ANY-DIMENSIONAL INVARIANT UNIVERSALITY

by

Shengtai Yao

A thesis submitted to Johns Hopkins University
in conformity with the requirements for the degree of
Master of Science in Engineering

Baltimore, Maryland
December, 2025

Abstract

Traditional supervised learning typically operates within fixed input-output dimensions. In contrast, the emerging research field of any-dimensional learning aims to approximate sequences of functions defined over arbitrary dimensions. By leveraging symmetries such as group action invariance and equivariance, these methods enable models trained in low-dimensional spaces to generalize effectively to high-dimensional settings. Building upon the recent transferability framework [1], which guarantees performance consistency across dimensions, this work addresses the fundamental challenge of expressivity. Specifically, under the constraint of group action invariance, we propose transferable architectures and establish their universality for three fundamental types of any-dimensional data: sets, graphs, and point clouds.

Thesis Reader

Mateo Díaz
Assistant Professor
Department of Applied Mathematics and Statistics
Johns Hopkins Whiting School of Engineering

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my thesis advisor, Mateo Díaz. Thank you for introducing me to this field and for your unwavering support along this journey. I am deeply appreciative of the countless hours you have spent helping me master the fundamentals, refine my ideas, polish my writing, and support me in more ways than I can list. Your patience and encouragement consistently gave me the confidence to keep moving forward.

I also extend my sincere appreciation to Eitan Levin. I am deeply grateful for your exceptional patience and for always being available to answer my many questions. Thank you for being a constant source of knowledge and inspiration. This thesis is impossible without your sharp insights and deep understanding.

And finally, I want to thank my friends and family. Thank you for your unconditional belief in me and for supporting me throughout this academic journey. You are always my warmest place to rest and recharge.

Table of Contents

Chapter 1

Introduction

Traditional supervised learning typically aims to learn mappings between fixed input and output dimensions using training examples of the same size. However, many modern learning tasks involve maps defined for arbitrary dimensions. For instance, graph parameters, particle system dynamics, and regularizers remain meaningful regardless of the number of nodes, particles, or variables. To tackle these problems, several recent any-dimensional architectures have been proposed. These architectures parameterize an infinite sequence of functions with varying input and output dimensions using only finitely many parameters. Examples include Graph Neural Networks (GNN) [2], DeepSets [3], and PointNet [4].

Following this, it is important to understand the specific properties of the model, data, and task that allow learning performance to generalize across varying dimensions. This phenomenon, termed transferability, was first studied in the literature on GNNs [5, 2, 6, 7] by leveraging the tools of graphon theory [8]. Recent work [1] proposed a general framework for transferability that extends beyond the context of GNNs. They provide a formal definition of the concept

and demonstrate that transferability is achieved whenever specific conditions of compatibility and continuity are satisfied. Given the theoretical framework, a natural but fundamental question is:

> Can any-dimensional models universally approximate functions defined across dimensions?

Universal approximation is a cornerstone of machine learning theory, originating from the classical universality results for fully connected neural networks [9]. However, for any-dimensional models, the challenge becomes more subtle. Such architectures must be not only universal, but also transferable across dimensions. Achieving this dual objective is equivalent to constructing a model that is both continuous and universal within the corresponding limit space.

Main contributions Next, we summarize our main contributions.

1. (*Universality for any-dimensional set models*) We investigate both zero-padding and duplication-padding compatibilities for permutation-invariant set functions. For each case, we propose transferable invariant models and establish their universality. Our architectures build based on DeepSets [3] and normalized DeepSets [10]. Furthermore, we extend the conditions for the universality of normalized DeepSets originally established in [10].

2. (*Universality for any-dimensional graph models*) We study duplication-padding compatibility and invariance under node permutation for graph functions, which naturally aligns with the graphon theory in [8]. We introduce Tensor Contraction Graphon Networks (TGN) and establish its

universality within the graphon space equipped with the δ_p distance. Furthermore, our architecture is capable of approximating all functions continuous with respect to the cut distance $\delta_\square$, representing an advancement over the architecture proposed in [11].

3. (*Universality for any-dimensional point cloud models*) We analyze duplication-padding compatibility alongside invariance under both point permutation and rotation action for point cloud functions. We propose an architecture based on Invariant Graphon Networks (IWN) [11] and prove its universality.

Outline Section 1.2 in this chapter reviews related work. Chapter 2 summarizes the theoretical framework for transferability across dimensions introduced in [1].Chapter **??** first formalizes our goal of universality and outlines the general proof sketch, and then presents specific instantiations for set, graph, and point cloud functions; for each case, we introduce a transferable model and prove its universality. Chapter 4 includes all the missing proofs from Chapter **??**. Finally, Chapter 5 concludes the paper and outlines the directions for future work.

1.1 Notation

We use $\mathbb{R}$ to denote the set of real numbers, $\mathbb{N}$ the set of positive integers, and $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. For any $n \in \mathbb{N}$, let $[n] := \{1, \dots, n\}$. We denote the orthogonal group of dimension k by $\mathrm{O}(k)$, and the symmetric group of permutations of n elements by S_n. On the domain $[0, 1]$, we denote the class of measure-preserving bijections by $S_{[0,1]}$, and the class of measure-preserving functions by $\overline{S}_{[0,1]}$. Let

$C(\mathcal{X}, \mathcal{Y})$ denote the class of continuous functions mapping from a topological space $\mathcal{X}$ to $\mathcal{Y}$. When $\mathcal{Y} = \mathbb{R}$, we simply write $C(\mathcal{X})$. For a normed vector space $\mathcal{X}$, $C_0(\mathcal{X})$ denotes the subspace of $C(\mathcal{X})$ consisting of functions vanishing at infinity. Let $\mathcal{B}(\mathcal{X}, \mathcal{Y})$ denote the space of bounded linear operators from a normed space $\mathcal{X}$ to $\mathcal{Y}$. For a measure space $(\mathcal{X}, \mu)$, let $L^p(\mathcal{X}; \mathbb{R}^k)$ denote the space of measurable functions $f : \mathcal{X} \to \mathbb{R}^k$ with finite L^p norm, i.e., $\|f\|_p := (\int_{\mathcal{X}} \|f(x)\|_{\mathbb{R}^k}^p d\mu(x))^{1/p} < \infty$, where $\|\cdot\|_{\mathbb{R}^k}$ is a fixed norm on $\mathbb{R}^k$. We denote by $\ell_p(\mathbb{R}^d)$ the sequence space of real-valued vector sequences $X = (X_{n:})_{n\in\mathbb{N}}$ satisfying $\sum_{n=1}^{\infty} \|X_{n:}\|_{\mathbb{R}^d}^p < \infty$. We use $\|\cdot\|_p$ to denote L^p or ℓ_p norm. Let $\mathcal{N}^{w,\phi}(d, h)$ denote the class of functions $f : \mathbb{R}^d \to \mathbb{R}^h$ represented by w-layer fully connected neural networks with activation function ϕ, where $d, h \in \mathbb{N}$. Let $\mathcal{F}_p(\mathbb{R}^d, \mathbb{R}^h)$ be the class of continuous functions satisfying the p-moment growth condition; that is, there exists a constant $C > 0$ such that $\|f(x)\| \le C(1 + \|x\|^p)$ for all $x \in \mathbb{R}^d$. Let $\mathcal{P}(\mathcal{X})$ be the set of probability measures on a space $\mathcal{X}$. We denote by $\mathcal{P}_p(\mathbb{R}^k)$ the subset of probability measures on $\mathbb{R}^k$ with finite p-th moment. For a measurable function $f : \mathcal{X} \to \mathcal{Y}$ and a measure $\mu \in \mathcal{P}(\mathcal{X})$, we denote the pushforward measure on $\mathcal{Y}$ by $f_{\#}\mu$, defined as $(f_{\#}\mu)(A) = \mu(f^{-1}(A))$ for all measurable sets $A \subseteq \mathcal{Y}$.

1.2 Related work

Expressivity of machine learning The universal approximation capability of fully connected neural networks for continuous functions in compact subsets was established in early seminal works [9, 12, 13, 14]. Recently, several studies have generalized these results to unbounded domains [15, 16, 17], provided

that the target continuous functions and network activation functions satisfy specific conditions. In the context of GNNs, representation capacity is known to be limited [18] and is typically characterized by the Weisfeiler-Lehman (WL) graph isomorphism test [19]. Numerous works have investigated the expressive power of various GNNs variants using the WL-test [18, 20, 21]. For invariant and equivariant machine learning, several architectures have been proven universal for approximating continuous invariant/equivariant maps [22, 23, 24, 25]. In contrast, the expressive capacity of any-dimensional learning models remains under-explored. While specific results exist [11, 10], these works were not developed within the formal framework of any-dimensional transferability.

Transferability across dimensions A critical phenomenon in any-dimensional learning is the preservation of output consistency across dimensions, provided the inputs remain structurally similar. In the context of GNNs, this property was introduced as transferability [2]. Related literature also refers to this as the convergence and stability of the GNN model [26]. This concept has since been further developed [27, 5] and widely studied in various GNN variants [28, 29, 30]. More recently, [1] proposed a unified framework for analyzing data defined across dimensions, extending beyond the graph domain. Building upon their previous work [31, 32], they introduce consistent sequences to define a limit space that contains input objects of arbitrary dimension. This framework enables the computation of distances between objects of differing dimensions.

Equivariant machine learning Our work naturally aligns with equivariant machine learning, a field that focuses on incorporating symmetry properties into

learning models. Initial investigations into this field primarily focus on graph convolutional networks [33, 34]. Subsequently, these approaches are unified under the framework of geometric machine learning [35], applying geometric principles to a wide variety of data structures, including sets [36, 37], graphs [38, 39], point clouds [40, 41]. Besides graph convolutions, there are also many other equivariant machine learning models expressed in terms of representation theory [42, 43, 44], canonicalization [45], and invariant theory [46, 47, 48], among others.

Dimension-free learning Numerous dimension-free learning models exist that are capable of processing inputs of arbitrary dimensions. In this work, our focus is on architectures that preserve specific symmetry properties across dimensions, as exemplified by GNNs, DeepSets, among others discussed previously. Beyond these, the landscape of dimension-free learning involves several other well-known paradigms. Convolutional Neural Networks (CNNs) [49] process arbitrary image sizes via translation invariant convolution filters. Recurrent Neural Networks (RNNs) [50] and Transformers [51] can handle arbitrary length of natural language sentences. Neural Operators [52, 53] are designed to learn mesh-free operators, allowing to work on grids of varying resolutions.

Chapter 2

Background on transferability across dimensions

In this chapter, we introduce the background on transferability across dimensions, relying on the theoretical developments detailed in [1]. In brief, the framework relies on two key conditions: (i) compatibility across dimensions, which permits the definition of cross-dimensional functions on an extended limit space, and (ii) continuity within this limit space, which enables meaningful comparisons between input objects of different dimensions.

First, we construct the limit space with the definition of a consistent sequence. A detailed formulation is provided in Definition 2.2. The direct sum of consistent sequence is defined in Definition 2.3.

Definition 2.1. *A **consistent sequence** $\mathbb{V} = \left\{(V_n)_{n\in\mathbb{N}}, \left(\varphi_{N,n}\right)_{n\preceq N}, (\mathrm{G}_n)_{n\in\mathbb{N}}\right\}$ is a sequence of finite dimensional vector spaces V_n, maps $\varphi_{N,n}$ and groups G_n acting linearly on V_n, indexed by a directed poset [1] $(\mathbb{N}, \preceq)$, such that for all $n \preceq N$, the group G_n is embedded into G_N, and $\varphi_{N,n} : V_n \hookrightarrow V_N$ is a linear, G_n-equivariant embedding.*

[1] The directed poset is a partial order on $\mathbb{N}$ where every two elements have an upper bound.

Definition 2.2 (Detailed version of Definition 2.1)**.** *A* ***consistent sequence*** *of group representations over directed poset* $(\mathbb{N}, \preceq)$ *is*

$$\mathbb{V} = \left\{(V_n)_{n\in\mathbb{N}}, \left(\varphi_{N,n}\right)_{n\preceq N}, (\mathrm{G}_n)_{n\in\mathbb{N}}\right\},$$

where,

1. *(**Groups**)* (G_n) *is a sequence of groups indexed by* $\mathbb{N}$ *such that whenever* $n \preceq N, \mathrm{G}_n$ *is embedded into* G_N *via an injective group homomorphism* $\theta_{N,n} : \mathrm{G}_n \to \mathrm{G}_N$*, where*

$$\theta_{i,i} = \mathrm{id}_{\mathrm{G}_i} \quad \text{for all } i \in \mathbb{N},$$

$$\theta_{k,j} \circ \theta_{j,i} = \theta_{k,i} \quad \text{whenever } i \preceq j \preceq k \text{ in } \mathbb{N}.$$

2. *(**Vector spaces**)* (V_n) *is a sequence of (finite-dimensional, real) vector spaces indexed by* $\mathbb{N}$*, such that each* V_n *is a* G_n*-representation, and whenever* $n \preceq N, V_n$ *is embedded into* V_N *through a linear embedding* $\varphi_{N,n} : V_n \hookrightarrow V_N$*, where*

$$\varphi_{i,i} = \mathrm{id}_{V_i} \quad \text{for all } i \in \mathbb{N},$$

$$\varphi_{k,j} \circ \varphi_{j,i} = \varphi_{k,i} \quad \text{whenever } i \preceq j \preceq k \text{ in } \mathbb{N}.$$

3. *(**Equivariance**) Every* $\varphi_{N,n}$ *is* G_n*-equivariant, i.e.,*

$$\varphi_{N,n}(g \cdot v) = \theta_{N,n}(g) \cdot \varphi_{N,n}(v) \text{ for all } g \in \mathrm{G}_n, v \in V_n.$$

Definition 2.3. *The* d*-th direct sum of* $\mathbb{V}$ *is defined as*

$$\mathbb{V}^{\oplus d} := \left\{\left(V_n^{\oplus d}\right), \left(\varphi_{N,n}^{\oplus d}\right), (\mathrm{G}_n)\right\},$$

where $V_n^{\oplus d}$ *denotes the direct sum of* d *copies of* V_n *and* $\varphi_{N,n}^{\oplus d} : V_n^{\oplus d} \to V_N^{\oplus d}$ *is defined by applying* $\varphi_{N,n}$ *to each component. The group* G_n *acts on* $V_n^{\oplus d}$ *by simultaneously acting on every copy of* V_n*, i.e.* $g \cdot (v_1, \ldots, v_d) := (g \cdot v_1, \ldots, g \cdot v_d)$.

We provide two types of consistent sequences that will be the focus of this work in Example 2.4.

Example 2.4. *Let* $V_n = \mathbb{R}^n$*, with the symmetric group* S_n *acting by coordinate permutation.*

1. *(**Zero-padding embedding**) The sequence is indexed by* $\mathbb{N}$ *with the standard ordering* $\leq$*, with the vector spaces embedding map* $\varphi_{N,n} : V_n \hookrightarrow V_N$*,*

$$\varphi_{N,n}((x_1, \ldots, x_n)) = (x_1, \ldots, x_n, \underbrace{0, \ldots, 0}_{(N-n)\ 0\text{'s}}).$$

2. *(**Duplication-padding embedding**) The sequence is indexed by* $\mathbb{N}$ *with the divisibility ordering (*$n \preceq N$ *whenever* N *is divisible by* n*), with the vector spaces embedding map* $\varphi_{N,n} : V_n \hookrightarrow V_N$*,*

$$\varphi_{N,n}((x_1, \ldots, x_n)) = (\underbrace{x_1, \ldots, x_1}_{N/n \text{ copies}}, \ldots, \underbrace{x_n, \ldots, x_n}_{N/n \text{ copies}}).$$

We are now able to define a common space that contains objects of any dimensions, identified through the embedding equivalence. We further introduce a notion of function compatibility, which ensures that mappings between such spaces are well defined.

