon'/C$ gives the desired bound.

Step 5. Combining error bounds. The total approximation error decomposes via the triangle inequality into the factorization error from *Step 1* and the realization error from *Steps 2-4*:

$$\sup_{a \in K} \|\mathcal{G}^\dagger(a) - \mathcal{G}(a)\|_{\mathcal{U}} \leq \underbrace{\sup_{a \in K} \|\mathcal{G}^\dagger(a) - \tilde{\mathcal{G}}(a)\|_{\mathcal{U}}}_{< \varepsilon/2 \text{ by Step 1 (eq. 4.35)}} + \underbrace{\sup_{a \in K} \|\tilde{\mathcal{G}}(a) - \mathcal{G}(a)\|_{\mathcal{U}}}_{\text{realization error from Steps 2-4}} \quad (4.51)$$

The realization error arises from three independent approximations, each controlled by the error tolerance $\delta > 0$: the encoding of functionals $c_j \in \mathcal{A}^*$ via integral kernels (*Step 2*, (4.38)), the approximation of the continuous transformation φ via a neural network embedded in the NO architecture (*Step 3*), and the reconstruction of basis elements $\beta_j \in \mathcal{U}$ via neural network approximations (*Step 4*, (4.45)). Since each of these three approximation errors can be made arbitrarily small by choosing δ sufficiently small independently of one another, there exists a $\delta > 0$ such that their combined contribution is bounded by $\varepsilon/2$. Together with Equation 4.35, this yields

$$\sup_{a \in K} \|\mathcal{G}^\dagger(a) - \mathcal{G}(a)\|_{\mathcal{U}} < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon \quad (4.52)$$

which completes the proof outline. $\square$

4.3.2 Relaxing Compactness to Boundedness for Sobolev Spaces

Remark 4.3.3 (On the Role of Compactness). Compactness of the function set $K \subset \mathcal{A}$ plays a central role in the proof of Theorem 4.3.1 as several steps in the proof directly rely on it. Compactness guarantees total boundedness, which provides a finite ε -net of representative functions over the set K required to obtain a uniform truncation of Schauder

basis expansion via the Approximation Property (Definition 4.2.1). Additionally, the application of the classic Universal Approximation Theorem [1] throughout Steps 2-4 of the proof requires the input set K to be compact.

As argued in Remark 3.3.5, compactness of a set $K \subset \mathcal{A}$ in an infinite-dimensional space is a strong assumption, as by the Riesz' Theorem a closed and bounded ball in infinite dimensional space is not compact [18, Thm. 2.3.2, p. 60]. In Theorem 3.3.10 we extended the Discretization Invariance results of Kovachki et al [9] from compact to bounded sets in Sobolev spaces with enough regularity. However, the Universal Approximation guarantees of Theorem 4.3.1 are only provided for compact sets on Sobolev spaces $W^{k,p}(D)$ with $p \in [1, \infty)$ and $kp > d$. Similar to our approach in the proof of Theorem 3.3.10, we can relax the compactness assumption to boundedness in Sobolev spaces with enough regularity by invoking the Sobolev Embedding Theorem combined with the Morrey Inequality [15, Thm. 13, p. 302], followed by the Ascoli-Arzelà Theorem [18, Thm. 5.4.2, p. 196] to obtain a pre-compact set in $C(\bar{D})$, as presented in the subsequent corollary. However, this approach comes at a cost of imposing additional assumption of uniform continuity of the target operator $\mathcal{G}^\dagger$ in $C(\bar{D})$ with sup-norm topology on the input space $\mathcal{A}$, thus establishing a tradeoff.

Remark 4.3.4. The following corollary extends the UAT guarantee by Theorem 4.3.1 [9, Thm. 11] to bounded sets on Sobolev input spaces with explicit regularity assumptions. This result is original to this thesis.

Corollary 4.3.5 (Discretization-Invariant UAT for bounded sets on Sobolev Spaces). *Let $D \subset \mathbb{R}^d$ and $D' \subset \mathbb{R}^{d'}$ be bounded Lipschitz domains. Let $\mathcal{A}$ be a Sobolev space $W^{k_1,p_1}(D)$ for some $1 \leq p_1 < \infty$, $k_1 \in \mathbb{N}$ with $k_1 p_1 > d$, and let the output space $\mathcal{U}$ be one of $L^{p_2}(D')$ or $W^{k_2,p_2}(D')$ for some $1 \leq p_2 < \infty$, $k_2 \in \mathbb{N}$ or $C(\bar{D}')$. Let the target operator be $\mathcal{G}^\dagger : \mathcal{A} \rightarrow \mathcal{U}$. Assume additionally that $\mathcal{G}^\dagger$ is **uniformly continuous in $C(\bar{D})$ with sup-norm topology** on $\mathcal{A}$. Let the activation function σ in NO_T (Definition 2.2.27) be globally Lipschitz and non-polynomial. Then for any **bounded** set $K \subset \mathcal{A}$ and $0 < \varepsilon \leq 1$, there exist $T \in \mathbb{N}$ and a Neural Operator $\mathcal{G} \in \text{NO}_T$ such that*

$$\sup_{a \in K} \|\mathcal{G}^\dagger(a) - \mathcal{G}(a)\|_{\mathcal{U}} \leq \varepsilon. \quad (4.53)$$

Proof. Given a bounded open domain $D \subset \mathbb{R}^d$ with a Lipschitz continuous boundary and the Sobolev space $\mathcal{A} = W^{k_1,p_1}(D)$ with $k_1 > d/p_1$, by Sobolev Embedding Theorem [15, Cor. E.14, p. 302] and Morrey's Inequality [15, Thm E.13, p. 302] we obtain continuous embedding $\mathcal{A} \hookrightarrow C^{0,\alpha}(\bar{D})$ with $\alpha = \left\lceil \frac{d}{p_1} \right\rceil + 1 - \frac{d}{p_1} \in (0, 1]$ (for $d/p_1 \notin \mathbb{Z}$). Furthermore, the Morrey's Inequality guarantees that there exists a constant $C > 0$ such that

$$\|a\|_{C^{0,\alpha}(\bar{D})} \leq C \|a\|_{W^{k_1,p_1}}, \quad \forall a \in \mathcal{A} \quad (4.54)$$

Given that K is bounded, there exists a uniform bound $M := \sup_{a \in K} \|a\|_{\mathcal{A}}$, which, combined with the equation above implies that all $a \in K$ are *uniformly bounded* in $C^{0,\alpha}(\bar{D})$:

$$\sup_{x \in \bar{D}} |a(x)| \leq CM, \quad \forall a \in K \quad (4.55)$$

By the definition of Hölder continuity [19, Ch. 5.1, p. 254], there exists a constant L such that for all $a \in C^{0,\alpha}(\bar{D})$ and $x, y \in \bar{D}$ it holds

$$|a(x) - a(y)| \leq L \|x - y\|^\alpha \quad (4.56)$$

Note, that L depends on a . The smallest such L is the Hölder seminorm of a defined as

$$[a]_{C^{0,\alpha}(\bar{D})} := \sup_{\substack{x, y \in \bar{D} \\ x \neq y}} \frac{|a(x) - a(y)|}{\|x - y\|^\alpha} \quad (4.57)$$

and the Hölder norm is respectively

$$\|a\|_{C^{0,\alpha}(\bar{D})} := \|a\|_{C(\bar{D})} + [a]_{C^{0,\alpha}(\bar{D})} \quad (4.58)$$

[19, Ch. 5.1, p. 254]. Due to $\sup \|a\|_{W^{k_1,p_1}} \leq M$ and (4.55) we obtain the uniform bound

$$|a(x) - a(y)| \leq [a]_{C^{0,\alpha}(\bar{D})} \|x - y\|^\alpha \quad (4.59)$$

$$\leq \|a\|_{C^{0,\alpha}(\bar{D})} \|x - y\|^\alpha \quad (4.60)$$

$$\leq C \|a\|_{W^{k_1,p_1}} \|x - y\|^\alpha \quad (4.61)$$

$$\leq CM \|x - y\|^\alpha \quad (4.62)$$

Hence, for each $\varepsilon > 0$ we can find $\delta > 0$ with $\delta = (\varepsilon/CM)^{1/\alpha}$ such that for all $x, y \in \bar{D}$ with $\|x - y\| < \delta$ it holds

$$|a(x) - a(y)| \leq CM \|x - y\|^\alpha < CM \delta^\alpha = \varepsilon, \quad \forall a \in K \quad (4.63)$$

which implies that the bounded set K is *equicontinuous* in $C^{0,\alpha}(\bar{D})$.

The closed and bounded domain $\bar{D} \subset \mathbb{R}^d$ with $d < \infty$ is compact by the Heine-Borel Theorem. Pointwise boundedness of $K \subset C^{0,\alpha}(\bar{D})$ follows from uniform boundedness shown in (4.55). Due to the inclusion $C^{0,\alpha}(\bar{D}) \subset C(\bar{D})$, the set K is equicontinuous and pointwise bounded in $C(\bar{D})$, and by Ascoli-Arzelà Theorem [18, 5.4.2, p.196] the set K is precompact in $C(\bar{D})$ on the compact domain $\bar{D}$. Since $C(\bar{D})$ is a complete Banach space in the supremum norm, the closure $\bar{K}$ is compact.

We have shown $\bar{K}$ is compact in $C(\bar{D})$ and $K \subseteq \bar{K}$. By hypothesis $\mathcal{G}^\dagger$ is uniformly continuous in $C(\bar{D})$ with sup-norm topology over $\mathcal{A}$. However, the compact set $\bar{K} \subset C(\bar{D})$ may contains points $C(\bar{D}) \setminus W^{k_1,p_1}(D)$ where $\mathcal{G}^\dagger$ is not defined. Since K is dense in its closure $\bar{K}$, the output space $\mathcal{U}$ is a Banach space and hence complete, and $\mathcal{G}^\dagger$ is uniformly continuous on K , the continuous operator $\tilde{\mathcal{G}}^\dagger : \bar{K} \rightarrow \mathcal{U}$ extends $\mathcal{G}^\dagger$ uniquely onto the closure $\bar{K}$ [18, Thm. 1.5.2, p. 23].

Applying Theorem 4.3.1 with input space $\mathcal{A} = C(\bar{D})$ and compact set $\bar{K}$ gives us a Neural Operator $\mathcal{G} \in NO_T$ with

$$\sup_{a \in \bar{K}} \|\tilde{\mathcal{G}}^\dagger(a) - \mathcal{G}(a)\|_{\mathcal{U}} \leq \varepsilon. \quad (4.64)$$

Since $K \subseteq \bar{K}$ and $\mathcal{G}^\dagger(a) = \tilde{\mathcal{G}}^\dagger(a)$ for all $a \in K$, the bound from the corollary statement (4.53) follows. $\square$

In Theorem 3.3.10 we have extended the Discretization Invariance result for Neural Operators from compact to bounded sets on Sobolev spaces with continuous true operator $\mathcal{G}^\dagger : C(\bar{D}) \rightarrow \mathcal{U}$. The result of Corollary 4.3.5 guarantees that the Neural Operator $\mathcal{G} \in NO_T$ that we are discretizing is able to approximate the true operator $\mathcal{G}^\dagger$ on those bounded sets under an additional assumption of the continuity of $\mathcal{G}^\dagger$. The total error can be decomposed into two main components, the approximation error and discretization error. We were able to derive explicit error rates for the discretization error in Corollary 3.4.3. A natural question is whether we can achieve the same for the approximation error. We address this in the following remark.

Remark 4.3.6 (On non-constructive nature of Universal Approximation Theorem). As discussed in Remark 4.3.3, the proof of Theorem 4.3.1 relies on the Approximation Property of Banach spaces (Definition 4.2.1) as well as the classical Universal Approximation Theorem [1], both of which are non-constructive proofs. The proof of Theorem 4.2.7, which

guarantees the factorization of the Neural Operator in [Theorem 4.3.1](#), utilizes the Approximation Property of input and output Banach spaces to find finite truncations of a Schauder basis. As per [Definition 4.2.1](#), the Approximation Property only guarantees the existence of finite-rank approximations of the identity on compact sets but provides no bound on the rank required for a given $\varepsilon > 0$. Similarly, the classic Universal Approximation Theorem in its standard formulations (Cybenko (1989) [1], Hornik (1989) [2], Pinkus (1999) [23]) guarantees that a continuous real valued function can be approximated via single layer neural networks of arbitrary width, but none of these provide explicit convergence rates. Consequently, due to the reliance on two non-constructive results, our [Corollary 4.3.5](#) can only guarantee *existence* of the approximation for bounded sets on Sobolev spaces under additional assumptions, without providing explicit convergence rates.

