Journal of Machine Learning Research 25 (2024) 1-57 Submitted 8/22; Revised 7/24; Published 8/24

Deep Neural Network Approximation of Invariant Functions through Dynamical Systems

Qianxiao Li QIANXIAO@NUS.EDU.SG
Department of Mathematics
Institute for Functional Intelligent Materials
National University of Singapore
10 Lower Kent Ridge Road, Singapore, 119076

Ting Lin LINTINGSMS@PKU.EDU.CN
School of Mathematical Sciences
Peking University
5 Yiheyuan Road, Beijing, China, 100871

Zuowei Shen MATZUOWS@NUS.EDU.SG
Department of Mathematics
National University of Singapore
10 Lower Kent Ridge Road, Singapore, 119076

Editor: Joan Bruna

Abstract

We study the approximation of functions which are invariant with respect to certain permutations of the input indices using flow maps of dynamical systems. Such invariant functions include the much studied translation-invariant ones involving image tasks, but also encompasses many permutation-invariant functions that find emerging applications in science and engineering. We prove sufficient conditions for universal approximation of these functions by a controlled dynamical system, which can be viewed as a general abstraction of deep residual networks with symmetry constraints. These results not only imply the universal approximation for a variety of commonly employed neural network architectures for symmetric function approximation, but also guide the design of architectures with approximation guarantees for applications involving new symmetry requirements.

Keywords: Deep learning, approximation theory, dynamical systems, control, invariance

1. Introduction

Deep learning enabled significant progress in computer vision and natural language processing, arguably due to its ability to exploit structures of the data. For example, convolution neural networks (CNN) target translational symmetry (Azulay and Weiss, 2018), whereas recurrent neural networks (RNN) account for causal structures and time-homogeneity (Li et al., 2021, 2022). Recently, we witness increasing interest to apply machine learning to problems arising in science and engineering (Goh et al., 2017; Butler

et al., 2018; Noé et al., 2020). Here, very different structures present themselves in the underlying data. For instance, in modelling structure-property relationships involving atomic systems (Butler et al., 2018), the input data often comes in the form of a list of atoms with their respective property descriptions, and the goal is to predict some macroscopic quantities depending on these input specifications. Examples of such properties include energy, elasticity constants, etc. In these cases, there is symmetry with respect to permutations on the list of atoms, and sometimes on a subset of the properties, e.g. 3D atomic coordinates, via a choice of coordinate axes. These transformations can be viewed as specific subgroups of the symmetric group on the coordinates of the feature vectors, and such transformation leaves the macroscopic property invariant. In the computational chemistry literature, current methods to tackle this rely on graphical representations to induce invariance (Xie and Grossman, 2018; Chen et al., 2019), but such methods have limited applications if atomic properties are not spatial coordinates, e.g. correlation functions (Yeomans, 1992). Hence, it is desirable to devise methods to approximate functions that are invariant under the action of a specified subgroup of the permutation group on the feature indices.

It is the purpose of this work to investigate the role of deep learning in approximating functions that are invariant under the action of permutation subgroups on input indices. This includes the CNN as a special case, with the subgroup being the group of translations. However, to address new challenges arising in scientific applications, it is necessary to generalize the theory to accommodate other types of symmetries. There is an interesting interaction between deep neural networks and symmetry that is worth noting. On the one hand, enforcing a certain type of symmetry on any model hypothesis space necessarily restricts its approximation flexibility. On the other hand, in function approximation applications the goal is to develop a hypothesis space that has the power to approximate arbitrary functions. This dilemma leads to considerable challenges in building hypothesis spaces to approximate symmetric function families, especially when the symmetry group under consideration is complex (Bogatskiy et al., 2020; Finzi et al., 2020). In this regard, deep learning offers an attractive solution to this problem. The distinguishing feature of deep learning, compared with traditional hypothesis spaces, is the presence of compositional structures. A deep neural network represents a function in compositional form:

$$\boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{g} \circ \boldsymbol{F}_T \circ \cdots \circ \boldsymbol{F}_1(\boldsymbol{x}), \tag{1}$$

where each $\boldsymbol{F}_i$ is a shallow neural network layer and $\boldsymbol{g}$ is the last layer used to match output dimensions. The crucial point is that each $\boldsymbol{F}_i$ and $\boldsymbol{g}$ can be made simple, yet it can yield a complex $\boldsymbol{F}$ by simply increasing the number of layers T. Suppose now that we want to build an $\boldsymbol{F}$ which is invariant under a transformation $\mathtt{t}$ on the input data. Deep learning can accomplish this by simply choosing $\boldsymbol{F}_i$ to be equivariant, i.e. $\boldsymbol{F}_i(\mathtt{t}(\boldsymbol{x})) = \mathtt{t}(\boldsymbol{F}_i(\boldsymbol{x}))$ and $\boldsymbol{g}$ to be invariant, i.e. $\boldsymbol{g}(\mathtt{t}(\boldsymbol{x})) = \boldsymbol{g}(\boldsymbol{x})$. Check that this implies that $\boldsymbol{F}$ is $\mathtt{t}$ invariant. These symmetry restrictions may force each $\boldsymbol{F}_i$ and $\boldsymbol{g}$ to be simple, but this now no longer necessarily limits approximation power of the deep neural network, which builds complexity through increasing T. In other words, by building complexity

through composition, deep learning may circumvent the contradicting requirements of symmetry and flexibility.

Let us make concrete the preceding idea via some examples. For ease of exposition we will proceed informally while detailed analyses are deferred to Sections 4.2-4.3. Now, consider approximating a function f that is invariant under full permutation of its feature indices, i.e., for any permutation σ of $\{1, \dots, n\}$, we have

$$f(x_{\sigma(1)}, x_{\sigma(2)}, \cdots, x_{\sigma(n)}) = f(x_1, \cdots, x_n). \tag{2}$$

Such symmetries arise often in scientific applications, e.g. when f is the total energy of an interacting n-particle system and x_i is the descriptor (e.g. position) of the i^{th} particle. Clearly, the total energy should not depend on the order at which these particle features are presented. Typically, f can be rather complicated, so finding a good hypothesis space that approximates it yet retains the permutation symmetry may be challenging. On the other hand, it is easy to find some simple (nonlinear) functions that respects permutation symmetry. For example, consider the function

$$\boldsymbol{x} \mapsto \sum_{i=1}^{n} g(x_i), \tag{3}$$

where g is some nonlinear scalar function. We can also account for pairwise interactions through

$$\boldsymbol{x} \mapsto \sum_{i=1}^{n} h_1(x_i) + \sum_{i,j=1}^{n} h_2(x_i, x_j). \tag{4}$$

Clearly, both (3) and (4) are permutation invariant. However, neither of them have strong approximation power, in the sense that we cannot hope that they can approximate any symmetric invariant functions on $\mathbb{R}^n$, even if g, h_1, h_2 are well chosen. Yet, we can use such simple functions as building blocks to achieve universal approximation. The key idea is to also consider equivariant mappings, i.e. those that commutes with permutations. A simple example would be

$$\boldsymbol{x} \mapsto [x_1 + h(x_1, \bar{\boldsymbol{x}}), x_2 + h(x_2, \bar{\boldsymbol{x}}), \dots, x_n + h(x_n, \bar{\boldsymbol{x}})], \tag{5}$$

where $\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Again, we cannot expect such a map alone to approximate an arbitrary transformation from $\mathbb{R}^n$ to itself. However, we may consider the compositional form

$$f = g \circ \boldsymbol{f}_K \circ \cdots \circ \boldsymbol{f}_1. \tag{6}$$

where g may take the forms (3) or (4), and $\boldsymbol{f}_k$ may take the form (5) (different for each k).

Then, the question we are concerned with is whether (6) can indeed approximate any permutation invariant function by taking K sufficiently large. Due to the vast number of possible choices for the forms of $g, \boldsymbol{f}_k$, we shall not restrict our analysis to a specific choice. Rather, we seek *sufficient* conditions for any such choices to achieve universal approximation.

Moreover, we will go beyond the full permutation group and consider instead an arbitrary (transitive) subgroup of it. This is because in many applications, invariance is restricted to a proper subgroup of permutations. An example is product symmetric invariance, which finds applications in computational chemistry. For a data point $\boldsymbol{x} \in \mathbb{R}^{m\times n}$ written as a matrix

$$\boldsymbol{x} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & \cdots & \ddots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}, \tag{7}$$

we are interested in functions f that are invariant after one permutes any row or column of the data entry. More precisely, let σ be a permutation of $\{1, 2, \cdots, m\}$, τ be a permutation of $\{1, 2, \cdots, n\}$, then f satisfies

$$f\left(\begin{bmatrix} x_{\sigma(1)\tau(1)} & x_{\sigma(1)\tau(2)} & \cdots & x_{\sigma(1)\tau(n)} \\ x_{\sigma(2)\tau(1)} & \cdots & \ddots & x_{\sigma(2)\tau(n)} \\ \vdots & \vdots & \vdots & \vdots \\ x_{\sigma(m)\tau(1)} & x_{\sigma(m)\tau(2)} & \cdots & x_{\sigma(m)\tau(n)} \end{bmatrix}\right) = f\left(\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & \cdots & \ddots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}\right) \tag{8}$$

One may check that this is a proper subgroup of the full permutation group on mn objects. Now, constructing an architecture to achieve universal approximation under this symmetry is a non-trivial task, and few, if any, results are known. We will apply our theory to construct several types of residual block-based deep architecture to handle product permutations (Section 4.3). A final example is translational symmetry corresponding to the shift subgroup, for which convolutional type architectures has been devised to achieve approximation (see Section 4.1).

Overall, the question we consider here can be abstracted as follows. Given a transitive permutation subgroup G of the symmetric group, we derive conditions for which

$$f = g \circ \boldsymbol{f}_K \circ \cdots \boldsymbol{f}_1, \qquad \boldsymbol{f}_k = \boldsymbol{x} + \boldsymbol{g}_k(\boldsymbol{x}) \tag{9}$$

is a universal approximator. Here g is G invariant and $\boldsymbol{g}_k$ are G equivariant. A significant feature is that in our theory $g, \boldsymbol{g}_k$ can be as simple as possible, in the sense that we have no need to require, for example, that each $\boldsymbol{g}_k$ is a wide-enough neural network. In many applications, only parsimonious nonlinearity is required to achieve approximation through composition (increasing K).

The approximation of functions through composition has been studied in a number of recent works. For example, non-asymptotic optimal approximation rates for fully connected deep ReLU networks are obtained in Shen et al. (2022); Lu et al. (2021); Shen et al. (2019). In Zhang et al. (2022); Shen et al. (2021a,b), the approximation of fully connected deep network beyond the ReLU is discussed. In particular, a family of simple activation functions is designed in Shen et al. (2021a), whose corresponding neural networks can approximate an arbitrary continuous function with arbitrary accuracy by

a fixed size network. Here, our focus is on the interaction of composition and symmetry. We note that while there are a number of results in the literature on universal approximation of symmetric functions, they mostly focus on specific architectures and symmetry groups (Cohen and Welling, 2016; Bogatskiy et al., 2020; Finzi et al., 2020; Yarotsky, 2022; Sannai et al., 2019; Maron et al., 2019; Ravanbakhsh et al., 2017; Ravanbakhsh, 2020; Zaheer et al., 2017; Zweig and Bruna, 2021). The results in this paper are of a different nature. Our goal is to give general sufficient conditions for any architecture to achieve universal approximation under any symmetry constraints induced by suitable permutation subgroups. Thus, the results can be used to deduce universal approximation results for a variety of architectures and symmetry groups in essentially the same way, including convolutional neural networks, permutation-invariant networks, etc. We illustrate this in Section 4, where we show that our results immediately imply universal approximation of residual versions of popular architectures used to approximate symmetric functions. More importantly, these sufficient conditions can be used to guide the development of new architectures to accommodate new symmetry structures that may arise in future applications. We will give illustrations of this in the realm of property prediction for crystalline and amorphous materials (Tian et al., 2022).

To study this problem mathematically, we will adopt the continuous idealization of deep learning first introduced in E (2017); Haber and Ruthotto (2017); Li et al. (2018) and Chen et al. (2018). In particular, we derive general sufficient conditions for the approximation of functions invariant to the aforementioned symmetries through composition - now in the form of dynamics. These results expands upon those in Li et al. (2023) by incorporating symmetry considerations. A benefit of the dynamical systems approach is that we can consider the approximation problem with limited width. Many works focusing on the approximation theory of neural networks rely on using sufficiently large width. Often, this technique will lead to widths going to infinity to achieve accurate approximation. However, in practical deep residual architectures, increasing the depth at constant width leads to performance improvements (He et al., 2016), and our results can be used to analyze the origins of such gains.

The paper is organized as follows. In Section 2, we introduce notation and prior work on the analysis of approximation by dynamical hypothesis spaces, which sets the stage for the analysis in this paper. We then present the main approximation result in Section 3 and its applications in Section 4. The proofs of these results are presented in Section A, with auxiliary results found in the appendices.

2. Preliminaries

In this section, we introduce notations, definitions and present some known previous results relevant to the main results in this paper. Section 2.1 recalls main results in Li et al. (2023) on the approximation theory of flow maps without symmetry considerations. Section 2.2 introduces terminologies in group theory used to describe discrete invariance and equivariance. Based on these two concepts, in Section 2.3 we provide some ele-

mentary results on universal approximation property under invariance and equivariance, which reveal challenges one encounters in formulating a general approximation result.

Throughout this paper, we adopt the following notations:

1. We use boldface letters $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}$ for points in the Euclidean space $\mathbb{R}^n$. For scalars such as the component of these vectors, we use non-bold letters x, y, z.
2. We use normal, non-bold letters like f, g, h and α, β, γ for scalar-valued functions, shortened as *functions*, and normal bold letters like $\boldsymbol{f}, \boldsymbol{g}, \boldsymbol{h}$ and $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}$ for vector-valued functions, shortened as *mappings*.
3. We use sans serif letters $\mathsf{G}, \mathsf{H}, \mathsf{K}$ to denote groups, and a, b to denote elements in these groups.
4. We use calligraphic font letters $\mathcal{A}, \mathcal{F}, \mathcal{G}$ to denote families of functions or mappings.
5. Unless otherwise stated, we adopt periodic boundary conditions when specifying vector or tensor indices. That is, if $\boldsymbol{x} \in \mathbb{R}^n$, then $x_{n+1} := x_1$, $x_{-1} := x_n$, and so on.

2.1 Dynamical Hypothesis Spaces

We first recall the problem formulation and main results in Li et al. (2023) relevant to the present analysis. The key problem investigated there is the approximation of functions through dynamical evolution. In particular, associated with an ordinary differential equation (ODE)

$$\dot{\boldsymbol{z}}(t) = \boldsymbol{f}_t(\boldsymbol{z}(t)), \qquad \boldsymbol{z}(0) = \boldsymbol{x}, \qquad t \in [0, T], \tag{10}$$

is the mapping $\boldsymbol{x} \mapsto \boldsymbol{z}(T)$, which can be used to approximate functions by choosing the vector field $\boldsymbol{f}_t$ from a family $\mathcal{F}$. We call $\mathcal{F}$ a *control family*, since it serves to control the dynamics of $\boldsymbol{z}$, as in the study of optimal control of differential equations (see e.g. Evans (1983)). To build the flow-induced hypothesis space, we first define the *flow map* for time-homogenous continuous dynamical systems.

Definition 1 (Flow map) *Let $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^n$ be Lipschitz. We define the flow map associated with $\boldsymbol{f}$ at time horizon T as $\boldsymbol{\phi}(\boldsymbol{f}, T)(\boldsymbol{x}) = \boldsymbol{z}(T)$, where $\dot{\boldsymbol{z}}(t) = \boldsymbol{f}(\boldsymbol{z}(t))$ with initial data $\boldsymbol{z}(0) = \boldsymbol{x}$. It is well-known (see e.g. Arnold (1973)) that the mapping $\boldsymbol{\phi}(\boldsymbol{f}, T)$ is Lipschitz for any real number T, and the inverse of $\boldsymbol{\phi}(\boldsymbol{f}, T)$ is $\boldsymbol{\phi}(-\boldsymbol{f}, T)$. In particular, the flow map is bi-Lipschitz.*

Based on the flow map, we can define the *attainable set* for a given control family $\mathcal{F}$, which contains the compositions of flow maps generated by dynamics driven by vector fields chosen from $\mathcal{F}$.

Definition 2 (Attainable set) *For a given control family $\mathcal{F}$ of Lipschitz mappings, we define its attainable set*

$$\mathcal{A}_{\mathcal{F}} := \{\boldsymbol{\phi}(\boldsymbol{f}_1, \tau_1) \circ \cdots \circ \boldsymbol{\phi}(\boldsymbol{f}_k, \tau_k) : \boldsymbol{f}_1, \cdots, \boldsymbol{f}_k \in \mathcal{F}, \tau_j \geq 0, k \geq 1\}. \tag{11}$$

The attainable set builds complexity through compositions of flow maps, and can be used to approximate mappings in $\mathbb{R}^n$. However, often in applications we aim to approximate relationships whose range is not $\mathbb{R}^n$. An example is scalar regression problems, where we aim to construct approximations of functions from $\mathbb{R}^n$ to $\mathbb{R}$. Thus, to achieve approximation we require an additional composition with a terminal family of functions to fix the range. This gives rise to the following dynamical hypothesis space on which we study approximation properties.

Definition 3 (Dynamical hypothesis space) *Given a control family $\mathcal{F}$ and a terminal family $\mathcal{G}$ of functions from $\mathbb{R}^n$ to $\mathbb{R}$, both assumed Lipschitz, the dynamical hypothesis space $\mathcal{H}_{ode}$ is defined as*

$$\mathcal{H}_{ode} = \mathcal{H}_{ode}(\mathcal{F}, \mathcal{G}) := \{\boldsymbol{g} \circ \boldsymbol{\varphi} : \boldsymbol{g} \in \mathcal{G}, \boldsymbol{\varphi} \in \mathcal{A}_{\mathcal{F}}\}. \tag{12}$$

The key approximation problem seeks conditions on $\mathcal{F}$ and $\mathcal{G}$ that induce the density of $\mathcal{H}_{\text{ode}}$ in appropriate function spaces. This is also known as the universal approximation property in the machine learning literature. To establish such results in appropriate generality, the concept of *well functions* was introduced in Li et al. (2023) in order to provide sufficient conditions to achieve universal approximation. Here, we recall its definition.

Definition 4 (Well function) *We say a Lipschitz function $h : \mathbb{R}^n \to \mathbb{R}$ is a well function if there exists a bounded open convex set $\Omega \subset \mathbb{R}^n$ such that*

$$\{\boldsymbol{x} \in \mathbb{R}^n : h(\boldsymbol{x}) = 0\} = \overline{\Omega} \tag{13}$$

Moreover, we say that a vector valued function $\boldsymbol{h} : \mathbb{R}^n \to \mathbb{R}^{n'}$ is a well function if each of its component $h_i : \mathbb{R}^n \to \mathbb{R}$ is a well function in the sense above. Specifically, a Lipschitz function $h : \mathbb{R} \to \mathbb{R}$ is a one-dimensional well function if $\{x : h(x) = 0\}$ is a non-degenerate closed interval.

We now state the main density result in Li et al. (2023) concerning the dynamical hypothesis space for $n \geq 2$. In the following, for any collection $\mathcal{F}$ of functions on $\mathbb{R}^d$, we denote by $\mathrm{CH}(\mathcal{F})$ its convex hull and $\overline{\mathrm{CH}}(\mathcal{F})$ its closure in the topology of compact convergence.

Theorem 5 (Main result in Li et al. (2023)) *Suppose $n \geq 2$. Let $\boldsymbol{F} : \mathbb{R}^n \to \mathbb{R}^m$, $m \geq 1$, be continuous. If the control family $\mathcal{F}$ and the terminal family $\mathcal{G}$ are both Lipschitz and satisfy*

1. *For any compact set $K \subset \mathbb{R}^n$, there exists $\boldsymbol{g} \in \mathcal{G}$ such that $\boldsymbol{F}(K) \subset \boldsymbol{g}(\mathbb{R}^n)$.*
2. *$\mathcal{F}$ is restricted affine invariant. That is, $\boldsymbol{f} \in \mathcal{F}$ implies $D\boldsymbol{f}(A \cdot + \boldsymbol{b}) \in \mathcal{F}$, where $\boldsymbol{b} \in \mathbb{R}^n$ is any vector, and D, A are any $n \times n$ diagonal matrices such that the entries of D are ± 1 or 0, and the entries of A are smaller than or equal to 1 in absolute value.*

3. $\overline{\mathrm{CH}}(\mathcal{F})$ *contains a well function.*

Then for any $p \in [1, \infty)$*, compact* $K \subset \mathbb{R}^n$ *and* $\varepsilon > 0$*, there exists* $\hat{\boldsymbol{F}} \in \mathcal{H}_{ode}$ *such that* $\|\boldsymbol{F} - \hat{\boldsymbol{F}}\|_{L^p(K)} \leq \varepsilon$.

The straightforward but important application of Theorem 5 in deep learning is when $\mathcal{F} = \mathcal{F}_\sigma := \{V\sigma(W \cdot +b)\}$, where σ is a nonlinear activation function, such as ReLU or the Sigmoid function. This then corresponds to a universal approximation theorem of deep residual neural networks[1]. through composing fixed-width residual layers, which are building blocks of residual neural networks (He et al., 2016). However, we note that Theorem 5 makes no explicit reference to neural networks and can be viewed as a general result on approximation of functions by the flow maps of dynamical systems.