Definition 2.5. *Define* V_∞ *as the disjoint union* $\bigsqcup V_n$ *modulo an equivalence*

relation

$$V_\infty := \bigsqcup_n V_n/\sim,$$

where $v \sim \varphi_{N,n}(v)$ *whenever* $n \preceq N$*, and denote* $[v] \in V_\infty$ *as the equivalence class of* $v \in V_n$*.* G_∞ *and its element* $[g]$ *are defined analogously with the group embedding map* $\theta_{N,n}$*.*

Definition 2.6. *Let* $\mathbb{V} = \left\{(V_n), \left(\varphi_{N,n}\right), (\mathrm{G}_n)\right\}$ *and* $\mathbb{U} = \left\{(U_n), \left(\psi_{N,n}\right), (\mathrm{G}_n)\right\}$ *be two consistent sequences indexed by* $(\mathbb{N}, \preceq)$*. A sequence of maps* $(f_n : V_n \to U_n)$ *is* ***compatible*** *with respect to* $\mathbb{V}, \mathbb{U}$ *if* $f_N \circ \varphi_{N,n} = \psi_{N,n} \circ f_n$ *for all* $n \preceq N$*, and each* f_n *is* G_n*-equivariant.*

It has been established in [1] that compatible maps are precisely those sequences that admit an extension to the limit, as formally stated in Proposition 2.7.

Proposition 2.7. *Let* $\mathbb{V} = \left\{(V_n), \left(\varphi_{N,n}\right), (\mathrm{G}_n)\right\}$ *and* $\mathbb{U} = \left\{(U_n), \left(\psi_{N,n}\right), (\mathrm{G}_n)\right\}$ *be two consistent sequences. A sequence of maps* $(f_n : V_n \to U_n)$ *is compatible if and only if it extends to the limit; that is, there exists a* G_∞*-equivariant map*

$$f_\infty : V_\infty \to U_\infty$$

such that $f_n = f_\infty|_{V_n}$ *for all* n*.*

The next step is to endow V_∞ with a metric, allowing us to quantify how "close" two objects of different dimensions are.

Definition 2.8. *For a consistent sequence* $\mathbb{V}$*, a sequence of norms* $\left(\|\cdot\|_{V_n}\right)$ *on* V_n *is compatible if all the embeddings* $\varphi_{N,n}$ *and the* G_n*-actions are isometries. i.e., for all* $n \preceq N, x \in V_n, g \in \mathrm{G}_n$*,* $\left\|\varphi_{N,n}x\right\|_{V_N} = \|x\|_{V_n}$ *and* $\|g \cdot x\|_{V_n} = \|x\|_{V_n}$*.*

Similarly to compatible maps, compatible norms are precisely those that extend to the limit, that is, there exists a norm $\|\cdot\|_{V_\infty}$ on V_∞ such that for any n and $x \in V_n$, $\|x\|_{V_n} = \|x\|_{V_\infty}$, and the G_∞-action on V_∞ is an isometry with respect to $\|\cdot\|_{V_\infty}$; see Proposition 2.9 for a formal statement.

Proposition 2.9. *A sequence of metrics (d_n) on the spaces V_n is compatible if and only if it extends to a metric on the limit space. That is, there exists a metric d_∞ on V_∞ such that*

$$d_n(x, y) = d_\infty([x], [y]) \quad \text{for all } n \in \mathbb{N}, x, y \in V_n\,,$$

and the G_∞-action on V_∞ is an isometry with respect to d_∞, i.e.,

$$d_\infty(g \cdot x, g \cdot y) = d_\infty(x, y) \quad \text{for all } x, y \in V_\infty, g \in \mathrm{G}_\infty\,.$$

These constructions allow us to define a limit space that contains not only the equivalence classes of finite-dimensional objects, but also their limits.

Definition 2.10. *The **limit space** is the pair $\left(\overline{V_\infty}, \mathrm{G}_\infty\right)$ where $\overline{V_\infty}$ denotes the completion of V_∞ with respect to $\|\cdot\|_{V_\infty}$, endowed with the symmetrized metric*

$$\overline{\mathrm{d}}(x, y) := \inf_{g \in \mathrm{G}_\infty} \|g \cdot x - y\|_{V_\infty} \quad \text{for } x, y \in \overline{V_\infty}$$

This symmetrized metric is a pseudometric on $\overline{V_\infty}$ and a metric on the orbit closures of $\overline{V_\infty}$ under the action of G_∞. With this metric, we can consider functions on orbit closures that map "close" input objects to "close" outputs; this leads to the notion of transferability. The established theorem formally states that transferability of sequence of maps is exactly continuity in the limit space.

Definition 2.11. *Let $\mathbb{V}, \mathbb{U}$ be consistent sequences endowed with norms. A sequence of maps $(f_n : V_n \to U_n)$ is* ***continuously transferable*** *if there exists $f_\infty : \overline{V_\infty} \to \overline{U_\infty}$ that is continuous with respect to $\|\cdot\|_{V_\infty}, \|\cdot\|_{U_\infty}$, such that $f_n = f_\infty|_{V_n}$ for all n. Notice that if (f_n) is transferable and equivariant, then it must be compatible.*

Theorem 2.12. *Let $\mathbb{V}, \mathbb{U}$ be consistent sequences and let $(f_n : V_n \to U_n)$ be maps between them. For any sequence $(x_n \in V_n)$ converging to a limiting object $x \in \overline{V_\infty}$ in $\overline{\mathrm{d}}$, if (f_n) is continuously transferable, then $\overline{\mathrm{d}}\left(f_n\left(x_n\right), f_m\left(x_m\right)\right) \to 0$ as $n, m \to \infty$. Furthermore, if (x_n) converges to x at rate $R(n)$ and (f_n) is locally Lipschitz-transferable, then*

$$\overline{\mathrm{d}}\left(f_n\left(x_n\right), f_m\left(x_m\right)\right) \lesssim R(n) + R(m).$$

Chapter 3

Instantiations

In this chapter, we study several instantiations involving sets, graphs, and point clouds. We begin by restating our goal in a more detailed form and outlining a general proof strategy for each case. Section 3.1 addresses the case of sets, Section 3.2 focuses on graphs, and Section 3.3 discusses point clouds.

In this paper, we consider the sequence of real-valued functions $(f_n : V_n \to \mathbb{R})$, where each f_n is G_n-**invariant**.

First, our goal is to represent or approximate a transferable sequence. This is meaningful only when the models are themselves transferable. In other words, we aim to establish universality for any-dimensional models, which are corresponding to **continuous** functions in the orbit closure of the limit space. A crucial observation is that, when constructing such continuous machine-learning models, the choice of **activation function** plays a decisive role. We will later see, in the graph case, that commonly used pointwise nonlinearities in fact lead to discontinuous models. Thus, the activation function must be chosen carefully to guarantee equivariance, continuity, and universality. Ensuring all simultaneously is the central challenge.

Second, universality is typically established on a **compact** input domain, following Stone-Weierstrass arguments, which we recall below. The universal approximation property is meaningful only when the domain under consideration is rich enough to cover common practical cases. Fortunately, under invariance conditions, we can work within the orbit closure, which is a metric space, where the compact subsets are sufficiently rich. Moreover, in certain settings, such as set functions, one can even establish a necessary and sufficient condition for the compactness of the orbit closure.

Proposition 3.1 (Stone-Weierstrass Theorem ([54], Thm. 5.7))**.** *Let $C(S)$ be the sup-normed Banach space of all continuous real-valued functions on a compact Hausdorff space S. Let A be a closed subalgebra of $C(S)$ which contains a non-zero constant function. Then A is dense in $C(S)$ if and only if it separates points.*

Based on the two points above, the structure of each section will follow three main steps:

1. Characterize **compactness** within the orbit closure of the limit space.
2. Construct an any-dimensional model that is **continuous** on this space.
3. Establish **universality** via the Stone–Weierstrass theorem by showing that the proposed model forms a **subalgebra** and **separates points** in the orbit closure.

3.1 Set functions

3.1.1 Zero-padding compatibility

The zero-padding consistent sequence $\mathbb{V}_{\text{zero}} = \left\{(V_n), \left(\varphi_{N,n}\right), (\mathrm{G}_n)\right\}$ is defined as follows. The index set $\mathbb{N} = (\mathbb{N}, \leq)$ is the poset of natural numbers equipped with the standard ordering. For each $n \in \mathbb{N}$, let $V_n = \mathbb{R}^n$. For $n \leq N$, the zero-padding embedding is defined by

$$\varphi_{N,n} : \mathbb{R}^n \hookrightarrow \mathbb{R}^N,\ \varphi_{N,n}(x_1, \ldots, x_n) = (x_1, \ldots, x_n, \underbrace{0, \ldots, 0}_{(N-n)\ 0\text{'s}}).$$

The group $\mathrm{G}_n = \mathrm{S}_n$, which is the permutation group acting on $\mathbb{R}^n$ by permuting coordinates: $(g \cdot x)_i := x_{g^{-1}(i)}$ for $g \in \mathrm{S}_n$. The embedding of groups is given by, for $n \leq N$,

$$\theta_{N,n} : \mathrm{S}_n \mapsto \mathrm{S}_N,\ \theta_{N,n}(g) = \begin{bmatrix} g & 0 \\ 0 & I_{N-n} \end{bmatrix}.$$

For $p \in [1, \infty)$, each V_n is equipped with the ℓ_p-norms, i.e., $\|x\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$. By Proposition 2.9, this induced a norm on V_∞, also denoted as $\|\cdot\|_p$. Consequently, the limit space is

$$\overline{V_\infty} = \ell_p = \left\{x = (x_i)_{i=1}^\infty : \|x\|_p = \sum_{i=1}^\infty |x_i|^p < \infty\right\} \quad \text{for } p \in [1, \infty).$$

The orbit closure under G_∞-action, denoted as $\overline{O}\left(\overline{V_\infty}\right)$, can be identified as an ordered sequence in ℓ_p. The symmetric distance $\bar{d}_p(x, y) = \min_{\sigma \in \mathrm{G}_\infty} \|x - \sigma \cdot y\|_p$ defines a metric on $\overline{O}\left(\overline{V_\infty}\right)$.

Consider the d-th direct sum of $\mathbb{V}_{\text{zero}}$ as defined in Definition 2.3, $\mathbb{V}_{\text{zero}}^{\oplus d} = \left\{\left(\mathbb{R}^{n \times d}\right), \left(\varphi_{N,n}^{\oplus d}\right), (\mathrm{S}_n)\right\}$. Fix an arbitrary norm on $\mathbb{R}^d$, and we then define the

ℓ_p-norm on $\mathbb{R}^{n\times d}$ as

$$\|X\|_p = \left(\sum_{i=1}^{n} \|X_{i:}\|_{\mathbb{R}^d}^p\right)^{1/p} \qquad \text{for } p \in [1,\infty).$$

The constructions of the limit space and orbit closure are analogous.

We now introduce a description of compactness in the limit space $\ell_p\left(\mathbb{R}^d\right)$, as stated in the following proposition.

Proposition 3.2 (Compactness in $\ell_p(\mathbb{R}^d)$ ([55], Chap. 1, Ex. 6))**.** *Let $p \in [1,\infty)$, a set $K_p \subset \ell_p(\mathbb{R}^d)$ is compact if and only if it is*

- *closed,*
- *bounded:*
$$\sup_{X\in K_p} \|X\|_p = \sup_{X\in K_p} \left(\sum_{i=1}^{\infty} \|X_{i:}\|_{\mathbb{R}^d}^p\right)^{1/p} < \infty,$$
- *uniformly tail-controlled:*
$$\lim_{N\to\infty} \sup_{x\in K_p} \left(\sum_{n\geqslant N} \|X_{n:}\|_{\mathbb{R}^d}^p\right)^{1/p} = 0.$$

Example 3.3. *$K := \{(x_i)_{i=1}^{\infty} : x_i = 2^{-i}\sigma_i, \sigma_i \in \{0,1\}, \forall i \in \mathbb{N}\}$, $K \subset \ell_p(\mathbb{R})$ is compact.*

For a compact set $K_p \subset \ell_p\left(\mathbb{R}^d\right)$, let $\overline{O}\left(K_p\right)$ denote its orbit closure. Since the orbit map is clearly continuous, $\overline{O}\left(K_p\right)$ is compact in $\overline{O}\left(\overline{V_\infty}\right)$.Our goal is to use a continuous model on $\overline{O}(K_p)$ to approximate the space of continuous functions on it, denoted by $C\left(\overline{O}(K_p)\right)$.

Our next step is to seek a universal machine-learning model that remains continuous on the orbit closure. We consider the DeepSets architecture [3] of

the form

$$\text{DeepSets}(X) \;=\; \hat{\sigma}\Big(\sum_{i=1}^{\infty} \hat{\rho}(X_{i:})\Big), X \in K_p,$$

where $\hat{\rho} \in \mathcal{N}^{w,\phi}(d,h)$, $\hat{\sigma} \in \mathcal{N}^{w,\phi}(h,1)$, $h \in \mathbb{N}$. Unfortunately, this is not a continuous model, as the infinite sum may diverge for some sequences in K_p. Even if we impose the constraint $\hat{\rho}(\mathbf{0}_d) = \mathbf{0}_h$, the function $\hat{\rho}$ may decay arbitrarily slowly, and compactness alone is insufficient to guarantee convergence. Motivated by this observation, we slightly modify the DeepSets architecture so that the inner functions exhibit at least a p-moment decay rate around zero.

$$\text{DeepSets}_{\infty}^{\hat{\rho},\hat{\sigma}}(X) \;=\; \hat{\sigma}\Big(\sum_{i=1}^{\infty} \|X_{i:}\|_{\mathbb{R}^d}^{p} \cdot \hat{\rho}(X_{i:})\Big), X \in K_p, \tag{3.1}$$

where $\hat{\rho} \in \mathcal{N}^{w,\phi}(d,h)$, $\hat{\sigma} \in \mathcal{N}^{w,\phi}(h,1)$, $h \in \mathbb{N}$. Each intermediate space $\mathbb{R}^h$ is endowed with an arbitrary but fixed norm $\|\cdot\|_{\mathbb{R}^h}$. The following theorem shows that $\text{DeepSets}_{\infty}^{\rho,\sigma}$ is indeed a continuous model on the orbit closure, when ρ and σ are continuous. The proof is deferred to Section 4.1.1.

Theorem 3.4. *Let $h \in \mathbb{N}, \rho \in C(\mathbb{R}^d, \mathbb{R}^h)$, $\sigma \in C(\mathbb{R}^h, \mathbb{R})$, and $p \in [1,\infty)$. Then* $\text{DeepSets}_{\infty}^{\rho,\sigma}$ *in* (3.1) *defines a continuous map $\overline{O}(K_p) \to \mathbb{R}$, where $K_p \subset \ell_p(\mathbb{R}^d)$ is compact and $\overline{O}(K_p)$ is equipped with the metric of symmetric distance .*

Given the continuity condition, our next goal is to establish universality. The theorem is stated below, and its proof is deferred to Section 4.1.2, following the strategy of Stone's Weierstrass theorem.