4.3.3 Integration as the Mechanism for Discretization Invariant Approximation

Remark 4.3.7 (Continuous non-locality guarantees Discretization Invariance). In the *Step 2* of [Theorem 4.3.1](#) proof we defined the encoder operator $F : \mathcal{A} \rightarrow \mathbb{R}^J$ as a vector of functionals $F(a) := (c_1(a), \dots, c_J(a))$ with each $c_j \in \mathcal{A}^*$ and showed on the example of $\mathcal{A} = L^p(D)$ that by applying the Riesz Representation Theorem followed by the density argument of smooth functions with compact support $C_0^\infty(D)$ in $\mathcal{A}^* = L^q(D)$ we can represent each functional $c_j \in \mathcal{A}^*$ as an integration operator

$$c_j(a) \approx \int_D f_j(x) a(x) d\mu(x), \quad \forall a \in \mathcal{A} \quad (4.65)$$

where $f_j \in C_0^\infty(D)$ and μ is a Lebesgue measure. Hence, the functional c_j evaluates elements of the Banach space $a \in \mathcal{A}$ directly, and by continuity of the integral the expression in (4.65) is evaluated across the whole domain D , the property that we previously referred to as *continuous non-locality*.

Recall from [Theorem 4.3.1](#) proof that the continuous function f_j from (4.65) is further approximated by a neural network $\tilde{f}$ with a non-polynomial activation σ . The neural network $\tilde{f}_j$ is a function $\tilde{f}_j : \mathbb{R}^d \rightarrow \mathbb{R}$ defined as

$$\tilde{f}_j(x) = W_n \sigma(\dots W_1 \sigma(W_0 x + b_0) + b_1 \dots) + b_n \quad (4.66)$$

The parameters $\theta_j = \{W_0, b_0, \dots, W_n, b_n\}$ determine a function of the continuous variable $x \in \mathbb{R}^d$. Hence, the function $\tilde{f}_j$ can be evaluated at any $x \in D$, regardless of the choice of the discretization grid. When a practitioner chooses a grid $D_N = \{x_1, \dots, x_N\}$, they evaluate $\tilde{f}_j(x_k)$ at each grid point and compute $\sum_k w_k \tilde{f}_j(x_k) a(x_k)$. Choosing a different grid $D_{N'}$ only changes the points at which $\tilde{f}_j$ is evaluated when approximating the integral from (4.65). The parameters θ_j are independent of the choice of the discretization D_N because $\tilde{f}_j$ is defined pointwise on $\mathbb{R}^d$, not on a grid. Hence the grid only determines *where* $\tilde{f}_j$ is sampled, not *how* $\tilde{f}_j$ is defined.

Remark 4.3.8 (Pointwise evaluations as in Chen and Chen are a special case of Integral Kernel Operator). As stated in the previous remark, the key improvement that Integral Kernel based Neural Operator offers compared to the sensor based encoder approaches, as in Chen and Chen [3], DeepONet [4] or PCA-Net [6], is Discretization Invariance. We can demonstrate that Integral Kernel Operator based encoder is a generalization of the previous approaches. In [Theorem 4.1.1](#) we defined the encoder step as pointwise evaluations of the input function $a \in \mathcal{A}$ on a fixed grid $D_N := \{x_i\}_{i=1}^N$ as $F(a) := (a(x_1), \dots, a(x_N))$ (4.1.1). Formally, this can be expressed as

$$c_j(a) = \int_D \delta_{x_j}(x) a(x) d\mu(x) \quad (4.67)$$

where δ_{x_j} is a Dirac delta. However, δ_{x_j} is not a continuous function, it's a distribution, and it cannot be approximated via a neural network or integrated against a L^p function as [9, Lemma 32] proceeds, as outlined in *Step 2* of [Theorem 4.3.1](#) proof. In [9, Lemma 32] Kovachki et al. replace each δ_{x_j} with a smooth function with compact support $f_j \in C_0^\infty(D)$ (4.65), where each f_j can be defined as a mollifier $\varphi_r(\cdot - x_j) \in C_0^\infty(D)$ with support of radius $r > 0$. Each $f_j(x) = \varphi_r(x - x_j)$ can be seen as mollification of the Dirac mass δ_{x_j} and the expression for the functional $c_j \in \mathcal{A}^*$ becomes

$$c_j(a) = \int_D \varphi_r(x - x_j) a(x) d\mu(x) \quad (4.68)$$

As $r \rightarrow 0$, the mollified kernel φ_r collapses to a Dirac mass, and the expression above becomes the fixed grid pointwise evaluations. In the limiting behavior, the continuous non-locality degenerates to the discretized non-locality, where the approximating network parameters in (4.4) need to be retrained with every grid change D_N , hence the Discretization Invariance is lost.

Remark 4.3.9 (On the simplified Neural Operator architecture in the UAT proof). The proof of [Theorem 4.3.1](#) uses the simplified Neural Operator ([Definition 2.2.27](#)), which removes the lifting operator $\mathcal{P}$, the projection operator $\mathcal{Q}$, and the skip connection $\mathcal{W}$ from the full architecture ([Definition 2.2.24](#)). We examine whether these components would cause complications if retained.

The lifting operator $\mathcal{P}$ potentially introduces an issue in the proof. In *Step 2* of the proof, the encoder $F : \mathcal{A} \rightarrow \mathbb{R}^J$ is constructed from bounded linear functionals $c_j \in \mathcal{A}^*$, whose existence is guaranteed by the Approximation Property and whose integral representation follows from the Riesz Representation Theorem [18, Thm. 5.1.27]. If a nonlinear lifting $\mathcal{P}$ is composed before the first integral layer, the encoding becomes

$$c_j(\mathcal{P}(a)) = \int_D \kappa_j(x) \mathcal{P}(a(x)) dx \quad (4.69)$$

which is a nonlinear functional of a . Both the Approximation Property, which gives us linear finite-rank approximations of identity operator, and the Riesz Representation Theorem, which characterizes linear functionals, require the functionals $c_j \in \mathcal{A}^*$ to be *linear*. A linear lifting operator $\mathcal{P}(a) = Wa$ would preserve linearity, but would be redundant as it can be absorbed into the first Kernel Integral Operator.

The skip connection $\mathcal{W}$ and the projection $\mathcal{Q}$ do not introduce comparable complications. In *Step 3*, we defined the integration kernel in (4.41) as multiplication of a constant input vector $w \in \mathcal{R}^J$ by a spatially independent kernel $\kappa^{(t)}(x, y) = W_t/|D|$. For a constant input $v = w\mathbf{1}$, the skip connection gives $\mathcal{W}(w\mathbf{1})(x) = W \cdot w$, which is absorbed into the weight matrix of the classical neural network. The projection $\mathcal{Q}$ acts after the decoder and does not interfere with the linear reconstruction guaranteed by the AP.

Since the proof is existential, Kovachki et al. simply set $P = \text{Id}$, $Q = \text{Id}$, and $W = 0$ in every layer. The UAT then holds for the simplified architecture, and because the full architecture is strictly more expressive, as it contains the simplified architecture as the special case $P = \text{Id}$, $Q = \text{Id}$, $W = 0$, universality of the full architecture follows as an immediate corollary. What remains non-trivial, as noted by Kovachki et al. [9, Sec. 8.1], is explaining the *practical benefit* of these components. This would require quantitative approximation rates, which, as discussed in [Remark 4.3.6](#), is not achievable with the current proof approach due to the non-constructive nature of both the Approximation Property and the classical Universal Approximation Theorem that the proof relies on.

4.4 Synthesis: the Structure of Universal Approximation

In the preceding sections, we presented the Universal Approximation Theorem of Chen and Chen (1995) in [Theorem 4.1.1](#), which established sufficiency of the *encode-approximate-decode* strategy and implicitly introduced the core components of non-linearity, non-locality, and compactness. We then generalized these ideas through the framework of Kovachki et al. (2023) in [Theorem 4.3.1](#), which extended the result to general Banach spaces via the Approximation Property and provided the Integral Kernel parameterization that enabled the discretization invariance results of Chapter 3. In this section, we formalize the necessary requirements for Universal Approximation, distill the shared proof strategy into a unified template, and compare the Integral Kernel approach of Kovachki et al. with the minimal Averaging Neural Operator of Lanthaler et al. (2025) [\[10\]](#) to examine the tradeoff between architectural richness and theoretical minimality.

4.4.1 Necessary Components for Universal Approximation

From our analysis of Universal Approximation Theorems, presented in [Theorem 4.1.1](#), [Theorem 4.2.7](#) and [Theorem 4.3.1](#), we can deduce that there are three necessary components across all formulations that are required to guarantee the Universal Approximation of operators on infinite dimensional Banach Spaces: *non-locality* and *non-linearity* of the approximating function and *compactness* of the input set.

Non-linearity

Remark 4.4.1 (Non-linearity). A linear Neural Operator architecture $\hat{\mathcal{G}} : \mathcal{A} \rightarrow \mathcal{U}$ can only approximate linear operators, so any universal framework must incorporate a non-linear component. In the Neural Operator architecture, non-linearity enters through the activation function σ in the hidden network layers ([Definition 2.2.24](#)) and through the neural networks that parameterize the Lifting and Projection operators ([subsection 2.2.5](#)). This architectural choice is based on the classical Universal Approximation results by Cybenko (1989) [\[1\]](#) and Hornik et al. (1989) [\[2\]](#), which state that a single hidden layer with non-polynomial activation is sufficient for uniform approximation of continuous functions on compact subsets of $\mathbb{R}^d$. The underlying mechanism is the Stone-Weierstrass Theorem [\[18, Thm. 2.4.1, p. 61\]](#), as it can be shown that the closure of the span of ridge functions $x \mapsto \sigma(w \cdot x + b)$ with non-polynomial σ forms a subalgebra of $C(K, \mathbb{R})$ that separates points and contains the constants, and is therefore dense. This classical result is invoked at multiple points in both operator UAT proofs of [Theorem 4.1.1](#) and [Theorem 4.3.1](#), as it guarantees that the encoding functionals, the finite-dimensional transformation, and the output reconstruction components, as generalized in [Theorem 4.2.7](#), can each be implemented by neural networks to arbitrary accuracy.

Non-locality

Remark 4.4.2 (Non-locality). Operators between infinite-dimensional function spaces are in general non-local, meaning that the output at a point x of a domain $D \subset \mathbb{R}^d$ may depend on the values of the input function at points other than x . A pointwise operator $\hat{\mathcal{G}}(a)(x) := \sigma(Wa(x) + b)$ cannot learn such dependencies, regardless of the choice of a non-linear activation σ , weights W , or biases b . To illustrate this, consider the shift operator $\mathcal{G}^\dagger(a)(x) := a(x + h)$ for some fixed $h \neq 0$. Given two functions $a_1, a_2 \in \mathcal{A}$ that agree at x but differ at $x + h$, any pointwise operator satisfies $\hat{\mathcal{G}}(a_1)(x) = \hat{\mathcal{G}}(a_2)(x)$ because it only accesses the value $a(x)$, while $\mathcal{G}^\dagger(a_1)(x) = a_1(x + h) \neq a_2(x + h) = \mathcal{G}^\dagger(a_2)(x)$. No tuning of the parameters of $\hat{\mathcal{G}}$ can resolve this. Hence, any universal architecture must evaluate function $a \in \mathcal{A}$ non-locally [\[10, sec. 2.1\]](#).

By contrast, a Kernel Integral Operator $(Ka)(x) := \int_D \kappa(x, y) a(y) dy$ (def. 2.2.14) aggregates information about a across the entire domain D to produce the output at x , which is precisely why kernel integration is the natural parameterization for Neural Operator layers.