The main purpose of this paper is to derive similar results under symmetry constraints, where we aim to establish an analogue density result for group invariant dynamical hypothesis spaces. One may wonder if the following simple argument suffices: if we additionally require $\mathcal{F}$ to be equivariant and $\mathcal{G}$ to be invariant, then $\mathcal{H}_{\mathrm{ode}}$ is indeed invariant and thus if all conditions are Theorem 5 are satisfied, we deduce the universal approximation property. It turns out that we cannot translate Theorem 5 into invariant setting if the permutation group is transitive. For example, full permutation equivariance will force D, A to be multiples of the identity matrix. Hence, to make headway we will have to suitably relax the affine invariance condition. This will be the primary challenge in establishing the main results of this paper, and further highlights the competition between restrictions induced by symmetry and flexibility required for universal approximation.

2.2 Group Theory Notation

In this section, we fix some terminologies involving basic group theory. By $\mathsf{G} \leq \mathsf{H}$ we mean G is a subgroup of a group H.

Permutation Groups. Given a finite set S, a permutation group on S consists of some permutations on S which form a group under the composition. Without loss of generality, we can identify S with $\{n\} := \{1, \cdots, n\}$, and all groups considered here will be permutation groups. For fixed S, the group of all permutations is called the symmetric group, denoted as S. The identity element is denoted as e. We denote by $(i\ j)$ the transposition element in S that exchanges i and j, while keeping the others fixed.

Group Actions on Indices and Vectors. Given a permutation s on $\{n\}$, it is natural to describe how s acts on $\{n\}$. We use $i \mapsto \mathsf{s}(i)$ to denote this mapping. For a vector $\boldsymbol{x} = [x_1, \cdots, x_n] \in \mathbb{R}^n$, we define the action on $\mathbb{R}^n$ by $\mathsf{s}(\boldsymbol{x}) = [x_{\mathsf{s}(1)}, \cdots, x_{\mathsf{s}(n)}]$ and for a point set A, define $\mathsf{s}(A) := \{\mathsf{s}(\boldsymbol{x}) : \boldsymbol{x} \in A\}$. All the transformation in $\mathbb{R}^n$ considered in this paper are of this type.

1. While Theorem 5 is stated in the continuous setting, basic numerically analysis shows that the result carries over to discretized architectures, see Section D.

Transitivity. We call a permutation group G on $\{n\}$ *transitive* if for each $i, j \in \{n\}$, there exists a permutation $\mathsf{g} \in \mathsf{G}$ such that $\mathsf{g}(i) = j$.

Stabilizer. Given $\mathsf{G} \leq \mathsf{S}$, define the *stabilizer* of element i as $\mathrm{Stab}_i(\mathsf{G}) = \{\mathsf{g} \in \mathsf{G} : \mathsf{g}(i) = i\}$. A basic fact is that if $\mathsf{a}(i) = j$, then $\mathsf{a}\,\mathrm{Stab}_i(\mathsf{G})\mathsf{a}^{-1} = \mathrm{Stab}_j(\mathsf{G})$. Thus, for a transitive group G, all stabilizers are conjugate to each other.

Cross Section. Given $\mathsf{g} \in \mathsf{S}$, we define the *cross section*

$$Q_{\mathsf{g}} := \{\boldsymbol{x} \in \mathbb{R}^n : x_{\mathsf{g}^{-1}(1)} > x_{\mathsf{g}^{-1}(2)} \cdots > x_{\mathsf{g}^{-1}(n)}\}. \tag{14}$$

If we write $Q := Q_{\mathsf{e}}$, the cross section of the identity element keeping all indices fixed, then we have $Q_{\mathsf{g}} = \mathsf{g}(Q)$.

Transversal. Given $\mathsf{G} \leq \mathsf{S}$, and denote by $k := |\mathsf{S}|/|\mathsf{G}|$. We say a collection $A := \{\mathsf{a}_1, \mathsf{a}_2, \cdots, \mathsf{a}_k\} \subset \mathsf{S}$ is a (right) *transversal* with respect to G if

- $\mathsf{a}_i \mathsf{a}_j^{-1} \notin \mathsf{G}$ for $i \neq j$, and
- for any $\mathsf{b} \in \mathsf{S}$, there exists $\mathsf{a}_i \in A$ such that $\mathsf{a}_i \mathsf{b}^{-1} \in \mathsf{G}$.

We define the *cross section* with respect to A as

$$Q_A := \bigcup_{i=1}^{k} Q_{\mathsf{a}_i}. \tag{15}$$

Invariant Functions and Equivariant Mappings. We now give precise definitions of invariant and equivariant mappings, together with related concepts. Let G be a permutation subgroup. We say $f : \mathbb{R}^n \to \mathbb{R}$ is a G *invariant function* if for all $\mathsf{g} \in \mathsf{G}$,

$$f(\mathsf{g}(\boldsymbol{x})) = f(\boldsymbol{x}). \tag{16}$$

We say $\boldsymbol{\varphi} : \mathbb{R}^n \to \mathbb{R}^n$ is a G *equivariant mapping* if for all $\mathsf{g} \in \mathsf{G}$,

$$\boldsymbol{\varphi}(\mathsf{g}(\boldsymbol{x})) = \mathsf{g}(\boldsymbol{\varphi}(\boldsymbol{x})). \tag{17}$$

We introduce some examples of invariant functions and equivariant mappings to close this part. These motivate us to build approximation frameworks under such symmetry considerations, and serve as important examples for the application of our theory to deep learning subsequently.

Example 1 (Translation) *We first recall the definition of translation in one and multiple dimensions. For the one-dimensional case, we define* $\mathsf{T} = \{\mathsf{t}^1, \mathsf{t}^2, \cdots, \mathsf{t}^n\}$*, where* $\mathsf{t}(i) = i+1$ *is the shift operator in one dimension. Recall that the periodic condition is assumed. For the multidimensional case, we define the translation group as* $\mathsf{T}_{d_1} \times \cdots \times \mathsf{T}_{d_N}$*, for* $n = d_1 \cdots d_N$*, and*

$$\mathsf{t}_k([i_1, \cdots, i_N]) = [i_1, \cdots, \mathsf{t}(i_k), \cdots, i_N]. \tag{18}$$

Here, the ambient Euclidean space is identified as $\mathbb{R}^n := \mathbb{R}^{d_1} \times \cdots \times \mathbb{R}^{d_N}$*, and the corresponding multi-index* $[i_1, \cdots, i_N]$ *is a sequence of length* N *and* $i_s \in \{d_s\}$*.*

One of the most commonly used T *equivariant mapping is constructed by the convolution operation. Given* $\boldsymbol{w} = [w_1, \cdots, w_p]$*, the convolution operation*[2] *in one dimension is*

$$\boldsymbol{w} * \boldsymbol{x} := \left[\sum_{k=1}^{p} x_k w_k, \sum_{k=1}^{p} x_{k+1} w_k, \cdots, \sum_{k=1}^{p} x_{n-1+k} w_k \right]. \tag{19}$$

We also write down the definition of the general N*-dimensional case using the multi-index* $\boldsymbol{i} = (i_1, \cdots, i_N)$*,* $1 \le i_k \le d_k$*. The convolution operation in* N *dimensions is*

$$[\boldsymbol{w} * \boldsymbol{x}]_{\boldsymbol{i}} = \sum_{\boldsymbol{k}} x_{\boldsymbol{k}+\boldsymbol{i}} w_{\boldsymbol{k}}, \tag{20}$$

where the summation is over $k_j = 1, \ldots, d_j$ *and* $\boldsymbol{k} + \boldsymbol{i} := [k_1 + i_1 - 1, \cdots, k_N + i_N - 1]$*. The* $N = 2$ *case is most relevant to image applications.*

Example 2 (Full permutation symmetry) *In this example, we consider the symmetric group* S*. Some common* S *invariant functions include* $x_1 + x_2 + \cdots + x_n$ *and* $x_1 x_2 \cdots x_n$*.* S *equivariant mapping may be built from these, such as*

$$[x_1, x_2, \cdots, x_n] \mapsto [x_1 + x_2 + \cdots + x_n, \cdots, x_1 + x_2 + \cdots + x_n], \tag{21}$$

or

$$[x_1, x_2, \cdots, x_n] \mapsto [x_1 + x_2 x_3 \cdots x_n, x_2 + x_1 x_3 \cdots x_n, \cdots x_n + x_1 x_2 \cdots x_{n-1}]. \tag{22}$$

Functions respecting full permutation symmetry are often used for the study of physical systems whose attributes are set-like features with no order structures, e.g. a list of constituent atoms in a crystal lattice. These are also called set functions, which features in a variety of recent studies (Zweig and Bruna, 2021; Zaheer et al., 2017; Qi et al., 2017).

The results of this paper apply not only to the aforementioned symmetries, but also other types of partial permutation subgroups that naturally arise in scientific applications. We will discuss this in greater detail in Section 4.

2.3 Universal Approximation Property under Invariance

The approximation setting studied in this paper concerns the universal approximation property (UAP) under symmetry induced by a permutation group $\mathsf{G} \le \mathsf{S}$. The following definition makes this precise.

2. As is customary in the deep learning literature, this definition of convolution is in fact the *cross correlation* and differs from the classical convolution in the order of indices for the filters. Note that such conventions do not affect any approximation results, so we choose to stick to the usual deep learning convention.

Definition 6 (G UAP) *Let $\mathcal{H}$ be a family of G invariant functions from $\mathbb{R}^n \to \mathbb{R}^m$. $\mathcal{H}$ is said to possess the G **universal approximation property** (G UAP) in L^p sense if for any G invariant continuous function $\boldsymbol{F} : \mathbb{R}^n \to \mathbb{R}^m$, compact set $K \subset \mathbb{R}^n$ and $\varepsilon > 0$, there exists $\hat{\boldsymbol{F}} \in \mathcal{H}$ such that*

$$\|\boldsymbol{F} - \hat{\boldsymbol{F}}\|_{L^p(K)} \leq \varepsilon. \tag{23}$$

Similarly, let $\mathcal{A}$ be a family of G equivariant mappings from $\mathbb{R}^n \to \mathbb{R}^n$. $\mathcal{A}$ is said to possess G UAP if for any G equivariant continuous mapping $\boldsymbol{\varphi} : \mathbb{R}^n \to \mathbb{R}^n$, compact set $K \subset \mathbb{R}^n$ and $\varepsilon > 0$, there exists $\hat{\boldsymbol{\varphi}} \in \mathcal{A}$ such that

$$\|\boldsymbol{\varphi} - \hat{\boldsymbol{\varphi}}\|_{L^p(K)} \leq \varepsilon. \tag{24}$$

We highlight that symmetry constraints naturally limit approximation capabilities. More concretely, one can show that if a function $\boldsymbol{F}$ (resp. mapping $\boldsymbol{\varphi}$) can be approximated by G invariant functions (resp. equivariant mapping), then $\boldsymbol{F}$ (resp. $\boldsymbol{\varphi}$) itself is G invariant (resp. equivariant).

Therefore, it is meaningful to study how universal approximation can be obtained under this symmetry-constrained setting. Finally, we note that while we will mostly adopt the continuous setting as introduced in Section 2.1, the universal approximation results obtained can be used to deduce their counter-parts in the practical discrete setting via standard arguments in numerical analysis. We discuss this point in detail in Appendix D.

3. Main Results

In this section, we present our main result on the universal approximation of G invariant functions via dynamics driven by equivariant control families. All proofs are defered to Section A. Concretely, let us fix a transitive subgroup $\mathsf{G} \leq \mathsf{S}$ and consider a target function $\boldsymbol{F} : \mathbb{R}^n \to \mathbb{R}^m$ that is G invariant. For brevity, we will hereafter use invariant (resp. equivariant) to mean G invariant (resp. G equivariant). Our main results in this section gives sufficient conditions on a control family to induce the G universal approximation property. We start with some definitions.

Definition 7 (coor **operator)** *Let G be a transitive subgroup and $\mathcal{F}$ be a collection of equivariant mappings on $\mathbb{R}^n$. We define the* coor *operator*

$$\mathrm{coor}(\mathcal{F}) := \left\{ f_1 : \mathbb{R}^n \to \mathbb{R} \ : \ [f_1 \circ \mathsf{t}_1, \cdots, f_1 \circ \mathsf{t}_n] \in \mathcal{F}, \mathsf{t}_i \in \mathsf{G}, \mathsf{t}_i(1) = i \right\}. \tag{25}$$

The coor operator associates each family of equivariant mappings with a corresponding family of invariant mappings. We give a simple example when G is the full permutation group S. Given $\boldsymbol{x} = [x_1, \cdots, x_n]$, define $\Sigma(\boldsymbol{x}) = \sum_{i=1}^n x_i$. Consider $\mathcal{F} = \{a\boldsymbol{x} + b\Sigma(\boldsymbol{x})\mathbf{1} : a, b \in \mathbb{R}\}$, then we have $\mathrm{coor}(\mathcal{F}) = \{ax_1 + b\Sigma(\boldsymbol{x})\}$. In fact, the following shows that $\mathrm{coor}(\mathcal{F})$ characterizes $\mathcal{F}$, so we may interchangeably refer to either. This enables us to only consider invariant function families when constructing equivariant families of mappings.

Proposition 8 *Let* G *be a transitive groups, and* $\mathsf{H} = \mathrm{Stab}_1(\mathsf{G})$ *be its stabilizer. We then have*

1. *If* $\mathcal{F}$ *is* G *equivariant, then* $\mathrm{coor}(\mathcal{F})$ *is* H *invariant.*
2. *Conversely, suppose that* $\mathcal{G}$ *is* H *invariant. Then there is a unique* $\mathcal{F}$*, such that* $\mathcal{F}$ *is* G *equivariant and* $\mathrm{coor}(\mathcal{F}) = \mathcal{G}$*.*

Let us now introduce a class of *symmetric invariant well functions*, which plays a central role in our analyses of function approximation using composition or dynamics. We have recalled the definition of *well function* introduced in Li et al. (2023) to prove approximation results without symmetry constraints in Section 2.1, Definition 4. Here, we will modify the notion of well functions to incorporate symmetry considerations.

Definition 9 (Symmetric Invariant Well Functions) *Let* $\tau : \mathbb{R}^n \to \mathbb{R}$ *be a* n*-dimensional Lipschitz function. We call it a symmetric invariant well function if the following conditions hold:*

1. τ *is* S *invariant.*
2. *There exists a finite interval* $\mathbb{I} \subset \mathbb{R}$ *such that if* $\boldsymbol{x} \in \mathbb{I}^n$*, then* $\tau(\boldsymbol{x}) = 0$*.*
3. *Given* $i \in \{n\}$ *and* $b > 0$*, there exists* $a > 0$ *such that* $\tau([x_1, \cdots, x_n]) \neq 0$ *for all* $|x_i| > a$ *and* $|x_j| < b$ *for* $j \neq i$*.*

Remark 10 *It is easy to verify that* $[x_1, \cdots, x_n] \mapsto h(x_1 + x_2 + \cdots + x_n)$ *and* $[x_1, \cdots, x_n] \mapsto h(x_1) + h(x_2) + \cdots + h(x_n)$ *are both symmetric invariant well functions, provided that* h *is a well function in the sense of Definition 4.*

We note that this definition is close to, but more general than just requiring a well function (in the sense of Definition 4) to be S *invariant. An* S *invariant well function is also a symmetric invariant well function, but the converse does not hold in general. For example, for any one-dimensional well function* h*, we consider* $h(x_1 + \cdots + x_n)$ *again. This is a symmetric invariant well function in the sense of Definition 9, but it is not a well function since its zero set is unbounded. The current broader definition allows for easier application of our results to practical architectures.*

Intuitively, if a control family $\mathcal{F}$ contains a symmetric invariant well function, then $\mathcal{F}$ have some nonlinearity to approximate a broad family of nonlinear functions. The existence of well functions ensures that we can construct flow maps that can arrange points at will: the zero set (see propery 2 of Definition 9) keeps points in it stationary, whereas outside of this set points can be transported as desired by exploiting some form of affine invariance. The symmetry conditions ensures that the desired equivariance is preserved under these transformations. Additionally, it turns out that to approximate equivariant mappings, the control family should have a minimal resolution in the sense that it can distinguish points in different orbits. We will need some technical requirements in this direction as outlined below. For a function $u : \mathbb{R} \to \mathbb{R}$, we define the *coordinate zooming function* $u^{\otimes} : \mathbb{R}^n \to \mathbb{R}^n$ by $[u^{\otimes}]_i(\boldsymbol{x}) = u(\boldsymbol{x})$ for all $\boldsymbol{x}$.

Definition 11 (Perturbation Property) *Define the* similarity *of two points*

$$\overline{s}(\boldsymbol{x},\boldsymbol{y}) = |\{i : x_i = y_i\}| \tag{26}$$

if at least one of them is in general position. Otherwise, the similarity is defined as 0. We say that $\mathcal{F}$ satisfies the perturbation property *if for any $\boldsymbol{x},\boldsymbol{y}$ with $\overline{s}(\boldsymbol{x},\boldsymbol{y}) \neq 0$, there exists $\boldsymbol{f} \in \mathcal{F}$ and a coordinate zooming function $u^{\otimes}$, such that $x_i = y_i$ for some i, but $[\boldsymbol{f}(u^{\otimes}(\boldsymbol{x}))]_i \neq [\boldsymbol{f}(u^{\otimes}(\boldsymbol{y}))]_i$.*

For approximating functions with full permutation invariance, the perturbation property gives sufficient resolution. In the most general case, one requires a little more.

Definition 12 (Direct Connectivity) *We say two permutation a and b are directly $\mathcal{F}$ connected if $\mathsf{a} = (i\ j)\mathsf{b}$ for some $i,j \in \{n\}$, and there exists a $\boldsymbol{z} \in \partial Q_{\mathsf{a}} \cap \partial Q_{\mathsf{b}}$, and $\boldsymbol{f} \in \mathcal{F}$, such that $[\boldsymbol{f}(\boldsymbol{z})]_i \neq [\boldsymbol{f}(\boldsymbol{z})]_j$. We say two permutation a and b are $\mathcal{F}$ connected if there exists $\mathsf{g}_0 = a, \mathsf{g}_1, \cdots, \mathsf{g}_s = \mathsf{b}$, where g_i and g_{i-1} are directly $\mathcal{F}$ connected.*

Definition 13 (Resolving a Group) *We say $\mathcal{F}$ resolves G if $\mathcal{F}$ satisfies the perturbation property (Definition 11), and moreover, it is G transversally transitive: i.e., there exists a transversal A such any two distinct elements $\mathsf{a},\mathsf{b} \in A$ are $\mathcal{F}$ connected.*

Now, we are ready to state our main approximation result that applies to any transitive $\mathsf{G} \leq \mathsf{S}$.

Theorem 14 *Let $\mathsf{G} \leq \mathsf{S}$ be transitive, $n \geq 2$ and $\boldsymbol{F} : \mathbb{R}^n \to \mathbb{R}^m$ be continuous and G invariant. Suppose that the control family $\mathcal{F}$ is G equivariant and resolves G, while the terminal family $\mathcal{G}$ is G invariant, satisfying the following conditions:*

1. *For any compact $K \subset \mathbb{R}^n$, there exists a Lipschitz $\boldsymbol{g} \in \mathcal{G}$ such that $\boldsymbol{F}(K) \subset \boldsymbol{g}(\mathbb{R}^n)$.*
2. *$\mathcal{F}$ is scaling and translation invariant along $\mathbf{1}$, i.e., $\boldsymbol{f} \in \mathcal{F}$ implies $a\boldsymbol{f}(b\boldsymbol{x}+c\mathbf{1}) \in \mathcal{F}$ for any $a,b,c \in \mathbb{R}$.*
3. *$\overline{\mathrm{CH}}(\mathrm{coor}(\mathcal{F}))$ contains both a symmetric invariant well function τ, and a function $\boldsymbol{x} \mapsto h(x_1)$, where h is a one-dimensional well function.*

Then, $\mathcal{H}_{ode}(\mathcal{F},\mathcal{G})$ satisfies the G UAP. That is, for any $p \in [1,\infty)$, compact $K \subset \mathbb{R}^n$ and $\varepsilon > 0$, there exists a G invariant $\hat{\boldsymbol{F}} \in \mathcal{H}_{ode}(\mathcal{F},\mathcal{G})$ such that

$$\|\boldsymbol{F} - \hat{\boldsymbol{F}}\|_{L^p(K)} \leq \varepsilon. \tag{27}$$

Let us briefly sketch the main idea of the proof. The basic framework for approximating a invariant function F by composition is to first fix an invariant function g whose range covers that of F. Then, it remains to construct an equivariant mapping $\varphi : \mathbb{R}^n \to \mathbb{R}^n$ so that $F \approx g \circ \varphi$. One can show that under fairly general conditions, the existence of φ is guaranteed (Proposition 15).

Thus, it remains to determine sufficient conditions on the control family $\mathcal{F}$ such that its attainable set $A_{\mathcal{F}}$ can approximate arbitrary equivariant mappings. As in Li et al. (2023), one can reduce this problem to matching an arbitrary finite point set in $\mathbb{R}^n$ to another using flows of the dynamical system. There, the crucial property of well functions enables this: the zero set of the well function can keep certain points fixed, whereas the non-zero part can move points. This then achieves the required point matching property through repeated composition of carefully constructed flow maps.

The key difference when extending this argument to the symmetric setting is that the induced motion on each coordinate component of a point is no-longer independent. For example, if $\mathsf{G} = \mathsf{S}$, then all coordinates of each point will experience the same transformation. Hence, the requirement for arbitrary point set matching must be suitably relaxed. It turns out that we only need to be able to match point sets consisting of points belong to distinct G orbits, greatly relaxing the type of constructions required. This is because, within the same orbit, we can use the invariant property of g and the equivariant property of $\mathcal{F}$ to transport the point to the desired location. Theorem 18 establishes such conditions that are sufficient to induce universal approximation under symmetry settings.