Theorem 3.5 (Universality of DeepSets_{∞})**.** *Let $\mathcal{F}_{\text{DS}}$ denote the class of functions of the form* $\text{DeepSets}_{\infty}^{\hat{\rho},\hat{\sigma}}$ *in* (3.1)*, where $p \in [1,\infty)$, $\hat{\rho} \in \mathcal{N}^{w,\phi}(d,h)$ and*

$\hat{\sigma} \in \mathcal{N}^{w,\phi}(h,1)$ *for some* $h, w \in \mathbb{N}$*, and* ϕ *is a continuous non-polynomial activation function.* $K_p \subset \ell_p\left(\mathbb{R}^d\right)$ *is compact. Then* $\mathcal{F}_{\mathrm{DS}}$ *is dense in* $C\left(\overline{O}(K_p)\right)$ *with respect to the supremum norm.*

3.1.2 Duplication-padding compatibility

The duplication-padding consistent sequence $\mathbb{V}_{\mathrm{dup}} = \left\{(V_n), (\varphi_{N,n}), (\mathrm{G}_n)\right\}$ is defined as follows. The index set $(\mathbb{N}, \cdot \mid \cdot)$ is the set of natural numbers with divisibility partial order, where $n \preceq N$ if and only if $n \mid N$. For each $n \in \mathbb{N}$, $V_n = \mathbb{R}^n$. For $n \preceq N$, the duplication embedding is given by

$$\varphi_{N,n} : \mathbb{R}^n \hookrightarrow \mathbb{R}^N,\ \varphi_{N,n}((x_1, \ldots, x_n)) = (\underbrace{x_1, \ldots, x_1}_{N/n \text{ times}}, \ldots, x_n, \ldots, x_n),$$

The group $\mathrm{G}_n = \mathrm{S}_n$, which is the permutation group acting on $\mathbb{R}^n$ by permuting coordinates: $(g \cdot x)_i := x_{g^{-1}(i)}$ for $g \in \mathrm{S}_n$. The embedding of groups is given by, for $n \preceq N$,

$$\theta_{N,n} : \mathrm{S}_n \mapsto \mathrm{S}_N,\ \theta_{N,n}(g) = g \otimes I_{N/n}\,.$$

The space V_∞ can be identified with the space of step functions, whose discontinuity points are rational. More precisely, each equivalence class $[x] \in V_\infty$, where $x \in V_n$, corresponds to a step function

$$f_x : [0,1] \mapsto \mathbb{R},\ f_x(t) = x_i \text{ for } t \in \left(\frac{i-1}{n}, \frac{i}{n}\right],\ i = 1, \ldots, n.$$

For $p \in [1, \infty)$, each V_n is equipped with the normalized ℓ_p-norms, i.e., $\|x\|_{\overline{p}} = \left(\frac{1}{n}\sum_{i=1}^n |x_i|^p\right)^{1/p}$. Under the identification with step functions, this corresponds to $\|f_x\|_p = \left(\int_0^1 |f_x(t)|^p dt\right)^{1/p}$. By Proposition 2.9, this induced a norm

on V_∞. Consequently, for $p \in [1, \infty)$, the limit space is

$$\overline{V_\infty} = L^p([0,1]) = \left\{ f : [0,1] \to \mathbb{R} \text{ measurable} : \int_0^1 |f(t)|^p dt < \infty \right\}.$$

The G_∞-action on $\overline{V_\infty}$ can be interpreted as composition with a measure-preserving bijection, i.e.,

$$g \cdot f = f \circ \varphi_{g^{-1}},$$

where $\varphi_{g^{-1}} \in S_{[0,1]}$ is the corresponding measure-preserving bijection on $[0,1]$. Under this view, the orbit closure of $\overline{V_\infty}$ can be identified with $\mathcal{P}_p(\mathbb{R})$, the space of probability measures on $\mathbb{R}$ with finite p-th moment, endowed with Wasserstein p-distance. A detailed proof that the symmetric distance $\bar{d}_p$ coincides with the Wasserstein p-distance can be found in Proposition E.1 of [1].

Similarly, for the d-th direct sum of $\mathbb{V}_{\text{dup}}^{\oplus d}$, we fix an arbitrary norm on $\mathbb{R}^d$. The orbit closure can be identified with $\mathcal{P}_p\left(\mathbb{R}^d\right)$ endowed with the Wasserstein p-distance with respect to $\|\cdot\|_{\mathbb{R}^d}$.

We now introduce a description of compactness in the space of orbit closures, as stated in the following proposition.

Proposition 3.6 (Compactness in $\mathcal{P}_p\left(\mathbb{R}^d\right)$ with W_p ([56], Prop 7.1.5))**.** *For $p \in [1, \infty)$, a set $K \subset \mathcal{P}_p\left(\mathbb{R}^d\right)$ is compact if and only if it is*

- *closed,*
- *tight: $\forall \varepsilon > 0$, $\exists K_\varepsilon$ compact in $\mathbb{R}^d$, such that*

$$\mu\left(\mathbb{R}^d \backslash K_\varepsilon\right) \leq \varepsilon, \ \forall \mu \in K,$$

- *p-uniformly integrable:*

$$\lim_{R\to\infty}\sup_{\mu\in K}\int_{\|x\|_{\mathbb{R}^d}>R}\|x\|_{\mathbb{R}^d}^p\ \mathrm{d}\mu(x)=0.$$

Example 3.7. *It's easy to verify the set* $K=\left\{\mu\in\mathcal{P}_p(\mathbb{R}^d)\ \middle|\ (\mu)\subset\overline{B(0,R)}\right\}$ *is a compact set, where* $R>0$ *is a constant.*

For a compact set $K\subset\mathcal{P}_p\left(\mathbb{R}^d\right)$, which is a compact set in the space of orbit closures. We aim to approximate the space of continuous functions on it, denoted by $C(K)$.

We consider the normalized DeepSets architecture [10] defined by

$$\overline{\mathrm{DeepSets}}_\infty^{\hat{\rho},\hat{\sigma}}(\mu)=\hat{\sigma}\left(\int\hat{\rho}d\mu\right),\tag{3.2}$$

where $\hat{\rho}\in\mathcal{N}^{w,\phi}(d,h)$, $\hat{\sigma}\in\mathcal{N}^{w,\phi}(h,1)$, $h\in\mathbb{N}$, and the intermediate space $\mathbb{R}^h$ is endowed with a norm $\|\cdot\|_{\mathbb{R}^h}$. We denote by $\mathcal{F}_p(\mathbb{R}^d,\mathbb{R}^h)$ the class of continuous functions satisfying the p-moment growth condition, that is, there exists $M>0$ such that for any $x\in\mathbb{R}^d$,

$$\|f(x)\|_{\mathbb{R}^h}\le M(1+\|x\|_{\mathbb{R}^d}^p).$$

The following theorem establishes that $\overline{\mathrm{DeepSets}}_\infty^{\rho,\sigma}$ defines a continuous model when ρ,σ are continuous and ρ satisfies the p-moment growth condition.

Theorem 3.8. *Let* $h\in\mathbb{N},\rho\in\mathcal{F}_p(\mathbb{R}^d,\mathbb{R}^h)$, $\sigma\in C(\mathbb{R}^h,\mathbb{R})$, *and* $p\in[1,\infty)$. *Then* $\mathrm{DeepSets}_\infty^{\rho,\sigma}$ *in* (3.2) *defines a continuous map* $K\to\mathbb{R}$, *where* $K\subset\mathcal{P}_p(\mathbb{R}^d)$ *is a compact set equipped with the Wasserstein* p*-distance.*

This theorem follows directly from the fact that W_p-convergence is equivalent to weak convergence in $\mathcal{P}_p$, see [57], Def. 6.8 and Thm. 6.9.

Finally, we establish the universality of the normalized DeepSets architecture in (3.2). We note that the universality result in [10] applies only to measures with compact support (Example 3.7), which corresponds to a special case of our theorem. The proof of Theorem 3.9 can be found in Section 4.1.3.

Theorem 3.9 (Universality of $\overline{\text{DeepSets}}_\infty$)**.** *Let $\mathcal{F}_{\overline{\text{DS}}}$ denote the class of functions of the form $\overline{\text{DeepSets}}_\infty^{\hat{\rho},\hat{\sigma}}$ in (3.2), where $p \in [1,\infty)$, $\hat{\rho} \in \mathcal{N}^{w,\phi}(d,h)$ and $\hat{\sigma} \in \mathcal{N}^{w,\phi}(h,1)$ for some $h, w \in \mathbb{N}$, and ϕ is a Lipschitz continuous, non-polynomial activation function that is asymptotically polynomial at $\pm\infty$. $K \subset \mathcal{P}_p(\mathbb{R}^d)$ is compact. Then $\mathcal{F}_{\overline{\text{DS}}}$ is dense in $C(K)$ with respect to the supremum norm.*

We note that the Lipschitz continuity of the activation function ϕ ensures that $\mathcal{N}^{w,\phi}(d,h) \subset \mathcal{F}_p(\mathbb{R}^d, \mathbb{R}^h)$, for $p \in [1,\infty)$. The requirement that ϕ is asymptotically polynomial at $\pm\infty$ guarantees the approximation capability of neural networks on non-compact domains; see [15]. Representative activation functions meeting these conditions include ReLU, Softplus, and Sigmoid, among others.

3.2 Graph functions

The duplication-padding consistent sequence for graphs

$$\mathbb{V}_{\text{dup}}^G = \left\{(V_n), (\varphi_{N,n}), (\mathrm{G}_n)\right\}$$

is defined as follows. The index set $(\mathbb{N}, \cdot \mid \cdot)$ is the set of natural numbers with divisibility partial order, where $n \preceq N$ if and only if $n \mid N$. For each $n \in \mathbb{N}$,

$V_n = \mathbb{R}^{n\times n}_{\text{sym}}$. For $n \preceq N$, the duplication embedding is given by

$$\varphi_{N,n} : \mathbb{R}^{n\times n}_{\text{sym}} \hookrightarrow \mathbb{R}^{N\times N}_{\text{sym}},\ \varphi_{N,n}(A) = A \otimes \left(\mathbf{1}_{N/n}\mathbf{1}^{\top}_{N/n}\right).$$

The group is the symmetric group S_n and the group embedding is the same as the case of duplication-padding sets. S_n acts on V_n via $g \cdot A = gAg^{\top}$.

Similarly, the space V_∞ can be identified with the space of step graphon by

$$W_A : [0,1]^2 \mapsto \mathbb{R},\ W_A(x,y) = A_{ij} \text{ for } (x,y) \in \left(\frac{i-1}{n}, \frac{i}{n}\right] \times \left(\frac{j-1}{n}, \frac{j}{n}\right], i, j \in [n].$$

The symmetric group acts on the induced step graphon by

$$\sigma \cdot W_A = W_A^{\sigma^{-1}} := W_A(\sigma^{-1}(x), \sigma^{-1}(y)).$$

We endow each V_n with the normalized ℓ^p norm, which extends to a norm in V_∞. This extension coincides with the L^p norm on graphons

$$\|W\|_p := \left(\int_{S\times T} |W(x,y)|^p dxdy\right)^{1/p}.$$

The symmetric distance coincides with the δ_p distance, defined by

$$\overline{d}(W, W') = \delta_p(W, W') := \inf_{\varphi \in S_{[0,1]}} \| (U^{\varphi} - W) \|_p,$$

where $S_{[0,1]}$ denotes the set of all measure-preserving bijections $[0,1] \mapsto [0,1]$.

The orbit closure space can be identified with the space of symmetric measurable functions $[0,1]^2 \mapsto \mathbb{R}$, modulo the equivalence relation $W_1 \sim W_2$ whenever $\delta_p(W_1, W_2) = 0$. More details for the duplication consistent sequence for graphs can be found in [1].

Next, we define the graphon space as

$$\widetilde{\mathcal{W}_0} := \mathcal{W}_0/\sim$$

where $\mathcal{W}_0$ is the space of symmetric measurable functions $[0,1]^2 \mapsto [0,1]$, and $W_1 \sim W_2$ whenever $\delta_p(W_1, W_2) = 0$. It is difficult to explicitly characterize compactness in $\left(\widetilde{\mathcal{W}_0}, \delta_p\right)$, but we will see an example in Section 3.3. Our goal is to construct a model that is continuous and universal on a compact subset of $\left(\widetilde{\mathcal{W}_0}, \delta_p\right)$.

A direct construction is via the homomorphism density on graphons. A simple graph is a finite graph without loops and multi-edges. The homomorphism density of a simple graph $F = (V, E)$ in a graphon $W \in \widetilde{\mathcal{W}_0}$ is defined as

$$t(F, W) = \int_{[0,1]^V} \prod_{ij \in E} W\left(x_i, x_j\right) \prod_{i \in V} dx_i \ .$$

Proposition 3.10. *Define the class of functions*

$$\mathcal{HD} := \operatorname{span}\{\, t(F, \cdot) : F \text{ is a simple graph} \,\},$$

the linear span of homomorphism density functions of simple graphs, defined on $\left(\widetilde{\mathcal{W}_0}, \delta_p\right)$. *Then* $\mathcal{HD} \subset C(\widetilde{\mathcal{W}_0})$ *and is dense in* $C(\widetilde{\mathcal{W}_0})$.

The proof is straightforward by Stone-Weierstrass theorem: by the counting lemma (Lemma 10.22 in [8]), each $t(F, \cdot)$ is continuous with respect to $\delta_\square$ and therefore is continuous with respect to δ_p; moreover, $\mathcal{HD}$ is a subalgebra since $t\left(F_1, \cdot\right) t\left(F_2, \cdot\right) = t\left(F_1 \sqcup F_2, \cdot\right)$; and $\mathcal{HD}$ separates points because $\delta_p(U, W) = 0$ is equivalent to $t(F, U) = t(F, W)$ for every simple graph (Corollary 8.14 in [8]).

However, the model built directly from homomorphism densities is not tractable

in practice because it is not a finite-parameter model: universality requires computing homomorphism densities for infinitely many simple graphs. Instead, we aim to parametrize these homomorphism densities so that graph patterns can be learned from data. This is achieved through basic building blocks including linear maps and the nonlinear activation function.

First, we introduce the linear equivariant map as in [11]. For $k \in \mathbb{N}_0$, define a class of functions

$$\mathcal{G}_k := \{G : [0,1]^k \mapsto \mathbb{R} \text{ measurable bounded }\}/\sim_{\text{a.e.}},$$

where $G_1 \sim_{\text{a.e.}} G_2$ if $G_1 = G_2$ almost everywhere. The set of linear equivariant maps is defined as:

$$\text{LE}_{k\to l} := \{L \in \mathcal{B}(\mathcal{G}_k, \mathcal{G}_l) : L \text{ is equivariant}\},$$

where $\mathcal{B}(\cdot,\cdot)$ denotes the bounded linear operators, and L is called equivariant if for all measure-preserving maps $\varphi \in \bar{S}_{[0,1]}$,

$$L\,(G^\varphi) = L(G)^\varphi \text{ almost everywhere,}$$

where $G^\varphi(x_1, \cdots, x_k) := G(\varphi(x_1), \cdots, \varphi(x_k))$.

The linear maps in $\text{LE}_{k\to l}$ must be consistent with the almost everywhere equivalence relation in $\mathcal{G}_k$. To be specific, operations on sets of measure zero are not in $\text{LE}_{k\to l}$, for example, mapping $G(x,y)$ to $G'(x) := G(x,x)$. Moreover, note that equivariance here corresponds to any measure-preserving function, without the restriction to bijection. The group permutation action can be regarded as an invertible measure preserving function, whereas the duplication action can be

expressed as $\varphi(x) = Nx \bmod 1$, which is measure preserving but not invertible. Therefore, the linear equivariant map can be understood to be equivariant with respect to both the group action and the duplication action. There have been some studies on this class of linear equivariant maps.

Proposition 3.11 (Dimension of $\mathrm{LE}_{k\to l}$ ([11], Thm 4.2))**.** *Let $k, l \in \mathbb{N}_0$. Then, $\mathrm{LE}_{k\to l}$ is a finite-dimensional vector space of dimension*

$$\dim \mathrm{LE}_{k\to l} = \sum_{s=0}^{\min\{k,l\}} s! \binom{k}{s}\binom{l}{s} \leq \mathrm{bell}(k+l).$$

We provide a basis for this space and some examples of linear equivariant maps.