Compactness

Remark 4.4.3 (Compactness). All discussed formulations of UAT for operators on infinite-dimensional Banach spaces (Theorem 4.1.1, Theorem 4.2.7, Theorem 4.3.1) require the input set $K \subset \mathcal{A}$ to be compact. The fundamental reason is that each proof relies on reducing an operator between infinite-dimensional spaces to a finite-dimensional approximation problem, and compactness is what makes such a reduction possible. Total boundedness that follows from compactness of K guarantees that for any $\varepsilon > 0$, the behavior of all functions in K can be captured by finitely many representatives, which enables a uniform finite-rank approximation of the operator $\mathcal{G}^\dagger$ on the entire set K . The role of compactness for UAT was introduced in Remark 4.1.9 and discussed in more detail in Remark 4.3.3. Furthermore, as we show in Corollary 4.3.5, for bounded subsets of $W^{m,p}(D)$ with $mp > d$, the Morrey inequality [15, Thm. E.13, p. 302] provides the equicontinuity and uniform boundedness that is required by Ascoli-Arzelà Theorem [18, Thm. 5.4.2, p. 196], so that compactness can be replaced by of boundedness in Sobolev spaces at the cost of additional assumption of target operator continuity in $C(\bar{D})$ with sup norm topology over the input space.

4.4.2 The Proof Template

The goal of Universal Approximation for operators on Banach spaces is to show that, given a compact set $K \subset \mathcal{A}$ of input functions on a bounded domain $D \subset \mathbb{R}^d$ and a continuous operator $\mathcal{G}^\dagger : \mathcal{A} \rightarrow \mathcal{U}$, there exists a Neural Operator $\mathcal{G} \in \text{NO}_T$ (Definition 2.2.24) that approximates $\mathcal{G}^\dagger$ uniformly on K . Since $\mathcal{G}^\dagger$ is continuous and K is compact, the image $\mathcal{G}^\dagger(K) \subset \mathcal{U}$ is also compact, and as discussed in Remark 4.4.3, both the input and output sets admit finite-rank approximations. A general proof strategy, shared by all approaches we have presented, is to approximate $\mathcal{G}^\dagger$ by a finite-rank operator of the form

$$\tilde{\mathcal{G}}(a) := \sum_{j=1}^J \alpha_j(a) \beta_j, \quad (4.70)$$

where $\alpha_1, \dots, \alpha_J : \mathcal{A} \rightarrow \mathbb{R}$ are continuous functionals encoding the input and $\beta_1, \dots, \beta_J \in \mathcal{U}$ are fixed elements of the output space providing a reconstruction basis. The encoding

$$\mathcal{E}(a) := (\alpha_1(a), \dots, \alpha_J(a)) \in \mathbb{R}^J \quad (4.71)$$

compresses the input into a finite vector $v \in \mathbb{R}^J$, while the decoding

$$\mathcal{D}(v) := \sum_{j=1}^J v_j \beta_j \quad (4.72)$$

reconstructs an element of $\mathcal{U}$ as a linear combination. In general, the functionals α_j may be linear or nonlinear depending on the approach, we discuss this distinction in the following.

The proofs of Theorem 4.1.1 and Theorem 4.3.1 both realize the *encode-approximate-decode* strategy of (4.70), as do several other architectures discussed in Remark 4.2.9. The approaches differ in how each component is implemented, and as we discuss below, Lanthaler et al. (2025) [10] show that considerably simpler implementations than those of Theorem 4.3.1 are already sufficient for universality.

Encoding

Remark 4.4.4 (Encoding via the Approximation Property). In [Theorem 4.2.7](#), we showed that the Approximation Property ([Definition 4.2.1](#)) guarantees the existence of a finite-rank approximation of the identity operator I_J^A on a compact set K , which can be written as $I_J^A(a) = \sum_{j=1}^J c_j(a) \eta_j$ with $\{\eta_j\}_{j=1}^J \subset \mathcal{A}$ being a subset of a Schauder basis ([Definition 4.2.2](#)) of $\mathcal{A}$. Since the basis elements $\{\eta_j\}_{j=1}^J$ are fixed for all $a \in K$, the coefficients $c_j : \mathcal{A} \rightarrow \mathbb{R}$ are continuous linear functionals $c_j \in \mathcal{A}^*$, and the finite-dimensional encoding is given by $F_J^A(a) := (c_1(a), \dots, c_J(a)) \in \mathbb{R}^J$. The nonlinear transformation described by $\mathcal{G}^\dagger$ is factored into a separate component $\varphi \in C(\mathbb{R}^J, \mathbb{R}^{J'})$ as defined in (4.29), so that the full encoder from (4.71) is the composition $\mathcal{E}(a) := \varphi \circ F_J^A(a)$.

Architecturally, each linear functional c_j is realized as an Kernel Integral Operator ([Definition 2.2.14](#)). The Riesz Representation Theorem [[18](#), Thm 5.1.27, p. 176] expresses each c_j as integration against a kernel, and the density of $C_0^\infty(D)$ in the dual [[18](#), Thm. 5.1.23, p.174] allows approximation of this kernel by smooth functions, which in turn can be parameterized as neural networks via classical UAT. The resulting integral kernel $\kappa(x, y)$ evaluates $a \in K$ non-locally across the entire domain D .

Remark 4.4.5 (Encoding via Averaging). Lanthaler et al. take a different approach from [Theorem 4.2.7](#) and do not rely on the Approximation Property. Instead of discretizing the input space $\mathcal{A}$, they employ Fourier analysis on the output space directly, similar to the approach by Chen & Chen we discussed in [Lemma 4.1.5](#). Given a continuous operator $\mathcal{G}^\dagger : C^s(D) \rightarrow C^{s'}(D)$ on a bounded Lipschitz domain $D \subset \mathbb{R}^d$, the image of a compact set $K' := \mathcal{G}^\dagger(K) \subset C^{s'}(D)$ is compact. By the Stein Extension Theorem [[10](#), Lemma A.1], the set K' is extended to a compact set of periodic functions $K'_{\text{per}} \subset C_{\text{per}}^{s'}(B)$ on a bounding box $B \supset D$, which enables Fourier expansion. To guarantee uniform convergence of the Fourier series over the compact set in the $C^{s'}$ norm, each function in K'_{per} is mollified with parameter $\delta > 0$, resulting in a set of C^∞ functions that is uniformly bounded in $C^r(B)$ for any $r > s'$. By a classical result on the uniform convergence of Fourier series for smooth periodic functions [[24](#), Ch. 1, Sec. 3], if a set of periodic functions is uniformly bounded in $C^r(B)$ for some $r > s'$, then the Fourier series converges uniformly over that set in the $C^{s'}$ norm. Hence, Fourier series over the mollified set converges uniformly in the $C^{s'}$ norm, and the mollification error is controlled separately via [[10](#), Lemma A.2]. Hence, for large enough $J \in \mathbb{N}$, the truncated Fourier expansion approximates every output in $\mathcal{G}^\dagger(K)$ simultaneously. Denoting the retained Fourier basis elements as $\{\beta_j\}_{j=1}^J \subset L_{\text{per}}^2(B)$, the finite-rank approximation takes the form

$$\tilde{\mathcal{G}}(a) := \sum_{j=1}^J \langle E(\mathcal{G}^\dagger(a)), \beta_j \rangle_{L^2} \beta_j, \quad \forall a \in K, \quad (4.73)$$

where $E : C^{s'}(D) \rightarrow C_{\text{per}}^{s'}(B)$ is the Stein extension operator [[10](#), Lemma A.1]. Defining the encoding functionals as $\alpha_j(a) := \langle E(\mathcal{G}^\dagger(a)), \beta_j \rangle_{L^2}$, we recover the *encode-decode* formulation of (4.70). Note that the α_j are nonlinear functionals of the input a , since the nonlinear operator $\mathcal{G}^\dagger$ appears inside the inner product.

While the theoretical approach taken by Lanthaler et al. [[10](#)] is fundamentally different from the Approximation Property based approach of [Theorem 4.2.7](#), the resulting finite-rank formulation again takes the form of (4.70). Architecturally, the Averaging Neural Operator ([Definition 2.2.32](#)) implements each functional α_j through the Lifting network $R : \mathbb{R}^k \times D \rightarrow \mathbb{R}^{d_c}$ combined with a single Averaging operation. The network $R(a(y), y)$ absorbs the weight function β_j into its parameterization, so that the unweighted average $\frac{1}{|D|} \int_D R(a(y), y) dy$ recovers the weighted inner products of (4.73) [[10](#), Lemma A.9]. Since R is a continuous map between finite-dimensional Euclidean spaces, it can be approximated

by a neural network with non-polynomial activation to arbitrary accuracy via the classical UAT [1, 2].

Remark 4.4.6 (Non-locality in the encoding). Despite the different theoretical routes, all discussed approaches realize the encoder of (4.70) as a finite set of non-local functional evaluations over the compact set K , confirming the necessity of non-locality established in Remark 4.4.2. Chen and Chen (1995)[3] evaluate a at sensor points $\{x_1, \dots, x_m\} \subset D$, Kovachki et al. (2023)[9] integrate against learned kernels, and Lanthaler et al. (2025)[10] take inner products against Fourier basis functions. In each case, the encoding requires access to the values of $a \in \mathcal{A}$ across the domain D , not just at a single point, to produce consistent finite-dimensional encoding of the function a regardless of the inputs.

Decoding and Reconstruction

The objective of the Decoder operator $\mathcal{D}$ from (4.70) is, given an encoding vector $v \in \mathbb{R}^J$, to reconstruct an element $u \in \mathcal{U}$ as a linear combination of the basis elements $\{\beta_j\}_{j=1}^J \subset \mathcal{U}$, as defined in (4.72).

Remark 4.4.7 (Reconstruction of the output space). The decoder $\mathcal{D}(v) := \sum_{j=1}^J v_j \beta_j$ from (4.70) requires a finite basis $\{\beta_j\}_{j=1}^J \subset \mathcal{U}$ that can reconstruct any element in the compact image $\mathcal{G}^\dagger(K) \subset \mathcal{U}$ up to arbitrary accuracy. Kovachki et al. obtain such a basis from the Approximation Property of $\mathcal{U}$ as a truncation of a Schauder basis, as defined in (4.28), which guarantees existence but does not specify the basis elements explicitly, as outlined in the proof of Theorem 4.2.7. Lanthaler et al. (2025) [10] obtain the basis constructively as the retained Fourier basis functions $\{\beta_j\}_{j=1}^J \subset L_{\text{per}}^2(B)$ from the truncated expansion in (4.73). In both cases, compactness of $\mathcal{G}^\dagger(K)$ is what guarantees that finitely many basis elements suffice for a uniform approximation over the entire output set. Architecturally, Kovachki et al. realize the reconstruction through Kernel Integral Operators by approximating the truncated Schauder basis elements with neural networks via classic UAT [1] and parametrizing the integration kernel of the decoder κ^{out} (4.44) as a summation evaluated locally (4.45), while in the ANO (Definition 2.2.32) the reconstruction is handled by the projection network $Q : \mathbb{R}^{d_c} \times D \rightarrow \mathbb{R}^{k'}$, which approximates the linear combination $\sum v_j \beta_j(x)$ via classical UAT [1] with the spatial coordinate $x \in D$ provided through positional encoding.

4.4.3 The Tradeoff: Minimality vs. Quantitative Control

The proof strategies of Kovachki et al. and Lanthaler et al. arrive at the same conclusion, the universality of Neural Operators, but they are motivated by different questions. Kovachki et al. ask whether the empirically observed success of the Neural Operator architecture can be justified theoretically, and their proof shows that Integral Kernel Operators alone are sufficient to implement every component of the *encode-approximate-decode* factorization. Lanthaler et al. ask what minimal requirements on the Neural Operator structure are needed for universality to hold at all. The two perspectives are complementary in their nature but lead to different uses of the same operator architecture, and comparing them reveals a tradeoff between structural richness and theoretical minimality.