We emphasize that Theorem 14 is not a strict generalization of Theorem 5. While we may take $\mathsf{G} = \{\mathsf{e}\}$ to remove the symmetry constraints, the trivial subgroup $\{\mathsf{e}\}$ is not transitive, hence the result here do not apply. In the proof of this theorem, we in fact prove a more general result regarding the approximation of invariant functions by composing families of equivariant functions (Theorem 18). This result applies broadly to compositional hypothesis spaces and is not limited to ODE-type dynamical systems. In fact, it subsumes both Theorem 5 and Theorem 14, and may be viewed as an abstract sufficient condition for universal approximation through composition. We defer the detailed discussion of this result to Section A.2.

In the application of these results, one only need to check whether a network architecture of interest satisfies the above four conditions. Conditions 1 and 2 are easy to check. The other conditions require additional effort, but we will show in Section 4 that both known popular and useful novel architecture types are included in our results, and their verification are presented in Appendix C. Our main result applies to continuous-time dynamics, which corresponds to a continuum idealization of practical deep neural networks (E, 2017; Haber and Ruthotto, 2017; Li et al., 2018). We show in Appendix. D that the approximation properties can be naturally passed to discrete residual structures used in practice.

Necessity of some conditions in Theorem 14. We close this subsection by discussing the necessity of some conditions given in Theorem 14. In particular, we explain why we require at least two types of well functions — a symmetric invariant one and a one-dimensional ($\boldsymbol{x} \mapsto h(x_1)$) one — in Condition 3 of Theorem 14. Consider

$$\dot{\boldsymbol{x}} = \gamma(\boldsymbol{x})\mathbf{1}, \tag{28}$$

where γ is an arbitrary scalar function. In this case, $\mathcal{F} := \{\gamma(\boldsymbol{x})\mathbf{1} : \gamma \in C(\mathbb{R}^n)\}$. Check that $\mathcal{F}$ satisfies Condition 2 in Theorem 14 and $\overline{\mathrm{CH}}(\mathrm{coor}(\mathcal{F}))$ contains a symmetric well

function, but not $\boldsymbol{x} \mapsto h(x_1)$. Observe that any flow map $\boldsymbol{\varphi} \in \mathcal{A}_{\mathcal{F}}$ will satisfy $[\boldsymbol{\varphi}(\boldsymbol{x})]_i - [\boldsymbol{\varphi}(\boldsymbol{x})]_j = x_i - x_j$. As a result, $\mathcal{A}_{\mathcal{F}}$ cannot approximate every S equivariant function. Further, consider the terminal family as a single scalar function $g(\boldsymbol{x}) = \max x_i - \min x_i$, then for all $\boldsymbol{F} \in \mathcal{H}_{\text{ode}}(\mathcal{F}, \{\boldsymbol{g}\})$, it holds that $F(\boldsymbol{x}) = g(\boldsymbol{x})$. This shows the approximation property of invariant functions is therefore limited. Theorem 5 assumes a stronger affine invariance condition than Theorem 14. Thus, while γ in (28) can be taken as a well function, the associated control family $\mathcal{F}$ does not satisfy the restricted affine invariance property in Theorem 5.

On the other hand, if we only consider coordinate-wise well functions, then the following dynamics driven by a coordinate zooming function

$$\dot{\boldsymbol{x}} = u^{\otimes}(\boldsymbol{x}), \tag{29}$$

can be constructed to satisfy all other conditions of Theorem 14. In this case, $\mathcal{A}_{\mathcal{F}}$ only consists of coordinate-wise mappings, and hence cannot approximate every S equivariant function for $n \geq 2$. For instance, if we choose the terminal family as a single function $g(\boldsymbol{x}) = \max x_i$, then we can conclude that $\mathcal{H}_{\text{ode}}(\mathcal{F}, \{\boldsymbol{g}\})$ cannot approximate all S invariant functions.

Next, we consider the translation group $\mathsf{G} = \mathsf{T}$, the singleton terminal family $\mathcal{G} = \{g(\boldsymbol{x}) = x_1 + x_2 + x_3\}$, and the control family

$$\mathcal{F} = \{[x_1, x_2, x_3] \mapsto [ah(bs + cx_1 + d), ah(bs + cx_2 + d), ah(bs + cx_3 + d)] : a, b, c, d \in \mathbb{R}\}. \tag{30}$$

It is easy to see Condition 2 in Theorem 14 holds for $\mathcal{F}$. To verify Condition 3, we notice that, in this case

$$\text{coor}(\mathcal{F}) = \{ah(b(x_1 + x_2 + x_3) + c(x_1) + d)\}. \tag{31}$$

Clearly, $h(x_1) + h(-1 - x_1) \in \overline{\text{CH}}(\text{coor}(\mathcal{F}))$ is a one-dimensional well function. Further, it follows from Remark 10 that $h(x_1 + x_2 + x_3)$ is a symmetric invariant well function.

Therefore, Conditions 1-3 are satisfied for this control system. However, this control family can only produce S equivariant mappings. By Proposition 36, it suffices to construct a function which is T invariant but not S invariant. We provide a concrete example:

$$f(\boldsymbol{x}) = (x_1 - x_2)(x_2 - x_3)(x_3 - x_1), \tag{32}$$

which is T equivariant, but not S equivariant. This function cannot be approximated by $\mathcal{H}_{\text{ode}}(\mathcal{F}, \mathcal{G})$ defined above.

In this specific case, the failure to approximate (32) can be explained as follows. Recall the definition of cross section in (15). Observe that any flow map generated from $\mathcal{F}$ maps Q_{a} into itself. In other words, the flow does not have enough resolution to steer points across different cross sections. However, this motion is necessary if we want to achieve approximation when $\mathsf{G} \neq \mathsf{S}$.

Moreover, we remark that the condition of resolution is also necessary for the case that the output dimension equals to the input dimension and the terminal function

is the identity mapping. Suppose that for a transversal A, there exist $\mathtt{a}, \mathtt{b} \in A$ such that they are not $\mathcal{F}$ connected. Since $\mathtt{a}$ and $\mathtt{b}$ are not $\mathcal{F}$ connected, it then implies for $\boldsymbol{\varphi} \in \mathcal{A}_{\mathcal{F}}$, $\boldsymbol{\varphi}(Q_{\mathtt{a}}) \not\subseteq Q_{\mathtt{b}}$. We choose a smooth function $\boldsymbol{\xi} \in C_c^\infty(Q_{\mathtt{b}})$ such that there exists an interior domain $\Omega_{\mathtt{a}} \subseteq Q_{\mathtt{a}}$ such that $\boldsymbol{\xi}(\boldsymbol{x}) = \mathtt{b}\mathtt{a}^{-1}\boldsymbol{x}$ for $\boldsymbol{x} \in \Omega$. We then extend this function to the whole of $\mathbb{R}^n$ while still being equivariant. Then, $\boldsymbol{\xi}(\Omega_{\mathtt{a}}) \subset Q_{\mathtt{b}}$. This means $\boldsymbol{\varphi}$ cannot approximate $\boldsymbol{\xi}$, which leads to a contradiction.

Finally, we remark that the requirement that $\mathcal{F}$ resolves G may exclude certain cases. For example, if G is the alternating group, then no two elements in the transversal are directly connected by any $\mathcal{F}$ (see Definition 12), so no $\mathcal{F}$ can resolve it. This may be consistent with the observed difficulty in approximating functions that are invariant with respect to the alternating group (Yarotsky, 2022).

3.1 Sketch of the approximation framework and proof ideas

In this subsection, we summarize the proof ideas of our main theorem, i.e., Theorem 14. The complete proof is contained in Appendix A. Our method establishes a basic framework for the approximation theory of invariant or equivariant functions via composition or dynamics. Hereafter, we fix $\mathsf{G} \leq \mathsf{S}$ as a transitive subgroup, and G invariance (resp. G equivariance) will be shortened as invariance (resp. equivariance). For simplicity, we only consider the case $m = 1$. Namely, we only consider the scalar regression setting where the approximation target is a function $F : \mathbb{R}^n \to \mathbb{R}$. The case $m > 1$ follows similarly as in the construction in Li et al. (2023, Proposition 3.8), and is hence omitted for simplicity.

Our proof follows the following principle, which indicates that approximating an invariant function can be decomposed into two steps:

1. Choose a simple invariant function $g : \mathbb{R}^n \to \mathbb{R}$, for example, $g(\boldsymbol{x}) = \sum_{i=1}^n x_i$.

2. Construct an equivariant hypothesis set $\mathcal{A}$ consisting of equivariant mappings on $\mathbb{R}^n$ to itself with enough complexity.

Then, the set of functions $\{g \circ \boldsymbol{\varphi} : \boldsymbol{\varphi} \in \mathcal{A}\}$ can serve as universal approximators for invariant functions. In the context of neural networks, g will be the final layer, whereas $\boldsymbol{\varphi} \in \mathcal{A}$ consists of a stack of intermediate layers. Theorem 25 to be presented later will deal with how the required equivariant function $\boldsymbol{\varphi}$ can be constructed from compositions.

Proposition 15 stated below establishes that this two-step decomposition scheme is sufficient for approximation. In fact, this proposition reduces the problem to studying the approximation property of equivariant mappings.

Proposition 15 (Approximating Equivariant Mapping is Sufficient) *Suppose $F \in C(\mathbb{R}^n)$ is an invariant function. Let g be a Lipschitz continuous and invariant function with $F(\mathbb{R}^n) \subset g(\mathbb{R}^n)$. Then, for any compact $K \subset \mathbb{R}^n$ and $\varepsilon > 0$, there exists an equivariant mapping $\boldsymbol{\varphi} \in C(\mathbb{R}^n; \mathbb{R}^n)$ such that*

$$\|F - g \circ \boldsymbol{\varphi}\|_{L^p(K)} \leq \varepsilon. \tag{33}$$

Next, we discuss how to achieve universal approximation of G equivariant mappings through composition. We begin with the following definitions, which can be regarded as a summary and generalization of those introduced in Li et al. (2023). Roughly speaking, our strategy is to show that under mild conditions, the universal approximation property can be reduced to transporting a finite point set X to another finite point set Y, with X and Y having the same cardinality. However, this result does not hold if no additional requirement is imposed on X and Y. For example, for some $\mathsf{g} \in \mathsf{G}$, if $\mathsf{g}(\boldsymbol{x}) = \boldsymbol{x}'$, then the $\boldsymbol{f}(\boldsymbol{x}') = \mathsf{g}(\boldsymbol{f}(\boldsymbol{x}))$ provided that $\boldsymbol{f}$ is G equivariant, imposing a constraint on Y.

To this end, we introduce some notions on the position of points in a point set.

Definition 16 (General Position) *We say a vector $\boldsymbol{x}$ is in* general position *if all of its coordinates are distinct.*

Moreover, we introduce the concept of G *distinctness*, restricting the position of the finite point set X to obey such constraints.

Definition 17 (G Distinctness) *A point set $X = \{\boldsymbol{x}^1, \cdots, \boldsymbol{x}^M\}$ is called* G distinct*, if $\mathsf{g}(\boldsymbol{x}^i) = \boldsymbol{x}^j$ implies $\mathsf{g} = \mathsf{e}$ and $i = j$, with e being the identity element. In other words, a point set is G distinct if and only if the G orbit of the points are distinct.*

With these definitions in mind, our approach first proves the theorem below concerning a general compositional approximation under symmetry conditions, which gives general sufficient conditions for building a complex hypothesis spaces out of potentially simple ones through function composition. Recall that Q_A is defined in (15), and $\overline{Q_A}$ is the closure of Q_A.

Theorem 18 (Universal Approximation of Invariant Functions via Composition) *Let $F \in C(\mathbb{R}^n)$ be invariant. Suppose $\mathcal{G}$ is a family of invariant functions and $\mathcal{A}$ is a family of equivariant Lipschitz mappings with the following properties:*

1. *For any compact $K \subset \mathbb{R}^n$, there exists a Lipschitz $g \in \mathcal{G}$ such that $F(K) \subset g(\mathbb{R}^n)$.*
2. *$\mathcal{A}$ is closed under composition, i.e., if $\boldsymbol{f_1} \in \mathcal{A}$ and $\boldsymbol{f_2} \in \mathcal{A}$, then $\boldsymbol{f_1} \circ \boldsymbol{f_2} \in \mathcal{A}$.*
3. *(Coordinate zooming) For any increasing function $v \in C(\mathbb{R})$, compact interval $\mathbb{I} \subset \mathbb{R}$ and tolerance $\varepsilon > 0$, there exists $u : \mathbb{R} \to \mathbb{R}$ such that $u^{\otimes} \in \mathcal{A}$ and $\|u - v\|_{C(\mathbb{I})} \leq \varepsilon$.*
4. *(Point matching)*

   For $M > 0$, a transversal A of G, a G distinct point set $\{\boldsymbol{x}^1, \cdots, \boldsymbol{x}^M\}$ with $\boldsymbol{x}^i \in \overline{Q_A}$, another point set $\{\boldsymbol{y}^1, \cdots, \boldsymbol{y}^M\} \subset Q_A$ and tolerance $\varepsilon > 0$, there exists $\boldsymbol{f} \in \mathcal{A}$ such that

   (a) *For $\boldsymbol{x}^i$ in general position, we have $|\boldsymbol{f}(\boldsymbol{x}^i) - \boldsymbol{y}^i| \leq \varepsilon$.*

   (b) *For $\boldsymbol{x}^i$ not in general position, we have $|\boldsymbol{f}(\boldsymbol{x}^i)| \leq 1$.*

Then, for any compact set $K \subset \mathbb{R}^n$, tolerance $\varepsilon > 0$ and $p \in [1, \infty)$, there exists $\boldsymbol{\varphi} \in \mathcal{A}$ and $g \in \mathcal{G}$ such that $\|F - g \circ \boldsymbol{\varphi}\|_{L^p(K)} \leq \varepsilon$.

Theorem 18 tells us that the coordinate zooming property and point matching property are sufficient for universal approximation. Then, it suffices to show that under the assumption of Theorem 14, we can achieve the coordinate zooming property and point matching property. The role of composition is to make sure that the construction can be divided into that of the coordinate zooming mapping and a mapping with point matching property. Subsequently, we show that each of these can then be realized using constructions of appropriate dynamics – a continuum version of composition. This is the bulk of the technical parts of the proof, and the key steps are laid out in Appendix A. Note that here we only assume all mappings in $\mathcal{A}$ are Lipschitz, but this does not imply $\sup_{\boldsymbol{\phi} \in \mathcal{A}} \mathrm{Lip}(\boldsymbol{\phi}) < \infty$.

Remark 19 *Theorem 18 can be related in spirit to a part of the Stone–Weierstrass theorem. Both this result and the Stone–Weierstrass theorem provide sufficient conditions for universal approximation. Our results deal with compositional hypothesis spaces whereas the Stone-Weierstrass theorem applies to unital algebras. However, the Stone–Weierstrass theorem also provides necessary conditions for density, but this is not the case for our current results. Moreover, the point matching property is reminiscent of the point separation property in the Stone–Weierstrass theorem. Specifically, the point matching property enables the construction of piecewise constant approximants, while the point separation property enables the construction of appropriate "polynomials" in the algebra that can approximate a target continuous function from above and below.*

4. Applications and Discussions

In this section, we demonstrate how our proposed framework can be applied to obtain universal approximation results of a variety of deep learning architectures designed to capture or preserve symmetry. Some of these, such as the convolutional neural network, have been subjects under intense study. On the other hand, our framework can also be applied to study novel architectures for emerging machine learning applications.

Before diving into application cases, we first discuss some advantages of our theoretical approach. Unlike many other results on approximation theory of symmetric functions by neural networks (e.g. Yarotsky (2022)), the results here do not rely on the specification of an explicit network architecture, much like the flavor of those in Li et al. (2023). In fact, we show that a variety of different architectures satisfy the assumptions in our theory. Second, more symmetry scenarios can be handled under our theory, since the only assumption we made about the symmetry group G is transitivity. Besides shift invariance (convolutional networks) and full permutation invariance (set function networks), we can also study other useful symmetry structures, such as the product permutation invariant structures (See Section 4).

The main message in this paper is that the deep neural networks can be powerful from merely increasing its depth. This effect is studied by several papers without symmetry

considerations, e.g. Li et al. (2023); Cuchiero et al. (2020); Ruiz-Balet and Zuazua (2021); Tabuada and Gharesifard (2022); Shen et al. (2019, 2021b). However, for the symmetry-constrained case, this is highly non-trivial and not well-understood in the literature. For example, Yarotsky (2022, Section 2) proves that universal approximation can be achieved by a representation of symmetric polynomials, which corresponds to those neural networks with sufficient large width. If we do not limit the width, then the approximation result can be achieved by averaging techniques, (Yarotsky, 2022, Section 2). For example, if we have a neural network f_{NN} that approximates a G invariant function f on some cross section, but may not satisfy G invariance itself. Then, we perform the averaging transformation

$$f_{NN}^{avr}(\boldsymbol{x}) = \sum_{\mathsf{g}\in\mathsf{G}} f_{NN}(\mathsf{g}\boldsymbol{x}),$$

which yields a G invariant function that approximates f. To represent this averaged structure, however, one generally need to increase the width requirement by $|\mathsf{G}|$ times. Another way to achieve equivariance is by canonicalization, which is discussed by Kaba et al. (2023). Here, the canonicalization function $\mathsf{h} : \mathbb{R}^n \to \mathsf{G}$ satisfies G equivariance, namely, $\mathsf{h}(\mathsf{g}\boldsymbol{x}) = \mathsf{g}\mathsf{h}(\boldsymbol{x})$ for all $\mathsf{g} \in \mathsf{G}$. They proved that $\boldsymbol{\phi}(\boldsymbol{x}) = \mathsf{h}(\boldsymbol{x})\boldsymbol{f}(\mathsf{h}(\boldsymbol{x})^{-1}\boldsymbol{x})$ is also G equivariant. The equivariance condition on h can be relaxed if some other equivariance condition is imposed in $\boldsymbol{f}$, see Kaba et al. (2023). The key to canonicalization is to construct a suitable h, G valued and equivariant, which is also the main part of Kaba et al. (2023). In fact, if the function h is constructed, then any universal approximator can be modified via canonicalization. However, it would be rather challenging to construct h in general situations, which is also stressed in their paper.

In summary, the above methods (group averaging, canonicalization) work by imposing symmetry on (usually larger) hypothesis spaces already having the universal approximation property. On the contrary, the result herein establish sufficient conditions for universal approximation from a hypothesis space which already possesses the symmetry. Moreover, the result established herein ensure that the architecture has simple structure (limited width), and consequently the approximation power comes from the depth of neural networks. This is also convenient in practical implementations, where the simplicity of each layer (without having to average over a group) is desirable. Hence, the results here cannot be deduced from those for wide neural networks.

Throughout this section, σ denotes an activation function. The only assumption on σ is that $\overline{\mathrm{CH}}(\{w\sigma(a\cdot+b) : w, a, b, \in \mathbb{R}\})$ contains a one-dimensional well function h. This is verified in Li et al. (2023) that common activation functions like ReLu, Sigmoid and Tanh, satisfy this assumption. For convenience, we will also choose a function family $\mathcal{G}$ such that the covering Condition 1 in Theorem 14 is satisfied: $\mathcal{G}$ is a G invariant function family such that for any compact set $K \subset \mathbb{R}^m$, there exists $\boldsymbol{g} \in G$ such that $K \subset \boldsymbol{g}(\mathbb{R}^n)$. For example, for scalar regression problems ($m = 1$) we may choose $\mathcal{G} = \{g\}$ where $g(\boldsymbol{x}) = \sum_{i=1}^n x_i$ to satisfy both invariance and the range covering condition. Therefore, it suffices to check the perturbation assumption and G transversal transitivity on each case, to show that $\mathcal{F}$ resolves G (see Definition 13).

4.1 Shift Invariant Architectures

We first discuss the approximation of shift invariant functions by dynamical systems driven by shift equivariant control families. A prime example of such architectures is the residual convolutional neural network. We will show that our results immediately lead to universal approximation results for shift-invariant functions using certain types of convolutional neural networks.

We recall the definition of the translation group in one or more dimensions. The one-dimensional case is relevant to sequence modelling applications (e.g. Oord et al. (2016)) whereas the two/three dimensional cases apply to image and video analysis tasks. In one dimension, the translation group is generated by the shift operator t such that $\mathsf{t}(i) = i + 1$. We then define $\mathsf{T} = \mathsf{T}_{1D} = \{\mathsf{t}, \mathsf{t}^2, \cdots, \mathsf{t}^n\}$. In two dimensions, we consider the case where the n points are arranged in a 2D grid of sizes n_1, n_2 with $n = n_1 n_2$. In image applications, each coordinate in this grid refers to a pixel in an image. We may re-index the coordinates by the multi-index (i_1, i_2), where $i_1 = 1, 2, \cdots, n_1$ and $i_2 = 1, 2, \cdots, n_2$. The two dimensional translation group T_{2D} is generated by two shift operators:

$$\begin{aligned} \mathsf{t}_1((i_1, i_2)) = (i_1 + 1, i_2) \qquad \mathsf{t}_2((i_1, i_2)) = (i_1, i_2 + 1) \\ \mathsf{T}_{2D} := \{\mathsf{t}_1^{j_1} \circ \mathsf{t}_2^{j_2} : j_1 = 1, 2, \cdots, n_1, \quad j_2 = 1, 2, \cdots, n_2\}. \end{aligned} \tag{34}$$

The higher dimensional cases T_{kD} follow similarly.

In one dimension where an input is a sequence represented as a vector, the simplest continuous idealization of a convolutional neural network can be expressed as

$$\dot{\boldsymbol{x}}(t) = v(t)\sigma(\boldsymbol{w}(t) * \boldsymbol{x}(t) + b(t)\mathbf{1}), \tag{35}$$

where for each t, $v(t), b(t)$ are scalars, $\boldsymbol{w}(t), \boldsymbol{x}(t), \mathbf{1}$ are vectors and $*$ is the discrete convolution operation, as defined in (19) and (20).