Example 3.12. (Bases in $\mathrm{LE}_{2\to 3}$, $\dim \mathrm{LE}_{2\to 3} = 13$).

- *Binary embeddings:*

$$\begin{aligned} (L_1G)(x,y,z) &= G(x,y), & (L_2G)(x,y,z) &= G(y,x),\\ (L_3G)(x,y,z) &= G(x,z), & (L_4G)(x,y,z) &= G(z,x),\\ (L_5G)(x,y,z) &= G(y,z), & (L_6G)(x,y,z) &= G(z,y). \end{aligned}$$

- *Unary marginal embedding:*

$$\begin{aligned} (L_7G)(x,y,z) &= \textstyle\int_{[0,1]} G(x,t)\ dt, & (L_8G)(x,y,z) &= \textstyle\int_{[0,1]} G(t,x)\ dt,\\ (L_9G)(x,y,z) &= \textstyle\int_{[0,1]} G(y,t)\ dt, & (L_{10}G)(x,y,z) &= \textstyle\int_{[0,1]} G(t,y)\ dt,\\ (L_{11}G)(x,y,z) &= \textstyle\int_{[0,1]} G(z,t)\ dt, & (L_{12}G)(x,y,z) &= \textstyle\int_{[0,1]} G(t,z)\ dt, \end{aligned}$$

- *Scalar embedding:*

$$(L_{13}G)(x,y,z) = \int_{[0,1]^2} G(u,v)\ dudv.$$

Example 3.13. (Examples of linear equivariant maps).

1. $P_j : \mathcal{G}_k \mapsto \mathcal{G}_k$,

$$(P_j\ G)(x_1, \cdots, x_k) = G(x_2, \cdots, x_{j-1}, x_1, x_j, \cdots, x_k),\ 1 \le j \le k$$

$P_j \in \mathrm{LE}_{k\to k}$, *which permutes the* j*-th coordinate to the first position.*

2. $\mathrm{Lift} : \mathcal{G}_k \mapsto \mathcal{G}_{k+1}$,

$$(\mathrm{Lift}\ G)(x_0, x_1, \cdots, x_k) = G(x_1, \cdots, x_k).$$

$\mathrm{Lift} \in \mathrm{LE}_{k\to k+1}$, *which lifts the input with one redundant coordinate.*

3. $\mathcal{L}_0 : \mathcal{G}_k \mapsto \mathcal{G}_0$,

$$(\mathcal{L}_0\ G) = \int_{[0,1]^k} G\,(x_1, \cdots, x_k)\, dx_1 \cdots dx_k,\ k \ge 1,$$

$\mathcal{L}_0 \in \mathrm{LE}_{k\to 0}$, *which integrates over all coordinates.*

4. $L_1 \in \mathrm{LE}_{k\to l}, L_2 \in \mathrm{LE}_{m\to k}$, *then* $L_1 \circ L_2 \in \mathrm{LE}_{m\to l}$.

Second, we introduce the tensor contraction operator that serves as the nonlinear activation function in the network. The operator contracts over the first two coordinates across all input tensors: For any number m of input tensors, each with dimension $k_i \ge 2$, the operator

$$T : \prod_{i=1}^{m} \mathcal{G}_{k_i} \mapsto \mathcal{G}_{(\sum_{i=1}^{m} k_i)-2m+2}$$

$$(T\,(G_1, \ldots, G_m))\left(x, y, x_3^1, \ldots, x_{k_1}^1, \ldots, x_3^m, \ldots, x_{k_m}^m\right) = \prod_{i=1}^{m} G_i\left(x, y, x_3^i, \ldots, x_{k_i}^i\right).$$

We define i-th linear layer as

$$\mathcal{L}_i(G_i, W) := \begin{pmatrix} L_i^1(G_i) + b_i^1 \\ \vdots \\ L_i^{h_i}(G_i) + b_i^{h_i} \\ L_i^0(W) + b_i^0 \end{pmatrix},$$

where $G_i \in \mathcal{G}_{k_i}$, for some $k_i \in \mathbb{N}_0$; $L_i^j \in \mathrm{LE}_{k_i \to m_{ij}}$, $b_i^j \in \mathbb{R}$, for $j \in [h_i]$ and some $m_{ij} \geq 2$; $L_i^0 \in \mathrm{LE}_{2 \to m_{i0}}$, for some $m_{i0} \geq 2$; $W \in \widetilde{\mathcal{W}_0}$. We denote the map of i-th linear layer $\mathcal{L}_i^W$ as:

$$\mathcal{L}_i^W : \mathcal{G}_{k_i} \mapsto \prod\nolimits_{j=0}^{h_i} \mathcal{G}_{m_{ij}} \ , \mathcal{L}_i^W(G_i) = \mathcal{L}_i(G_i, W) \ .$$

We then contract them using T to get the input of $(i+1)$-th linear layer

$$T \circ \mathcal{L}_i(G_i, W) := G_{i+1} \in \mathcal{G}_{(\sum_{j=0}^{h_i} m_{ij}) - 2h_i} \ .$$

Suppose we have $l \geq 1$ layers. Let the input graphon be $G_1 = W \in \widetilde{\mathcal{W}_0}$. The final output tensor G_{l+1} is then passed through $\mathcal{L}_0$, which is defined in Example 3.13.3 as the integral over all coordinates. Thus, the overall architecture, we call it **Tensor Contraction Graphon Network (TGN)**, can be written as:

$$f : \widetilde{\mathcal{W}_0} \to \mathbb{R}, \ f(W) = \mathcal{L}_0 \circ T \circ \mathcal{L}_l^W \circ \cdots \circ T \circ \mathcal{L}_1^W(W). \tag{3.3}$$

More precisely, (3.3) is the following network:

$$W \xrightarrow{\mathcal{L}_1^W} \begin{pmatrix} L_1^1(W) + b_1^1 \\ \vdots \\ L_1^{h_1}(W) + b_1^{h_1} \\ L_1^0(W) + b_1^0 \end{pmatrix} \xrightarrow{T} G_2 \xrightarrow{\mathcal{L}_l^2} \cdots$$

$$\cdots G_l \xrightarrow{\mathcal{L}_l^W} \begin{pmatrix} L_l^1(G_l) + b_l^1 \\ \vdots \\ L_l^{h_l}(G_l) + b_l^{h_l} \\ L_l^0(W) + b_l^0 \end{pmatrix} \xrightarrow{T} G_{l+1} \xrightarrow{\mathcal{L}_0} f(W)$$

Theorem 3.14. *The tensor contraction graphon network, as defined in* (3.3)*, is well defined and continuous with respect to* $\left(\widetilde{\mathcal{W}_0}, \delta_p\right)$.

The result follows immediately from the observation that the network implements a homomorphism density of a multigraph, which is a continuous function with respect to the δ_p topology.

Theorem 3.15 (Universality with Tensor Contraction Graphon Networks)**.** *Let* $\mathcal{F}_d$ *be the class of functions which has the form*

$$\hat{f} : \widetilde{\mathcal{W}_0} \mapsto \mathbb{R}, \hat{f}(W) = \sum_{s=1}^{S} \alpha_s f_s(W),$$

where each f_s *has the form of* (3.3)*;* $\alpha_s \in \mathbb{R}; S \in \mathbb{N}$*, then* $\mathcal{F}_d$ *is dense in* $C\left(\widetilde{\mathcal{W}_0}\right)$.

We defer the proof of Theorem 3.15 to Section 4.2.1

3.3 Point cloud functions

Similar to the case of graphs, the duplication-padding consistent sequence for point clouds $\mathbb{V}^P_{\text{dup}} = \left\{(V_n), (\varphi_{N,n}), (\mathrm{G}_n)\right\}$ is defined as follows. The index set $(\mathbb{N}, \cdot \mid \cdot)$ is the set of natural numbers with divisibility partial order, where $n \preceq N$ if and only if $n \mid N$. For each $n \in \mathbb{N}$, $V_n = \mathbb{R}^{n\times k}$, which represents sets of n points in $\mathbb{R}^k$, and k is fixed. The group is $\mathrm{G}_n = \mathrm{S}_n \times \mathrm{O}(k)$, where S_n is the permutation group, and $\mathrm{O}(k)$ is the orthogonal group. The group action on V_n is defined as:

$$(g, h) \cdot X = gXh^\top.$$

For $n \preceq N$, the duplication embedding is given by

$$\varphi_{N,n} : \mathbb{R}^{n\times k} \hookrightarrow \mathbb{R}^{N\times k},\ \varphi_{N,n}(X) = X \otimes \mathbf{1}_{N/n},$$

and the group embedding is given by

$$\theta_{N,n} : \mathrm{S}_n \times \mathrm{O}(k) \hookrightarrow \mathrm{S}_N \times \mathrm{O}(k),\ \theta_{N,n}(g, h) = \left(g \otimes I_{N/n}, h\right).$$

Consider the Euclidean norm $\|\cdot\|_2$ on $\mathbb{R}^k$ which corresponds to the inner product preserved by elements of $\mathrm{O}(k)$. We equip each V_n with the normalized ℓ_p norm:

$$\|X\|_{\bar{p}} = \left(\frac{1}{n}\sum_{i=1}^{n} \|X_{i:}\|_2^p\right)^{1/p}, \quad p \in [1, \infty).$$

Similar to the case of sets, we can identify each matrix $X \in \mathbb{R}^{n\times k}$ with a step function $[0, 1] \mapsto \mathbb{R}^k$. Then the limit space can be identified with $\overline{V_\infty} =$

$L^p\left([0,1];\mathbb{R}^k\right)$, the symmetric distance is defined as

$$\bar{d}(f,g) = \inf_{h\in \mathrm{O}(k)} \inf_{\varphi \in S_{[0,1]}} \left(\int_0^1 \|h(f(t)) - g(\varphi(t))\|_2^p dt \right)^{1/p},$$

where $S_{[0,1]}$ is the set of measure preserving bijections. The orbit closure $\overline{O}(\overline{V_\infty})$ can be identified with $\overline{V_\infty}/\sim$, where $f \sim g$ if $\bar{d}(f,g) = 0$.

In another view, we can identify each matrix $X \in \mathbb{R}^{n\times k}$ with an empirical probability measure on $\mathbb{R}^k$. Then the orbit closure can be identified with the probability measures on $\mathbb{R}^k$ with the Wasserstein p distance under the $\mathrm{O}(k)$ action. The symmetric distance can be given as

$$\overline{\mathrm{d}}_p(f,g) = \inf_{h\in \mathrm{O}(k)} W_p\left(h\cdot \mu_f, \mu_g\right),$$

where $\mu_f = f_{\#}\lambda$, λ is the Lebesgue measure on $[0,1]$; the action of $\mathrm{O}(k)$ is $h\cdot\mu = h_{\#}\mu$.

Consider a set

$$K_f := \{f \in L^p\left([0,1];\mathbb{R}^k\right) : \sup_{t\in[0,1]} \|f(t)\|_2 \le R\},$$

where $R > 0$ is a positive constant, and we note that we will not distinguish functions that are almost everywhere equal, and we can identify K_f with a subset of $L^\infty\left([0,1];\mathbb{R}^k\right)$. Then consider the set of push forward measures by functions in K_f,

$$K_\mu = \{f_{\#}\lambda : f \in K_f\},$$

as shown in Example 3.7, $K_\mu \in \mathcal{P}_p\left(\mathbb{R}^k\right)$ is compact, and since the orbit map is continuous, we get a compact set in the orbit closure, denoted by $\overline{O}(K_f)$.

Let $p = 2$, define the covariance function W_f as

$$W_f : [0,1]^2 \mapsto [0,1],\ W_f(x,y) := \frac{1}{2R^2}\langle f(x), f(y)\rangle + \frac{1}{2}\ ,$$

which is a symmetric measurable function if $f \in K_f$. Then define the distance between W_f and W_g as

$$\delta_2\left(W_f, W_g\right) = \inf_{\varphi \in S_{[0,1]}} \left\|W_f - W_g^{\varphi}\right\|_{L^2},$$

where $W^{\varphi}(x,y) := W\left(\varphi(x), \varphi(y)\right)$.

We consider the Invariant Graphon Network (IWN) architecture [11]

$$\mathcal{IWN} : \mathcal{W}_0 \to \mathbb{R},$$

$$W \mapsto \sum_{s=1}^{S} L_s^{(2)}\left(\varrho\left(L_s^{(1)}(W) + b_s^{(1)}\right)\right) + b^{(2)},$$

where $S \in \mathbb{N}_0$, $L_s^{(1)} \in \mathrm{LE}_{2\to k_s}$, $L_s^{(2)} \in \mathrm{LE}_{k_s \to 0}$, $b_s^{(1)}, b^{(2)} \in \mathbb{R}$ for $k_s \in \mathbb{N}$, $s \in \{1, \ldots, S\}$, and ρ continuous, non-polynomial and acting pointwise. We input the covariance function into the IWN, and obtain our model for point clouds

$$\overline{O}(K_f) \to (\widetilde{W_0}, \delta_2) \to \mathbb{R} \tag{3.4}$$

$$f \mapsto W_f \mapsto \mathcal{IWN}(W_f), \tag{3.5}$$

and it is straightforward to verify that this mapping is well defined. We prove in Section 4.3.1 and 4.3.2 that this model is continuous and universal.

Theorem 3.16. *For $f, g \in K_f \subset L^2\left([0,1]; \mathbb{R}^k\right)$, the model in (3.4) is continuous with respect to the symmetric distance.*

Theorem 3.17 (Universality with Invariant Graphon Networks)**.** *Let $\mathcal{F}_{IWN}$ be the class of functions with the form of* (3.4)*, then $\mathcal{F}_{IWN}$ is dense in $C\left(\overline{O}(K_f)\right)$ with respect to $\|\cdot\|_\infty$.*

Chapter 4

Analysis

In this chapter, we present the detailed proofs of the theorem introduced in Chapter 3. In particular, we show that our model is continuous in the limit space and is universal to approximate continuous functions.

4.1 Missing proofs from Section 3.1

4.1.1 Proof of Theorem 3.4

Proof of Theorem 3.4. Since K_p is compact, there exists $M > 0$ such that

$$\sup_{X \in K_p} \left(\sum_{j=1}^{\infty} \|X_{j:}\|_{\mathbb{R}^d}^p \right)^{1/p} \leq M$$

Therefore,

$$\sup_{X \in K_p} \sup_{i \in \mathbb{N}} \|X_{i:}\|_{\mathbb{R}^d} = \sup_{X \in K_p} \sup_{i \in \mathbb{N}} \left(\|X_{i:}\|_{\mathbb{R}^d}^p \right)^{1/p} \leq \sup_{X \in K_p} \left(\sum_{j=1}^{\infty} \|X_{j:}\|_{\mathbb{R}^d}^p \right)^{1/p} \leq M.$$

Since ρ is continuous, it maps compact sets to compact sets, thus, there exists $M_\rho > 0$ such that

$$\sup_{X \in K_p} \sup_{i \in \mathbb{N}} \|\rho(X_{i:})\|_{\mathbb{R}^h} \leq M_\rho.$$

Define

$$S : K_p \to \mathbb{R}^h, S(X) = \sum_{i=1}^{\infty} \|X_{i:}\|_{\mathbb{R}^d}^p \rho(X_{i:})$$

$$\mathrm{S}_n : K_p \to \mathbb{R}^h, \mathrm{S}_n(X) = \sum_{i=1}^{n-1} \|X_{i:}\|_{\mathbb{R}^d}^p \rho(X_{i:}),\ n > 1$$

S_n is clearly continuous in K_p since the coordinate projection, ρ, and finite sums of continuous functions are all continuous. By the uniformly tail-controlled condition in Proposition 3.2,

$$\lim_{n\to\infty} \sup_{X\in K_p} \|\mathrm{S}_n(X) - S(X)\|_{\mathbb{R}^h} \le M_\rho \lim_{n\to\infty} \sup_{X\in K_p} \sum_{i\ge n} \|X_{i:}\|_{\mathbb{R}^d}^p = 0$$

Therefore, S_n uniformly converges to S. Then by the uniform limit theorem, i.e., a uniform limit of continuous functions is continuous, S is continuous in K_p. Moreover, since K_p is compact, then S is uniformly continuous:

$$\forall \varepsilon > 0, \exists \delta > 0 \text{ such that } \forall X, Y \in K_p,\ \|X - Y\|_p \le \delta \Rightarrow \|S(X) - S(Y)\|_{\mathbb{R}^h} \le \varepsilon.$$

Observe that S is G_∞-action invariant, then S defines a function on the space of orbit closure $\overline{O}(K_p)$. Choose any $[X], [Y] \in \overline{O}(K_p)$, $\bar{d}([X], [Y]) \le \delta$ implies that there exists $g \in \mathrm{G}_\infty$, such that $\|g \cdot X - Y\|_p \le \delta$. Then

$$\|S([X]) - S([Y])\|_{\mathbb{R}^h} = \|S(g \cdot X) - S(Y)\|_{\mathbb{R}^h} \le \varepsilon,$$

which yields that S is uniformly continuous $\overline{O}(K_p) \mapsto \mathbb{R}^h$.