Remark 4.4.8 (Architectural comparison). Both approaches operate within the Neural Operator framework of Definition 2.2.24, but differ in which roles are assigned to each component of the Neural Operator in the proof. Kovachki et al. [9] set the lifting $\mathcal{P}$ (Definition 2.2.21) and projection $\mathcal{Q}$ (Definition 2.2.22) to the identity and the skip connection $\mathcal{W}$ (Definition 2.2.14) to zero, as discussed in (Remark 4.3.9), and as we demonstrated in Theorem 4.3.1, each step of the *encode-approximate-decode* framework is implemented as a composition of the Kernel Integral Operators and the nonlinear activation. Each kernel

$\kappa^{(t)}(x, y)$ is parameterized as a neural network (4.39), (4.41), (4.44) and varies in both spatial arguments x and $y \in D$, which allows the integration operator to evaluate the contribution of $v(y)$ differently depending on the evaluation point x of the domain D .

Lanthaler et al. [10] take the opposite approach. The kernel integral is reduced to a spatial averaging function $\kappa(x, y) := |D|^{-1} I_{d_c \times d_c}$, which provides non-locality but is constant in the spatial position variable $x \in D$. In contrast to Kovachki et al., the lifting network $R(a(y), y)$ that implements the operator $\mathcal{P}$ does the encoding (Remark 4.4.5) and the projection network $Q(v, x)$ implementation of $\mathcal{Q}$ is responsible for the decoding (Remark 4.4.7), with the skip connection $Wv(x) + b$ responsible for passing the positional information between them. The resulting Averaging Neural Operator (Definition 2.2.32) is the minimal subarchitecture of the general Neural Operator that possesses the both required non-linearity and non-locality properties, discussed in Remark 4.4.1 and Remark 4.4.2. As a consequence, the expressivity of the ANO is controlled by the channel dimension d_c , i.e. the width of the Lifting operator (Definition 2.2.21), rather than by the complexity of the integral kernel [10, Cor. 3]. Lanthaler et al. show in [10, Cor. 1–5] that any implementation of Neural Operators, including the FNO (Definition 2.2.31) and the general Kernel Integral NO, can be viewed as the ANO, since any Neural Operator whose expansion basis includes a constant function contains the ANO as a special case.

Remark 4.4.9 (Quantitative Bounds and Discretization Invariance). The minimality of the ANO architecture, while theoretically elegant, limits the possibility of obtaining explicit convergence rates. In the framework of Kovachki et al., the encoding functionals $c_j \in \mathcal{A}^*$ are represented as integration against kernels $\kappa_j \in C(\bar{D} \times \bar{D})$ (4.39) via the Riesz Representation Theorem [18, thm. 5.1.27, p. 176], as per *Step 2* of the Theorem 4.3.1 proof, and the regularity of these kernels is determined by the input function space. When the continuous integral $\int_D \kappa(x, y) v(y) dy$ is replaced by a quadrature rule on N points, the discretization error can be controlled through the Morrey inequality [15, Thm. E.13, p. 302] applied to both κ and v , resulting in the algebraic convergence rates that we established in Corollary 3.4.3.

In the ANO (Definition 2.2.32), the Kernel Integral Operator is parametrized with the constant $\kappa(x, y) := |D|^{-1} I_{d_c \times d_c}$, and the only component that depends on the input space is the Lifting network $R(a(y), y)$. In the proof of [10, Lemma A.9], the encoder is constructed by requiring R to simulate weighted inner products $\frac{1}{|D|} \int_D a(y) \cdot \beta_{j,\delta}(y) dy$ via the unweighted average $\frac{1}{|D|} \int_D R(a(y), y) dy$, where $\beta_{j,\delta} := |D|(\beta_j * \rho_\delta)$ are mollified L^2 -basis functions formed via Fourier series [10, eq. A.7], as described in Remark 4.4.5. As the approximation accuracy $\varepsilon > 0$ decreases, finer mollification ($\delta \rightarrow 0$) via $\rho_\delta \in C_0^\infty$ and more basis functions ($J \rightarrow \infty$) are required, and the weight functions represented as mollified basis $\beta_{j,\delta}$ has increasingly higher oscillation frequency. The neural network R that simulates these weights is guaranteed to exist by the classical UAT [10, Lemma A.11], but its complexity, in particular its Lipschitz constant, is not bounded as a function of ε . The discretization error of the empirical encoder $\frac{1}{N} \sum_{i=1}^N R(a(x_i), x_i)$ therefore depends on the regularity of R , which is determined by the width, depth, and weight parameters of the approximating Neural Network. The classical UAT [1] guarantees the existence of such parameters that enable convergence for each $\varepsilon > 0$, but does not provide any explicit bounds. In contrast, the regularity of kernel $\kappa(x, y)$ in the Kovachki et al. approach is determined by the input function space $\mathcal{A}$, which is what enables the application of the Morrey inequality [15, Thm. E.13, p.302] in Corollary 3.4.3. Lanthaler et al. themselves identify quantitative error estimates for the ANO as an open problem [10, Sec. 5].

The complementary nature of the two described approaches connects directly to the structure of this thesis. Lanthaler et al. provide the theoretical confirmation that the three necessary components identified in subsection 4.4.1, namely *non-linearity* (Remark 4.4.1), *non-locality* (Remark 4.4.2), and *compactness* (Remark 4.4.3), are not only necessary but

also sufficient for Universal Approximation to hold. However, it is the expressivity of the Kernel Integral Operator structure of Kovachki et al. that allows the application of Functional-analytic tools such as the Morrey inequality to derive the explicit convergence rates for empirical Neural Operator architectures established in [Corollary 3.4.3](#).

Chapter 5

Conclusion

This thesis examined the theoretical foundations of Neural Operators from a functional-analytic perspective, focusing on Discretization Invariance and Universal Approximation. The main original contributions are an alternative proof of the Discretization Invariance Theorem that offers explicit algebraic convergence rates ([Theorem 3.3.10](#), [Corollary 3.4.3](#)) and an extension of the Universal Approximation Theorem to bounded subsets of Sobolev spaces under additional assumptions of operator regularity ([Corollary 4.3.5](#)).

Neural Operators

The Neural Operator architecture is motivated by the classical Green’s function, which solves linear PDEs via a global integral operator. The Kernel Integral Operator $\mathcal{K}$ defines a non-local component, which is necessary for well-posedness of the empirical operator $\hat{\mathcal{G}}$ defined on discretized grid ([Remark 2.2.3](#)), and its compactness ([Proposition 2.2.17](#)) necessitates introduction of the skip connection $\mathcal{W}$ to recover the local, high frequency behavior that integral operator otherwise smooths out. In defining these components we formalized some of the requirements otherwise left implicit in the source literature, in particular the at-most-linear-growth requirement on the activation function Σ without which the induced Nemytskii operator fails to map between L^p Lebesgue spaces ([Proposition 2.2.13](#)). All empirical architectures of Neural Operators studied in this thesis reflect the same structural components dictated by functional-analytic necessity: the non-local Kernel Integral, the non-compact Skip Connection, and the non-linear Activation Function.

Discretization Invariance

The explicit convergence rate established in [Corollary 3.4.3](#) decomposes the discretization error into two components: one governed by the Hölder regularity of the input function class, the other by the Lipschitz regularity of the learned integral kernel. The input regularity term is dominating, which means that the approximation accuracy of a discretized Neural Operator is constrained by the regularity of the PDE solution functions that are being approximated, not by architectural design choices of the model. Smoothing the learned kernel beyond Lipschitz continuity has diminishing returns. The resulting lower bound on the grid resolution $N \gtrsim \varepsilon^{-d/\alpha}$ for the target approximation error rate $\varepsilon > 0$ confirms that the curse of dimensionality is intrinsic to the Kernel Integral based model architecture.

These rates were enabled by an alternative proof approach for the Discretization Invariance Theorem of Kovachki et al. ([Theorem 3.3.2](#)). The original proof by Kovachki et al. requires compactness of the input set K , obtained via an ε -net construction that reduces the problem to finitely many representative functions, but only guarantees asymptotic convergence. The proposed alternative approach ([Theorem 3.3.10](#)) omits constructing the ε -net by instead utilizing the Hölder continuity that the Morrey embedding provides

on bounded subsets of $W^{k,p}(D)$ with $kp > d$, allowing us to uniformly bound the integrand variation on each partition cell of discretized grid directly. This enables establishing explicit rates, and additionally relaxes the compactness requirement on K to boundedness.

Universal Approximation

All Universal Approximation Theorems for operators on Banach spaces analyzed in this thesis rely on the *encode-approximate-decode* proof strategy. This approach is necessary due to the structural mismatch between the infinite-dimensional spaces on which operators act and the finite-dimensional neural networks used to approximate them. Functional analysis resolves this by providing tools that enable reducing the problems on infinite-dimensional Banach spaces to finite-dimensional approximations.

First, it guarantees existence of a basis structure on the relevant Banach spaces, either via the Approximation Property (Theorem 4.2.7) or via extension to periodic domain followed by the Fourier decomposition (Theorem 4.1.1, Remark 4.4.5), which enables approximating elements of input and output spaces as linear combinations of finitely many elements of truncated basis with arbitrary accuracy. The encoding step then reduces to finding the coefficients of such a representation, and can be expressed as evaluating continuous functionals on the input space. Second, the Riesz Representation Theorem allows representing these functionals as integration against kernels, which motivates the Kernel Integral component in the Neural Operator architecture. Hence, the problem of learning non-linear operators on Banach spaces effectively reduces to approximating a finite set of real valued functionals, which produce a finite dimensional encoding of the input maps.

In the decoding step the output elements are reconstructed from finite coefficient vectors as a linear combination of the truncated output basis. The reduction to finite dimensional encoding vectors allows applying the classic Universal Approximation guarantees [1, 2] to learn the non-linear transformation described by the target operator. The approximation error decomposes accordingly into an encoding error controlled by the input space discretization, an approximation error from the learned finite-dimensional transformation, and a decoding error determined by the expressiveness of the output basis.

The Universal Approximation guarantees rely on three requirements: *non-locality*, *non-linearity* and *compactness*. Non-locality is required because the encoding functionals must aggregate information about the input function across the entire domain to produce a constant finite-dimensional representation.

Non-linearity is introduced via the activation function and is grounded in the Stone-Weierstrass Theorem. The closure of ridge functions with non-polynomial activation is dense in $C(K, \mathbb{R})$, and this density result is invoked at every stage of the *encode-approximate-decode* factorization to guarantee that each component can be implemented by a neural network.

Compactness of the input function set K ensures that the finite-rank reduction works uniformly. Total boundedness, which follows from compactness, guarantees that for any $\varepsilon > 0$, the behavior of all functions in K can be described by finitely many representatives, so that a single set of truncated basis elements and encoding functionals can be defined for the entire input set uniformly.

The key theorems studied in this thesis instantiate the *encode-approximate-decode* strategy differently, but the differences amount to variations on the same structure. The fixed-grid approach of Chen & Chen (1995) [3] encodes via Dirac evaluations at sensor points, which can be viewed as a special case of the Kernel Integral encoding of Kovachki et al. (2023) [9] in the degenerate limit where the mollification radius collapses to zero. Lanthaler et al. (2025) [10] encode via inner products against Fourier basis functions absorbed into the lifting network, arriving at the same finite-rank formulation through a different route. The roles assigned to each component of the Neural Operator differ across approaches, but

the *encode-approximate-decode* template is identical.

Continuous non-locality via kernel integration is what enables Discretization Invariance. The Kernel Integral formulation of the non-local component allows defining the encoding independently of any particular grid, so that the same learned parameters remain valid across different grid resolutions. By contrast, when non-locality is simplified to pointwise evaluation, as in Chen & Chen [3], the architecture becomes grid-dependent and must be retrained for each discretization.

The relaxation of compactness to boundedness requirements on subsets of $W^{k,p}(D)$ with $kp > d$ under additional assumption of uniform continuity of the target operator in $C(\bar{D})$ with sup-norm topology, established in Corollary 4.3.5 via the Morrey embedding, establishes a tradeoff and extends the Universal Approximation guarantees to a subset of a wider input class of bounded sets of functions admitted by our Discretization Invariance results from Theorem 3.3.10. However, explicit approximation rates for the UAT could not be established, as both the Approximation Property and the classical UAT, on which the analyzed proofs rely, are non-constructive existence results that guarantee convergence without quantifying the required network complexity or the rate at which the error decreases.