In the general, high dimensional case, we define two variants of the idealization of residual convolution networks

$$\mathcal{F}_{*,1} := \{v\sigma(\boldsymbol{w} * \boldsymbol{x} + b\mathbf{1}) \ : \ \boldsymbol{w} \in \mathbb{R}^{d_1 \times d_2 \times \cdots d_n}, v, b \in \mathbb{R}\}, \tag{36}$$

and

$$\mathcal{F}_{*,2} := \{\boldsymbol{w} * \sigma(\boldsymbol{x} + b\mathbf{1}) \ : \ \boldsymbol{w} \in \mathbb{R}^{d_1 \times d_2 \times \cdots d_n}, b \in \mathbb{R}\}. \tag{37}$$

Note that these control families define a class of convolution-driven dynamics where the size of the convolutional filters is equal to the size of the input signal.

Now, we may apply Theorem 14 to obtain the following result.

Corollary 20 *$\mathcal{H}_{ode}(\mathcal{F}_{*,1}, \mathcal{G})$ and $\mathcal{H}_{ode}(\mathcal{F}_{*,2}, \mathcal{G})$ possesses the T_{kd} UAP.*

Proof For the perturbation assumption, without loss of generality, we assume that $x_1 = y_1$. It suffices to notice that for two point $\boldsymbol{x}$ and $\boldsymbol{y}$ such that they are T distinct, there exists i such that $x_i \neq y_i$. We choose $\boldsymbol{w} = \boldsymbol{e}_i$, whose value is 1 in the i coordinate,

and 0 in other coordinates. Then, we choose an appropriate b such that $\sigma(x_i + b) = 0$ while $\sigma(y_i + b) \neq 0$, therefore

$$[\boldsymbol{w} * \sigma(\boldsymbol{x} + b\mathbf{1})]_1 \neq [\boldsymbol{w} * \sigma(\boldsymbol{y} + b\mathbf{1})]_1 \tag{38}$$

and

$$[\sigma(\boldsymbol{w} * \boldsymbol{x} + b\mathbf{1})]_1 \neq [\sigma(\boldsymbol{w} * \boldsymbol{y} + b\mathbf{1})]_1. \tag{39}$$

Hence we show both $\mathcal{F}_{*,i}$ satisfy the perturbation assumption.

For transversal transitivity, we show that for each $\mathsf{a}, \mathsf{b} \in \mathsf{S}$, a and b are $\mathcal{F}$ directly connected. For simplicity, we consider the case when

$$Q_{\mathsf{a}} = \{x_1 > x_2 > \cdots > x_n\} \tag{40}$$

and

$$Q_{\mathsf{b}} = \{x_2 > x_1 > \cdots > x_n\}, \tag{41}$$

and the other cases are similar. In this case, we can choose $\boldsymbol{z} \in \partial Q_{\mathsf{a}} \cap \partial Q_{\mathsf{b}}$ such that $z_2 \neq z_3$, We choose $\boldsymbol{w} = \boldsymbol{e}_2$, whose value is 1 in second coordinate, and 0 in other coordinates. Then, by a similar approach as what we did in the proof of Lemma 31, we can prove that a and b are $\mathcal{F}$ directly connected. ■

The approximation ability of convolutional neural networks has been widely studied in the literature (Yang et al., 2022; Bolcskei et al., 2019; Bao et al., 2019; Zhou, 2020; Yarotsky, 2018, 2022; Petersen and Voigtlaender, 2020). Thus, we should compare these with the results presented here. First, we note that the shift invariance we discussed is in the index set. In contrast, Yang et al. (2022) uses a (non-symmetric) ReLU feed-forward neural network to approximate shift-invariance mapping in n dimensions via a wavelet-like construction. Also, Yarotsky (2022, Section 3) considered the limit of deep convolutional networks and their ability to approximate all translation equivariant functions. Both works consider the continuum limit: the target function is an equivariant function with respect to the continuous translation group defined on a continuous domain (say $\mathbb{R}^2$), and the approximation sequence contains functions defined on increasingly fine discretization of this domain, each respecting discrete translation symmetry in the sense as defined in this paper, but for a differently-sized group (corresponding to the number of discretization points). Thus, the approximation results are established in the asymptotic sense, when the discretization becomes finer and finer. In contrast, here we work directly with a fixed discrete grid and consider the approximation of equivariant functions by flows, both defined on this grid and respecting the same translation group on this grid. Hence, the only limit we consider is in the depth (time) direction.

Second, in each layer we use a periodic boundary condition so that the analyzed architecture is exactly shift-invariant. If zero boundary condition is adopted, then the architecture will not be strictly shift-invariant. As a consequence, the boundary condition will deteriorate the interior symmetry structure when the network is deep enough. For example, while Zhou (2020); Oono and Suzuki (2019); Bao et al. (2019) studied

approximation properties of convolutional neural networks, due to the aforementioned boundary conditions the approximation results are established in the absence of symmetry requirements. Third and most importantly, our approximation results do not require a specific architecture, and the restriction on the form of the controlled dynamics is minimal.

It is important to emphasize that the convolutional control families we consider here have filters sizes equal to the input or hidden state sizes, while requiring only 1 channel. Thus, this is different from many of the aforementioned settings where small filters or multiple channels may be considered. The reason we apply our results to this simple setting is because we focus on symmetry considerations, and these are in a sense minimal constructions of convolutional networks that can achieve universal approximation of shift-invariant function. However, the idea here can be generalized to filters with variable lengths. These require additional technical arguments and will be addressed in future work.

Lastly, we note that Petersen and Voigtlaender (2020) built a connection between fully connected neural networks and convolutional neural networks in order to translate approximation results for one to the other. There, the authors exploit a suitable transformation representative, which corresponds to the coor operator we introduced in this paper. In this sense, Theorem 14 extends the notion of representatives to deduce universal approximations under more general types of symmetries, including but not limited to shift symmetry. Most previous works on the approximation theory of CNNs focus on non-residual convolutional neural networks. To the best of our knowledge, Corollary 20 gives a first guarantee of universal approximation of shift-invariant functions using continuous-time (residual) CNNs.

4.2 Full Permutation Invariant Architectures

Besides the well-known convolutional neural networks, we demonstrate that our framework can be used to derive universal approximation results under other symmetry settings. Here, we focus on functions possessing full permutation invariance, i.e., those invariant to arbitrary rearrangement of the coordinates of their input vectors. These are also known as "set functions", since their outputs only depend on the collection of input coordinate values but not on their order of arrangement. Learning functions that possess such invariance structures is important in many applications, including population statistics, anomaly detection, cosmology (Zaheer et al., 2017). One can see that this invariance requirement is strong, and thus structures obeying such invariance tend to be limited in flexibility in approximation. As motivated earlier, composition or dynamics provides a way to expand complexity of hypothesis spaces while respecting possibly very restrictive invariances. Our goal is to prove a general sufficient result for universal approximation of full permutation invariant functions through dynamics, thereby providing a guidance to practical construction of neural network architectures respecting these symmetries.

In our framework, the symmetry group under consideration is the full permutation group, i.e., $\mathsf{G} = \mathsf{S}$. Recall that in this case, we only need to construct a control family

that verify Conditions 2 and 3 in Theorem 14. One can show that any of the following choices are sufficient:

$$\mathcal{F}_s := \{a\sigma(w\boldsymbol{x} + v\Sigma\boldsymbol{x} + b\mathbf{1}) : a, w, v, b \in \mathbb{R}\}, \tag{42}$$

$$\mathcal{F}_s := \{a\sigma(w\boldsymbol{x} + c\mathbf{1}) + b\Sigma(\sigma(v\boldsymbol{x} + d\mathbf{1})) : a, b, w, v, d \in \mathbb{R}\}, \tag{43}$$

where Σ is a matrix of all 1s, therefore $\Sigma\boldsymbol{x} = (x_1 + \cdots + x_n)\mathbf{1}$.

Most existing work on the universal approximation of permutation invariant functions by neural networks focus on shallow or non-residual architectures (Sannai et al., 2019; Maron et al., 2019; Ravanbakhsh, 2020). Thus, approximation there is achieved by allowing width to be large enough. In contrast, we investigate how to build approximators with symmetry through composition or dynamics. Our method shows that for a fixed width, many constructions of the layer architecture can achieve universal approximation in this manner. We highlight that our theory can handle the permutation invariance generated by both "intrinsic" equivariant layers in the sense discussed in Sannai et al. (2019), or by averaging in Ravanbakhsh (2020); Yarotsky (2022).

Concretely, we show that our results show that many popular permutation invariant architectures can achieve universal approximation at fixed width in its deep, residual form. In fact, the control families (42)-(43) can be viewed as residual versions of the DeepSets (Zaheer et al., 2017) and PointNet (Charles et al., 2017) architectures (see also Germain et al. (2022) for applications to PDEs), and we can verify that they indeed induce uniform approximation through composition. One caveat is that in these works, it is sometimes suggested to use the max operator in place of the Σ operator. Our results do not currently apply to this case. For example, we can show that the variant $\mathcal{F}_{s,\max} := \{v\sigma(a\boldsymbol{x} + b\max(\boldsymbol{x})\mathbf{1} + c\mathbf{1}) : v, a, b, c \in \mathbb{R}\}$, where $\max(\boldsymbol{x}) = \max_i x_i$ does not possess S UAP for arbitrary choice of terminal family satisfying the conditions of Theorem 14. Concretely, we may choose $m = 1$, $\mathcal{G} = \{g\}$, where $g(\boldsymbol{x}) = \max(\boldsymbol{x})$. Then, for any $\boldsymbol{x}, \boldsymbol{y} \in Q$ such that $x_1 > y_1$, one can show that for any $\boldsymbol{\varphi} \in \mathcal{A}_{\mathcal{F}_{s,\max}}$, it holds that $\boldsymbol{\varphi}(\boldsymbol{x}), \boldsymbol{\varphi}(\boldsymbol{y}) \in Q$ and $[\boldsymbol{\varphi}(\boldsymbol{x})]_1 > [\boldsymbol{\varphi}(\boldsymbol{y})]_1$. Therefore, this family cannot approximate functions such as $\min(\boldsymbol{x}) = \min_i x_i$. This argument does not rule out universal approximation for other choices of terminal families.

As a further application of our results, we can show that the Janossy pooling type of variants introduced in Wagstaff et al. (2022) can also induce universal approximation in its residual form. This corresponds to the following control families (order 1 and order 2 Janossy pooling, respectively)

$$\mathcal{F}_{s,Ja,1} := \{v\sigma(a\boldsymbol{x} + b\sum_{i=1}^{n}\phi(x_i) + c\mathbf{1}) : v, a, b, c \in \mathbb{R}\}, \tag{44}$$

$$\mathcal{F}_{s,Ja,2} := \{v\sigma(a\boldsymbol{x} + b\sum_{i=1}^{n}\phi(x_i, x_j) + c\mathbf{1}) : v, a, b, c \in \mathbb{R}\}, \tag{45}$$

where the ϕ in each case is some chosen scalar valued function. We can verify that $\mathcal{F}_{s,Ja,1}$ induces the S UAP if ϕ is Lipschitz and there exists a neighborhood U of $\phi(0)$

such that $\phi^{-1}(U)$ is bounded. In particular, any sigmoid function satisfies this condition. Similarly, $\mathcal{F}_{s,Ja,2}$ induces S UAP if ϕ is symmetric, Lipschitz and $\phi(z,0)$ satisfies the previous condition.

Thus, we arrive at the following corollary that shows that all these architectures in their residual form possesses S UAP, and checking them is a simple application of our results. The verification details are presented in Appendix C.

Corollary 21 *Choosing $\mathcal{F}_s$ as any of* (42)-(45)*, $\mathcal{H}_{ode}(\mathcal{F}_s, \mathcal{G})$ possesses the S UAP.*

Proof Here we only provide the proof of $\mathcal{F}_{s,1}$ and $\mathcal{F}_{s,2}$, and defer that of Janossy pooling to Appendix C. Since any transversal only contains one element, it suffices to show $\mathcal{F}_{s,i}$ satisfies the perturbation assumption. For $\mathcal{F}_{s,1}$, if $\boldsymbol{x}$ and $\boldsymbol{y}$ are S distinct, then there exists a function $u^{\otimes} \in \mathcal{F}$ such that

$$u(x_1) + u(x_2) \cdots + u(x_d) \neq u(y_1) + u(y_2) \cdots + u(y_d). \tag{46}$$

For $\mathcal{F}_{s,1}$, we set $a = 1, w = 0$, and choose a suitable b as before such that

$$\sigma(\mathbf{\Sigma} u^{\otimes}(\boldsymbol{x}) + b\mathbf{1}) \neq \sigma(\mathbf{\Sigma} u^{\otimes}(\boldsymbol{y}) + b\mathbf{1}). \tag{47}$$

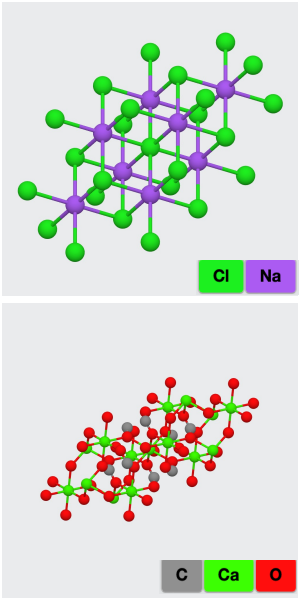
For $\mathcal{F}_{s,2}$, we set $a = w = 0, b = 1$, and the rest is similar to $\mathcal{F}_{s,1}$. ■

Remark 22 *It is known that when the hidden dimension is too small, then universal approximation through PointNet/DeepSet and their variants are not possible, see Wagstaff et al. (2022, Theorem 20). However, we consider in this paper residual structures, which do not appear to have such a constraint. It is also interesting to see if the above results can be extended to probabilistic settings such as those considered in Bloem-Reddy and Teh (2020) and the combination of sum-based and LSTM-based symmetric architectures proposed in Vinyals et al. (2015).*

4.3 Product Permutation Invariant Architectures

Besides the full symmetric group, one is often interested in functions which are invariant to certain subgroups of S distinct from those generated by simple shifts. These symmetry requirements arise naturally in a number of applications in computational chemistry and materials science. As an example, to specify a crystal lattice consisting of a number of atomic species, a convenient representation is in the form of the crystallographic information file (CIF) (Hall et al., 1991). Examples of partial features under the CIF representations of crystal lattices are shown in Fig 1. Observe that the rows and columns can be rearranged without affecting the essential structure it represents. Thus, quantities (energy, band-gap) that depend on such structures are invariant to these permutations as well. This can be effectively used for property prediction (Xie and Grossman, 2018; Chen et al., 2019) or inverse design (Ren et al., 2020).

If we flatten the input data into a vector of dimension n, such permutations forms a subgroup of S. We call this group a *product permutation group* and functions invariant



	X-coord	Y-coord	Z-coord
Na	0	0	0
Cl	0.5	0.5	0.5

	X-coord	Y-coord	Z-coord
Ca	0	0	0
Ca	0.5	0.5	0.5
C	0.25	0.25	0.25
C	0.75	0.75	0.75
O	0.0073	0.4927	0.75
O	0.25	0.9927	0.5073
O	0.4927	0.75	0.0073
O	0.5073	0.25	0.9927
O	0.75	0.0073	0.4927
O	0.9927	0.5073	0.25

Figure 1: Structure of sodium chloride (top) and calcium carbonate (bottom) in CIF form. Coordinates are fractional, relative to the crystal lattice vectors and atoms are listed in each unit cell. Observe that the rows and columns can be rearranged without affecting the structure it represents, since the order of the atoms and the order of the relative coordinate axes do not matter, provided the property of interest are rotation and reflection symmetric. Thus, quantities (density, energy, band-gap) that depend on such structures are invariant to product permutation in two dimensions, at least on these coordinate features. Graphics are obtained from the materials project (Persson, 2014a,b).

with respect to it are collectively referred to as *product permutation invariant functions*. We now give their precise definitions.

Consider the direct product of permutation acting on a double index (i, j)

$$(\mathsf{a}, \mathsf{b})(i, j) = (\mathsf{a}(i), \mathsf{b}(j)). \tag{48}$$

The collection of all (a, b) forms the $2D$ product permutation group, denoted as $\mathsf{S}_{n_1} \times \mathsf{S}_{n_2}$, or simply S_{2D}, which is a subgroup of S where $n = n_1 n_2$. Properties such as crystal structure induced ones are invariant with respect to this permutation group if the data is presented as in Figure 1.

Now, a simple way to build a control family that is equivariant with respect to the $2D$ product permutation group is

$$\mathcal{F}_{s,2D,1} = \{v\sigma(w_0\boldsymbol{x} + w_{r,1}\Sigma_{r,1}\boldsymbol{x} + w_{c,1}\Sigma_{c,1}\boldsymbol{x} + c\mathbf{1}) : v, w_0, w_{r,1}, w_{c,1}, c \in \mathbb{R}\}. \tag{49}$$

Here we assume that $\boldsymbol{x} \in \mathbb{R}^{n_1 \times n_2}$. The operators $\Sigma_{r,1}$ and $\Sigma_{c,1}$ denote the row and column sum of a tensor $\boldsymbol{x}$, namely, $\Sigma_{r,1}\boldsymbol{x}, \Sigma_{c,1}\boldsymbol{x} \in \mathbb{R}^{n_1 \times n_2}$ are defined by

$$[\Sigma_{r,1}\boldsymbol{x}]_{i,j} = \sum_{j'} x_{i,j'}, \text{ and} \tag{50}$$

$$[\Sigma_{c,1}\boldsymbol{x}]_{i,j} = \sum_{i'} x_{i',j}. \tag{51}$$

Furthermore, one can consider the second order variant

$$\mathcal{F}_{s,2D,2} = \{v\sigma(w_0\boldsymbol{x} + w_{r,1}\Sigma_{r,1}\boldsymbol{x} + w_{c,1}\Sigma_{c,1}\boldsymbol{x} + w_{r,2}\Sigma_{r,2}\boldsymbol{x} + w_{c,2}\Sigma_{c,2}\boldsymbol{x} + c\mathbf{1}) : \\ v, w_0, w_{r,1}, w_{r,2}, w_{c,1}, w_{c,2}, c \in \mathbb{R}\}, \tag{52}$$

where

$$[\Sigma_{r,2}\boldsymbol{x}]_{i,j} = \sum_{j',j''} x_{i,j'}x_{i,j''}, \tag{53}$$

$$[\Sigma_{c,2}\boldsymbol{x}]_{i,j} = \sum_{i',i''} x_{i',j}x_{i'',j}. \tag{54}$$

Clearly, it holds that $\mathcal{F}_{s,2D,1} \subset \mathcal{F}_{s,2D,2}$., and hence universal approximation of the former implies that of the latter. We can use our results to deduce the following UAP property for these architectures.

Corollary 23 *The dynamical hypothesis spaces $\mathcal{H}_{ode}(\mathcal{F}_{s,2D,1}, \mathcal{G})$ and thus $\mathcal{H}_{ode}(\mathcal{F}_{s,2D,2}, \mathcal{G})$ possess the S_{2D} UAP.*

Proof See Appendix C. ■

The above result can be generalized to higher dimensions. Consider the kD product permutation group $\mathsf{S}_{kD} := \mathsf{S}_{n_1} \times \mathsf{S}_{n_2} \times \mathsf{S}_{n_k}$, where the element $\mathbf{a} = (\mathsf{a}_1, \cdots, \mathsf{a}_k)$ acts on a multi-index $\boldsymbol{i} = (i_1, \cdots, i_k)$ by

$$\mathbf{a}(\boldsymbol{i}) = (\mathsf{a}_1(i_1), \cdots, \mathsf{a}_k(i_k)). \tag{55}$$

We first extend $\Sigma_{r,1}$ and $\Sigma_{c,1}$ to higher dimensions. Define

$$[\Sigma_{s,1}\boldsymbol{x}]_{\boldsymbol{i}} = \sum_{\boldsymbol{j}:j_s=i_s} x_{\boldsymbol{j}}, \tag{56}$$

and similarly,

$$[\Sigma_{s,2}\boldsymbol{x}]_{\boldsymbol{i}} = \sum_{\boldsymbol{j},\boldsymbol{j}':j_s=j'_s=i_s} x_{\boldsymbol{j}}x_{\boldsymbol{j}'}. \tag{57}$$

Next, we define

$$\mathcal{F}_{s,kD,1} = \{v\sigma(w_0\boldsymbol{x} + \sum_{s=1}^{k} w_{s,1}\Sigma_{s,1}\boldsymbol{x} + c\mathbf{1}) : v, w_0, w_{s,1}, c \in \mathbb{R}\}, \tag{58}$$

and

$$\mathcal{F}_{s,kD,2} = \{v\sigma(w_0\boldsymbol{x} + \sum_{s=1}^{k} w_{s,1}\Sigma_{s,1}\boldsymbol{x} + w_{s,2}\Sigma_{s,2}\boldsymbol{x} + c\mathbf{1}) : v, w_0, w_{s,1}, w_{s,2}, c \in \mathbb{R}\}. \tag{59}$$

Then, an analogous approximation results below holds for high-dimensional product permutation symmetric functions. The proof is identical to the $2D$ case and is hence omitted.

Corollary 24 *The dynamical hypothesis spaces $\mathcal{H}_{ode}(\mathcal{F}_{s,kD,1}, \mathcal{G})$ and thus $\mathcal{H}_{ode}(\mathcal{F}_{s,kD,2}, \mathcal{G})$ possess the S_{kD} UAP for $k \geq 2$.*

Proof See Appendix C. ■

To the best of our knowledge, there is no study on the approximation theory of residual architectures under product permutation invariance. However, as discussed earlier, such data symmetry structures feature in a wide variety of applications, especially in computational chemistry and physics. One way to deal with these symmetries in the practical literature, at least in the case of lattice structures, is to appeal to graphical representations (Xie and Grossman, 2018; Chen et al., 2019), which extracts product permutation invariant features that can be subsequently used to predict desired material properties. Here, the network architecture proposed are more general in the sense that it does not rely on having an explicit graphical representation, i.e. the data need not contain spatial coordinate information. Instead, deep neural networks are constructed to satisfy these symmetries in a intrinsic manner. Corollaries 23 and 24 establishes a first basic approximation guarantee of these networks for modelling functions symmetric under product permutation.