Finally, $\mathrm{DeepSets}_\infty^{\rho,\sigma} = \sigma \circ S$, and σ is continuous, then the composition $\mathrm{DeepSets}_\infty^{\rho,\sigma}$ defines a continuous map $\overline{O}\left(K_p\right) \to \mathbb{R}$.

□

4.1.2 Proof of Theorem 3.5

We first introduce some notations, define:

$$\mathcal{F}_{\text{cont}} := \left\{ f(x) = \sigma\Big(\sum_{i=1}^{\infty} \|X_{i:}\|^p \rho(X_{i:}) \Big) \;\Big|\; h \in \mathbb{N},\ \rho \in C\big(\mathbb{R}^d, \mathbb{R}^h\big),\ \sigma \in C\big(\mathbb{R}^h, \mathbb{R}\big) \right\},$$

which is just the class of functions of the form $\text{DeepSets}_{\infty}^{\rho,\sigma}$ in (3.1), where $\rho \in C\big(\mathbb{R}^d, \mathbb{R}^h\big)$ and $\sigma \in C\big(\mathbb{R}^h, \mathbb{R}\big)$ for some $h \in \mathbb{N}$.

The proof of Theorem 3.5 proceeds in two main steps. We first show that $\mathcal{F}_{\text{cont}}$ is dense in $C\left(\overline{O}(K_p)\right)$ by following the Stone-Weierstrass argument. We then establish that $\mathcal{F}_{\text{DS}}$ is dense in $\mathcal{F}_{\text{cont}}$. And then by the transitivity of density, $\mathcal{F}_{\text{DS}}$ is therefore dense in $C\left(\overline{O}(K_p)\right)$. The proof is organized into the following three lemmas.

Lemma 4.1. *$\mathcal{F}_{\text{cont}}$ is a subalgebra of $C\left(\overline{O}(K_p)\right)$.*

Proof of Lemma 4.1. To show $\mathcal{F}_{\text{cont}}$ is a subalgebra, it suffices to verify that it is closed under scalar multiplication, addition, and pointwise multiplication.

1. Closure under scalar multiplication is trivial.

2. Closure under addition:

   Let $f_1, f_2 \in \mathcal{F}_{\text{cont}}$,then

   $$f_1(X) = \sigma_1\left(\sum_{i=1}^{\infty} \|X_{i:}\|^p \rho_1\left(X_{i:}\right) \right),\ f_2(X) = \sigma_2\left(\sum_{i=1}^{\infty} \|X_{i:}\|^p \rho_2\left(X_{i:}\right) \right),$$

   for some $\rho_1 \in C(\mathbb{R}^d, \mathbb{R}^{h_1})$, $\rho_2 \in C(\mathbb{R}^d, \mathbb{R}^{h_2})$, $\sigma_1 \in C(\mathbb{R}^{h_1}, \mathbb{R})$, $\sigma_2 \in C(\mathbb{R}^{h_2}, \mathbb{R})$, $h_1, h_2 \in \mathbb{N}$. Then define:

   $$\begin{cases} \rho_0 : \mathbb{R}^d \to \mathbb{R}^{h_1+h_2}, \rho_0(V) = (\rho_1(V), \rho_2(V)) \\ \sigma_0 : \mathbb{R}^{h_1+h_2} \to \mathbb{R}, \sigma_0\left(u_1, u_2\right) = \sigma_1\left(u_1\right) + \sigma_2\left(u_2\right) \end{cases},$$

which are both continuous. Then we get

$$(f_1 + f_2)(X) = \sigma_0\left(\sum_{i=1}^{\infty} \|X_{i:}\|^p \rho_1(X_{i:}), \sum_{i=1}^{\infty} \|X_{i:}\|^p \rho_2(X_{i:})\right)$$

$$= \sigma_0\left(\sum_{i=1}^{\infty} \|X_{i:}\|^p \rho_0(X_{i:})\right),$$

which implies closure under addition, $f_1 + f_2 \in \mathcal{F}_{\text{cont}}$.

3. Closure under pointwise multiplication:

   Define:

$$\begin{cases} \rho_0' : \mathbb{R}^d \to \mathbb{R}^{h_1+h_2}, \rho_0'(V) = (\rho_1(V), \rho_2(V)) \\ \sigma_0' : \mathbb{R}^{h_1+h_2} \to \mathbb{R}, \sigma_0'(u_1, u_2) = \sigma_1(u_1) \cdot \sigma_2(u_2) \end{cases},$$

   which are both continuous. Then we get

$$(f_1 \cdot f_2)(X) = f_1(X) \cdot f_2(X) = \sigma_0'\left(\sum_{i=1}^{\infty} \|X_{i:}\|^p \rho_1(X_{i:}), \sum_{i=1}^{\infty} \|X_{i:}\|^p \rho_2(X_{i:})\right)$$

$$= \sigma_0'\left(\sum_{i=1}^{\infty} \|X_{i:}\|^p \rho_0'(X_{i:})\right),$$

   which implies closure under pointwise multiplication, $f_1 \cdot f_2 \in \mathcal{F}_{\text{cont}}$.

□

Lemma 4.2. *$\mathcal{F}_{\text{cont}}$ separates points in $\overline{O}(K_p)$, and contains a nonzero constant function.*

We first claim that points lying in different orbits must differ in one of their first finitely many coordinates. More precisely,

Claim 4.3. *For $X, Y \in K_p \subset \ell_p(\mathbb{R}^d)$ compact, if for any $\varepsilon > 0$, the multisets*

$$\{X_{i:} : \|X_{i:}\|_{\mathbb{R}^d} > \varepsilon\} = \{Y_{i:} : \|Y_{i:}\|_{\mathbb{R}^d} > \varepsilon\},$$

then

$$\bar{d}(X,Y) = \inf_{g\in \mathrm{G}_\infty} \|g\cdot X - Y\|_p = \inf_{g\in \mathrm{G}_\infty} \left(\sum_{i=1}^{\infty} \|X_{g^{-1}(i):} - Y_{i:}\|_{\mathbb{R}^d}^p\right)^{1/p} = 0.$$

Proof of Claim 4.3. Denote the index sets

$$I_x(\varepsilon) := \{i \in \mathbb{N} : \|X_{i:}\|_{\mathbb{R}^d} > \varepsilon\}\,; I_y(\varepsilon) = \{i \in \mathbb{N} : \|Y_{i:}\|_{\mathbb{R}^d} > \varepsilon\}\,.$$

By Proposition 3.2, $\sup_{x\in K_p} \left(\sum_{i=1}^{\infty} \|X_{i:}\|_{\mathbb{R}^d}^p\right)^{1/p} < \infty$, then $I_x(\varepsilon)$ and $I_y(\varepsilon)$ are both finite sets, and $|I_x(\varepsilon)| = \left|I_y(\varepsilon)\right|$. Then there exists $\sigma_\varepsilon \in G_\infty$, such that $X_{\sigma_\varepsilon^{-1}(i):} = Y_{i:}, \quad \forall i \in I_y(\varepsilon)$.

Define

$$f_\varepsilon(i) := \|X_{i:}\|_{\mathbb{R}^d}^p \mathbf{1}_{\{V:\ \|V\|_{\mathbb{R}^d}\le\varepsilon\}}(X_{i:}), \quad g_0(i) := \|X_{i:}\|_{\mathbb{R}^d}^p\,; \quad T_x(\varepsilon) := \sum_{i=1}^{\infty} f_\varepsilon(i),$$

where $\mathbf{1}$ denotes the indicator function. As $\varepsilon \to 0, f_\varepsilon(i) \to 0$ pointwise, and since $f_\varepsilon(i) \le g_0(i)$, and $\sum_{i=1}^{\infty} \|X_{i:}\|_{\mathbb{R}^d}^p < \infty$, then by the dominated convergence theorem, $T_x(\varepsilon) \to 0$.

Similarly, we can define $T_y(\varepsilon) := \sum_{\|Y_{i:}\|_{\mathbb{R}^d}\le\varepsilon} \|Y_{i:}\|_{\mathbb{R}^d}^p, T_y(\varepsilon) \to 0$ as $\varepsilon \to 0$.

Therefore, $\forall\delta > 0, \exists\varepsilon > 0$ small enough, such that $T_x(\varepsilon) + T_y(\varepsilon) \le \delta^p 2^{1-p}$.

$$
\begin{aligned}
\bar{d}(X,Y)^p &\le \|\sigma_\varepsilon \cdot X - Y\|_p^p \\
&= \sum_{i \in I_y(\varepsilon)} \left\| X_{\sigma_\varepsilon^{-1}(i):} - Y_{i:} \right\|_{\mathbb{R}^d}^p + \sum_{i \notin I_y(\varepsilon)} \left\| X_{\sigma_\varepsilon^{-1}(i):} - Y_{i:} \right\|_{\mathbb{R}^d}^p \\
&\le \sum_{i \notin I_y(\varepsilon)} 2^{p-1} \left(\left\| X_{\sigma^{-1}(i):} \right\|_{\mathbb{R}^d}^p + \|Y_{i:}\|_{\mathbb{R}^d}^p \right) \\
&\le 2^{p-1} \left(T_x(\varepsilon) + T_y(\varepsilon) \right) \\
&\le \delta^p,
\end{aligned}
$$

which means for any $\delta > 0, \bar{d}(X,Y) \le \delta$, implies $\bar{d}(X,Y) = 0$. □

Proof of Lemma 4.2. Let $\sigma \equiv 1$, which implies that $\mathcal{F}_{\text{cont}}$ contains a nonzero constant function.

As for the separating points, by Claim 4.3, for $X, Y \in K_p$ and $\bar{d}(X,Y) > 0$, then there exists $\varepsilon > 0$, such that the finite multisets

$$
\{X_{i:} : \|X_{i:}\|_{\mathbb{R}^d} > \varepsilon\} \neq \{Y_{i:} : \|Y_{i:}\|_{\mathbb{R}^d} > \varepsilon\}.
$$

More formally, denote

$$
m_x(v) := \#\{i : X_{i:} = v, \|X_{i:}\|_{\mathbb{R}^d} > \varepsilon\}; \quad m_y(v) := \#\{i : Y_{i:} = v, \|Y_{i:}\|_{\mathbb{R}^d} > \varepsilon\},
$$

then the inequality of the two finite multisets can be written as

$$
\exists v \in \mathbb{R}^d, \text{ such that } m_x(v) \neq m_y(v).
$$

Since the two multisets are finite, then

$$
\gamma := \min_{u \in Q} \|u - v\|_{\mathbb{R}^d} > 0,
$$

where $Q := \{u \in \mathbb{R}^d : u = X_{i:} \text{ or } u = Y_{i:}, i \in \mathbb{N},\ \|u\|_{\mathbb{R}^d} > \varepsilon,\ u \neq v\}$.

We denote $B_v^\eta := \{V \in \mathbb{R}^d : \|V - v\|_{\mathbb{R}^d} < \eta\}$. By Urysohn's Lemma ([58], Lemma 15.6), there exists a continuous function $\rho : \mathbb{R}^d \to \mathbb{R}$, such that $\rho\left(\overline{B}_v^{\eta/3}\right) = 1$ and $\rho\left(\mathbb{R}^d / B_v^{\eta/2}\right) = 0$. Let $h = 1$ and $\sigma(u) = u$, then,

$$\sigma\left(\sum_{i=1}^{\infty} \|X_{i:}\|_{\mathbb{R}^d}^p \rho\left(X_{i:}\right)\right) = \|v\|_{\mathbb{R}^d}^p m_x(v),$$

$$\sigma\left(\sum_{i=1}^{\infty} \|Y_{i:}\|_{\mathbb{R}^d}^p \rho\left(Y_{i:}\right)\right) = \|v\|_{\mathbb{R}^d}^p m_y(v).$$

Using the fact that $m_x(v) \neq m_y(v)$, we conclude that $\mathcal{F}_{\text{cont}}$ separates points in $\overline{O}\left(K_p\right)$.

□

Lemma 4.4. *$\mathcal{F}_{\text{DS}}$ is dense in $\mathcal{F}_{\text{cont}}$.*

Proof of Lemma 4.4. The argument relies on the classical universal approximation theorem (UAT) for fully connected neural networks [13], which can be stated as: for any $\eta > 0$, and any $\rho \in C\left(\mathbb{R}^d, \mathbb{R}^h\right)$, there exists $\hat{\rho} \in \mathcal{N}^{w,\phi}(d, h)$ such that

$$\sup_{\|V\|_{\mathbb{R}^d} \leq R} \|\rho(V) - \hat{\rho}(V)\|_{\mathbb{R}^h} \leq \eta, \tag{4.1}$$

where $w \in \mathbb{N}$ and ϕ continuous, and non-polynomial; $R > 0$ is some positive constant.

Given any function $f \in \mathcal{F}_{\text{cont}}$, it admits the form

$$f(X) = \sigma\left(\sum_{i=1}^{\infty} \|X_{i:}\|_{\mathbb{R}^d}^p \rho(X_{i:})\right),\ X \in K_p,$$

for some $h \in \mathbb{N}$, $\sigma \in C(\mathbb{R}^h, \mathbb{R})$, $\rho \in C(\mathbb{R}^d, \mathbb{R}^h)$.

By compactness of K_p in Proposition 3.2, we set

$$M_x := \sup_{x \in K_p} \sum_{i=1}^{\infty} \|X_{i:}\|_{\mathbb{R}^d}^p.$$

Since ρ is continuous, it maps compact sets to compact sets, so define

$$M_\rho := \sup_{\|V\|_{\mathbb{R}^d} \le M_x} \|\rho(V)\|_{\mathbb{R}^h}.$$

Then for every $X \in K_p$, $\left\| \sum_{i=1}^{\infty} \|X_{i:}\|_{\mathbb{R}^d}^p \rho(X_{i:}) \right\|_{\mathbb{R}^h} \le M_x M_\rho$.