Limitations and Future Work

The explicit convergence rates established in Corollary 3.4.3 are derived for a single-layer Neural Operator and define an upper bound on the discretization error. Two theoretical directions follow. The first is the extension to multi-layer architectures with Lipschitz continuous kernels and activations, where the question is how the discretization error accumulates through layer composition. The second concerns the tightness of the obtained lower bound on the mesh refinement $N \gtrsim \varepsilon^{-d/\alpha}$. The fact that input regularity governs the rate (Remark 3.4.4) is independently supported by the FNO-specific aliasing analysis of Lanthaler et al. (2024) [13], which suggests the dependence is a structural feature of the discretization problem and motivates further quantitative study.

The error decomposition also suggests an empirical research direction. Since the bound identifies input regularity as the dominant source of discretization error, Corollary 3.4.3 suggests that Discretization Invariance of a model should degrade precisely when the input function class in a Sobolev space fails to satisfy the regularity threshold $kp > d$. This offers a falsifiable explanation for the failures of Discretization Invariance reported in the literature. The hypothesis is that Fourier Neural Operators should retain resolution independence on problems with sufficiently regular solutions and lose it on problems where the regularity condition is violated. This hypothesis can be tested through controlled ablation studies by varying the spatial dimension d , the Sobolev regularity of the input function class, and the grid resolution N , while measuring the empirical convergence rate against the predicted exponent. The results would provide a direct connection between the functional-analytic theory developed in this thesis and the empirical behavior of Neural Operator architectures.

Bibliography

- [1] George Cybenko. “Approximation by superpositions of a sigmoidal function”. In: *Mathematics of Control, Signals, and Systems* 2 (1989), pp. 303–314.
- [2] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. “Multilayer feedforward networks are universal approximators”. In: *Neural Networks* 2.5 (1989), pp. 359–366. ISSN: 0893-6080. DOI: [https://doi.org/10.1016/0893-6080\(89\)90020-8](https://doi.org/10.1016/0893-6080(89)90020-8). URL: <https://www.sciencedirect.com/science/article/pii/0893608089900208>.
- [3] T. Chen and H. Chen. “Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its application to dynamical systems”. In: *IEEE Transactions on Neural Networks* 6.4 (1995), pp. 911–917. DOI: 10.1109/72.392253.
- [4] Lu Lu, Pengzhan Jin, and George Em Karniadakis. “DeepONet: Learning nonlinear operators for identifying differential equations based on the universal approximation theorem of operators”. In: *Nature Machine Intelligence* 3.3 (2021), pp. 218–229. DOI: 10.1038/s42256-021-00302-5.
- [5] Nicholas H. Nelsen and Andrew M. Stuart. “The Random Feature Model for Input-Output Maps between Banach Spaces”. In: *SIAM Journal on Scientific Computing* 43.5 (2021), A3212–A3243. DOI: 10.1137/20M133957X.
- [6] Kaushik Bhattacharya et al. “Model reduction and neural networks for parametric PDEs”. In: *The SMAI Journal of Computational Mathematics* 7 (2021), pp. 121–157. DOI: 10.5802/smai-jcm.74.
- [7] Zongyi Li et al. “Fourier Neural Operator for Parametric Partial Differential Equations”. In: *arXiv preprint arXiv:2010.08895* (2020).
- [8] Nikola Kovachki, Samuel Lanthaler, and Siddhartha Mishra. “On Universal Approximation and Error Bounds for Fourier Neural Operators”. In: *Journal of Machine Learning Research* 22.290 (2021), pp. 1–76.
- [9] Nikola B. Kovachki et al. “Neural Operator: Learning Maps Between Function Spaces With Applications to PDEs”. In: *Journal of Machine Learning Research* 24.89 (2023), pp. 1–97.
- [10] Samuel Lanthaler, Zongyi Li, and Andrew M. Stuart. “Nonlocality and Nonlinearity Implies Universality in Operator Learning”. In: *Constructive Approximation* 62 (2025), pp. 261–303. DOI: 10.1007/s00365-025-09718-3.
- [11] Zhi-Qin John Xu, Lulu Zhang, and Wei Cai. “On understanding and overcoming spectral biases of deep neural network learning methods for solving PDEs”. In: *Journal of Computational Physics* 530 (2025). Review article on spectral bias and mitigation strategies for DNN-based PDE solvers, p. 113905. ISSN: 0021-9991. DOI: 10.1016/j.jcp.2025.113905. URL: <https://doi.org/10.1016/j.jcp.2025.113905>.

- [12] S. Qin et al. “Toward a Better Understanding of Fourier Neural Operators: Analysis and Improvement from a Spectral Perspective”. In: *arXiv preprint* (2024). Analyzes spectral bias in FNOs and proposes SpecB-FNO. arXiv: 2404.07200 [cs.LG]. URL: <https://arxiv.org/abs/2404.07200>.
- [13] Samuel Lanthaler, Andrew M. Stuart, and Margaret Trautner. “Discretization Error of Fourier Neural Operators”. In: *arXiv preprint arXiv:2405.02221* (2024).
- [14] Simone Brugiapaglia, Nicola Rares Franco, and Nicholas H. Nelsen. “A short tour of operator learning theory: Convergence rates, statistical limits, and open questions”. In: *arXiv preprint* (2026). arXiv: 2603.00819 [math.NA].
- [15] Delio Mugnolo and Joachim Kerner. *Partielle Differentialgleichungen*. Kurstext (Modul 61218). Hagen: FernUniversität in Hagen, 2025.
- [16] Zongyi Li et al. “Neural Operator: Graph Kernel Network for Partial Differential Equations”. In: *arXiv preprint arXiv:2003.03485* (2020).
- [17] Michael Renardy and Robert C. Rogers. *An Introduction to Partial Differential Equations*. 2nd ed. Vol. 13. Texts in Applied Mathematics. See Section 10.3.4. New York: Springer-Verlag, 2004, pp. xvi+434. ISBN: 0-387-00444-0.
- [18] Delio Mugnolo, Joachim Kerner, and Hafida Laasri. *Funktionalanalysis*. Kurstext (Kurs 01245). Hagen: FernUniversität in Hagen, 2023.
- [19] Lawrence C. Evans. *Partial Differential Equations*. 2nd ed. American Mathematical Society, 2010.
- [20] Robert E. Megginson. *An Introduction to Banach Space Theory*. New York: Springer, 1998. ISBN: 978-0387984315.
- [21] Helmut H. Schaefer. *Topological Vector Spaces*. Vol. 3. Graduate Texts in Mathematics. New York: Springer-Verlag, 1971. ISBN: 0-387-98726-6.
- [22] Z. Ciesielski and J. Domsta. “Construction of an orthonormal basis in $C^m(I^d)$ and $W_p^m(I^d)$ ”. In: *Studia Mathematica* 41.2 (1972), pp. 211–224. DOI: 10.4064/sm-41-2-211-224.
- [23] Allan Pinkus. “Approximation theory of the MLP model in neural networks”. In: *Acta Numerica* 8 (1999), pp. 143–195. DOI: 10.1017/S0962492900002919.
- [24] Dunham Jackson. *The Theory of Approximation*. Vol. 11. American Mathematical Society, 1930.
- [25] Werner Kirsch. *Maß und Integrationstheorie*. Kurstext (Kurs 01145). Hagen: FernUniversität in Hagen, 2024.

Appendix A

Neural Operator Properties

As specified in [Remark 2.2.2](#), the Discretization Invariance and Universal Approximation Theorems by Kovachki et al. [9] present their results on $C(\bar{D})$, $L^p(D)$ and $W^{k,p}$ spaces with $p \in [1, \infty)$ and $D \subset \mathbb{R}^d$. Hence, in the following sections we extend the definitions of the theoretical form of the Neural Operator [Definition 2.2.27](#) and its components, Kernel Integral Operator $\mathcal{K}$ ([Definition 2.2.14](#) and Nemytskii Operator induced by the Activation function Σ ([Definition 2.2.12](#)) to $C(\bar{D})$ with sup-norm and $W^{k,p}$ with $p \in [1, \infty)$ spaces. Note, that the requirements on kernel function regularity of the Kernel Integral operator differ dependent on the topology.

A.1 Neural Operators on $C(\bar{D})$ with sup-norm

Proposition A.1.1 (Induced Nemytskii operator Σ is well-defined on $C(\bar{D})$ with sup-norm). *Let $\sigma : \mathbb{R}^{d_v} \rightarrow \mathbb{R}^{d_v}$ be a continuous non-linear activation function. The map Σ defined by $[\Sigma v](x) = \sigma(v(x))$ is a well-defined non-linear bounded Nemytskii operator on $C(\bar{D}; \mathbb{R}^{d_v})$.*

Proof. Well-definiteness. As a composition of continuous functions, the function $[\Sigma v](x) = \sigma(v(x))$ is continuous for all $v \in C(\bar{D}; \mathbb{R}^{d_v})$. By Heine-Borel Theorem, the closed and bounded set $\bar{D} \subset \mathbb{R}^d$ is compact. The image $v(\bar{D})$ is compact by [18, Prop. 1.8.20, p. 32], and thus the image $[\Sigma v](\bar{D}) = \sigma(v(\bar{D}))$ is also compact and therefore bounded, which implies that the supremum norm is finite for all $x \in \bar{D}$:

$$\|\Sigma v\|_\infty = \sup_{x \in \bar{D}} |\sigma(v(x))| < \infty, \quad \forall v \in C(\bar{D}; \mathbb{R}^{d_v}) \quad (\text{A.1})$$

It follows, that the induced operator Σ is a well defined operator on $C(\bar{D}; \mathbb{R}^{d_v})$ topology with $\|\cdot\|_\infty$ norm.

Boundedness. Let $B_R := \{v \in C(\bar{D}) : \|v\|_\infty \leq R\}$ with $R < \infty$ be a bounded set. Then for any $x \in \bar{D}$, $\|v(x)\| \leq R$, i.e. the image $v(\bar{D})$ lies in a closure of a bounded ball $\overline{B(0, R)}$ for all $v \in B_R$. A closed and bounded ball or radius R is compact in $\mathbb{R}^d$. By the Extreme Value Theorem [18, Thm. 1.8.25, p.33], the continuous function σ achieves its maximum on the compact set $\overline{B(0, R)}$. For any $v \in B_R$ it holds

$$\|\Sigma v\|_\infty = \sup_{x \in \bar{D}} |\sigma(v(x))| \leq \max_{\|y\|_\infty \leq R} |\sigma(y)| < \infty \quad (\text{A.2})$$

The operator Σ maps bounded sets to bounded sets in $C(\bar{D}; \mathbb{R}^{d_v})$ and is therefore bounded.