5. Conclusion and Future Work

In this paper, we provide a sufficient condition for a wide variety of residual-type neural networks to possess the universal approximation property under symmetry settings. Our results apply to architectures that can be finitely wide and achieve expressivity by increasing depth. The conditions we present can be readily checked to guarantee expressivity of both existing and new architectures used in applications where symmetry is important.

These results, along with the non-symmetric version (Li et al., 2023), parallels classical studies in approximation theory, such as the Stone–Weierstrass theorem. Specifically, we provide a deep version of the separation property, which enables us to prove

the universal approximation property. An interesting future direction is to determine necessary conditions for universal approximation through deep networks, or more generally, through maps induced by dynamical systems. Some initial results on interpolation, rather than approximation, are derived in Cheng et al. (2023).

Lastly, the symmetry addressed in this paper is restricted to *discrete symmetry* involving permutation subgroups on the coordinate indices. In practice, continuous symmetries (such as rotation, scaling, etc.) have important applications in image processing, robotics and other scientific domains. Approximation results addressing such symmetries are worthy of future study.

Acknowledgments

We are grateful for discussions with Isaac P. S. Tian and Tonio Buonassisi on the applications of the theory developed to materials modelling.

Q.L. is supported by the National Research Foundation, Singapore under the NRF fellowship (project No. NRF-NRFF13-2021-0005). T.L. is partially supported by The Elite Program of Computational and Applied Mathematics for PhD Candidates in Peking University. Z.S. is supported under the Distinguished Professorship of National University of Singapore.

Appendix A. Technical Details of the Main Result

This section includes the postponed proofs in Section 3 following the sketch shown there. For the convenience of the reader, we summarize the notation that will be used in the appendix, together with the corresponding definitions found earlier for easy reference.

$\boldsymbol{\phi}(\boldsymbol{f}, T)$	flow map	Definition 1
$\mathcal{A}_{\mathcal{F}}$	attainable set	Definition 2
$\mathcal{H}_{ode}$	dynamical hypothesis space	Definition 3
σ	well function	Definition 4
Q_{g}, Q_A	cross section	(15)
coor	coor operator	Definition 7

We now aim to prove Theorem 25 below, whose corollary is Theorem 14. Recall the definition of resolving a group from Definition 13. We hereafter take a fixed transversal A of G.

Theorem 25 (UAP from Dynamical Systems) *Suppose that $\overline{\mathcal{F}} := \overline{\mathrm{CH}}(\mathcal{F})$ is a Lipschitz control family resolving G, and*

1. *there exists a well function σ, such that $\mathcal{F}(\sigma) = \{a\sigma(bx_1 + c) : a, b, c \in \mathbb{R}\} \subset \mathrm{coor}(\overline{\mathcal{F}})$;*
2. *there exists a symmetric invariant well function τ such that $\pm\tau \in \mathrm{coor}(\overline{\mathcal{F}})$.*

Also, suppose that for any compact $K \subset \mathbb{R}^n$, there exists a Lipschitz $g \in \mathcal{G}$ such that $F(K) \subset g(\mathbb{R}^n)$. Then, for $1 \leq p < \infty$, and any G invariant mapping $F \in C(\mathbb{R}^n; \mathbb{R})$, compact region K and tolerance $\varepsilon > 0$, there exists $\hat{F} \in \mathcal{H}_{ode}$ such that $\|F - \hat{F}\|_{L^p(K)} \leq \varepsilon$.

The role of each section, proposition and lemma is as follows:

- Section A.1 reduces the problem to studying the approximation property of equivariant mappings. The main result in this subsection (i.e., Proposition 15) is based on a level-set argument, but attention on its symmetry structure, which is shown in Lemma 27.
- Section A.2 is the fundamental framework, and the main theorem (i.e. Theorem 18) tells us that the coordinate zooming property and point matching property are sufficient for universal approximation.
- Now, it suffices to show that under the assumption of Theorem 14. We can achieve the coordinate zooming property and point matching property.

  For coordinate zooming property, we prove it with the help of Lemma 29 (which is a direct consequence of the previous work, see Theorem 30).

  The point matching property is the most technical part in this manuscript, and we split it into four steps, and each step is associated with a lemma.

- Lemma 31 is purely technical and is used to prove the subsequent lemmas. It shows one can move points into a general position as much as possible. This lemma is the only one that requires the perturbation property (Definition 11). Note that the perturbation property is necessary for Lemma 31.
- Lemma 32 indicates that points in general position can be partially reordered. This is similar to Lemma 4.13 in Li et al. (2023), but the main difficulty here comes from symmetry. The partial reordering should conform to the given symmetry.
- Actually, using these two lemmas one can already prove the case when $\mathsf{G} = \mathsf{S}$, we do not need the concept of resolvence (Definition 13).

The remaining two lemmas concern the general case.

- Lemma 33 develops the main idea dealing with general symmetry structure. It shows that under the assumption that $\mathcal{F}$ resolves G, that we can move a single point across cross sections. This is necessary to achieve approximation in the general case.
- Lemma 34 is technical, it uses Lemma 33 recursively to establish an induction argument that implies the general result.

A.1 Reducing the Problem to Approximating Equivariant Mappings

We begin with proving the following result by the level-set argument.

Proposition 26 (Approximating Equivariant Mapping is Sufficient) *Suppose $F \in C(\mathbb{R}^n)$ is an invariant function. Let g be a Lipschitz continuous and invariant function with $F(\mathbb{R}^n) \subset g(\mathbb{R}^n)$. Then, for any compact $K \subset \mathbb{R}^n$ and $\varepsilon > 0$, there exists an equivariant mapping $\boldsymbol{\varphi} \in C(\mathbb{R}^n; \mathbb{R}^n)$ such that*

$$\|F - g \circ \boldsymbol{\varphi}\|_{L^p(K)} \leq \varepsilon. \tag{33}$$

This is the main result in this subsection, and the remaining part of this subsection is to prove this proposition. To this end, the following lemma is required. It tells us that a G invariant set can be decomposed into the disjoint union of cross sections, up to a small set. Recall the definition of the cross section Q_A as introduced in (15):

$$Q_A = \cup_{\mathsf{a} \in A} Q_{\mathsf{a}}, \qquad \text{where } Q_{\mathsf{a}} = \mathsf{a}(Q). \tag{60}$$

Lemma 27 (Partitioning the Space via Cross Sections) *Given $\mathsf{G} \leq \mathsf{S}$, the following holds for any G transversal A:*

1. *For two distinct $\mathsf{g}, \mathsf{g}' \in \mathsf{G}$, we have $\mathsf{g}(Q_A) \cap \mathsf{g}'(Q_A) = \emptyset$.*
2. *$\mathbb{R}^n \setminus \bigcup_{\mathsf{g} \in \mathsf{G}} \mathsf{g}(Q_A)$ is the set of points that are not in general position, and thus of zero Lebesgue measure.*

Consequently, if K is a G invariant set,(i.e., $\mathsf{g}(K) = K$ for all $\mathsf{g} \in \mathsf{G}$), then $K \backslash \cup_{\mathsf{g}\in\mathsf{G}}(K \cap \mathsf{g}(Q_A))$ is of zero Lebesgue measure.

Proof Suppose $\boldsymbol{x} \in \mathsf{g}(Q_A) \cap \mathsf{g}'(Q_A)$. Then, by definition there exists $\mathsf{a} \neq \mathsf{a}' \in A$, such that $\boldsymbol{x} \in Q_{\mathsf{ga}} \cap Q_{\mathsf{g}'\mathsf{a}'}$. The structure of Q_{a} immediately tells us that $\mathsf{ga} = \mathsf{g}'\mathsf{a}'$. Since A is a transversal of G, we have $\mathsf{g} = \mathsf{g}'$ and $\mathsf{a} = \mathsf{a}'$, leading to a contradiction.

On the other hand, we claim that if $\boldsymbol{x}$ is in general position, then there exists g such that $\boldsymbol{x} \in \mathsf{g}(Q_A)$. The construction is straightforward: since $\boldsymbol{x} \in Q_{\mathsf{a}}$ for some $\mathsf{a} \in \mathsf{S}$, then by the choice of A, there exists a decomposition $\mathsf{a} = \mathsf{gb}$, such that $\mathsf{b} \in A$ and $\mathsf{g} \in \mathsf{G}$. Therefore, we have $\boldsymbol{x} \in \cup_{\mathsf{g}\in\mathsf{G}}\mathsf{g}(Q_A)$. The reverse inclusion holds trivially. ■

Now we are ready to prove Proposition 15. The idea is simple. We first obtain $\boldsymbol{u}$ without the equivariance constraint, this is done by Li et al. (2023, Theorem 3.8). The core of the following proof is to modify the $\boldsymbol{u}$ into our desired $\boldsymbol{f}$.

Proof (Proof of Proposition 15) Without loss of generality, we assume that K is a G invariant set, otherwise we can enlarge K to make it G invariant. Choose the G transversal A arbitrarily, it follows from Lemma 27 that $K = \cup_{\mathsf{g}\in\mathsf{G}}(K \cap \mathsf{g}(Q_A))$ up to a measure zero set. Define $\varepsilon' := \frac{\varepsilon}{|\mathsf{G}|(1+\operatorname{Lip} g)}$, by results in Li et al. (2023, Theorem 3.8), for any $\varepsilon' > 0$ there exists $\boldsymbol{u}$ such that

$$\|F - g \circ \boldsymbol{u}\|_{L^p(K)} \leq \varepsilon'. \tag{61}$$

Note that $\boldsymbol{u}$ here is not necessarily equivariant, otherwise we are done. Now we attempt to find $\boldsymbol{f}$ by some kind of "equivariantization" on $\boldsymbol{u}$ as explained below. Since $\boldsymbol{u}$ is in L^p, we consider a compact set $O \subset Q_A$ such that $\|\boldsymbol{u}\|_{L^p(Q_A\backslash O)} \leq \varepsilon'$. Take a smooth truncation function $\chi \in C^\infty(\mathbb{R}^d)$, whose value is in $[0, 1]$, such that $\chi|_O = 1$ and $\chi|_{Q_A^c} = 0$.

For $\boldsymbol{x} \in \mathsf{g}(Q_A)$ with $\mathsf{g} \in \mathsf{G}$, define $\boldsymbol{f}(\boldsymbol{x}) = \mathsf{g}(\tilde{\boldsymbol{u}}(\mathsf{g}^{-1}(\boldsymbol{x})))$, where $\tilde{\boldsymbol{u}} = \chi\boldsymbol{u}$ is smoothed truncated version of $\boldsymbol{u}$. Since different $\mathsf{g}(Q_{\mathsf{G}})$ are disjoint, the value of $\boldsymbol{f}$ is unique in $\cup_{\mathsf{g}\in\mathsf{G}}(K \cap \mathsf{g}(Q_{\mathsf{G}}))$. We set $\boldsymbol{f}(\boldsymbol{x}) = 0$ in the complement of $\cup_{\mathsf{g}\in\mathsf{G}}(K \cap \mathsf{g}(Q_A))$. The truncation function χ ensures that $\boldsymbol{f}$ vanishes on the boundary of Q_A, therefore $\boldsymbol{f}$ is continuous, and direct verification shows that $\boldsymbol{f}$ is G equivariant.

It then suffices to estimate $\|F - g \circ \boldsymbol{f}\|_{L^p}$, since both F and $g \circ \boldsymbol{f}$ are equivariant, it is natural and helpful to restrict our estimation on $K \cap Q_A$, since

$$\|F - g \circ \boldsymbol{f}\|_{L^p(K)} = |\mathsf{G}|\ \|F - g \circ \boldsymbol{f}\|_{L^p(K\cap Q_A)}. \tag{62}$$

To estimate the error on $K \cap Q_A$, we first estimate the error $\|\boldsymbol{u} - \boldsymbol{f}\|_{L^p(K\cap Q_A)}$. Since $\boldsymbol{u}$ and $\boldsymbol{f}|_{Q_A} = \tilde{\boldsymbol{u}}$ coincide on O, we have

$$\begin{aligned} \|\boldsymbol{u} - \boldsymbol{f}\|_{L^p(K\cap Q_A)} &= \|\boldsymbol{u} - \tilde{\boldsymbol{u}}\|_{L^p(K\cap Q_A)} \\ &\leq \|\boldsymbol{u}\|_{L^p(Q_A\backslash O)} = \varepsilon'. \end{aligned} \tag{63}$$

The inequality holds from χ takes value in $[0, 1]$. Since g is Lipschitz, we have $\|g \circ \boldsymbol{u} - g \circ \boldsymbol{f}\|_{L^p(K\cap Q_A)} \leq \operatorname{Lip} g\varepsilon'$, yielding that $\|F - g \circ \boldsymbol{f}\|_{L^p(K\cap Q_A)} \leq (1 + \operatorname{Lip} g)\varepsilon'$. We finally

have $\|F - g \circ \boldsymbol{f}\|_{L^p(K)} \leq (1 + \operatorname{Lip} g)|\mathsf{G}|\varepsilon' = \varepsilon$. ■

A.2 Universal Approximation under Symmetry using Compositions

In this section, we prove the following result.

Theorem 18 (Universal Approximation of Invariant Functions via Composition) *Let $F \in C(\mathbb{R}^n)$ be invariant. Suppose $\mathcal{G}$ is a family of invariant functions and $\mathcal{A}$ is a family of equivariant Lipschitz mappings with the following properties:*

1. *For any compact $K \subset \mathbb{R}^n$, there exists a Lipschitz $g \in \mathcal{G}$ such that $F(K) \subset g(\mathbb{R}^n)$.*
2. *$\mathcal{A}$ is closed under composition, i.e., if $\boldsymbol{f_1} \in \mathcal{A}$ and $\boldsymbol{f_2} \in \mathcal{A}$, then $\boldsymbol{f_1} \circ \boldsymbol{f_2} \in \mathcal{A}$.*
3. *(Coordinate zooming) For any increasing function $v \in C(\mathbb{R})$, compact interval $\mathbb{I} \subset \mathbb{R}$ and tolerance $\varepsilon > 0$, there exists $u : \mathbb{R} \to \mathbb{R}$ such that $u^{\otimes} \in \mathcal{A}$ and $\|u - v\|_{C(\mathbb{I})} \leq \varepsilon$.*
4. *(Point matching)*

   For $M > 0$, a transversal A of G, a G distinct point set $\{\boldsymbol{x}^1, \cdots, \boldsymbol{x}^M\}$ with $\boldsymbol{x}^i \in \overline{Q_A}$, another point set $\{\boldsymbol{y}^1, \cdots, \boldsymbol{y}^M\} \subset Q_A$ and tolerance $\varepsilon > 0$, there exists $\boldsymbol{f} \in \mathcal{A}$ such that

   (a) *For $\boldsymbol{x}^i$ in general position, we have $|\boldsymbol{f}(\boldsymbol{x}^i) - \boldsymbol{y}^i| \leq \varepsilon$.*

   (b) *For $\boldsymbol{x}^i$ not in general position, we have $|\boldsymbol{f}(\boldsymbol{x}^i)| \leq 1$.*

Then, for any compact set $K \subset \mathbb{R}^n$, tolerance $\varepsilon > 0$ and $p \in [1, \infty)$, there exists $\boldsymbol{\varphi} \in \mathcal{A}$ and $g \in \mathcal{G}$ such that $\|F - g \circ \boldsymbol{\varphi}\|_{L^p(K)} \leq \varepsilon$.

Remark 28 *Here, we make some remarks about the point matching property, which is the most technical part in this paper. Simply speaking, we use the point matching property (a) to bound the approximation error. For technical reasons, however, the points that are not in general position should be taken into consideration, and cannot be ignored even if they are contained in a set of 0 Lebesgue measure. The difficulty here is that for any equivariant function $\boldsymbol{\varphi}$, the image $\boldsymbol{\varphi}(\boldsymbol{x})$ is also not in general position if $\boldsymbol{x}$ itself is not. As a result, we cannot send it to an arbitrary $\boldsymbol{y}$ as in the statement of the point matching property. This shows why we need to treat the points not in general position separately.*

The reason why we cannot ignore these points is as follows. Our error estimate is based on a decomposition of the approximation error to three parts (see Step 3 in the proof): 1) error on a union of hyper-cubes centered at the points in general position; 2) error on a union of hyper-cubes centered at the points not in general position; and 3) error on the remaining set. Even if the volume of the domain in 2) goes to zero in the limit of vanishing hyper-cube sizes and increasing number of points, one must rule out the possibility that while trying to reduce the approximation error in 1), one

causes a rise in the approximation error in 2) so much so that the limit of 2) does not become arbitrarily small in this limit. Now, this can be resolved if we place some uniform regularity assumption on $\mathcal{A}$, e.g. $\sup_{\boldsymbol{f}\in\mathcal{A}} \operatorname{Lip} \boldsymbol{f} < \infty$. However, for many applications, including the flow-based hypothesis considered in this paper, this does not hold. Condition 4(b) is a much weaker assumption that achieves the same effect.

Proof Without loss of generality, we assume $K = [-\kappa, \kappa]^n$ is a hyper-cube centered at the origin. Select Q_A for arbitrary G transversal A. Since K is G invariant, the decomposition $K = \cup_{\mathsf{g}\in\mathsf{G}}\mathsf{g}(K \cap \mathsf{g}(Q_A))$ holds up to a measure zero set, according to Lemma 27.

Observe that $K \cap Q_A$ is a polyhedron. In the following, we denote by $\boldsymbol{\lambda}$ the Lebesgue measure.

Step 1 By Proposition 15, we can find an invariant $g \in \mathcal{G}$ and an equivariant $\boldsymbol{\varphi} \in C(\mathbb{R}^n, \mathbb{R}^n)$ such that

$$\|F - g \circ \boldsymbol{\varphi}\|_{L^p(K)} \leq \varepsilon/2. \tag{64}$$

Now we consider a piecewise constant approximant $\boldsymbol{\varphi}_0$. Given a scale $\delta > 0$, consider the grid $\delta\mathbb{Z}^n$ with size δ. Let $\boldsymbol{i} = [i_1, \cdots, i_n] \in \mathbb{Z}^n$ be a multi-index, and $\chi_{\boldsymbol{i}}$ be the indicator of the cube

$$\square_{\boldsymbol{i},\delta} := [i_1\delta, (i_1+1)\delta] \times \cdots \times [i_n\delta, (i_n+1)\delta]. \tag{65}$$

Since $\boldsymbol{\varphi}$ is in $L^p(K)$, by standard approximation theory $\boldsymbol{\varphi}$ can be approximated by an equivariant piecewise constant (and G equivariant) function

$$\boldsymbol{\varphi}_0(\boldsymbol{x}) = \sum_{\boldsymbol{i}} \boldsymbol{y}_{\boldsymbol{i}} \chi_{\boldsymbol{i}}(\boldsymbol{x}), \tag{66}$$

where

$$\boldsymbol{y}_{\boldsymbol{i}} = \boldsymbol{\lambda}(\square_{\boldsymbol{i},\delta})^{-1} \int_{\square_{\boldsymbol{i},\delta}} \boldsymbol{\varphi}(\boldsymbol{x}) d\boldsymbol{x} \tag{67}$$

is the local average value of $\boldsymbol{\varphi}$ in $\square_{\boldsymbol{i},\delta}$. Then, we have

$$\|\boldsymbol{\varphi} - \boldsymbol{\varphi}_0\|_{L^p(K)} \leq \omega_{\boldsymbol{\varphi}}(\delta)[\boldsymbol{\lambda}(K)]^{1/p} \to 0 \tag{68}$$

as $\delta \to 0$, where $\omega_{\boldsymbol{\varphi}}$ is the modulus of continuity (restricted in the region K), i.e.,

$$\omega_{\boldsymbol{\varphi}}(\delta) := \sup_{|\boldsymbol{x}-\boldsymbol{y}|\leq\delta} |\boldsymbol{\varphi}(\boldsymbol{x}) - \boldsymbol{\varphi}(\boldsymbol{y})| \tag{69}$$

for $\boldsymbol{x}$ and $\boldsymbol{y}$ in K and $\boldsymbol{\lambda}(K)$ is the Lebesgue measure of K. Since g is G invariant, we can replace $\boldsymbol{y}_i$ by arbitrary $\mathsf{g}(\boldsymbol{y}_i)$ for $\mathsf{g} \in \mathsf{G}$. Therefore, we can assume without loss of generality that $\boldsymbol{y} \in Q_A$. Thus,

$$\begin{aligned} \|F - g \circ \boldsymbol{\varphi}_0\|_{L^p(K)} \leq\ & \|F - g \circ \boldsymbol{\varphi}\|_{L^p(K)} + \|g \circ \boldsymbol{\varphi} - g \circ \boldsymbol{\varphi}_0\|_{L^p(K)} \\ \leq\ & \varepsilon/2 + \operatorname{Lip}(g)\omega_{\boldsymbol{\varphi}}(\delta)[\boldsymbol{\lambda}(K)]^{1/p}. \end{aligned} \tag{70}$$

Choose suitable $\delta > 0$ such that the right-hand side of (70) is smaller than ε.