Applying the UAT (4.1) to σ, for any $\varepsilon > 0$, there exists $\hat{\sigma} \in \mathcal{N}^{w,\phi}(h, 1)$, such that

$$\sup_{\|u\|_{\mathbb{R}^h} \le M_x M_\rho} |\sigma(u) - \hat{\sigma}(u)| \le \frac{\varepsilon}{2}.$$

Since $\hat{\sigma}$ is continuous, there exists $\delta > 0$, such that

$$\|u - v\|_{\mathbb{R}^h} \le \delta \Rightarrow |\hat{\sigma}(u) - \hat{\sigma}(v)| \le \frac{\varepsilon}{2}.$$

Applying the UAT again to ρ, there exists $\hat{\rho} \in \mathcal{N}^{w,\phi}(d, h)$ such that

$$\sup_{\|V\|_{\mathbb{R}^d} \le M_x} \|\rho(V) - \hat{\rho}(V)\|_{\mathbb{R}^h} \le \frac{\delta}{M_x}.$$

Consequently,

$$\sup_{x \in K_p} \left\| \sum_{i=1}^{\infty} \|X_{i:}\|_{\mathbb{R}^d}^p (\rho(X_{i:}) - \hat{\rho}(X_{i:})) \right\|_{\mathbb{R}^h} \le M_x \cdot \frac{\delta}{M_x} = \delta.$$

Therefore,

$$\sup_{x\in K_p}\left|\sigma\left(\sum_{i=1}^{\infty}\|X_{i:}\|_{\mathbb{R}^d}^p\rho(X_{i:})\right)-\hat{\sigma}\left(\sum_{i=1}^{\infty}\|X_{i:}\|_{\mathbb{R}^d}^p\hat{\rho}(X_{i:})\right)\right|$$

$$\leq \sup_{x\in K_p}\left|\sigma\left(\sum_{i=1}^{\infty}\|X_{i:}\|_{\mathbb{R}^d}^p\rho(X_{i:})\right)-\hat{\sigma}\left(\sum_{i=1}^{\infty}\|X_{i:}\|_{\mathbb{R}^d}^p\rho(X_{i:})\right)\right|+$$

$$\left|\hat{\sigma}\left(\sum_{i=1}^{\infty}\|X_{i:}\|_{\mathbb{R}^d}^p\rho(X_{i:})\right)-\hat{\sigma}\left(\sum_{i=1}^{\infty}\|X_{i:}\|_{\mathbb{R}^d}^p\hat{\rho}(X_{i:})\right)\right|$$

$$\leq \frac{\varepsilon}{2}+\frac{\varepsilon}{2}=\varepsilon,$$

which shows that $\mathcal{F}_{\mathrm{DS}}$ uniformly approximates every function in $\mathcal{F}_{\mathrm{cont}}$, and therefore $\mathcal{F}_{\mathrm{DS}}$ is dense in $\mathcal{F}_{\mathrm{cont}}$. □

Proof of Theorem 3.5. By Proposition 3.1, the Stone–Weierstrass Theorem, and Lemmas 4.1 and 4.2, we conclude that $\mathcal{F}_{\mathrm{cont}}$ is dense in $C\left(\overline{O}(K_p)\right)$. Combining this with Lemma 4.4 and the transitivity of density, it follows that $\mathcal{F}_{\mathrm{DS}}$ is dense in $C\left(\overline{O}(K_p)\right)$. □

4.1.3 Proof of Theorem 3.9

We first define

$$\mathcal{F}_{\overline{\mathrm{cont}}} := \left\{ f(\mu)=\sigma\left(\int \rho\, d\mu\right) \mid h\in\mathbb{N},\ \rho\in\mathcal{F}_p\left(\mathbb{R}^d,\mathbb{R}^h\right),\ \sigma\in C\left(\mathbb{R}^h,\mathbb{R}\right)\ \right\}.$$

The proof strategy of Theorem 3.9 is similar to Theorem 3.5. We decompose the proof into three auxiliary lemmas.

Lemma 4.5. *$\mathcal{F}_{\overline{\mathrm{cont}}}$ is a subalgebra of $C(K)$.*

Proof of Lemma 4.5. To show $\mathcal{F}_{\overline{\text{cont}}}$ is a subalgebra, it suffices to verify that it is closed under scalar multiplication, addition and pointwise multiplication.

1. Closure under scalar multiplication is trivial.

2. Closure under addition:

   Let $f_1, f_2 \in \mathcal{F}_{\overline{\text{cont}}}$, $f_1(\mu) = \sigma_1\left(\int \rho_1 d\mu\right)$ and $f_2(\mu) = \sigma_2\left(\int \rho_2 d\mu\right)$, for some $\rho_1 \in \mathcal{F}_p(\mathbb{R}^d, \mathbb{R}^{h_1}), \rho_2 \in \mathcal{F}_p(\mathbb{R}^d, \mathbb{R}^{h_2})$, $\sigma_1 \in C(\mathbb{R}^{h_1}, \mathbb{R}), \sigma_2 \in C(\mathbb{R}^{h_2}, \mathbb{R})$. We choose positive constants $M_1 > 0$, $M_2 > 0$ such that for any $x \in \mathbb{R}^d$,

   $$\|\rho_1(x)\|_{\mathbb{R}^{h_1}} \leq M_1(1 + \|x\|_{\mathbb{R}^d}^p), \quad \|\rho_2(x)\|_{\mathbb{R}^{h_2}} \leq M_2(1 + \|x\|_{\mathbb{R}^d}^p).$$

   Then define:

   $$\begin{cases} \rho_0 : \mathbb{R}^d \to \mathbb{R}^{h_1+h_2}, \rho_0(x) = (\rho_1(x), \rho_2(x)) \\ \sigma_0 : \mathbb{R}^{h_1+h_2} \to \mathbb{R}, \sigma_0(u_1, u_2) = \sigma_1(u_1) + \sigma_2(u_2) \end{cases},$$

   which are both continuous.

   $$\begin{aligned} \|\rho_0(x)\|_{\mathbb{R}^{h_1+h_2}} &= \left\|\left(\rho_1(x), \mathbf{0}_{h_2}\right) + \left(\mathbf{0}_{h_1}, \rho_2(x)\right)\right\|_{\mathbb{R}^{h_1+h_2}} \\ &\leq \left\|\left(\rho_1(x), \mathbf{0}_{h_2}\right)\right\|_{\mathbb{R}^{h_1+h_2}} + \left\|\left(\mathbf{0}_{h_1}, \rho_2(x)\right)\right\|_{\mathbb{R}^{h_1+h_2}}. \end{aligned}$$

   Since the sets

   $$\left\{(y, \mathbf{0}_{h_2}) \in \mathbb{R}^{h_1+h_2} : \|y\|_{\mathbb{R}^{h_1}} \leq 1\right\}, \left\{(\mathbf{0}_{h_1}, y) \in \mathbb{R}^{h_1+h_2} : \|y\|_{\mathbb{R}^{h_2}} \leq 1\right\}$$

   are both compact in $\mathbb{R}^{h_1+h_2}$, and the norm is continuous, then there exists finite positive constant $A > 0$, $B > 0$, such that,

   $$A = \sup_{\|y\|_{\mathbb{R}^{h_1}} \leq 1} \|(y, \mathbf{0}_{h_2})\|_{\mathbb{R}^{h_1+h_2}}, \quad B = \sup_{\|y\|_{\mathbb{R}^{h_2}} \leq 1} \|(\mathbf{0}_{h_1}, y)\|_{\mathbb{R}^{h_1+h_2}}.$$

Thus,

$$\|\rho_0(x)\|_{\mathbb{R}^{h_1+h_2}} \leq A\|\rho_1(x)\|_{\mathbb{R}^{h_1}} + B\|\rho_2(x)\|_{\mathbb{R}^{h_2}} \leq (AM_1 + BM_2)(1 + \|x\|_{\mathbb{R}^d}^p),$$

which yields that ρ_0 still satisfies p-moment growth condition.

Then we get

$$(f_1 + f_2)(\mu) = \sigma_0\left(\int \rho_1 d\mu \, , \int \rho_2 d\mu\right) = \sigma_0\left(\int \rho_0 d\mu\right),$$

where $\sigma_0 \in C(\mathbb{R}^{h_1+h_2}, \mathbb{R})$ and $\rho_0 \in \mathcal{F}_p(\mathbb{R}^d, \mathbb{R}^{h_1+h_2})$, which implies closure under addition.

3. Closure under pointwise multiplication

Define:

$$\begin{cases} \rho_0' : \mathbb{R}^d \to \mathbb{R}^{h_1+h_2}, \rho_0'(x) = (\rho_1(x), \rho_2(x)) \\ \sigma_0' : \mathbb{R}^{h_1+h_2} \to \mathbb{R}, \sigma_0'(u_1, u_2) = \sigma_1(u_1) \cdot \sigma_2(u_2) \end{cases},$$

which are both continuous, and $\rho_0' \in \mathcal{F}_p(\mathbb{R}^d, \mathbb{R}^{h_1+h_2})$ shown above. Then we get

$$(f_1 \cdot f_2)(\mu) = f_1(\mu) \cdot f_2(\mu) = \sigma_0'\left(\int \rho_1 d\mu \, , \int \rho_2 d\mu\right) = \sigma_0'\left(\int \rho_0' d\mu\right),$$

which implies closure under pointwise multiplication, $f_1 \cdot f_2 \in \mathcal{F}_{\overline{\text{cont}}}$.

□

Lemma 4.6. *$\mathcal{F}_{\overline{\text{cont}}}$ separates points in K, and contains a nonzero constant function.*

Proof of Lemma 4.6. $\mathcal{F}_{\overline{\text{cont}}}$ contains a nonzero constant function is trivial. We next show that $\mathcal{F}_{\overline{\text{cont}}}$ separates points in K.

Suppose $\mu, \nu \in K \subset \mathcal{P}_p\left(\mathbb{R}^d\right)$, and $\mu \neq \nu$. Then there exist a Borel set $B \subset \mathbb{R}^d$ such that $\mu(B) \neq \nu(B)$. Without loss of generality, assume $\mu(B) > \nu(B)$. Let $\delta := \mu(B) - \nu(B) > 0$.

Since $\mathbb{R}^d$ is a metric space, then any measure in $\mathcal{P}_p(\mathbb{R}^d)$ is regular [59] (Thm 7.1.7). Therefore, by the definition of regularity, for any $\varepsilon > 0$, there exists a closed set $E \subset B$ such that

$$\mu(B) \leq \mu(E) + \varepsilon,$$

and an open set $O \supset B$ such that

$$\nu(B) \geq \nu(O) - \varepsilon.$$

E and $\mathbb{R}^d \backslash O$ are disjoint closed set in $\mathbb{R}^d$. Then by Urysohn's lemma [58] (Lemma 15.6) , there exists a continuous function $\phi : \mathbb{R}^d \to [0, 1]$, with $\phi(E) = 1$ and $\phi(\mathbb{R}^d \backslash O) = 0$.

Then,

$$\int \phi \, d\mu \geq \mu(E) \geq \mu(B) - \varepsilon,$$

$$\int \phi \, d\nu \leq \nu(O) \leq \nu(B) + \varepsilon.$$

Let $\varepsilon = \frac{\delta}{3} > 0$, then,

$$\int \phi \, d\mu - \int \phi \, d\nu \geq \mu(B) - \nu(B) - 2\varepsilon = \frac{\delta}{3} > 0.$$

Recall

$$\mathcal{F}_{\overline{\text{cont}}} := \left\{ f(\mu) = \sigma\left(\int \rho d\mu\right) \mid h \in \mathbb{N}, \rho \in \mathcal{F}_p\left(\mathbb{R}^d, \mathbb{R}^h\right), \sigma \in C\left(\mathbb{R}^h, \mathbb{R}\right) \right\}.$$

Choose $h = 1, \sigma(u) = u$, and $\rho = \phi \in C\left(\mathbb{R}^d\right)$. ρ satisfies $|\rho(x)| \leq 1 \leq (1 + \|x\|^p_{\mathbb{R}^d})$,

which implies $\rho \in \mathcal{F}_p\left(\mathbb{R}^d, \mathbb{R}^h\right)$.

Then it follows that for any $\mu, \nu \in K \subset \mathcal{P}_p\left(\mathbb{R}^d\right)$ with $\mu \neq \nu$, there exists a function $f \in \mathcal{F}_{\overline{\text{cont}}}$ such that $f(\mu) \neq f(\nu)$.

□

Lemma 4.7. *$\mathcal{F}_{\overline{\text{DS}}}$ is dense in $\mathcal{F}_{\overline{\text{cont}}}$.*

Proof of Lemma 4.7. For any function $f \in \mathcal{F}_{\overline{\text{cont}}}$, we can write

$$f(\mu) = \sigma\left(\int \rho d\mu\right), \quad \mu \in K,$$

for some $h \in \mathbb{N}$, $\rho \in \mathcal{F}_p(\mathbb{R}^d, \mathbb{R}^h)$, and $\sigma \in C\left(\mathbb{R}^h, \mathbb{R}\right)$. There exists a positive constant $M > 0$ such that for any $x \in \mathbb{R}^d$,

$$\|\rho(x)\|_{\mathbb{R}^h} \leq M(1 + \|x\|_{\mathbb{R}^d}^p).$$

By Theorem 3.8, the map $\mu \mapsto \int \rho d\mu$ is continuous. Thus, its image of the compact set K is also compact. In particular, there exists $M_0 > 0$ such that

$$\sup_{\mu \in K} \left\| \int \rho d\mu \right\|_{\mathbb{R}^h} \leq M_0.$$

Applying the UAT in (4.1) to σ, we obtain: for any $\varepsilon > 0$, there exists $\hat{\sigma} \in \mathcal{N}^{w,\phi}(h, 1)$ such that:

$$\sup_{\mu \in K} \left| \sigma\left(\int \rho d\mu\right) - \hat{\sigma}\left(\int \rho d\mu\right) \right| \leq \frac{\varepsilon}{2}.$$

Since $\hat{\sigma}$ is continuous, there exists $\delta > 0$ such that

$$\|u - v\|_{\mathbb{R}^h} \leq \delta \Rightarrow |\hat{\sigma}(u) - \hat{\sigma}(v)| \leq \frac{\varepsilon}{2}.$$

Since K is compact, it is tight and p-uniformly integrable, by Proposition

3.6. Let $\overline{B}_R := \{x \in \mathbb{R}^d : \|x\|_{\mathbb{R}^d} \leq R\}$ denote the closed ball of radius R, then

$$\exists\ R_1 > 0, \text{ such that } \sup_{\mu \in K} \mu\left(\mathbb{R}^d \backslash \overline{B}_{R_1}\right) \leq \frac{\delta}{4M} \quad \text{(tight)},$$

$$\exists\ R_2 > 0, \text{ such that } \sup_{\mu \in K} \int_{\mathbb{R}^d \backslash \overline{B}_{R_2}} \|x\|_{\mathbb{R}^d}^p \, \mathrm{d}\mu \leq \frac{\delta}{4M} \quad (p\text{-uniformly integrable}).$$

Let $R := \max\{R_1, R_2\}$, then

$$\begin{aligned}
\sup_{\mu \in K} \int_{\mathbb{R}^d \backslash \overline{B}_R} \|\rho(x)\|_{\mathbb{R}^h} d\mu(x) &\leq \sup_{\mu \in K} \int_{\mathbb{R}^d \backslash \overline{B}_R} M(1 + \|x\|_{\mathbb{R}^d}^p) d\mu \\
&\leq M \sup_{\mu \in K} \mu(\mathbb{R}^d \backslash \overline{B}_R) + M \sup_{\mu \in K} \int_{\mathbb{R}^d \backslash \overline{B}_R} \|x\|_{\mathbb{R}^d}^p d\mu \\
&\leq M \sup_{\mu \in K} \mu(\mathbb{R}^d \backslash \overline{B}_{R_1}) + M \sup_{\mu \in K} \int_{\mathbb{R}^d \backslash \overline{B}_{R_2}} \|x\|_{\mathbb{R}^d}^p d\mu \\
&\leq \frac{\delta}{4} + \frac{\delta}{4} = \frac{\delta}{2}
\end{aligned}$$