Continuity. Observe a convergent sequence $(v_n)_{n=1}^\infty \subset C(\bar{D})$ with $v_n \rightarrow v$. Since convergent sequences are bounded, $\|v_n\|_\infty \leq R$, $\forall n \in \mathbb{N}$ and $\|v\|_\infty \leq R$. By Heine-Borel Theorem, the closed and bounded ball $\overline{B(0, R)} \subset \mathbb{R}^{d_v}$ is compact, and by Heine-Cantor Theorem the continuous function σ is uniformly continuous on $\overline{B(0, R)} \subset \mathbb{R}^{d_v}$. It follows,

that for any $\varepsilon > 0$ there exists $\delta > 0$ such that for any $y, y' \in \overline{B(0, R)}$ with $\|y - y'\| < \delta$ it holds $\|\sigma(y) - \sigma(y')\| < \varepsilon$. Since $v_n \rightarrow v$ uniformly, there exists $N \in \mathbb{N}$ such that $\sup_{x \in \bar{D}} \|v_n(x) - v(x)\| < \delta$. For all $n \geq N$ it holds:

$$\|\Sigma v - \Sigma v'\|_\infty \leq \sup_{x \in \bar{D}} |\sigma(v(x)) - \sigma(v'(x))| < \varepsilon \quad (\text{A.3})$$

Hence, the induced Nemytskii operator Σ is continuous on $C(\bar{D})$ in sup-norm. $\square$

Definition A.1.2 (Kernel Integral Operator $\mathcal{K}$ on $C(\bar{D})$ with sup-norm). Let $D \subset \mathbb{R}^d$ be a bounded measurable domain. Let $\mathcal{V} := C(\bar{D})$. Given a parameterized continuous kernel function $\kappa_\phi \in C(\overline{D \times D}; \mathbb{R}^{d_v \times d_v})$, the *Kernel Integral Operator* $\mathcal{K}_\phi : \mathcal{V} \rightarrow \mathcal{V}$ is defined as:

$$[\mathcal{K}_\phi v](x) = \int_D \kappa_\phi(x, y) v(y) dy, \quad \forall x \in D, \quad (\text{A.4})$$

Proposition A.1.3 (Kernel Integral Operator $\mathcal{K}$ is a well defined and bounded operator on $C(\bar{D})$ with sup-norm). *The Kernel Integral Operator $\mathcal{K}_\phi : \mathcal{V} \rightarrow \mathcal{V}$ is a well defined and bounded operator on $C(\bar{D})$ space with sup-norm topology.*

Proof. By Heine-Borel Theorem, the closed and bounded set $\bar{D} \subset \mathbb{R}^d$ is compact, and $\bar{D} \times \bar{D}$ is compact as a product of compact sets. By [18, Prop. 1.8.20, p. 32], the image $\kappa_\phi(\bar{D}, \bar{D})$ is compact and therefore bounded. Thus, the kernel has a finite supremum norm:

$$\|\kappa\|_\infty = \sup_{x \in \bar{D}} \sup_{y \in \bar{D}} |\kappa(x, y)| < \infty \quad (\text{A.5})$$

Well-definiteness. To show $\mathcal{K}_\phi$ is well-defined, we must prove that for any $v \in \mathcal{V}$, the image $\mathcal{K}_\phi v$ is continuous on $\bar{D}$. For $v \in C(\bar{D})$ with $\|v\|_\infty = 0$ the result follows directly, hence we assume $\|v\|_\infty \neq 0$. Since $\bar{D} \times \bar{D}$ is compact, κ_ϕ is uniformly continuous. For any $\varepsilon > 0$, there exists a $\delta > 0$ such that for all $x, x' \in \bar{D}$ and all $y \in \bar{D}$ with $|x - x'| < \delta$, we have $|\kappa_\phi(x, y) - \kappa_\phi(x', y)| < \varepsilon/|D|\|v\|_\infty$. It follows, that for any $x, x' \in \bar{D}$ with $|x - x'| < \delta$:

$$|[\mathcal{K}_\phi v](x) - [\mathcal{K}_\phi v](x')| \leq \int_D |\kappa_\phi(x, y) - \kappa_\phi(x', y)| |v(y)| dy \quad (\text{A.6})$$

$$\leq \int_D \frac{\varepsilon}{|D|\|v\|_\infty} \|v\|_\infty dy = \varepsilon. \quad (\text{A.7})$$

hence $\mathcal{K}_\phi v \in C(\bar{D})$ and the operator $\mathcal{K}_\phi$ is well-defined on $C(\bar{D})$ with $\|\cdot\|_\infty$.

Boundedness. To prove that the operator $\mathcal{K}_\phi$ is bounded, we first show the pointwise boundedness. For any $x \in \bar{D}$ it holds:

$$|[\mathcal{K}_\phi v](x)| = \left| \int_D \kappa_\phi(x, y) v(y) dy \right| \quad (\text{A.8})$$

$$\leq \int_D \|\kappa_\phi\|_\infty \|v\|_\infty dy \quad (\text{A.9})$$

$$= |\bar{D}| \|\kappa_\phi\|_\infty \|v\|_\infty < \infty \quad (\text{A.10})$$

Taking supremum over $x \in \bar{D}$:

$$\|\mathcal{K}_\phi v\|_\infty = \sup_{x \in \bar{D}} |[\mathcal{K}_\phi v](x)| \leq |\bar{D}| \|\kappa_\phi\|_\infty \|v\|_\infty < \infty \quad (\text{A.11})$$

Thus, the operator $\mathcal{K}_\phi$ is a well defined and bounded operator on $C(\bar{D})$ space with sup-norm. $\square$

Having established the necessary properties of the components, we state the definition of the theoretical form of Neural Operator (Definition 2.2.27) in the $C(\bar{D})$ with sup-norm topology.

Definition A.1.4 (Neural Operator (theoretical form), [9, Thm. 11]). Let $D \subset \mathbb{R}^d$ be a bounded domain. Let the input space $\mathcal{A} = C(\bar{D}; \mathbb{R}^{d_a})$ and the output target space $\mathcal{U} = C(\bar{D}; \mathbb{R}^{d_u})$ be spaces of continuous functions with $\|\cdot\|_\infty$ norm.

The *theoretical Neural Operator* $\mathcal{G} : \mathcal{A} \rightarrow \mathcal{U}$ is defined as

$$\mathcal{G} = \Sigma_T \circ \mathcal{K}_{\phi,T} \circ \cdots \circ \Sigma_1 \circ \mathcal{K}_{\phi,1} \quad (\text{A.12})$$

for $T \in \mathbb{N}$, where each layer consists only of a Kernel Integral Operator $\mathcal{K}_{\phi,t}$ with continuous kernel $\kappa_{\phi,t} \in C(\bar{D} \times \bar{D}; \mathbb{R}^{d_v \times d_v})$, followed by a Nemytskii operator Σ_t induced by a non-linear continuous activation function σ_t .

A.2 Neural Operators on Sobolev spaces $W^{k,p}$ with $kp > d$

Remark A.2.1 (Domain notation). Throughout this thesis the bounded domain is denoted as $D \subset \mathbb{R}^d$. In the following propositions we instead write Ω for the same domain, to avoid collision with the weak derivative operators D^α and the Jacobian $D\sigma$ that appear throughout this section.

Induced Nemytskii Operator Σ on $W^{1,p}$ space

Proposition A.2.2 (Well-definedness and boundedness of Σ on $W^{1,p}$). Let $\Omega \subset \mathbb{R}^d$ be a bounded domain. Let σ act componentwise via a scalar activation $\rho \in C^1(\mathbb{R}; \mathbb{R})$, i.e. $\sigma(y) = (\rho(y_1), \dots, \rho(y_d))$, with bounded derivative $M := \sup_{t \in \mathbb{R}} |\rho'(t)| < \infty$. Then the induced activation operator Σ defined by $[\Sigma v](x) = \sigma(v(x))$ is a well-defined non-linear bounded Nemytskii operator on $W^{1,p}(\Omega; \mathbb{R}^d)$ with $1 \leq p < \infty$, and there exists a constant $C = C(M, |\Omega|, \rho(0)) > 0$ such that for all $v \in W^{1,p}(\Omega; \mathbb{R}^d)$ it holds

$$\|\Sigma v\|_{W^{1,p}} \leq C(1 + \|v\|_{W^{1,p}}). \quad (\text{A.13})$$

Note, that a bounded Jacobian implies at most linear growth, so by Proposition 2.2.13 the L^p component is already proven and it remains to prove the properties for the weak derivative.

Remark A.2.3 (Limitations). Proposition A.2.2 only proves well definiteness and boundedness of Σ on $W^{1,p}$. The proof of the respective properties of Σ defined on $W^{k,p}$ with $k \geq 2, p \in [1, \infty)$ is out of scope for this thesis.

Proof. Since $|\rho'| \leq M$ on $\mathbb{R}$, by applying the Mean Value Theorem we obtain the linear growth bound $|\sigma(y)| \leq |\sigma(0)| + M|y|$, so σ satisfies the hypotheses of Proposition 2.2.13 and it follows, that $\Sigma v \in L^p(\Omega; \mathbb{R}^d)$ with

$$\|\Sigma v\|_{L^p}^p \leq 2^{p-1}(|\sigma(0)|^p |\Omega| + M^p \|v\|_{L^p}^p) < \infty \quad (\text{A.14})$$

Since $\rho \in C^1(\mathbb{R})$ with $|\rho'| \leq M$ and $\mu(\Omega) < \infty$, the Sobolev chain rule [18, Thm. 5.3.17] applies to each component $v^i \in W^{1,p}(\Omega)$, and the weak partial derivatives of $\sigma \circ v$ are defined as

$$\partial_j(\Sigma v)^i = \rho'(v^i) \partial_j v^i, \quad i, j \in \{1, \dots, d\} \quad (\text{A.15})$$

It remains to show that these derivatives are in $L^p(\Omega)$. By (A.15) and $|\rho'| \leq M$, we bound the L^p norm of the weak derivative:

$$\|\partial_j(\Sigma v)^i\|_{L^p} = \|\rho'(v^i) \partial_j v^i\|_{L^p} \leq M \|\partial_j v^i\|_{L^p}, \quad i, j \in \{1, \dots, d\} \quad (\text{A.16})$$

Combining (A.14) and (A.16), we obtain

$$\|\Sigma v\|_{W^{1,p}}^p = \|\Sigma v\|_{L^p}^p + \sum_{j=1}^d \|\partial_j(\Sigma v)\|_{L^p}^p \quad (\text{A.17})$$

$$\leq 2^{p-1}(|\sigma(0)|^p|\Omega| + M^p\|v\|_{L^p}^p) + M^p\|v\|_{W^{1,p}}^p \quad (\text{A.18})$$

$$\leq C^p(1 + \|v\|_{W^{1,p}})^p < \infty \quad (\text{A.19})$$

for a constant $C = C(M, |\Omega|, \rho(0))$. By taking p -th roots we obtain the inequality (A.13), which implies that $\Sigma v \in W^{1,p}(\Omega; \mathbb{R}^d)$. Therefore, Σ is a well-defined bounded Nemytskii operator on $W^{1,p}(\Omega; \mathbb{R}^d)$. $\square$

Kernel Integral Operator $\mathcal{K}$ on $W^{k,p}$ space

Proposition A.2.4 (Kernel Integral Operator $\mathcal{K}_\phi$ is a well-defined and bounded operator on $W^{k,p}(\Omega)$). *Let $\Omega \subset \mathbb{R}^d$ be bounded, let $1 \leq p < \infty$, and let $k \in \mathbb{N}_0$. Assume the kernel satisfies $\kappa_\phi \in C^k(\bar{\Omega} \times \bar{\Omega})$. Then the Kernel Integral Operator $\mathcal{K}_\phi : W^{k,p}(\Omega) \rightarrow W^{k,p}(\Omega)$, is a well-defined and bounded operator on $W^{k,p}(\Omega)$.*

Proof. Throughout, D_x^α denotes differentiation in the first variable, and we write $N_k := \#\{\alpha \in \mathbb{N}_0^d : |\alpha| \leq k\} = \binom{k+d}{d}$ for the number of multi-indices of order at most k .