Step 2 Let $\boldsymbol{p_i} = \boldsymbol{i}\delta = [i_1\delta, \cdots, i_n\delta]$ be the a vertex of $\square_{\boldsymbol{i},\delta}$. Define $\mathcal{I}$ as the maximal subset of $\{\boldsymbol{i} : \boldsymbol{p_i} \in \overline{Q_A}\}$ such that $\mathcal{P} := \{\boldsymbol{p_i} : \boldsymbol{i} \in \mathcal{I}\}$ is G distinct. By the maximal property, and the definition of G distinctness (see Definition 17), we know that if $p_{\boldsymbol{j}} \in \overline{Q_A}$ with some $\boldsymbol{j} \notin \mathcal{I}$, then there must exist $\mathsf{g} \in \mathsf{G}$ and $\boldsymbol{i} \in \mathcal{I}$ such that $\boldsymbol{p_i} = \mathsf{g}(\boldsymbol{p_j})$.

Given $\varepsilon > 0$, by the point matching property (Condition 4) we can find $\boldsymbol{f}$ such that

- For $\boldsymbol{p_i}$ in general position, $|\boldsymbol{f}(\boldsymbol{p_i}) - \boldsymbol{y_i}| \le \varepsilon$ for all $\boldsymbol{p_i} \in \mathcal{P}$.
- For $\boldsymbol{p_i}$ not in general position, $|\boldsymbol{f}(\boldsymbol{p_i})| \le 1$.

Then, by the extremeness of $\mathcal{P}$, the inequality holds true for all $\boldsymbol{p_i}$ such that $\square_{\boldsymbol{i},\delta} \subset K$.

For $\alpha \in (0,1)$, define the shrunken cube

$$\square^{\alpha}_{\boldsymbol{i},\delta} := [i_1\delta, (i_1+\alpha)\delta] \times \cdots \times [i_n\delta, (i_n+\alpha)\delta], \tag{71}$$

and define $K^{\alpha} = \bigcup_{\square_{\boldsymbol{i},\delta} \subset K} \square^{\alpha}_{\boldsymbol{i},\delta}$ to be a subset of K. Given $\beta > 0$, we now use the coordinate zooming property (Condition 3) to find $u^{\otimes} \in \mathcal{A}$ such that

$$u([ih, (i+\alpha h)]) \subset [ih, (i+\frac{\beta}{n}\delta)] \text{ for } i \in \{i_s : s = 1, \ldots, n; \boldsymbol{i} \in \mathcal{I}\}. \tag{72}$$

To do this, we first construct a piecewise linear function $\tilde{u}$ such that

$$\tilde{u}|_{[i\delta,(i+\alpha)\delta]}(x) = i + \frac{\beta}{2n}\delta, \tag{73}$$

by setting

$$\tilde{u}|_{[(i+\alpha)\delta,(i+1)\delta]}(x) = (x - (i-\alpha)\delta)/(1-\alpha) + i + \frac{\beta}{2n}\delta \tag{74}$$

explicitly, and select $\varepsilon < \frac{\beta}{3n}\delta$. Then we use Coordinate zooming property with respect to $v = \tilde{u}$ and ε, to obtain u, such that $\|u - \tilde{u}\|_{C(\mathbb{I})} \le \varepsilon$, yielding the condition (72) holds.

Therefore, we have

$$|\boldsymbol{f}(u^{\otimes}(\boldsymbol{x})) - y_{\boldsymbol{i}}| \le 2\varepsilon \text{ for } \boldsymbol{x} \in \square^{\alpha}_{\boldsymbol{i},\delta}, \tag{75}$$

where $\boldsymbol{p}_i$ is in general position, and

$$|\boldsymbol{f}(u^{\otimes}(\boldsymbol{x}))| \le 1 + \varepsilon \text{ for } \boldsymbol{x} \in \square^{\alpha}_{\boldsymbol{i},\delta}, \tag{76}$$

where $\boldsymbol{p}_i$ is not in general position.

These two estimates (75) and (76) will be useful in the final step.

Step 3 We are ready to estimate the error $\|F - g\circ\boldsymbol{f}\circ u^{\otimes}\|_{L^p(K)}$. Since $\|F - g\circ\boldsymbol{\varphi}\|_{L^p(K)} \le \varepsilon$, it suffices to estimate $\|g\circ\boldsymbol{\varphi} - g\circ\boldsymbol{f}\circ u^{\otimes}\|_{L^p(K)} \le \mathrm{Lip}(g)\|\boldsymbol{\varphi} - \boldsymbol{f}\circ u^{\otimes}\|_{L^p(K)}$.

The estimation is split into three parts,

$$\begin{aligned} K \setminus K^{\alpha}, \\ K_1^{\alpha} = \bigcup_{\boldsymbol{p_i} \text{ in general position}} \square^{\alpha}_{\boldsymbol{i},\delta}, \\ K_2^{\alpha} = \bigcup_{\boldsymbol{p_i} \text{ not in general position}} \square^{\alpha}_{\boldsymbol{i},\delta}. \end{aligned} \tag{77}$$

Notice that $K^\alpha = \cup \square^\alpha_{\boldsymbol{i},\delta}$.

For K_1^α, from (75) in the end of Step 2, we have $\|\boldsymbol{f} \circ u^\otimes - \boldsymbol{\varphi}_0\|_{L^\infty(K_1^\alpha)} \leq 2\varepsilon$, and thus

$$\|\boldsymbol{f} \circ u^\otimes - \boldsymbol{\varphi}_0\|_{L^p(K_1^\alpha)} \leq 2\varepsilon[\boldsymbol{\lambda}(K^\alpha)]^{1/p} \leq 2\varepsilon[\boldsymbol{\lambda}(K)]^{1/p}. \tag{78}$$

For K_2^α, note that if $\boldsymbol{p_i}$ does not in general position, then all points in $\square_{\boldsymbol{i},\delta}$ will be close to a hyperplane $\Gamma_{ij} := \{\boldsymbol{x} : x_i = x_j\}$ for some distinct i, j, the distance from those points to Γ_{ij} will be small than $\sqrt{n}\delta$. Therefore, the Lebesgue measure of $K_2^\alpha \subset K_2$ will be smaller than that of all points whose distance to the union of hyperplanes Γ_{ij} is less than $\sqrt{n}\delta$, which is $O(\delta)$. Thus, we have

$$\begin{aligned} \|\boldsymbol{f} \circ u^\otimes - \boldsymbol{\varphi}_0\|_{L^p(K_2^\alpha)} \leq & \ (1 + \varepsilon + \|\boldsymbol{\varphi}_0\|_{C(K)})O(\delta) \\ \leq & \ (1 + \varepsilon + \|\boldsymbol{\varphi}\|_{C(K)})O(\delta). \end{aligned} \tag{79}$$

The last line holds since $\|\boldsymbol{\varphi}_0\|_{C(K)} \leq \|\boldsymbol{\varphi}\|_{C(K)}$ by construction. Here, the constant 1 on the right-hand side comes from the bound on the points not in general position, namely, Condition 4.(b), used in (76). If we do not impose such a condition, the right-hand side of (79) could go to infinity, leading to the stability issue.

For $K \setminus K^\alpha$, we have

$$\begin{aligned} \|\boldsymbol{f} \circ u^\otimes - \boldsymbol{\varphi}_0\|_{L^p(K\setminus K^\alpha)} &\leq (\|\boldsymbol{f}\|_{C(K)} + \|\boldsymbol{\varphi}\|_{C(K)}) \ \boldsymbol{\lambda}(K \setminus K^\alpha)^{1/p} \\ &\leq (\|\boldsymbol{f}\|_{C(K)} + \|\boldsymbol{\varphi}\|_{C(K)})(1 - \alpha^d)^{1/p}[\boldsymbol{\lambda}(K)]^{1/p}. \end{aligned} \tag{80}$$

We first choose δ sufficiently small such that the right hand side of (79) is not greater than ε, then choose α such that $1-\alpha$ is sufficiently small, and $(\|\boldsymbol{f}\|_{C(K)} + \|\boldsymbol{\varphi}\|_{C(K)})(1 - \alpha^d)^{1/p} \leq \varepsilon$.

Hence, the total error is

$$\|F - g \circ \boldsymbol{f} \circ u^\otimes\| \leq 3\varepsilon[\boldsymbol{\lambda}(K)]^{1/p} + 2\varepsilon. \tag{81}$$

Combining two estimates we conclude the result (with $3\varepsilon[\boldsymbol{\lambda}(K)]^{1/p} + 2\varepsilon$ replacing ε.) ■

A.3 Results on Dynamical Hypothesis Spaces.

Before going to the proofs of the main results, we first prove the following auxiliary lemma, which says that the presence of well function is enough to guarantee the coordinate zooming property, see Condition 3 in Theorem 18. This is also the core part of the previous paper (Li et al., 2023).

Lemma 29 (Well Function Achieves Coordinate Zooming Property) *For a one-dimensional function $\sigma \in C(\mathbb{R})$, define its control family under affine invariance as follows:*

$$\mathcal{F}(\sigma) := \{x \mapsto w\sigma(ax + b) : w, a, b \in \mathbb{R}\}. \tag{82}$$

Here, x is a one-dimensional variable. If there exists a one-dimensional well function σ such that $\mathcal{F}(\sigma) \subset \mathrm{coor}(\mathcal{F})$, then $\mathcal{A}_\mathcal{F}$ has the coordinate zooming property.

Proof This follows immediately from the definition of coordinate zooming property and Theorem 30. ■

Theorem 30 (Main result in Li et al. (2023), one-dimensional case) *For $F : \mathbb{R} \to \mathbb{R}$ being continuous and increasing, if the one-dimensional control family $\mathcal{F}$ satisfies*

1. *For any compact interval I there exists a Lipschitz $g \in \mathcal{G}$ such that $F(I) \subset g(\mathbb{R})$.*
2. *$\mathcal{F}$ is affine invariant.*
3. *$\overline{\mathrm{CH}}(\mathcal{F})$ contains a well function.*

Then for any compact interval $I \subset \mathbb{R}^n$ and $\varepsilon > 0$, there exists $\hat{F} \in \mathcal{H}_{ode}(\mathcal{F}, \mathcal{G})$ such that $\|F - \hat{F}\|_{L^\infty(I)} \leq \varepsilon$.

For convenience, the following proof will assume $\mathcal{F}$ (instead of $\overline{\mathrm{CH}}(\mathcal{F})$) satisfies the above conditions. This turns out to be sufficient, since one can prove that the control family generated by $\mathcal{F}$ and $\overline{\mathrm{CH}}(\mathcal{F})$ have the same closure under the topology of compact convergence. (See Proposition 39).

We first state without proof two lemmas that will be used to prove Theorem 25.

These two lemmas will be proved later.

Lemma 31 (Perturbation Lemma) *Suppose that the point set $\{\boldsymbol{x}^1, \cdots, \boldsymbol{x}^m\}$ is G distinct and $\mathcal{F}$ satisfies the perturbation property (Definition 11). Then, there exists a mapping $\boldsymbol{\beta} \in \mathcal{A}_{\mathcal{F}}$ so that $\boldsymbol{y}^i = \boldsymbol{\beta}(\boldsymbol{x}^i)$, $i = 1, \ldots, m$,satisfy: if $(i, j) \neq (k, l)$ but*

$$y_i^j = y_k^l, \tag{83}$$

then both $\boldsymbol{y}^j$ and $\boldsymbol{y}^l$ are not in general position. A point set $\{\boldsymbol{y}^i\}$ satisfying this condition will subsequently be called well-perturbed.

Lemma 32 (Partially Ordering Lemma) *Suppose $\mathcal{A} = \mathcal{A}_{\mathcal{F}}$ is generated by a G equivariant Lipschitz control family $\mathcal{F}$, and there exists a symmetric invariant well function $\tau(\boldsymbol{x})$, such that $\pm\tau \in \mathrm{coor}(\mathcal{F})$. Let A be a transversal. Consider a finite point set $X = \{\boldsymbol{x}^1, \cdots, \boldsymbol{x}^M\} \subset \overline{Q_A}$ such that X is well-perturbed. Then, there exists $\boldsymbol{\beta} \in \mathcal{A}$ such that $\boldsymbol{\beta}(X) \subset \overline{Q_A}$ is partially ordered. Moreover, we may require that*

1. *$\boldsymbol{\beta}(X)$ is well perturbed, as defined in Lemma 31.*
2. *If $\boldsymbol{x} \in X$ is in general position, so is $\boldsymbol{\beta}(\boldsymbol{x})$.*

In fact, when $\mathsf{S} = \mathsf{G}$, we can prove Theorem 25 for full permutation case. To obtain the general result, we further need the following two lemmas. The following lemma tells us what we can obtain from the resolvence.

Lemma 33 *Suppose $\mathcal{F}$ satisfies the conditions in Theorem 25. Then for each $\boldsymbol{x} \in Q_A$, there exists $\boldsymbol{\gamma} \in \mathcal{A}_{\mathcal{F}}$ such that $\boldsymbol{\gamma}(\boldsymbol{x}) \in Q$.*

The following lemma performs mathematical induction with the help of Lemma 33.

Lemma 34 *Suppose that $\{\boldsymbol{x}^1, \cdots, \boldsymbol{x}^m\}$ are partially ordered, and that $\mathcal{F}$ satisfies the conditions in Theorem 25. If for some $\boldsymbol{x}^s$, there exists $\boldsymbol{\zeta}^\circ \in \mathcal{A}_{\mathcal{F}}$ such that $\boldsymbol{\zeta}^\circ(\boldsymbol{x}^s) \in Q$ Then we can find $\boldsymbol{\zeta} \in \mathcal{A}_{\mathcal{F}}$ such that*

1. *$\boldsymbol{\zeta}(\boldsymbol{x}^1), \cdots, \boldsymbol{\zeta}(\boldsymbol{x}^m)$ are partially ordered.*
2. *For $i \neq s$, if $\boldsymbol{x}^i$ is in the cross section Q_{a}, then so is $\boldsymbol{\zeta}(\boldsymbol{x}^i)$.*
3. *$\boldsymbol{\zeta}(\boldsymbol{x}^s) \in Q$.*

Assuming these lemmas, now we are ready to prove Theorem 25.

Proof (Proof of Theorem 25) The proof is divided into several steps.

Step 1 We only need to show that the point matching property (see Condition 3 of Theorem 18) holds for such $\mathcal{F}$. Without loss of generality, we assume that

$$X = \{\boldsymbol{x}^1, \boldsymbol{x}^2, \cdots, \boldsymbol{x}^M\}, \tag{84}$$

where $\boldsymbol{x}^i$ $(i = 1, \cdots, m)$ are in general position, and $\boldsymbol{x}^i$ $(i = m+1, \cdots, M)$ are not in general position. By the perturbation lemma (Lemma 31), we can find $\boldsymbol{\alpha} \in \mathcal{A}_{\mathcal{F}}$ such that $\boldsymbol{\alpha}(X)$ is well-perturbed. Thus applying the partially ordering lemma (Lemma 32), we can find $\boldsymbol{\beta} \in \mathcal{A}_{\mathcal{F}}$ such that $\widetilde{X} = \boldsymbol{\beta}(\boldsymbol{\alpha}(X)) \subset Q_A$ is partially ordered and meets the following two requirements as stated in Lemma 32. In the rest of the proof, we will use $\widetilde{X}$ to replace X.

Step 2 First, we assert that, there exists $\boldsymbol{\beta} \in \mathcal{A}_{\mathcal{F}}$ such that $\boldsymbol{\beta}$ can send all points in general position into Q, that is,

1. For $\boldsymbol{x}$ in general position, $\boldsymbol{\beta} \circ \boldsymbol{\alpha}(\boldsymbol{x}) \in Q$.
2. $\boldsymbol{\beta} \circ \boldsymbol{\alpha}(X)$ is partially ordered.

Now let us deal with the assertion, by Lemma 27 we can find $\boldsymbol{\zeta}_1^\circ \in \mathcal{A}_{\mathcal{F}}$ such that $\boldsymbol{\zeta}_1^\circ(\boldsymbol{x}^1) \in Q$. By Lemma 34, we can modify this $\widehat{\boldsymbol{\zeta}_1}$ to $\boldsymbol{\zeta}_1$, such that

1. $\boldsymbol{\zeta}_1(\boldsymbol{x}^1), \cdots, \boldsymbol{\zeta}_1(\boldsymbol{x}^m)$ is partially ordered.
2. For $i \neq s$, if $\boldsymbol{x}^i$ is in the same cross section Q_{a}, then so is $\boldsymbol{\zeta}_1(\boldsymbol{x}^i)$.
3. $\boldsymbol{\zeta}_1(\boldsymbol{x}^1) \in Q$.

Sequentially, we can find $\boldsymbol{\zeta}_2, \cdots, \boldsymbol{\zeta}_m$ to map all points in general position in Q, therefore

$$\boldsymbol{\beta} = \boldsymbol{\zeta}_m \circ \boldsymbol{\zeta}_{m-1} \circ \cdots \circ \boldsymbol{\zeta}_1 \tag{85}$$

satisfies the assertion.

Step 3 With a little abuse of notations, we use $\boldsymbol{x}^i$ to denote $\boldsymbol{\beta} \circ \boldsymbol{\alpha}(\boldsymbol{x}^i)$. Now, the conditions in Lemma 32 read,

1. For $\boldsymbol{x}$ in general position, $\boldsymbol{x} \in Q$.
2. X is partially ordered.

We now are ready to prove the point matching property. We first consider an ideal case to illustrate our idea to the proof: if all the points in X are in general position (namely, $m = M$), the proof will be straightforward. We can assume that the destination point set

$$Y = \{\boldsymbol{y}^1, \boldsymbol{y}^2, \cdots, \boldsymbol{y}^m\}, \tag{86}$$

where $\boldsymbol{y}^i (i = 1, \cdots, m)$ are in general position. Again, by the construction of Lemma 32, we can find $\boldsymbol{\gamma} \in \mathcal{A}_{\mathcal{F}}$ such that $\widetilde{Y} = \boldsymbol{\gamma}(Y) \subset Q$ is partially ordered. We may set $u^{\otimes}$ as a coordinate zooming function such that

$$|u(\tilde{x}_j^i) - \tilde{y}_j^i| \leq \varepsilon / (\operatorname{Lip} \boldsymbol{\gamma}^{-1} \cdot \operatorname{Lip}(\boldsymbol{\beta} \circ \boldsymbol{\alpha})). \tag{87}$$

Therefore, the mapping $\boldsymbol{\rho} = \boldsymbol{\gamma}^{-1} \circ u^{\otimes} \circ \boldsymbol{\beta} \circ \boldsymbol{\alpha}$ can make sure that $|\boldsymbol{\rho}(\boldsymbol{x}^i) - \boldsymbol{y}^i| \leq \varepsilon$.

Step 4 Now we consider the general case, which needs a slight modification on the destination point set Y: We consider adding a point into Y. Define

$$Y_+ = \{\boldsymbol{y}^1, \cdots, \boldsymbol{y}^m, \boldsymbol{y}^{m+1} = \boldsymbol{z} := \boldsymbol{0}\}, \tag{88}$$

where $\boldsymbol{y}^i (i = 1, \cdots, m)$ is in general position, and no coordinate value of them are 0. This Y_+, by definition, is well-perturbed. Therefore, by the partially ordering lemma (Lemma 32), we can find $\boldsymbol{\gamma} \in \mathcal{A}_{\mathcal{F}}$ such that $\widetilde{Y}_+ = \boldsymbol{\gamma}(\boldsymbol{\alpha}(Y_+)) \subset \overline{Q}$ is partially ordered. In this case, we may also set $u^{\otimes}$ as a coordinate zooming function such that $|u(\tilde{x}_j^i) - \tilde{y}_j^i| \leq \varepsilon / (\operatorname{Lip} \boldsymbol{\gamma}^{-1} \cdot \operatorname{Lip}(\boldsymbol{\beta} \circ \boldsymbol{\alpha}))$. However, the issue we encounter here is that the value of $\tilde{y}_j^i$ have not been defined for $i = m + 1, \cdots, M$.

To resolve this, we now determine the value of $\tilde{y}_j^i$ for $i = m + 1, \cdots, M$. Denote by $\tilde{\boldsymbol{z}} := \boldsymbol{\gamma}(\boldsymbol{z}) = z\boldsymbol{1}$. Due to the S equivariance, we require

1. $|y_j^i - \tilde{z}| \leq \varepsilon / (\operatorname{Lip} \boldsymbol{\gamma}^{-1} \cdot \operatorname{Lip}(\boldsymbol{\beta} \circ \boldsymbol{\alpha}))$
2. $\tilde{y}_j^i \geq \tilde{y}_{j'}^{i'}$ if and only if $\tilde{x}_j^i \geq \tilde{x}_{j'}^{i'}$.

Therefore, the mapping $\boldsymbol{\rho} = \boldsymbol{\gamma}^{-1} \circ u^{\otimes} \circ \boldsymbol{\beta} \circ \boldsymbol{\alpha}$ ensures that $|\boldsymbol{\rho}(\boldsymbol{x}^i) - \boldsymbol{y}^i| \leq \varepsilon$ for $i = 1, 2, \cdots, m$, and $|\boldsymbol{\rho}(\boldsymbol{x}^i)| \leq \varepsilon$ for $i = m + 1, \cdots, M$. ■

A.4 Proofs of Lemmas

The rest of this section aims to prove lemmas 31, 32, 33 and 34.

A.4.1 Proof of Lemma 31: Perturbation Lemma

We recall the definition of the perturbation property from Definition 13. Define the *similarity* of two points, at least one of which is in general position, as

$$\overline{s}(\boldsymbol{x}, \boldsymbol{y}) = |\{(i, i') : x_i = y_{i'}\}|. \tag{89}$$

When both points are not in the general position, the similarity is defined as zero. Assuming the perturbation property is satisfied, $\overline{s}(\boldsymbol{x}, \boldsymbol{y}) \neq 0$ implies the existence of $\boldsymbol{f} \in \mathcal{F}$ and a coordinate zooming function $u^{\otimes}$ such that

$$[\boldsymbol{f}(u^{\otimes}(\boldsymbol{x}))]_i \neq [\boldsymbol{f}(u^{\otimes}(\boldsymbol{y}))]_{i'} \tag{90}$$

holds for some i such that $x_i = y_{i'}$.