Since the sets $\overline{B}_R$ and $\overline{\mathbb{R}^d \backslash \overline{B}_{2R}}$ are both closed, Urysohn's Lemma implies that there exists a continuous function

$$\phi : \mathbb{R}^d \to [0, 1] \text{ such that } \phi(x) = 1 \text{ for } x \in \overline{B}_R, \quad \phi(x) = 0 \text{ for } x \in \overline{\mathbb{R}^d \backslash \overline{B}_{2R}}.$$

Then,

$$\phi \cdot \rho(x) := \phi(x)\rho(x) \in C_0\left(\mathbb{R}^d, \mathbb{R}^h\right).$$

Moreover,

$$\sup_{\mu \in K} \int \|\phi \cdot \rho - \rho\|_{\mathbb{R}^h} d\mu \leq \sup_{\mu \in K} \int_{\mathbb{R}^d \backslash \overline{B}_R} \|\rho(x)\|_{\mathbb{R}^h} d\mu(x) \leq \frac{\delta}{2}.$$

When approximating functions in $C_0(\mathbb{R}^d)$, a stronger universal approximation theorem is available [15]. It states that if the activation function ϕ is

continuous, nonpolynomial, and asymptotically polynomial at $\pm\infty$, then

$$C_0\left(\mathbb{R}^d\right) \subset \overline{\mathcal{N}^{w,\phi}(d,1)}.$$

The extension to vector-valued functions is immediate. For any $f \in C_0\left(\mathbb{R}^d, \mathbb{R}^h\right)$, write $f = (f_1, f_2, \ldots, f_h)$, where each $f_i \in C_0\left(\mathbb{R}^d\right)$ vanishes at infinity. Then we can construct h independent networks to approximate each output dimension. Concatenating these networks yields a vector-valued neural network. Consequently,

$$C_0\left(\mathbb{R}^d, \mathbb{R}^h\right) \subset \overline{\mathcal{N}^{w,\phi}(d,h)}.$$

Since $\phi \cdot \rho \in C_0\left(\mathbb{R}^d, \mathbb{R}^h\right)$, there exists $\hat{\rho} \in \mathcal{N}^{w,\phi}(d,h)$ such that

$$\sup_{x\in\mathbb{R}^d} \|\hat{\rho}(x) - \phi\cdot\rho(x)\| \leq \frac{\delta}{2}.$$

Therefore,

$$\begin{aligned}\left\|\int (\rho - \hat{\rho}) d\mu\right\|_{\mathbb{R}^h} &\leq \int \|\phi\cdot\rho - \rho\|_{\mathbb{R}^h} d\mu + \int \|\phi\cdot\rho - \hat{\rho}\|_{\mathbb{R}^h} d\mu \\ &\leq \frac{\delta}{2} + \frac{\delta}{2} = \delta,\end{aligned}$$

and hence

$$\sup_{\mu\in K}\left|\hat{\sigma}\left(\int \rho d\mu\right) - \hat{\sigma}\left(\int \hat{\rho} d\mu\right)\right| \leq \frac{\varepsilon}{2}.$$

We define $\hat{f} = \hat{\sigma}(\int \hat{\rho} d\mu) \in \mathcal{F}_{\overline{\mathrm{DS}}}$, and estimate the approximation error:

$$\begin{aligned}
&\sup_{\mu \in K} \left| \sigma\left(\int \rho d\mu\right) - \hat{\sigma}\left(\int \hat{\rho} d\mu\right) \right| \\
&\leq \sup_{\mu \in K} \left| \sigma\left(\int \rho d\mu\right) - \hat{\sigma}\left(\int \rho d\mu\right) \right| + \sup_{\mu \in K} \left| \hat{\sigma}\left(\int \rho d\mu\right) - \hat{\sigma}\left(\int \hat{\rho} d\mu\right) \right| \\
&\leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon,
\end{aligned}$$

which concludes that $\mathcal{F}_{\overline{\mathrm{DS}}}$ is dense in $\mathcal{F}_{\overline{\mathrm{cont}}}$.

□

Proof of Theorem 3.9. By Proposition 3.1, the Stone–Weierstrass Theorem, and Lemmas 4.5 and 4.6, we conclude that $\mathcal{F}_{\overline{\mathrm{cont}}}$ is dense in $C(K)$. Combining this with Lemma 4.7 and the transitivity of density, it follows that $\mathcal{F}_{\overline{\mathrm{DS}}}$ is dense in $C(K)$. □

4.2 Missing proofs from Section 3.2

4.2.1 Proof of Theorem 3.15

Proof of Theorem 3.15. It suffices to prove that, for any simple function $F = (V, E)$, the homomorphism density $t(F, \cdot)$ can be represented in the form (3.3).

The proof proceeds by induction on the number of edges. For the base case, consider $F = (V, E)$ with $|E| = 1$. In this case, we define

$$\mathcal{L}_1^W = \begin{pmatrix} L_1^1(W) + b_1^1 \\ L_1^0(W) + b_1^0 \end{pmatrix}, \quad L_1^1(W) = 0,\ b_1^1 = 1,\ L_1^0(W) = W,\ b_1^0 = 0.$$

Then $T \circ \mathcal{L}_1^W(W) = W$, which, after integrating by $\mathcal{L}_0$, yields exactly the homomorphism density.

Induction Hypothesis: Assume that for any simple graph $F = (V, E)$ with $|E| = m$, $m \geq 1$ (regardless of the number of vertices), the following representation holds:

$$\prod_{(i,j)\in E} W(x_i, x_j) \;=\; T \circ \mathcal{L}_l^W \circ \cdots \circ T \circ \mathcal{L}_1^W(W),$$

for some integer l and layers $\{\mathcal{L}_i^W\}_{i=1}^l$.

Then consider a simple graph $F = (V, E)$ with $|E| = m + 1$. Choose an edge $e \in E$, and let $F_e = (V_e, E_e)$ be the graph obtained by removing e. We have $|E_e| = m$ and denote $|V_e| = n$, so by the induction hypothesis, F_e admits the desired representation, i.e.,

$$T \circ \mathcal{L}_l^W \circ \cdots \circ T \circ \mathcal{L}_1^W(W) = G_l(x_1, \cdots, x_n) = \prod_{(i,j)\in E_e} W\left(x_i, x_j\right).$$

Here we note that the output may be equal to the homomorphism density, but the output tensor can contain more coordinates, as in the lift operation defined in Example 3.13.2. However, since we are allowed to permute them as in Example 3.13.1, we can, without loss of generality, assume that the output tensor has exactly $|V_e|$ coordinates.

Depending on the endpoints of e, the vertex set satisfies one of the following cases:

1. $|V_e| = |V| - 2$: both endpoints of e are new (the new edge forms a separate component on two new vertices).

We can add one layer as:

$$G_l \to \begin{pmatrix} \mathrm{Lift}_2(G_l) \\ W \end{pmatrix} \xrightarrow{T} G_{l+1},$$

where $\mathrm{Lift}_2 \in \mathrm{LE}_{n\to n+2}$ is defined as:

$$\mathrm{Lift}_2 : \mathcal{G}_n \mapsto \mathcal{G}_{n+2},\ (\mathrm{Lift}_2\ G)\ (x_0, y_0, x_1, \cdots, x_n) = G\ (x_1, \cdots, x_n)\ .$$

And then take the integral over all coordinates of G_{l+1}, which is just the homomorphism density function with respect to F.

2. $|V_e| = |V| - 1$: exactly one endpoint of e is new (the new edge attaches a new vertex to the existing graph), and suppose e connects the vertex $p \in V_e$ and a new vertex $n+1$. We can add one layer as:

$$G_l \to \begin{pmatrix} \mathrm{LP}(G_l) \\ W \end{pmatrix} \xrightarrow{T} G_{l+1},$$

where $\mathrm{LP} \in \mathrm{LE}_{n\to n+1}$ is defined as:

$$\mathrm{LP} : \mathcal{G}_n \mapsto \mathcal{G}_{n+1}, (\mathrm{LP}\ \ G)\ (x_0, x_1, \cdots, x_n) = G\left(x_2, \cdots, x_p, x_1, x_{p+1}, \cdots, x_n\right).$$

3. $|V_e| = |V|$: both endpoints of e already belong to V_e (the new edge connects two existing vertices), and suppose e connects vertices $p, q \in V_e$. Then, we can add one layer as:

$$G_l \to \begin{pmatrix} \mathrm{P}_{pq}(G_l) \\ W \end{pmatrix} \xrightarrow{T} G_{l+1},$$

where $\mathrm{P}_{pq} \in \mathrm{LE}_{n\to n}$ is defined as (assume $p < q$):

$$\mathrm{P}_{pq} : \mathcal{G}_n \mapsto \mathcal{G}_n,$$

$$\left(\mathrm{P}_{pq}\ G\right)\ (x_1, \cdots, x_n) = G\left(x_3, \cdots, x_{p+1}, x_1, x_{p+2}, \cdots, x_{q+1}, x_2, x_{q+2}, \cdots, x_n\right).$$

Therefore, by induction, we complete the proof. □

4.3 Missing proofs from Section 3.3

4.3.1 Proof of Theorem 3.16

Proof of Theorem 3.16 . By Lemma 4.5 in [11], $\mathcal{IWN}$ is continuous with respect to δ_2 distance. It suffices to show that $f \mapsto W_f$ is continuous with symmetric distance in its input and δ_2 distance in its range.

For $f, g \in K_f = \left\{ f \in L^2\left([0,1]; \mathbb{R}^k\right) : \sup_{t\in[0,1]} \|f(t)\|_2 \le R \right\}$,

$$
\begin{aligned}
&\delta_2(W_f, W_g)^2 \\
&= \inf_{\varphi\in S_{[0,1]}} \left(\int_{[0,1]^2} \left| W_f(x,y) - W_g(\varphi(x),\varphi(y)) \right|^2 dxdy \right) \\
&= \frac{1}{4R^4} \inf_{\varphi\in S_{[0,1]}} \left(\int_{[0,1]^2} |\langle f(x), f(y)\rangle - \langle g(\varphi(x)), g(\varphi(y))\rangle|^2 dxdy \right) \\
&= \frac{1}{4R^4} \inf_{\varphi\in S_{[0,1]}} \left(\int_{[0,1]^2} |\langle f(x) - g(\varphi(x)), f(y)\rangle + \langle g(\varphi(x)), f(y) - g(\varphi(y))\rangle|^2 dxdy \right) \\
&\le \frac{1}{4R^2} \inf_{\varphi\in S_{[0,1]}} \left(\int_{[0,1]^2} \left(\|f(x) - g(\varphi(x))\|_2 + \|f(y) - g(\varphi(y))\|_2 \right)^2 dxdy \right) \\
&\le \frac{1}{2R^2} \inf_{\varphi\in S_{[0,1]}} \left(\int_{[0,1]^2} \left(\|f(x) - g(\varphi(x))\|_2^2 + \|f(y) - g(\varphi(y))\|_2^2 \right) dxdy \right) \\
&= \frac{1}{R^2} \inf_{\varphi\in S_{[0,1]}} \left(\int_0^1 \|f(x) - g(\varphi(x))\|_2^2 dx \right)
\end{aligned}
$$

Since for any $h \in \mathrm{O}(k)$,

$$W_{h\cdot f} = \langle h \cdot f(x), h \cdot f(y)\rangle = \langle f(x), f(y)\rangle = W_f,$$

therefore,

$$\delta_2(W_f, W_g)^2 = \delta_2(W_{h\cdot f}, W_g)^2 \leq \frac{1}{R^2} \inf_{\varphi \in S_{[0,1]}} \left(\int_0^1 \|h \cdot f(x) - g(\varphi(x))\|_2^2 dx \right).$$

Then

$$\delta_2(W_f, W_g) \leq \frac{1}{R} \inf_{h \in \mathrm{O}_k} \inf_{\varphi \in S_{[0,1]}} \left(\int_0^1 \|h \cdot f(x) - g(\varphi(x))\|_2^2 dx \right)^{1/2} = \frac{1}{R} \overline{d}(f, g),$$

which implies that the covariance map $f \mapsto W_f$ is continuous.

□

4.3.2 Proof of Theorem 3.17

The proof relies on the correspondence between $\overline{O}(K_f)$ and $(\widetilde{W_0}, \delta_2)$ under the action of orthogonal group, together with the universality of the $\mathcal{IWN}$ model on compact subsets of $(\widetilde{\mathcal{W}_0}, \delta_2)$ [11]. We first state and prove several auxiliary lemmas.

Lemma 4.8. *For $f, g \in K_f \subset L^2\left([0,1]; \mathbb{R}^k\right)$, and W_f,W_g are the corresponding covariance functions, then $\overline{d}(f, g) = 0$ if and only if $\delta_2(W_f, W_g) = 0$.*

We first claim that

Claim 4.9. *Let $f, g \in L^2\left([0,1], \mathbb{R}^k\right)$. Suppose that*

$$\langle f(x), f(y) \rangle = \langle g(x), g(y) \rangle \quad \text{for almost every } x, y \in [0,1].$$

Then there exists an orthogonal transformation $h \in \mathrm{O}(k)$ such that $g = hf$ almost everywhere.

Proof of Claim 4.9. First, we can identify $L^2\left([0,1], \mathbb{R}^k\right)$ with $\mathcal{L}\left(L^2([0,1]), \mathbb{R}^k\right)$,

which is the set of bounded linear operators between two Hilbert spaces $L^2([0,1]) \mapsto \mathbb{R}^k$, equipped with the Hilbert–Schmidt norm $\|\cdot\|_{HS}$.

To specify, each $X \in L^2\left([0,1], \mathbb{R}^k\right)$ can be written in components as

$$X = (\varphi_1, \ldots, \varphi_k)^\top, \quad \varphi_i \in L^2([0,1]),$$

which defines a bounded linear map $X : L^2([0,1]) \to \mathbb{R}^k$ by

$$X\varphi = \left(\langle \varphi_1, \varphi \rangle, \ldots, \langle \varphi_k, \varphi \rangle\right)^\top.$$

Conversely, by Riesz Representation Theorem [60] (Thm 3.4), each bounded linear operator $X \in \mathcal{L}\left(L^2([0,1]), \mathbb{R}^k\right)$ can be written as this form for some $\varphi_1, \cdots, \varphi_k \in L^2([0,1])$.

Define $\mathcal{V}_k := \operatorname{span}\{\varphi_1, \ldots, \varphi_k\}$. Then X vanishes on the orthogonal complement $\mathcal{V}_k^\perp$. Since $X : \mathcal{V}_k \to \mathbb{R}^k$ is a linear map between finite-dimensional vector spaces, it admits a singular value decomposition [61] (Thm. 6.26).

Thus, there exist non-negative singular values $\sigma_1, \ldots, \sigma_k \in \mathbb{R}_{\geq 0}$, orthonormal bases $\{v_1, \ldots, v_k\} \subset \mathbb{R}^k$, and $\{u_1, \ldots, u_k\} \subset L^2([0,1])$ such that

$$X = \sum_{i=1}^{k} \sigma_i \langle u_i, \cdot \rangle v_i.$$

Moreover, for each $\sigma_i > 0$, the u_i is an eigenvector of the self-adjoint operator X^*X corresponding to the eigenvalue σ_i^2. For indices i with $\sigma_i = 0$, the vectors u_i can be chosen to complete the set $\{u_1, \ldots, u_k\}$ into an orthonormal basis of $\mathcal{V}_k$.