By the Heine–Borel Theorem, the closed and bounded set $\bar{\Omega} \subset \mathbb{R}^d$ is compact, and $\bar{\Omega} \times \bar{\Omega}$ is compact as a product of compact sets. Since $\kappa_\phi \in C^k(\bar{\Omega} \times \bar{\Omega})$, each partial derivative $D_x^\alpha \kappa_\phi$ with $|\alpha| \leq k$ is continuous on $\bar{\Omega} \times \bar{\Omega}$. By [18, Prop. 1.8.20, p. 32] its image is compact and therefore bounded. Hence the kernel has finite C^k -bound

$$M := \max_{|\alpha| \leq k} \sup_{x \in \bar{\Omega}} \sup_{y \in \bar{\Omega}} |D_x^\alpha \kappa_\phi(x, y)| < \infty. \quad (\text{A.20})$$

Well-definiteness. Let $v \in W^{k,p}(\Omega)$. Since Ω is bounded, $v \in L^p(\Omega) \subset L^1(\Omega)$. We first show that $\mathcal{K}_\phi v$ admits classical derivatives up to order k , obtained by differentiating under the integral. For every $|\alpha| \leq k$ the integrand is dominated independently of x ,

$$|D_x^\alpha \kappa_\phi(x, y) v(y)| \leq M |v(y)| \quad (\text{A.21})$$

with $M|v| \in L^1(\Omega)$, so the Theorem on Differentiation under the Integral Sign [25, Satz 6.1.12] applies and we obtain

$$D^\alpha[\mathcal{K}_\phi v](x) = \int_{\Omega} D_x^\alpha \kappa_\phi(x, y) v(y) dy, \quad |\alpha| \leq k. \quad (\text{A.22})$$

It remains to verify that each such derivative is continuous on $\bar{\Omega}$. For v with $\|v\|_{L^1(\Omega)} = 0$ the claim is trivial, hence assume $\|v\|_{L^1(\Omega)} \neq 0$ and fix $|\alpha| \leq k$. As $D_x^\alpha \kappa_\phi$ is continuous on the compact set $\bar{\Omega} \times \bar{\Omega}$, it is uniformly continuous. For any $\varepsilon > 0$ there is $\delta > 0$ such that for all $x, x' \in \bar{\Omega}$ and all $y \in \bar{\Omega}$ with $|x - x'| < \delta$,

$$|D_x^\alpha \kappa_\phi(x, y) - D_{x'}^\alpha \kappa_\phi(x', y)| < \frac{\varepsilon}{\|v\|_{L^1(\Omega)}}. \quad (\text{A.23})$$

It follows for any $x, x' \in \bar{\Omega}$ with $|x - x'| < \delta$,

$$|D^\alpha[\mathcal{K}_\phi v](x) - D^\alpha[\mathcal{K}_\phi v](x')| \leq \int_{\Omega} |D_x^\alpha \kappa_\phi(x, y) - D_{x'}^\alpha \kappa_\phi(x', y)| |v(y)| dy \quad (\text{A.24})$$

$$\leq \frac{\varepsilon}{\|v\|_{L^1(\Omega)}} \int_{\Omega} |v(y)| dy = \varepsilon. \quad (\text{A.25})$$

Thus $D^\alpha[\mathcal{K}_\phi v] \in C(\bar{\Omega})$ for every $|\alpha| \leq k$, so $\mathcal{K}_\phi v \in C^k(\bar{\Omega}) \subset W^{k,p}(\Omega)$. In particular the weak derivatives of $\mathcal{K}_\phi v$ coincide with the classical ones and are in $L^p(\Omega)$. Hence $\mathcal{K}_\phi$ is well-defined on $W^{k,p}(\Omega)$.

Boundedness. Fix $v \in W^{k,p}(\Omega)$ and $|\alpha| \leq k$. For any $x \in \bar{\Omega}$, Hölder's inequality in the y -variable with conjugate exponent q with $\frac{1}{p} + \frac{1}{q} = 1$ gives the pointwise bound

$$|D^\alpha[\mathcal{K}_\phi v](x)| = \left| \int_{\Omega} D_x^\alpha \kappa_\phi(x, y) v(y) dy \right| \quad (\text{A.26})$$

$$\leq \|D_x^\alpha \kappa_\phi(x, \cdot)\|_{L^q(\Omega)} \|v\|_{L^p(\Omega)} \quad (\text{A.27})$$

$$\leq M |\Omega|^{1/q} \|v\|_{L^p(\Omega)}. \quad (\text{A.28})$$

Raising to the p -th power and integrating over Ω with $1 + p/q = p$,

$$\|D^\alpha \mathcal{K}_\phi v\|_{L^p(\Omega)}^p = \int_{\Omega} |D^\alpha[\mathcal{K}_\phi v](x)|^p dx \leq M^p |\Omega|^{1+p/q} \|v\|_{L^p(\Omega)}^p \quad (\text{A.29})$$

$$= M^p |\Omega|^p \|v\|_{L^p(\Omega)}^p. \quad (\text{A.30})$$

Summing over all $|\alpha| \leq k$ and taking p -th roots,

$$\|\mathcal{K}_\phi v\|_{W^{k,p}(\Omega)} = \left(\sum_{|\alpha| \leq k} \|D^\alpha \mathcal{K}_\phi v\|_{L^p(\Omega)}^p \right)^{1/p} \leq N_k^{1/p} M |\Omega| \|v\|_{L^p(\Omega)}. \quad (\text{A.31})$$

Finally, since $\|v\|_{L^p(\Omega)} \leq \|v\|_{W^{k,p}(\Omega)}$, we conclude

$$\|\mathcal{K}_\phi v\|_{W^{k,p}(\Omega)} \leq N_k^{1/p} M |\Omega| \|v\|_{W^{k,p}(\Omega)} < \infty. \quad (\text{A.32})$$

Thus $\mathcal{K}_\phi$ is a well-defined and bounded operator on $W^{k,p}(\Omega)$. $\square$

Remark A.2.5. The bound in (A.31) is expressed in terms of the weaker norm $\|v\|_{L^p(\Omega)}$ rather than $\|v\|_{W^{k,p}(\Omega)}$. Hence $\mathcal{K}_\phi$ in fact acts as a smoothing operator and defines a bounded operator $\mathcal{K}_\phi : L^p(\Omega) \rightarrow W^{k,p}(\Omega)$, and boundedness of the operator $\mathcal{K}_\phi : W^{k,p}(\Omega) \rightarrow W^{k,p}(\Omega)$ follows as an immediate corollary obtained from $\|v\|_{L^p(\Omega)} \leq \|v\|_{W^{k,p}(\Omega)}$.

Theoretical form of the Neural Operator NO_T on $W^{k,p}$ space

Definition A.2.6 (Neural Operator (theoretical form), [9, Thm. 11]). Let $\Omega \subset \mathbb{R}^d$ be a bounded domain. Let the input space $\mathcal{A} = W^{k,p}(\Omega; \mathbb{R}^{d_a})$ and the output target space $\mathcal{U} = W^{k,p}(\Omega; \mathbb{R}^{d_u})$ be Sobolev spaces with $p \in [1, \infty)$ and $kp > d$, $k \geq 1$.

The *theoretical Neural Operator* $\mathcal{G} : \mathcal{A} \rightarrow \mathcal{U}$ is defined as

$$\mathcal{G} = \Sigma_T \circ \mathcal{K}_{\phi,T} \circ \cdots \circ \Sigma_1 \circ \mathcal{K}_{\phi,1} \quad (\text{A.33})$$

for $T \in \mathbb{N}$, where each layer consists only of a Kernel Integral Operator $\mathcal{K}_{\phi,t}$ with continuously differentiable kernel $\kappa_{\phi,t} \in C^k(\bar{\Omega} \times \bar{\Omega}; \mathbb{R}^{d_v \times d_v})$, followed by an activation Nemytskii operator Σ_t induced by a non-linear activation function $\sigma_t \in C^k(\mathbb{R}^{d_v})$.

Appendix B

Discretization Invariance Proof

This section presents a detailed review of the proof of the Discretization Invariance Theorem 3.3.2, presented by Kovachki et. al. (2023) [9, Thm. 8, Appx. E]. A concise proof outline can be found in section 3.3. Given there are several steps in the proof of Theorem 3.3.2 that require clarification, we present quotes from the proof that require attention and address the points inline. Additionally, we append the omitted steps in derivations of equations presented in the proof.

Proof of Theorem 3.3.2. The proof begins with the statement:

"Without loss of generality, we will assume that $D = D'$ and, by continuous embedding, that $\mathcal{A} = \mathcal{U} = C(\bar{D})$." [9, Appx. E]

Invoking the continuous embedding requires explicit conditions on the parameter and solution spaces $\mathcal{A}$ and $\mathcal{U}$. Kovachki et al. define $\mathcal{A}$ and $\mathcal{U}$ as general Banach spaces. Such an embedding into $C(\bar{D})$ does not hold for $L^p(D)$, so we take $\mathcal{A}$ and $\mathcal{U}$ to be Sobolev spaces and invoke the Sobolev Embedding Theorem [18, Thm. 5.3.28, p. 193] under the corresponding regularity assumptions. Specifically, we assume $\mathcal{A}, \mathcal{U} \subseteq W^{k,p}(D)$ with $k > d/p$, for a bounded $D \subset \mathbb{R}^d$ with Lipschitz boundary. The "without loss of generality" statement therefore holds only under these additional regularity assumptions.

*"Furthermore, note that, **by continuity**, it suffices to show the result for the single layer" [9, Appx. E]*

$$NO = \left\{ f \mapsto \sigma_1 \left(\int_D \kappa(\cdot, y) f(y) dy + b \right) : \right. \quad (B.1)$$

$$\left. \begin{aligned} \kappa &\in N_{n_1}(\sigma_2; \mathbb{R}^d \times \mathbb{R}^d), \\ b &\in N_{n_2}(\sigma_2; \mathbb{R}^d), \quad n_1, n_2 \in \mathbb{N} \end{aligned} \right\}$$

The "by continuity" statement is not precise as stacking arbitrary amount of NO layers may lead to accumulation of the approximation error. Hence, continuity alone is insufficient to control the propagation of the discretization error through the the arbitrary multilayer NO operator and would require establishing quantitative bounds, for instance via Lipschitz continuity or modulus of continuity bounds. We address this limitation in the subsection 3.3.2.

*"Let $K \subset \mathcal{A}$ be a **compact set** and $(D_j)_{j=1}^\infty$ be a discrete refinement of D ." [9]*

Stating that $K \subset \mathcal{A}$ is a compact set is a non-trivial assumption for the set of parameter functions of PDEs. Given that the authors assume $\mathcal{A} \hookrightarrow C(\bar{D})$ earlier in the proof, for $K \subset \mathcal{A}$ to be compact in $C(\bar{D})$ further requirements must be satisfied. By the Ascoli-Arzelà Theorem [18, Thm. 5.4.2., p. 196], to guarantee the existence of a precompact set K in $C(\bar{D})$, the set of functions K must be *uniformly bounded* and *equicontinuous*, which excludes families of functions with arbitrarily steep gradients or high-frequency oscillations. This should be stated explicitly as a regularity assumption on the problem class. We address this limitation in the alternative proof that we propose in subsection 3.3.2.

The authors further define the partition of the domain D as follows:

"To each discretization D_j (of D) associate partitions $P_j^{(1)}, \dots, P_j^{(j)} \subseteq D$ which are pairwise disjoint, each contains a single, unique point of D_j , each has positive Lebesgue measure, and [9, Appx. E]"

$$\bigcup_{k=1}^j P_j^{(k)} = D$$

Hence, each partition cell $P_j^{(k)}$ contains exactly one point y_k of the discretization D_j , with $k \in \{1, \dots, j\}$. This pairing of a partition P_j with a discretization D_j allows the definition of a discrete approximation $\hat{\mathcal{G}}_j : \mathbb{R}^{jd} \times \mathbb{R}^j \rightarrow \mathcal{U}$ of an operator $\mathcal{G} \in NO$ with parameters κ, b as a Riemann sum

$$\hat{\mathcal{G}}_j(y_1, \dots, y_j, w_1, \dots, w_j) = \sigma_1 \left(\sum_{k=1}^j \kappa(\cdot, y_k) w_k |P_j^{(k)}| + b(\cdot) \right), \quad (\text{B.2})$$

where $w_k = a(y_k)$ are the pointwise evaluations of the parameter function $a \in \mathcal{A}$ at the discretization points $y_k \in D_j$, and $|\cdot|$ denotes the Lebesgue measure of a partition cell.

This concludes the setup. The two regularity assumptions invoked above now play their roles in the convergence argument. The embedding $\mathcal{A} \hookrightarrow C(\bar{D})$ makes the pointwise evaluations $a|_{D_j}$ well defined, so that the Riemann sum $\hat{\mathcal{G}}_j$ approximates the kernel integral operator. The compactness of $K \subset \mathcal{A}$ supplies, via total boundedness, the finite ε -net $\{a_n\}_{n=1}^N \subset K$ that reduces the estimate over all of K to a finite family of representatives. The remainder of the proof combines these two ingredients in a three- ε argument.