Proof (Proof of Lemma 31) We first extend the similarity $\overline{s}$ in (89) to a set, namely, define

$$\overline{s}(Y) = \sum_{j \neq j'} \overline{s}(\boldsymbol{y}^j, \boldsymbol{y}^{j'}). \tag{91}$$

The goal of perturbation lemma is to show there exists $\boldsymbol{\beta} \in \mathcal{A}_{\mathcal{F}}$ such that $\overline{s}(\boldsymbol{\beta}(X)) = 0$. Consider $\boldsymbol{\beta} \in \mathcal{A}_{\mathcal{F}}$ which minimizes $\overline{s}(\boldsymbol{\beta}(X))$, and for convenience we set $Y = \boldsymbol{\beta}(X)$, which means $\boldsymbol{y}^j = \boldsymbol{\beta}(\boldsymbol{x}^j)$. We assert that the quantity $\overline{s}(Y)$ should be zero.

Otherwise, there exists J, J' such that $\overline{s}(\boldsymbol{y}^J, \boldsymbol{y}^J) > 0$. By definition of the perturbation property, we can find $\boldsymbol{f} \in \mathcal{F}$ and $u^{\otimes}$ such that at for some I and I', $[\boldsymbol{f}(u^{\otimes}(\boldsymbol{y}^J))]_I \neq [\boldsymbol{f}(u^{\otimes}(\boldsymbol{y}^{J'}))]_{I'}$ but $y_I^J = y_{I'}^{J'}$.

Then, consider $\boldsymbol{\gamma}_t = \boldsymbol{\phi}(\boldsymbol{f}, t) \circ u^{\otimes}$. For sufficiently small t, observe that no new (i, i', j, j') will meet the requirement

$$[\boldsymbol{\gamma}_t(\boldsymbol{y}^j)]_i = [\boldsymbol{\gamma}_t(\boldsymbol{y}^{j'})]_{i'}, \qquad \text{and} \qquad y_i^j \neq y_{i'}^{j'}. \tag{92}$$

Therefore, the quantity

$$\mathcal{S}(t) := \sum_{j \neq j'} \overline{s}(\boldsymbol{\gamma}_t(\boldsymbol{y}^j), \boldsymbol{\gamma}_t(\boldsymbol{y}^{j'})) \tag{93}$$

will be non-increasing for sufficiently small t. That is, for sufficiently small $t > 0$, it holds that $\mathcal{S}(t) \leq \mathcal{S}(0)$. Notice that the argument holds for general $\boldsymbol{f}$ and u. Now, we use the perturbation property to conclude the proof by showing $\mathcal{S}(t) < \mathcal{S}(0)$ for sufficiently small $t > 0$. Since the flow map $\boldsymbol{\gamma}_t$ satisfies for some $y_i^J = y_{i'}^{J'}$, but

$$[\boldsymbol{\gamma}_t(\boldsymbol{y}^J)]_i \neq [\gamma_t(\boldsymbol{y}^{J'})]_{i'}. \tag{94}$$

Therefore, we can ensure that for sufficient small t,

$$\overline{s}(\boldsymbol{\gamma}_t(\boldsymbol{y}^J), \boldsymbol{\gamma}_t(\boldsymbol{y}^{J'})) < \overline{s}(\boldsymbol{y}^J, \boldsymbol{y}^{J'}), \tag{95}$$

Thus, it contradicts to the minimal choice of $\boldsymbol{\beta}$, since $\overline{s}(\boldsymbol{\gamma}_t \circ \boldsymbol{\beta}(X)) < \overline{s}(\boldsymbol{\beta}(X))$. This immediately yields that the minimal value of $\overline{s}(\boldsymbol{\beta}(X))$ is 0.

■

A.4.2 Proof of Lemma 32: Partially Ordering Lemma

In this proof, two kind of mappings will be used. The first kind is the symmetric invariant well function τ with zero interval $\mathbb{I}$ such that $\tau \in \mathrm{coor}(\mathcal{F})$. Since τ is S invariant, by the definition of the coor operator, the tensor-product mapping $\boldsymbol{\tau} \in \mathcal{F}$ such that $\boldsymbol{\tau}(\boldsymbol{x}) = \tau^{\otimes}(\boldsymbol{x}) = [\tau(\boldsymbol{x}), \tau(\boldsymbol{x}), \cdots, \tau(\boldsymbol{x})]$ is in $\mathcal{F}$. Recall that Lemma 32 has two additional requirements:

1. $\boldsymbol{\beta}(X)$ is well-perturbed, as defined in Lemma 31.
2. If $\boldsymbol{x} \in X$ is in general position, so is $\boldsymbol{\beta}(\boldsymbol{x})$.

Consider the following dynamical system for $\boldsymbol{x}$:

$$\frac{d}{dt}\boldsymbol{x} = \boldsymbol{\tau}(\boldsymbol{x}). \tag{96}$$

Clearly, the flow map $\boldsymbol{\phi}(\boldsymbol{\tau}, T)$ of $\boldsymbol{\tau}$ at any time horizon T satisfies $[\boldsymbol{\phi}(\boldsymbol{\tau}, T)\boldsymbol{x}]_i - [\boldsymbol{\phi}(\boldsymbol{\tau}, T)\boldsymbol{x}]_j = x_i - x_j$. Hence, the second requirement is always satisfied if we use (96) in our construction.

The other kind is the collection of all coordinate zooming functions. By Lemma 29, these are contained in $\overline{\mathcal{A}_{\mathcal{F}}}$. These preserve the order of all coordinates and thus the second requirement is also satisfied. Since we will only use these two kinds of mappings in our subsequent constructions, it suffices to prove that a composition of these two types of functions can be constructed to satisfy the well perturbed property.

Proof (Proof of Lemma 32)

Step 1 Consider $\mathcal{B}$, defined as the family of mappings consisting of all composition of $\boldsymbol{\phi}(\boldsymbol{\tau}, t)$ and coordinate zooming functions.

$$\begin{aligned}\mathcal{B} := \{&\boldsymbol{f}_1 \circ \boldsymbol{f}_2 \circ \cdots \circ \boldsymbol{f}_p : \\ &\text{each } \boldsymbol{f}_i \in \mathcal{A} \text{ is either } \boldsymbol{\phi}(\boldsymbol{\tau}, t_i) \text{ for some } t_i \text{ or a coordinate zooming function.}\}\end{aligned} \tag{97}$$

Obviously, all mappings in $\mathcal{B}$ will map each Q_{a} into itself. As a consequence, if $\boldsymbol{x}$ is not in general position, then $\boldsymbol{\alpha}(\boldsymbol{x})$ is not either. Thus, we may assign an order in $\{m\}$ such that $i \gg j$ if and only if

- $i > j$, if $\boldsymbol{x}^i$ and $\boldsymbol{x}^j$ are in general position, or
- $\boldsymbol{x}^i$ is not in general position, while $\boldsymbol{x}^j$ is in general position.

By Proposition 38, it holds that $\mathcal{B} \subset \overline{\mathcal{A}_{\mathcal{F}}}$. We first show that there exists $\boldsymbol{\beta} \in \mathcal{B}$ that $[\boldsymbol{\beta}(\boldsymbol{x}^i)]_1 > [\boldsymbol{\beta}(\boldsymbol{x}^j)]_1$ if $i \gg j$. For convenience, $i \ll j$ means $j \gg i$. By the definition of being well-perturbed, $i \ll j$ means that $x_1^i \neq x_1^j$.

Now, let us consider

$$b_{\min} = \min \left|\left\{(i,j) : i \ll j, [\boldsymbol{\beta}(\boldsymbol{x}^i)]_1 < [\boldsymbol{\beta}(\boldsymbol{x}^j)]_1, \boldsymbol{\beta} \in \mathcal{B} \text{ and } \boldsymbol{\beta}(X) \text{ is well perturbed.}\right\}\right| \tag{98}$$

and $\boldsymbol{\beta}$ achieves this minimum. It suffices to show $b_{\min} = 0$. Suppose not, take I and J such that $I \ll J$ and $[\boldsymbol{\beta}(\boldsymbol{x}^I)]_1 < [\boldsymbol{\beta}(\boldsymbol{x}^J)]_1$, and I and J are taken to minimize $[\boldsymbol{\beta}(\boldsymbol{x}^j)]_1 - [\boldsymbol{\beta}(\boldsymbol{x}^i)]_1$ under the aforementioned conditions. We show that there are no other $[\boldsymbol{\beta}(\boldsymbol{x}^k)]_1$ between them. In fact, if $[\boldsymbol{\beta}(\boldsymbol{x}^I)]_1 < [\boldsymbol{\beta}(\boldsymbol{x}^k)]_1 < [\boldsymbol{\beta}(\boldsymbol{x}^J)]_1$, then either $I < k$ or $k < J$ is satisfied, contradicting with the minimal choice of I and J.

We now give a direct construction to show that $\boldsymbol{\beta}$ is not minimal. With a little abuse of notation, We also use X to denote this new $\boldsymbol{\beta}(X)$. Since X is in a general position now, we have either $\min(x_k^J) < \min(x_k^I)$ or $\min(x_k^J) > \min(x_k^I)$.

Our construction is based on discussing them separately. By the definition of symmetric invariant well function, we can find $a > 1$ such that if $|x^I| > a$ and $x^j \in \mathbb{I}$ then $|\tau(\boldsymbol{x})| > \delta$. Since τ is continuous, we may assume that $|\tau(\boldsymbol{x})| \le \Delta$ for $|\boldsymbol{x}| \le 2a$ for some Δ.

- If $x_K^J := \min_k(x_k^J) < \min_k(x_k^I)$. We first choose a coordinate zooming function $u^{\otimes} \in \mathcal{A}$ such that

  1. $$|u(x_1^J) - u(x_1^I)| \le \varepsilon^2 \text{ and } |u(x_1^k) - u(x_1^l)| \ge \varepsilon \tag{99}$$
     for sufficiently small $\varepsilon < 1$. The detailed value of ε will be determined later.
  2. $$|u(x_i^j)| < 2a - 1, \forall i, j, \text{ and } u(x_i^J) > -a - 1, \forall j. \tag{100}$$
  3. $$(u(x_j^I) - \varepsilon, u(x_j^I) + \varepsilon) \subset \mathbb{I}, \forall j. \tag{101}$$

  With a slight abuse of notation, we use $\boldsymbol{x}^i$ to denote $u^{\otimes}(\boldsymbol{x}^i)$ for $i = 1, 2, \cdots, M$. Then, the flow $\boldsymbol{\phi}_{\boldsymbol{\tau}}$ and condition (101) ensure that $\boldsymbol{\phi}(\pm\boldsymbol{\tau}, t)(\boldsymbol{x}^I) = \boldsymbol{x}^I$ by the definition of $\mathbb{I}$. Next, we choose $\hat{\boldsymbol{\tau}} = \tau$ or $-\tau$ such that $[\boldsymbol{\phi}(\hat{\boldsymbol{\tau}}, \boldsymbol{x}^J, t)]_1$ is decreasing in t. By the construction of u in (100), when $t < \frac{\varepsilon}{\Delta}$, we have $[\boldsymbol{\phi}(\hat{\boldsymbol{\tau}}, t)\boldsymbol{x}^J]_1 - x_1^J \le \delta t$. Hence, we achieve our goal if $\delta t > \varepsilon^2$. Set $t \in (2\varepsilon^2\delta^{-1}, 3\varepsilon^2\delta^{-1})$ to be determined. We have $t \in \frac{\varepsilon}{\Delta}$ for $\varepsilon < \delta/(4\Delta)$. Since the set of t's such that $\boldsymbol{\phi}(\hat{\boldsymbol{\tau}}, t)X$ is in a general position is open and contains $t = 0$, such a t must be found, by condition (99).

- If $\min_k(x_k^J) > \min_k(x_k^I) =: x_K^I$. The arguments are similar except we exchange the role of I and J, and change decreasing into increasing.

In either scenario, we deduce that $\boldsymbol{\beta}$ is not minimal, thus $\boldsymbol{\beta}(X)$ must be partially ordered.

Step 2 Substituting X by $\boldsymbol{\beta}(X)$, we now assume that $x_1^i > x_1^j$ if $i \ll j$. We now show that there exists $\boldsymbol{\gamma} \in \mathcal{B}$ such that $\boldsymbol{\gamma}(X)$ is partially ordered. Suppose there exists $\boldsymbol{\alpha} \in \mathcal{B}$ such that

$$\boldsymbol{\alpha}(\boldsymbol{x}^1) \succ \boldsymbol{\alpha}(\boldsymbol{x}^2) \succ \cdots \boldsymbol{\alpha}(\boldsymbol{x}^I) \succ \boldsymbol{\alpha}(\boldsymbol{x}^k) \tag{102}$$

for $k = I + 1, \cdots, M$, while

$$[\boldsymbol{\alpha}(\boldsymbol{x}^1)]_1 > [\boldsymbol{\alpha}(\boldsymbol{x}^2)]_1 > \cdots [\boldsymbol{\alpha}(\boldsymbol{x}^m)]_1. \tag{103}$$

We give a direct construction of $\boldsymbol{\alpha}'$ such that the above inequality holds for $I+1$, provided $\boldsymbol{x}^{I+1}$ is in general position. Using this argument recurrently will yield the desired result.

Consider the case $I = 1$, and we also use $\boldsymbol{x}^i$ to denote $\boldsymbol{\alpha}(\boldsymbol{x}^i)$. Consider a coordinate zooming function such that only $u(x_1^1) > a$ is outside the interval $\mathbb{I}$. Then, the flow map $\boldsymbol{\gamma}_t = \boldsymbol{\phi}(\boldsymbol{\tau}, t) \circ u^{\otimes}$ will only move $\boldsymbol{x}^1$ while keeping other $\boldsymbol{x}^i$ fixed. At some time horizon $T < \infty$, we have $\boldsymbol{\gamma}_t(\boldsymbol{x}^1) \succ \boldsymbol{\gamma}_t(\boldsymbol{x}^i)$ for $i \neq 1$. Otherwise, we will obtain both $\boldsymbol{\gamma}_t(\boldsymbol{x}^1)$ is unbounded and $\min_i[\boldsymbol{\gamma}_t(\boldsymbol{x}^1)]_i \leq x_1^2$, which is a contradiction since $[\boldsymbol{\gamma}_t(\boldsymbol{x}^1)]_i - [\boldsymbol{\gamma}_t(\boldsymbol{x}^1)]_{i'}$ is a constant.

For the other I, we apply the same technique to find $\boldsymbol{\zeta} \in \mathcal{A}_{\mathcal{F}}$ such that for $\boldsymbol{\beta} = \boldsymbol{\zeta} \circ \boldsymbol{\alpha}$, we have $\boldsymbol{\beta}(\boldsymbol{x}^I) \succ \boldsymbol{\beta}(\boldsymbol{x}^k)$ for $k > I$. The only difference is that we need to modify $\boldsymbol{\zeta}$ such that it preserves the order of $\boldsymbol{x}^1$ to $\boldsymbol{x}^{I-1}$.

To do this, we need to move $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^{I-1}$ to an appropriate location so that modifications to the remaining coordinates do not induce a change of ordering for these coordinates. More precisely, suppose that $\boldsymbol{x}^{I-1}$ is in general position, and moreover we assume that $\boldsymbol{x}^{I-1} \in Q_{\mathsf{a}}$. Since $\boldsymbol{\zeta}$ must be a bijection, implying that we can choose $\boldsymbol{y}^{I-1} \in Q_{\mathsf{a}}$ such that $\boldsymbol{\zeta}(\boldsymbol{y}^{I-1}) \succ \boldsymbol{\beta}(\boldsymbol{x}^{I-1})$. Then, we choose a coordinate zooming function $u^{\otimes}$ to send $\boldsymbol{\alpha}(\boldsymbol{x}^{I-1})$ to $\boldsymbol{y}^{I-1}$. Let $\tilde{\boldsymbol{\beta}} = \boldsymbol{\zeta} \circ u^{\otimes} \circ \boldsymbol{\alpha}$. Then, the function $\tilde{\boldsymbol{\beta}}$ satisfies that $\tilde{\boldsymbol{\beta}}(\boldsymbol{x}^{I-1}) \succ \tilde{\boldsymbol{\beta}}(\boldsymbol{x}^I)$.

For the general case, using the same technique, we can sequentially determine the destination of $\boldsymbol{x}^{I-1}, \cdots, \boldsymbol{x}^1$, denoted as $\boldsymbol{y}^{I-1}, \cdots \boldsymbol{y}^1$, to fulfill the requirement.

■

To close this section, we give the proofs of Lemmas 33 and 34 that was used to prove Theorem 25.

A.4.3 Proof of Lemma 33

Proof (Proof of Lemma 33) It suffices to show that if a and b are $\mathcal{F}$ directly connected, then for given point $\boldsymbol{x} \in Q_{\mathsf{a}}$, there exists $\boldsymbol{\alpha} \in \mathcal{A}_{\mathcal{F}}$ such that $\boldsymbol{\alpha}(\boldsymbol{x}) \in Q_{\mathsf{b}}$. By direct connectivity there exists i, j such that $\boldsymbol{x} \in Q_{\mathsf{a}}$ we have $x_i > x_j$, and for $\boldsymbol{x} \in Q_{\mathsf{b}}$ we have $x_i < x_j$.

Suppose $\boldsymbol{z} \in \partial Q_{\mathsf{a}} \cap \partial Q_{\mathsf{b}}$ and $\boldsymbol{f} \in \mathcal{F}$ satisfies the assumption on direct connectivity, that is,

$$[\boldsymbol{f}(\boldsymbol{z})]_i \neq [\boldsymbol{f}(\boldsymbol{z})]_j. \tag{104}$$

By the continuity of $\boldsymbol{f}$, we also have $[\boldsymbol{f}(\boldsymbol{y})]_i \neq [\boldsymbol{f}(\boldsymbol{y})]_j$ for $\boldsymbol{y}$ in some neighborhood U of $\boldsymbol{z}$.

Let $\varepsilon < 1$ be a small constant whose value will be determined later, such that the ball $B(\boldsymbol{z}, 3\varepsilon) \in U$. Choose $\boldsymbol{y} \in Q_{\mathsf{a}}$ such that $y_i - y_j < \varepsilon^2/2$ while for other $(k, l) \neq (i, j)$, we require that $|y_k - y_l| > \varepsilon - \varepsilon^2/2$.

Consider a coordinate zooming function that $|u(x_k) - y_k| \leq \varepsilon^2/2$. Hence we have $u^{\otimes}(\boldsymbol{x}) \in B(\boldsymbol{z}, 2\varepsilon)$, and

1. $|u(x_i) - u(x_j)| \leq \varepsilon^2$,

2. For other pairs $(k,l) \neq (i,j)$, we have $|u(x_k) - u(x_l)| \geq \varepsilon$.

For $\boldsymbol{y} \in B(\boldsymbol{z}, 2\varepsilon)$ we can choose $a > 0$ such that $[\boldsymbol{f}(\boldsymbol{y})]_i > [\boldsymbol{f}(\boldsymbol{y})]_j + a$, and $|\boldsymbol{f}(\boldsymbol{y})| \leq \frac{1}{a}$, e.g. take $a < \min(\frac{1}{2}([\boldsymbol{f}(\boldsymbol{y})]_i - [\boldsymbol{f}(\boldsymbol{y})]_j), (|\boldsymbol{f}(\boldsymbol{y})| + 1)^{-1})$. Then, consider the following dynamics

$$\frac{d}{dt}\boldsymbol{y} = -\boldsymbol{f}(\boldsymbol{y}). \tag{105}$$

Simple estimation yields that the flow $\boldsymbol{\phi}(-\boldsymbol{f}, t)(u^{\otimes}(\boldsymbol{x})) \in Q_{\mathsf{b}}$ for $t \in (\frac{\varepsilon^2}{a}, \frac{a}{4}\varepsilon)$. We choose ε sufficiently small such that the interval is non-empty, and choose the mapping as $\boldsymbol{\phi}(-\boldsymbol{f}, t) \circ \boldsymbol{u}^{\otimes} \in \mathcal{A}_{\mathcal{F}}$.

■

A.4.4 Proof of Lemma 34

Proof (Proof of Lemma 34)

We consider the simplest case, when $\boldsymbol{\zeta}^{\circ} := \boldsymbol{\phi}(\boldsymbol{f}, T)$ is the flow map of some $\boldsymbol{f} \in \mathcal{F}$. The general case can be obtained by repeating the following argument. Denote by

$$\delta := \min\{|x_j^i - x_{j'}^{i'}| : x_j^i \neq x_{j'}^{i'}.\} \tag{106}$$

and there exists t_0 such that

$$\sup_{0 \leq t \leq t_0} \mathrm{Lip}[\boldsymbol{\phi}(\boldsymbol{f}, t) - id] \leq \frac{\delta}{4}. \tag{107}$$

Hence, consider $\boldsymbol{y} = \boldsymbol{\phi}(\boldsymbol{f}, t_0)\boldsymbol{x}$, then we have

$$|y_j^i - y_{j'}^{i'}| \geq \frac{\delta}{4}, \tag{108}$$

if $y_i^j \neq y_{i'}^{j'}$.

Therefore, $\boldsymbol{y}^1, \cdots, \boldsymbol{y}^m$ is partially ordered, and we have if $\boldsymbol{x}^i \in Q_{\mathsf{a}}$ then $\boldsymbol{y}^i \in Q_{\mathsf{a}}$ since the order of each coordinate does not change. Now consider the coordinate zooming function u, such that

$$|u(y_j^i) - x_j^i| \leq \frac{\delta}{8} \tag{109}$$

for $i \neq s$, and

$$|u(y_j^s) - y_j^s| \leq \varepsilon \tag{110}$$

for some small ε whose value will be determined later.