Let X_f be the linear operator defined by function $f \in L^2([0,1], \mathbb{R}^k)$, i.e.

$$X_f\varphi = \left(\langle f_1, \varphi \rangle, \ldots, \langle f_k, \varphi \rangle\right)^\top,$$

and the adjoint operator

$$X_f^* v = \sum_{i=1}^{k} v_i f_i.$$

Similar definition for X_g and X_g^*. Then consider the self-adjoint operator, for $\varphi \in L^2([0,1])$,

$$\begin{aligned}(X_f^* X_f \varphi)(x) = \sum_{i=1}^{k} \langle f_i, \varphi \rangle f_i(x) &= \sum_{i=1}^{k} f_i(x) \int_0^1 f_i(y)\varphi(y)dy \\ &= \int_0^1 \langle f(x), f(y) \rangle \varphi(y)dy.\end{aligned}$$

Since $\langle f(x), f(y) \rangle = \langle g(x), g(y) \rangle$ almost everywhere, then

$$X_f^* X_f = X_g^* X_g \ .$$

Therefore, we can choose the same orthonormal bases $\{u_1, \ldots, u_k\} \subset L^2([0,1])$ for X_f and X_g, such that

$$X_f = \sum_{i=1}^{k} \sigma_i \langle u_i, \cdot \rangle v_i, \quad X_g = \sum_{i=1}^{k} \sigma_i \langle u_i, \cdot \rangle w_i,$$

where $\{v_1, \ldots, v_k\}$ and $\{w_1, \ldots, w_k\}$ are both the orthonormal bases in $\mathbb{R}^k$. Then there exists $h \in \mathrm{O}(k)$ such that $h(v_i) = w_i$, which implies

$$hX_f = h(\sum_{i=1}^{k} \sigma_i \langle u_i, \cdot \rangle v_i) = \sum_{i=1}^{k} \sigma_i \langle u_i, \cdot \rangle w_i = X_g.$$

For any $\omega \in \mathbb{R}^k$,

$$\langle g(\cdot), w \rangle = X_g^* w = \left(hX_f\right)^* w = X_f^* h^\top w = \left\langle f(\cdot), h^\top w \right\rangle = \langle hf(\cdot), w \rangle,$$

yields $g = hf$ in $L^2([0,1], \mathbb{R}^k)$. □

Proof of Lemma 4.8. First, by Theorem 3.16, if $\overline{d}(f,g) = 0$, then $\delta_2(W_f, W_g) = 0$. We only need to show if $\delta_2(W_f, W_g) = 0$, then $\overline{d}(f,g) = 0$.

Since $\delta_2\left(W_f, W_g\right) = 0$, by Corollary 8.14 in [8], there exists measure-preserving maps $\varphi, \psi \in \bar{S}_{[0,1]}$ such that $W_f^{\varphi} = W_g^{\psi}$ almost everywhere, which is just

$$\langle f(\varphi(x)), f(\varphi(y))\rangle = \langle g(\psi(x)), g(\psi(y))\rangle \quad \text{for almost every } x, y \in [0,1].$$

By Claim 4.9, there exists $h \in \mathrm{O}(k)$ such that $hf^{\varphi} = g^{\psi}$, where $f^{\varphi}(x) = f(\varphi(x))$.

$$\begin{aligned}
\overline{d}(f,g) &= \inf_{h\in \mathrm{O}(k)} \inf_{\varphi \in S_{[0,1]}} \left(\int_0^1 \|h(f(t)) - g(\varphi(t))\|_2^2 dt \right)^{1/2} \\
&= \inf_{h\in \mathrm{O}(k)} \inf_{\varphi,\psi \in \overline{S}_{[0,1]}} \left(\int_0^1 \|h(f(\varphi(t))) - g(\psi(t))\|_2^2 dt \right)^{1/2} \\
&\leq 0,
\end{aligned}$$

which yields $\overline{d}(f,g) = 0$. Thus, $\overline{d}(f,g) = 0 \Leftrightarrow \delta_2\left(W_f, W_g\right) = 0$.

□

Lemma 4.10 ([8], Cor 8.14 and 10.34)**.** *$U, W : [0,1]^2 \mapsto \mathbb{R}$ are bounded symmetric measurable functions, the following are equivalent:*

1. $\delta_p(U,W) = 0, \quad p \in [1,\infty)$.
2. $\delta_{\square}(U,W) = 0$.
3. *For any simple graph F, the homomorphism densities are equal, i.e.,* $t(F,U) = t(F,W)$.

Proof of Theorem 3.17. We consider the class of functions $\mathcal{F}_h$ of the form

$$f \mapsto W_f \mapsto t\left(F, W_f\right),$$

where F is a simple graph and t denotes the homomorphism density. By Lemmas 4.8 and 4.10, the class $\mathcal{F}_h$ separates points. Analogous to Proposition 3.10, $\mathcal{F}_h$ is dense in $C\left(\overline{O}\left(K_f\right)\right)$.

By Theorem 3.16, and since the input $\overline{O}\left(K_f\right)$ is compact, the image of $\overline{O}\left(K_f\right)$ is a compact subset of $\left(\widetilde{\mathcal{W}_0}, \delta_2\right)$. The $\mathcal{IWN}$ model is universal on compact subsets of $\left(\widetilde{\mathcal{W}_0}, \delta_2\right)$, and thus can approximate arbitrarily well any homomorphism density of simple graphs [11]. Consequently, $\mathcal{F}_{\mathrm{IWN}}$ is dense in $C\left(\overline{O}\left(K_f\right)\right)$.

□

Chapter 5

Conclusion and future work

In conclusion, we introduce transferable models and establish their universality for three types of any-dimensional invariant models, operating on sets, graphs, and point clouds, respectively. Future work includes finding transferable graph models in cut distance, which ensures better combinatorial properties. Another promising research direction is to explore the expressivity of any-dimensional equivariant models.

References

[1] Eitan Levin, Yuxin Ma, Mateo Díaz, and Soledad Villar. "On Transferring Transferability: Towards a Theory for Size Generalization". In: *arXiv preprint arXiv:2505.23599* (2025).

[2] Luana Ruiz, Luiz Chamon, and Alejandro Ribeiro. "Graphon neural networks and the transferability of graph neural networks". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 1702–1712.

[3] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. "Deep sets". In: *Advances in neural information processing systems* 30 (2017).

[4] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. "Pointnet: Deep learning on point sets for 3d classification and segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 652–660.

[5] Sohir Maskey, Ron Levie, and Gitta Kutyniok. "Transferability of graph neural networks: an extended graphon approach". In: *Applied and Computational Harmonic Analysis* 63 (2023), pp. 48–83.

[6] Ron Levie. "A graphon-signal analysis of graph neural networks". In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 64482–64525.

[7] Ron Levie, Wei Huang, Lorenzo Bucci, Michael Bronstein, and Gitta Kutyniok. "Transferability of spectral graph convolutional neural networks". In: *Journal of Machine Learning Research* 22.272 (2021), pp. 1–59.

[8] László Lovász. *Large networks and graph limits*. American Mathematical Soc., 2012.

[9] George Cybenko. "Approximation by superpositions of a sigmoidal function". In: *Mathematics of control, signals and systems* 2.4 (1989), pp. 303–314.

[10] Christian Bueno and Alan Hylton. “On the representation power of set pooling networks”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 17170–17182.

[11] Daniel Herbst and Stefanie Jegelka. “Higher-order graphon neural networks: Approximation and cut distance”. In: *arXiv preprint arXiv:2503.14338* (2025).

[12] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. “Multilayer feedforward networks are universal approximators”. In: *Neural networks* 2.5 (1989), pp. 359–366.

[13] Moshe Leshno, Vladimir Ya Lin, Allan Pinkus, and Shimon Schocken. “Multilayer feedforward networks with a nonpolynomial activation function can approximate any function”. In: *Neural networks* 6.6 (1993), pp. 861–867.

[14] Kurt Hornik. “Approximation capabilities of multilayer feedforward networks”. In: *Neural networks* 4.2 (1991), pp. 251–257.

[15] Teun DH van Nuland. “Noncompact uniform universal approximation”. In: *Neural Networks* 173 (2024), p. 106181.

[16] Ariel Neufeld and Philipp Schmocker. “Universal approximation results for neural networks with non-polynomial activation function over non-compact domains”. In: *arXiv preprint arXiv:2410.14759* (2024).

[17] Ahmed Abdeljawad and Thomas Dittrich. “Weighted Sobolev Approximation Rates for Neural Networks on Unbounded Domains”. In: *arXiv preprint arXiv:2411.04108* (2024).

[18] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. “How powerful are graph neural networks?” In: *arXiv preprint arXiv:1810.00826* (2018).

[19] Andrei Leman and Boris Weisfeiler. “A reduction of a graph to a canonical form and an algebra arising during this reduction”. In: *Nauchno-Technicheskaya Informatsiya* 2.9 (1968), pp. 12–16.

[20] Haggai Maron, Heli Ben-Hamu, Hadar Serviansky, and Yaron Lipman. “Provably powerful graph networks”. In: *Advances in neural information processing systems* 32 (2019).

[21] Christopher Morris, Martin Ritzert, Matthias Fey, William L Hamilton, Jan Eric Lenssen, Gaurav Rattan, and Martin Grohe. "Weisfeiler and leman go neural: Higher-order graph neural networks". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. 01. 2019, pp. 4602–4609.

[22] Haggai Maron, Ethan Fetaya, Nimrod Segol, and Yaron Lipman. "On the universality of invariant networks". In: *International conference on machine learning*. PMLR. 2019, pp. 4363–4371.

[23] Nicolas Keriven and Gabriel Peyré. "Universal invariant and equivariant graph neural networks". In: *Advances in neural information processing systems* 32 (2019).

[24] Dmitry Yarotsky. "Universal approximations of invariant maps by neural networks". In: *Constructive Approximation* 55.1 (2022), pp. 407–474.

[25] Akiyoshi Sannai, Yuuki Takai, and Matthieu Cordonnier. "Universal approximations of permutation invariant/equivariant functions by deep neural networks". In: *arXiv preprint arXiv:1903.01939* (2019).

[26] Nicolas Keriven, Alberto Bietti, and Samuel Vaiter. "Convergence and stability of graph convolutional networks on large random graphs". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 21512–21523.

[27] Luana Ruiz, Luiz FO Chamon, and Alejandro Ribeiro. "Transferability properties of graph neural networks". In: *IEEE Transactions on Signal Processing* 71 (2023), pp. 3474–3489.

[28] Matthieu Cordonnier, Nicolas Keriven, Nicolas Tremblay, and Samuel Vaiter. "Convergence of message passing graph neural networks with generic aggregation on random graphs". In: *GSP 2023-6th Graph Signal Processing workshop*. 2023, pp. 1–3.

[29] Luana Ruiz, Ningyuan Teresa Huang, and Soledad Villar. "A spectral analysis of graph neural networks on dense and sparse graphs". In: *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2024, pp. 9936–9940.

[30] Chen Cai and Yusu Wang. "Convergence of invariant graph networks". In: *International Conference on Machine Learning*. PMLR. 2022, pp. 2457–2484.

[31] Eitan Levin and Venkat Chandrasekaran. "Free descriptions of convex sets". In: *arXiv preprint arXiv:2307.04230* (2023).

[32] Eitan Levin and Mateo Díaz. "Any-dimensional equivariant neural networks". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2024, pp. 2773–2781.

[33] Taco Cohen and Max Welling. "Group equivariant convolutional networks". In: *International conference on machine learning*. PMLR. 2016, pp. 2990–2999.

[34] Risi Kondor and Shubhendu Trivedi. "On the generalization of equivariance and convolution in neural networks to the action of compact groups". In: *International conference on machine learning*. PMLR. 2018, pp. 2747–2755.

[35] Michael M Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. "Geometric deep learning: Grids, groups, graphs, geodesics, and gauges". In: *arXiv preprint arXiv:2104.13478* (2021).

[36] Nimrod Segol and Yaron Lipman. "On universal equivariant set networks". In: *arXiv preprint arXiv:1910.02421* (2019).

[37] Haggai Maron, Heli Ben-Hamu, Nadav Shamir, and Yaron Lipman. "Invariant and equivariant graph networks". In: *arXiv preprint arXiv:1812.09902* (2018).

[38] Vıctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. "E (n) equivariant graph neural networks". In: *International conference on machine learning*. PMLR. 2021, pp. 9323–9332.

[39] Luana Ruiz, Fernando Gama, and Alejandro Ribeiro. "Graph neural networks: Architectures, stability, and transferability". In: *Proceedings of the IEEE* 109.5 (2021), pp. 660–682.

[40] Fabian Fuchs, Daniel Worrall, Volker Fischer, and Max Welling. "Se (3)-transformers: 3d roto-translation equivariant attention networks". In: *Advances in neural information processing systems* 33 (2020), pp. 1970–1981.

[41] Ben Blum-Smith, Ningyuan Huang, Marco Cuturi, and Soledad Villar. "A Galois theorem for machine learning: Functions on symmetric matrices and point clouds via lightweight invariant features". In: *arXiv preprint arXiv:2405.08097* (2024).

[42] Risi Kondor. "N-body networks: a covariant hierarchical neural network architecture for learning atomic potentials". In: *arXiv preprint arXiv:1803.01588* (2018).

[43] Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. "Tensor field networks: Rotation-and translation-equivariant neural networks for 3d point clouds". In: *arXiv preprint arXiv:1802.08219* (2018).

[44] Mario Geiger and Tess Smidt. "e3nn: Euclidean neural networks". In: *arXiv preprint arXiv:2207.09453* (2022).

[45] Sékou-Oumar Kaba, Arnab Kumar Mondal, Yan Zhang, Yoshua Bengio, and Siamak Ravanbakhsh. "Equivariance with learned canonicalization functions". In: *International Conference on Machine Learning*. PMLR. 2023, pp. 15546–15566.

[46] Soledad Villar, David W Hogg, Kate Storey-Fisher, Weichi Yao, and Ben Blum-Smith. "Scalars are universal: Equivariant machine learning, structured like classical physics". In: *Advances in neural information processing systems* 34 (2021), pp. 28848–28863.

[47] Ben Blum-Smith and Soledad Villar. "Machine learning and invariant theory". In: *Notices of the American Mathematical Society* 70.08 (2023), pp. 1–1.

[48] Soledad Villar, Weichi Yao, David W Hogg, Ben Blum-Smith, and Bianca Dumitrascu. "Dimensionless machine learning: Imposing exact units equivariance". In: *Journal of Machine Learning Research* 24.109 (2023), pp. 1–32.

[49] Mahdi Hashemi. "Enlarging smaller images before inputting into convolutional neural network: zero-padding vs. interpolation". In: *Journal of Big Data* 6.1 (2019), pp. 1–13.

[50] Richard Socher, Cliff C Lin, Chris Manning, and Andrew Y Ng. "Parsing natural scenes and natural language with recursive neural networks". In: *Proceedings of the 28th international conference on machine learning (ICML-11)*. 2011, pp. 129–136.

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

[52] Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. "Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators". In: *Nature machine intelligence* 3.3 (2021), pp. 218–229.

[53] Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. "Fourier neural operator for parametric partial differential equations". In: *arXiv preprint arXiv:2010.08895* (2020).

[54] Walter Rudin. "Functional analysis". In: (1991).

[55] Joseph Diestel. *Sequences and series in Banach spaces*. Vol. 92. Springer Science & Business Media, 2012.

[56] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer, 2005.

[57] Cédric Villani et al. *Optimal transport: old and new*. Vol. 338. Springer, 2008.

[58] Stephen Willard. *General topology*. Courier Corporation, 2012.

[59] Vladimir I Bogachev. *Measure theory*. Springer, 2007.

[60] John B Conway. *A course in functional analysis*. Springer, 2019.

[61] Stephen H Friedberg, Arnold J Insel, and Lawrence E Spence. *Linear algebra*. Prentice Hall, 1997.