Kovachki et al. use the compactness of $K \subset \mathcal{A}$ to establish a uniform bound on the parameter functions:

$$\sup_{a \in K} \|a\|_{\mathcal{U}} \leq M. \quad (\text{B.3})$$

Since $\mathcal{A} = \mathcal{U} = C(\bar{D})$, this bound is the supremum norm, and the Sobolev embedding gives $\|a\|_{\infty} \leq C \|a\|_{W^{k,p}}$ [18, Thm. 5.3.28, p. 193]. We absorb the embedding constant C into M to stay consistent with the original notation. The omitted bound on the integral operator then follows by the triangle inequality:

$$\left| \int_D \kappa(x, y) a(y) dy \right| \leq \int_D |\kappa(x, y)| |a(y)| dy \quad (\text{B.4})$$

$$\leq \|\kappa\|_{C(\bar{D} \times \bar{D})} \|a\|_{\infty} \int_D 1 dy \quad (\text{B.5})$$

$$\leq M |D| \|\kappa\|_{C(\bar{D} \times \bar{D})}. \quad (\text{B.6})$$

Hence the continuous pre-activation expression $\int_D \kappa(x, y) a(y) dy + b(x)$ is bounded by

$$\tilde{R} := M |D| \|\kappa\|_{C(\bar{D} \times \bar{D})} + \|b\|_{\infty}. \quad (\text{B.7})$$

The discrete pre-activation expression $\sum_{k=1}^j \kappa(x, y_k) a(y_k) |P_j^{(k)}| + b(x)$ satisfies the same bound, since

$$\sum_{k=1}^j |\kappa(x, y_k)| |a(y_k)| |P_j^{(k)}| \leq \|\kappa\|_{C(\bar{D} \times \bar{D})} \|a\|_{\infty} \sum_{k=1}^j |P_j^{(k)}| \quad (\text{B.8})$$

$$= M |D| \|\kappa\|_{C(\bar{D} \times \bar{D})} \quad (\text{B.9})$$

by $\bigcup_{k=1}^j P_j^{(k)} = D$. Kovachki et al. instead bound the two expressions jointly:

$$\begin{aligned} \sup_{x \in \bar{D}} \sup_{j \in \mathbb{N}} \left| \int_D \kappa(x, y) a(y) dy + \sum_{k=1}^j \kappa(x, y_k) a(y_k) |P_j^{(k)}| + 2b(x) \right| \\ \leq 2(M|D| \|\kappa\|_{C(\bar{D} \times \bar{D})} + \|b\|_{C(\bar{D})}) =: R. \end{aligned} \quad (\text{B.10})$$

which has a factor of 2 from summing both expressions under a single supremum. Since $\tilde{R} = R/2$, we have $[-\tilde{R}, \tilde{R}] \subset [-R, R]$, so uniform continuity of σ_1 on $[-R, R]$ in the next step holds on the tighter interval as well.

Kovachki et al. further state

*"Hence we need only consider σ_1 as a map $[-R, R] \rightarrow \mathbb{R}$. Thus, by **uniform continuity**, there exists a **modulus of continuity** $\omega : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}_{\geq 0}$ which is continuous, non-negative, and non-decreasing on $\mathbb{R}_{\geq 0}$, satisfies $\omega(z) \rightarrow \omega(0) = 0$ as $z \rightarrow 0$ and*

$$|\sigma_1(z_1) - \sigma_1(z_2)| \leq \omega(|z_1 - z_2|) \quad \forall z_1, z_2 \in [-R, R]. \quad (\text{B.11})$$

" [9]

Since the domain $[-R, R]$ is compact and σ_1 is continuous, the restriction $\sigma_1 : [-R, R] \rightarrow \mathbb{R}$ is uniformly continuous by the Heine-Cantor Theorem [18, Prop. 1.8.20, p. 32], which is what guarantees the modulus of continuity ω in (B.11). This modulus is the mechanism that turns a bound on the pre-activation error into a bound on the operator error. A discretization grid D_j fine enough to bring the Riemann sum within δ of the kernel integral results through ω in an approximation error below ε between $\mathcal{G} \in \mathcal{NO}$ and its discretization $\hat{\mathcal{G}}_j$. Formally, to show that there exists $Q(\varepsilon) \in \mathbb{N}$ such that for all $m > Q$

$$\sup_{a \in K} \|\hat{\mathcal{G}}(D_m, a|_{D_m}) - \mathcal{G}(a)\|_{\mathcal{U}} < \varepsilon, \quad (\text{B.12})$$

it suffices to show that for all $m > Q$

$$\sup_{a \in K} \sup_{x \in \bar{D}} \left| \int_D \kappa(x, y) a(y) dy - \sum_{k=1}^m \kappa(x, y_k) a(y_k) |P_m^{(k)}| \right| < \delta. \quad (\text{B.13})$$

We use δ for the pre-activation bound and ε for the operator bound to keep the two components of the final argument distinct, since [9] denote both by ε .

In order to show that (B.13) holds, the authors rely on compactness of $K \subset \mathcal{A}$, which guarantees the existence of an ε -net $\{a_n\}_{n=1}^N \subset K$ such that for any $a \in K$ there exists $n \in \{1, \dots, N\}$ with

$$\|a - a_n\|_{C(\bar{D})} \leq \frac{\delta}{4|D| \|\kappa\|_{C(\bar{D} \times \bar{D})}}, \quad (\text{B.14})$$

where the denominator is chosen so that the continuous and discrete approximation errors below reduce to multiples of $\delta/4$. This is central to the proof, as it allows the discretization of the infinite-dimensional set $K \subset \mathcal{A}$ by a finite family of N representative functions $\{a_n\}_{n=1}^N \subset K$.

"Since (D_j) is a discrete refinement, **by convergence of Riemann sums**, we can find some $q \in \mathbb{N}$ such that for any $t \geq q$, we have

$$\sup_{x \in \bar{D}} \left| \sum_{k=1}^t \kappa(x, y_k) |P_t^{(k)}| - \int_D \kappa(x, y) dy \right| < |D| \|\kappa\|_{C(\bar{D} \times \bar{D})} \quad (\text{B.15})$$

where $D_t = \{y_1, \dots, y_t\}$ [9]. "

The right-hand side here is the constant $|D| \|\kappa\|_{C(\bar{D} \times \bar{D})}$, which does not vanish as $t \rightarrow \infty$. The bound (B.15) is therefore not itself a convergence statement, despite the phrasing "by convergence of Riemann sums". It is a uniform estimate on the kernel quadrature, used later in (B.23). The actual Riemann convergence the proof relies on appears separately in (B.16), where the integrand $\kappa(x, \cdot) a_n(\cdot)$ is fixed and continuous on the compact domain $\bar{D}$, and the quadrature error can be bounded by $\delta/4$.

For the next estimate, the authors fix a single representative a_n from the ε -net and bound the quadrature error for that function. Since a_n is continuous and κ is continuous, the product $\kappa(x, \cdot) a_n(\cdot)$ is continuous, hence Riemann integrable, on the compact domain $\bar{D}$, so the Riemann sums converge to the integral as the grid spacing goes to zero. This convergence is non-quantitative, as it defines a threshold p_n for each a_n but no rate, which is what makes the argument an existence proof rather than a constructive one. We address this in subsection 3.3.2.

"Similarly, we can find $p_1, \dots, p_N \in \mathbb{N}$ such that, for any $t_n \geq p_n$, we have

$$\sup_{x \in \bar{D}} \left| \sum_{k=1}^{t_n} \kappa(x, y_k^{(n)}) a_n(y_k^{(n)}) |P_{t_n}^{(k)}| - \int_D \kappa(x, y) a_n(y) dy \right| < \frac{\delta}{4} \quad (\text{B.16})$$

where $D_{t_n} = \{y_1^{(n)}, \dots, y_{t_n}^{(n)}\}$. Let $m \geq \max\{q, p_1, \dots, p_N\}$ and denote $D_m = \{y_1, \dots, y_m\}$. " [9]

The threshold p_n and grid D_{t_n} depend on the representative a_n . Taking $m \geq \max\{q, p_1, \dots, p_N\}$ produces a single refinement D_m fine enough that (B.15) and (B.16) hold simultaneously for every a_n , $n \in \{1, \dots, N\}$. The finiteness of the net is what makes this maximum well defined.

Next, the authors derive two inequalities based on (B.14), used in the final estimate. We fill in the omitted derivation steps.

Applying the triangle inequality for integrals and bounding the integrand by its supremum norm gives

$$\sup_{x \in \bar{D}} \left| \int_D \kappa(x, y) (a(y) - a_n(y)) dy \right| \leq \sup_{x \in \bar{D}} \int_D |\kappa(x, y)| |a(y) - a_n(y)| dy \quad (\text{B.17})$$

$$\leq \|\kappa\|_{C(\bar{D} \times \bar{D})} \|a - a_n\|_{C(\bar{D})} \int_D 1 dy \quad (\text{B.18})$$

$$= |D| \|\kappa\|_{C(\bar{D} \times \bar{D})} \|a - a_n\|_{C(\bar{D})}. \quad (\text{B.19})$$

For the discrete part, we factor out the function difference, then add and subtract the integral $\int_D \kappa(x, y) dy$, applying (B.15) to the first resulting term and the bound $|\int_D \kappa(x, y) dy| \leq$

$|D| \|\kappa\|_{C(\bar{D} \times \bar{D})}$ to the second:

$$\sup_{x \in \bar{D}} \left| \sum_{k=1}^m \kappa(x, y_k) (a_n(y_k) - a(y_k)) |P_m^{(k)}| \right| \quad (\text{B.20})$$

$$\leq \|a_n - a\|_{C(\bar{D})} \sup_{x \in \bar{D}} \left| \sum_{k=1}^m \kappa(x, y_k) |P_m^{(k)}| \right| \quad (\text{B.21})$$

$$\leq \|a_n - a\|_{C(\bar{D})} \left(\sup_{x \in \bar{D}} \left| \sum_{k=1}^m \kappa(x, y_k) |P_m^{(k)}| - \int_D \kappa(x, y) dy \right| + \sup_{x \in \bar{D}} \left| \int_D \kappa(x, y) dy \right| \right) \quad (\text{B.22})$$

$$\leq 2 |D| \|\kappa\|_{C(\bar{D} \times \bar{D})} \|a_n - a\|_{C(\bar{D})}. \quad (\text{B.23})$$

We now assemble the final estimate. For any $a \in K$, choose a representative a_n from the ε -net. Adding and subtracting both the continuous integral $\int_D \kappa(x, y) a_n(y) dy$ and the discrete sum of a_n , the triangle inequality divides the quadrature error into three components: the continuous approximation error (B.19), the Riemann quadrature error for a_n (B.16), and the discrete approximation error (B.23).

$$\begin{aligned} & \sup_{x \in \bar{D}} \left| \int_D \kappa(x, y) a(y) dy - \sum_{k=1}^m \kappa(x, y_k) a(y_k) |P_m^{(k)}| \right| \\ & \leq \underbrace{\sup_{x \in \bar{D}} \left| \int_D \kappa(x, y) (a(y) - a_n(y)) dy \right|}_{\text{continuous approximation error}} \\ & \quad + \underbrace{\sup_{x \in \bar{D}} \left| \int_D \kappa(x, y) a_n(y) dy - \sum_{k=1}^m \kappa(x, y_k) a_n(y_k) |P_m^{(k)}| \right|}_{\text{Riemann quadrature error for } a_n} \\ & \quad + \underbrace{\sup_{x \in \bar{D}} \left| \sum_{k=1}^m \kappa(x, y_k) (a_n(y_k) - a(y_k)) |P_m^{(k)}| \right|}_{\text{discrete approximation error}} \end{aligned} \quad (\text{B.24})$$

$$\begin{aligned} & \leq \underbrace{|D| \|\kappa\|_{C(\bar{D} \times \bar{D})} \|a - a_n\|_{C(\bar{D})}}_{< \delta/4} + \underbrace{\frac{\delta}{4}}_{< \delta/4} + \underbrace{2 |D| \|\kappa\|_{C(\bar{D} \times \bar{D})} \|a - a_n\|_{C(\bar{D})}}_{< 2\delta/4} \\ & < \frac{\delta}{4} + \frac{\delta}{4} + \frac{2\delta}{4} = \delta. \end{aligned} \quad (\text{B.25})$$

The first and third terms are bounded using (B.19), (B.23), and the net estimate (B.14); the middle term by (B.16). Applying the modulus of continuity ω from (B.11) to this pre-activation bound, we conclude that for every $\varepsilon > 0$ there exists $Q(\varepsilon) \in \mathbb{N}$ such that for all $m > Q$

$$\sup_{a \in K} \|\hat{\mathcal{G}}(D_m, a|_{D_m}) - \mathcal{G}(a)\|_{C(\bar{D})} < \varepsilon. \quad (\text{B.26})$$

This completes the proof. $\square$