Therefore, we can apply $\boldsymbol{\phi}(\boldsymbol{f}, t_0)$ for $u^{\otimes}(\boldsymbol{y}^1), u^{\otimes}(\boldsymbol{y}^2), \cdots, u^{\otimes}(\boldsymbol{y}^m)$, and get the similar result. This iteration will break down after $[T/t_0] + 1$ iterations, and we denote by $\boldsymbol{\zeta}$ the final composited iteration. It suffices to control the error on $\boldsymbol{x}^s$, by a telescope expansion we have

$$|[\boldsymbol{\zeta}(\boldsymbol{x}^s)]_j - x_j^s| \leq \varepsilon L([T/t_0] + 1), \tag{111}$$

where $L = \sup_{0\le t\le t_0} \operatorname{Lip} \boldsymbol{\phi}(\boldsymbol{f}, t) \le 1 + \frac{\delta}{4}$. Choose $\varepsilon < 1$ such that the right-hand side is not greater than $\frac{1}{4}\min_{i\neq j} |x_i^s - x_j^s|$, which guarantees that $\boldsymbol{\zeta}(\boldsymbol{x}^s)$ and $\boldsymbol{x}^s$ are in the same Q_{a}. ■

Appendix B. Basic Properties of Equivariant Mappings

In this section, we collect a number of auxiliary technical results required in the proofs of the main results.

B.1 Closure Properties of Invariant and Equivariant Mappings

Proposition 35 *If $\mathcal{F}$ is G equivariant, then so is $\mathcal{A}_{\mathcal{F}}$.*

Proof It suffices to show that the flow map of a equivariant mapping is equivariant, since function composition preserves equivariance. Consider the ODE system $d\boldsymbol{x}/dt = \boldsymbol{f}(\boldsymbol{x})$. The Picard iteration sequence will converge to the solution of ODE, since $\boldsymbol{f}$ is Lipschitz. Clearly each function on Picard iteration is equivariant, then it follows from the fact that $\boldsymbol{x}$ is Lipschitz that the flow map is also equivariant. Precisely, given $\boldsymbol{x}_0(t) = \boldsymbol{0}$, define the Picard iteration as follows:

$$\boldsymbol{x}_n(t) = \boldsymbol{x}_0 + \int_0^t \boldsymbol{f}(\boldsymbol{x}_{n-1}(s))ds \tag{112}$$

for $n = 1, 2, \cdots$. Define the flow maps for the Picard iterates as

$$\boldsymbol{\phi}_n(\boldsymbol{f}, t) := \boldsymbol{x}_0 \mapsto \boldsymbol{x}_n(t). \tag{113}$$

By induction, each $\boldsymbol{\phi}_n(\boldsymbol{f}, t)$ is G equivariant, and $\boldsymbol{\phi}_n(\boldsymbol{f}, t) \to \boldsymbol{\phi}(\boldsymbol{f}, t)$ uniformly in any compact set $K \subset \mathbb{R}^n$, therefore by Proposition 36, the mapping $\boldsymbol{\phi}(\boldsymbol{f}, t)$ is G equivariant. ■

Proposition 36 (Closure Property of Invariant Functions) *Suppose $\boldsymbol{F} \in C(\mathbb{R}^n, \mathbb{R}^m)$ and for any compact $K \subset \mathbb{R}^n$, tolerance $\varepsilon > 0$, there exists a G invariant function $\hat{\boldsymbol{F}} \in C(\mathbb{R}^n)$ such that $\|\boldsymbol{F} - \hat{\boldsymbol{F}}\|_{L^p(K)} < \varepsilon$. Then $\boldsymbol{F}$ is G invariant.*

Moreover, suppose $\boldsymbol{\varphi} \in C(\mathbb{R}^n; \mathbb{R}^n)$ and for any compact $K \subset \mathbb{R}^n$, tolerance $\varepsilon > 0$, there exists a G equivariant mapping $\hat{\boldsymbol{\varphi}}$ such that $\|\boldsymbol{\varphi} - \hat{\boldsymbol{\varphi}}\|_{L^p(K)} < \varepsilon$. Then $\boldsymbol{\varphi}$ is G equivariant.

Proof Given $\mathsf{g} \in \mathsf{G}$, compact set K and $\varepsilon > 0$, we choose $\hat{\boldsymbol{F}}$ as stated. Since $\hat{\boldsymbol{F}}$ is G invariant, we have

$$\begin{aligned} \|\boldsymbol{F}(\boldsymbol{x}) - \boldsymbol{F}(\mathsf{g}(\boldsymbol{x}))\|_{L^p(K)} &\le \|\boldsymbol{F}(\boldsymbol{x}) - \hat{\boldsymbol{F}}(\boldsymbol{x})\|_{L^p(K)} + \|\hat{\boldsymbol{F}}(\boldsymbol{x}) - \hat{\boldsymbol{F}}(\mathsf{g}(\boldsymbol{x}))\|_{L^p(K)} \\ &\quad + \|\hat{\boldsymbol{F}}(\mathsf{g}(\boldsymbol{x})) - \boldsymbol{F}(\mathsf{g}(\boldsymbol{x}))\|_{L^p(K)} \\ &\le 2\varepsilon. \end{aligned} \tag{114}$$

This holds because g is measure-preserving, so $\|\hat{\boldsymbol{F}}\circ\mathsf{g}-\boldsymbol{F}\circ\mathsf{g}\|_{L^p(K)} = \|\hat{\boldsymbol{F}}-\boldsymbol{F}\|_{L^p(K)}$. Note that both $\boldsymbol{F}$ and $\boldsymbol{F}\circ\mathsf{g}$ are continuous, and $\varepsilon > 0$ is arbitrary, yielding that $\boldsymbol{F} = \boldsymbol{F}\circ\mathsf{g}$. Since this holds for all $\mathsf{g}\in\mathsf{G}$, we conclude that $\boldsymbol{F}$ must be G invariant.

The proof of the second part is similar. Given $\mathsf{g}\in\mathsf{G}$, and compact set K and $\varepsilon > 0$, we choose $\hat{\boldsymbol{\varphi}}$ as stated. Then we have

$$\begin{aligned}\|\mathsf{g}\circ\boldsymbol{\varphi}-\boldsymbol{\varphi}\circ\mathsf{g}\|_{L^p(K)} \leq\ & \|\mathsf{g}\circ\boldsymbol{\varphi}-\mathsf{g}\circ\hat{\boldsymbol{\varphi}}\|_{L^p(K)} + \|\mathsf{g}\circ\hat{\boldsymbol{\varphi}}-\hat{\boldsymbol{\varphi}}\circ\mathsf{g}\|_{L^p(K)} + \|\hat{\boldsymbol{\varphi}}\circ\mathsf{g}-\boldsymbol{\varphi}\circ\mathsf{g}\|_{L^p(K)}\\ \leq\ & 2\varepsilon,\end{aligned} \tag{115}$$

since $\hat{\boldsymbol{\varphi}}$ is G equivariant. Note that both $\mathsf{g}\circ\boldsymbol{\varphi}$ and $\boldsymbol{\varphi}\circ\mathsf{g}$ are continuous, and ε can be arbitrarily chosen, yielding that $\boldsymbol{\varphi}\circ\mathsf{g} = \mathsf{g}\circ\boldsymbol{\varphi}$. Hence $\boldsymbol{\varphi}$ must be G equivariant. ■

An immediate consequence is that if $\mathsf{G}\leq\mathsf{H}$, and our hypothesis space consists of H invariant functions, then functions that are G invariant but not H invariant cannot be approximated to arbitrary precision. Although we only consider finite permutation groups here, similar arguments yield that the same limitation arises for continuous groups, raising a significant problem in designing equivariant/invariant neural networks, e.g. under SO or SE symmetry (Weiler et al., 2018). Generally, this suggests that the construction of invariant and equivariant architectures will be much more challenging if the structure of G is complex, and using composition gives a convenient way to build complexity while preserving symmetry.

Using $\mathrm{coor}(\mathcal{F})$ we can make the connection between well function and coordinate zooming function. See Lemma 29 below.

B.2 Convex Hull Property of Equivariant Mapping

In this subsection, we show some basic properties of the convex hull closure. These propositions imply that Condition 2 in Theorem 14 can be reduced to requiring $\mathrm{coor}(\mathcal{F})$ to contains a symmetric invariant well function. The argument goes as follows: by Proposition 37, the conditions on Theorem 25 holds for $\overline{\mathrm{CH}}(\mathcal{F})$. Whence, by Theorem 25 and Proposition 15, we can deduce that the assertion in Theorem 14 holds for $\overline{\mathrm{CH}}(\mathcal{F})$. Applying Proposition 39, we obtain that the assertion also holds for $\mathcal{F}$.

Proposition 37 *Given a group* G*, the* coor *operator commutes with* $\overline{\mathrm{CH}}$ *operator, that is,*

$$\mathrm{coor}(\overline{\mathrm{CH}}(\mathcal{F})) = \overline{\mathrm{CH}}(\mathrm{coor}(\mathcal{F})). \tag{116}$$

Proof Notice that by Proposition 36 we know that $\overline{\mathrm{CH}}(\mathcal{F})$ is equivariant. The commutation between coor and CH is obvious. We only prove that

$$\mathrm{coor}(\overline{\mathcal{F}}) = \overline{\mathrm{coor}(\mathcal{F})}. \tag{117}$$

Suppose $f_1\in\mathrm{coor}(\overline{\mathcal{F}})$, which means there exists $\boldsymbol{f}\in\overline{\mathcal{F}}$ such that $\mathrm{coor}(\boldsymbol{f}) = f_1$. For any compact set K and tolerance ε, by the definition of $\overline{\mathcal{F}}$, there exists $\boldsymbol{g}\in\mathcal{F}$ such that $\|\boldsymbol{f}-\boldsymbol{g}\|_{C(K)}\leq\varepsilon$. Hence we have $\|f_1-\mathrm{coor}(\boldsymbol{g})\|_{C(K)}\leq\varepsilon$, yielding that $f_1\in\overline{\mathrm{coor}(\mathcal{F})}$.

Conversely, if $f_1 \in \overline{\mathrm{coor}(\mathcal{F})}$, then for any compact set K (assumed to be invariant) and tolerance $\varepsilon > 0$, we can find $g_1 \in \mathrm{coor}(\mathcal{F})$, then by the definition of coor operator, we can construct $\boldsymbol{f}$ and $\boldsymbol{g} \in \mathcal{F}$, such that $\|\boldsymbol{f} - \boldsymbol{g}\|_{C(K)} \leq \varepsilon$. Therefore $f_1 \in \mathrm{coor}(\overline{\mathcal{F}})$. Hence the equality in (117) holds.

■

Proposition 38 *If $\boldsymbol{f}, \boldsymbol{g} \in \overline{\mathcal{A}_{\mathcal{F}}}$, then so does $\boldsymbol{f} \circ \boldsymbol{g}$.*

Proof This follows immediate from the Lipschitz property of $\mathcal{F}$. ■

Proposition 39 *For any $\hat{\boldsymbol{\varphi}} \in \mathcal{A}_{\overline{\mathrm{CH}(\mathcal{F})}}$, compact set $K \in \mathbb{R}^n$ and tolerance $\varepsilon > 0$, we can find a mapping $\boldsymbol{\varphi} \in \mathcal{A}_{\mathcal{F}}$, such that $\|\boldsymbol{\varphi} - \hat{\boldsymbol{\varphi}}\|_{L^p(K)} \leq \varepsilon$.*

Proof The proof of the one-dimensional is found in Li et al. (2023), and the proof of higher-dimensional cases are analogous, hence omitted. ■

Appendix C. Verification of the Proposed Architectures

In this section, we provide a direct verification that the proposed architectures in Section 4 satisfy the assumptions of our universal approximation theorems, thereby guarantee their UAP for their respective symmetry groups G. Clearly, Condition 1 and 2 of Theorem 25 are all satisfied as a consequence of Remark 10. It suffices to check the perturbation assumption and G transversal transitivity on each case, to show that $\mathcal{F}$ resolves G (see Definition 13). For convenience, we recall the definition of each $\mathcal{F}$ and G, and σ denotes a well function. The following facts that hold for any well function σ will be frequently used:

1. For given x and $y \in \mathbb{R}$, we can find $b \in \mathbb{R}$ such that $\sigma(x + b) = 0$ but $\sigma(y + b) \neq 0$.
2. If $x_1, \cdots, x_n$ and $y_1, \cdots, y_n$ satisfy that
$$\sum_{i=1}^{n} \sigma(px_i + q) = \sum_{i=1}^{n} \sigma(py_i + q) \tag{118}$$
for all $p, q \in \mathbb{R}$, then it holds that $\{x_i : i = 1, 2, \cdots, n\} = \{y_i : i = 1, 2, \cdots n\}$. Hereafter, the sets are understood as multi-sets.

C.1 Shift Invariant Cases

We show that each $\mathcal{F}_{*,i}$ resolves T, where

$$\begin{aligned} \mathcal{F}_{*,1} :=& \{v\sigma(\boldsymbol{w} * \boldsymbol{x} + b\mathbf{1}) : \boldsymbol{w} \in \mathbb{R}^{d_1 \times \cdots \times d_n}, v, b \in \mathbb{R}\}, \\ \mathcal{F}_{*,2} :=& \{\boldsymbol{w} * \sigma(v\boldsymbol{x} + b\mathbf{1}), \boldsymbol{w} \in \mathbb{R}^{d_1 \times \cdots \times d_n}, v, b \in \mathbb{R}\}. \end{aligned} \tag{119}$$

Proof (Proof of Corollary 20) For the perturbation assumption, without loss of generality, we assume that $x_1 = y_1$. It suffices to notice that for two point $\boldsymbol{x}$ and $\boldsymbol{y}$ such that they are T distinct, there exists i such that $x_i \neq y_i$. We choose $\boldsymbol{w} = \boldsymbol{e}_i$, whose value is 1 in the i coordinate, and 0 in other coordinates. Then, we choose an appropriate b such that $\sigma(x_i + b) = 0$ while $\sigma(y_i + b) \neq 0$, therefore

$$[\boldsymbol{w} * \sigma(\boldsymbol{x} + b\mathbf{1})]_1 \neq [\boldsymbol{w} * \sigma(\boldsymbol{y} + b\mathbf{1})]_1 \tag{120}$$

and

$$[\sigma(\boldsymbol{w} * \boldsymbol{x} + b\mathbf{1})]_1 \neq [\sigma(\boldsymbol{w} * \boldsymbol{y} + b\mathbf{1})]_1. \tag{121}$$

Hence we show both $\mathcal{F}_{*,i}$ satisfy the perturbation assumption.

For transversal transitivity, we show that for each $\mathsf{a}, \mathsf{b} \in \mathsf{S}$, a and b are $\mathcal{F}$ directly connected. For simplicity, we consider the case when

$$Q_{\mathsf{a}} = \{x_1 > x_2 > \cdots > x_n\} \tag{122}$$

and

$$Q_{\mathsf{b}} = \{x_2 > x_1 > \cdots > x_n\}, \tag{123}$$

and the other cases are similar. In this case, we can choose $\boldsymbol{z} \in \partial Q_{\mathsf{a}} \cap \partial Q_{\mathsf{b}}$ such that $z_2 \neq z_3$, We choose $\boldsymbol{w} = \boldsymbol{e}_2$, whose value is 1 in second coordinate, and 0 in other coordinates. Then, by a similar approach as what we did in the proof of Lemma 31, we can prove that a and b are $\mathcal{F}$ directly connected.

■

C.2 Full Permutation Invariant Cases

We show that each $\mathcal{F}_{s,i}$ $(i = 1, 2)$ resolves S, where

$$\mathcal{F}_{s,1} := \{a\sigma(w\boldsymbol{x} + v\Sigma\boldsymbol{x} + b\mathbf{1}) : a, w, v, b \in \mathbb{R}\}, \tag{124}$$

$$\mathcal{F}_{s,2} := \{a\sigma(w\boldsymbol{x} + c\mathbf{1}) + b\Sigma(\sigma(v\boldsymbol{x}) + d\mathbf{1}) : a, b, c, d, w, v \in \mathbb{R}\}. \tag{125}$$

Proof (Proof of Corollary 21) Since any transversal only contains one element, it suffices to show $\mathcal{F}_{s,i}$ satisfies the perturbation assumption. For $\mathcal{F}_{s,1}$, if $\boldsymbol{x}$ and $\boldsymbol{y}$ are S distinct, then there exists a function $u^{\otimes} \in \mathcal{F}$ such that

$$u(x_1) + u(x_2) \cdots + u(x_d) \neq u(y_1) + u(y_2) \cdots + u(y_d). \tag{126}$$

For $\mathcal{F}_{s,1}$, we set $a = 1, w = 0$, and choose a suitable b as before such that

$$\sigma(\boldsymbol{\Sigma} u^{\otimes}(\boldsymbol{x}) + b\mathbf{1}) \neq \sigma(\boldsymbol{\Sigma} u^{\otimes}(\boldsymbol{y}) + b\mathbf{1}). \tag{127}$$

For $\mathcal{F}_{s,2}$, we set $a = w = 0, b = 1$, and the rest is similar to $\mathcal{F}_{s,1}$.

■

Additionally, we prove the following result for Janossy pooling (with respect to a Lipschitz function ϕ) for completeness.

Proposition 40 *The following control family*

$$\mathcal{F}_{s,Ja,1} = \{v\sigma(a\boldsymbol{x} + b\sum_{i=1}^{n}\phi(cx_i + d) + c\mathbf{1})\} \tag{128}$$

contains a symmetric well function and satisfies the perturbation property provided that there exists a neighborhood U of $\phi(0)$, such that $\phi^{-1}(U)$ is bounded.

Proof We first prove that the control family (128) contains a symmetric invariant well function, by showing the following function $f = \sigma(k[\frac{1}{n}\sum_{i=1}^{n}\phi(x_i) - \phi(0)])$ for a suitable k is a symmetric invariant well function. Without loss of generality, we assume that σ is a well function such that $\sigma(z) = 0$ for $|z| < r$ and $\sigma(z) \neq 0$ for $|z| > R$. The assumption on ϕ now implies that there exists M and δ such that if $|z| > M$ then $|\phi(z) - \phi(0)| > \delta$. Thus, we take $k = R/\delta$, so that f is a symmetric well function with $a = r/(k\,\mathrm{Lip}\,\phi)$ and $m = M$.

We now prove the perturbation property. It suffices to show that if $\boldsymbol{x}$ and $\boldsymbol{y}$ are S distinct, then there exists a function $u^{\otimes}$ such that

$$\phi(u(x_1)) + \phi(u(x_2)) + \cdots + \phi(u(x_n)) \neq \phi(u(y_1)) + \phi(u(y_2)) + \cdots + \phi(u(y_n)). \tag{129}$$

We prove it by contradiction. Suppose that

$$\phi(u(x_1)) + \phi(u(x_2)) + \cdots + \phi(u(x_n)) = \phi(u(y_1)) + \phi(u(y_2)) + \cdots + \phi(u(y_n)) \tag{130}$$

holds for all coordinate zooming functions but $x_1 > y_1$. Then, we can carefully choose u such that $u(x_2), \cdots, u(x_n)$ and $u(y_1), \cdots, u(y_n) \in B(u(0), \varepsilon)$ for some small ε, such that $\phi(u(x_1)) \in U$. However, since the value of $u(x_1)$ can be arbitrarily chosen, the assumption yields the contradiction. As a result, the control family (128) satisfies the perturbation property. ■

C.3 Product Permutation Invariant Cases

We prove that

$$\mathcal{F}_{s,2D,1} = \{v\sigma(w_0\boldsymbol{x} + w_{r,1}\Sigma_{r,1}\boldsymbol{x} + w_{c,1}\Sigma_{c,1}\boldsymbol{x} + c\mathbf{1}) : w_0, w_{r,1}, w_{c,1}, c \in \mathbb{R}\} \tag{131}$$

resolves S_{2D}, and the higher dimensional cases are similar.

Proof (Proof of Corollary 23)

We now check the perturbation property. As before, it suffices to prove that there exists (i_1, j_1) and (i_2, j_2), real numbers c_1, c_2, and an increasing function u, such that $x_{i_1,j_1} = y_{i_2,j_2}$, but

$$c_1\sum_{i'=1}^{d_1}u(x_{i',j_1}) + c_2\sum_{j'=1}^{d_2}u(x_{i_1,j'}) \neq c_1\sum_{i'=1}^{d_1}u(y_{i',j_2}) + c_2\sum_{j'=1}^{d_2}u(y_{i_2,j'}). \tag{132}$$

Suppose that, for all (i_1, j_1) and (i_2, j_2) such that $x_{i_1,j_1} = y_{i_2,j_2}$, we have

$$c_1 \sum_{i'=1}^{d_1} u(x_{i',j_1}) + c_2 \sum_{j'=1}^{d_2} u(x_{i_1,j'}) = c_1 \sum_{i'=1}^{d_1} u(y_{i',j_2}) + c_2 \sum_{j'=1}^{d_2} u(y_{i_2,j'}). \tag{133}$$

holds for all choice of c_1, c_2, and increasing functions u. It follows from (133) that

$$\sum_{i'=1}^{d_1} u(x_{i',j_1}) = \sum_{i'=1}^{d_1} u(y_{i',j_2}) \tag{134}$$

and

$$\sum_{j'=1}^{d_2} u(x_{i_1,j'}) = \sum_{j'=1}^{d_2} u(y_{i_2,j'}). \tag{135}$$

Since u is an arbitrary increasing function, we deduce that the equalities

$$\{x_{i,j_1} : i = 1, 2, \cdots, d_1\} = \{y_{i,j_2} : i = 1, 2, \cdots, d_1\} \tag{136}$$

and

$$\{x_{i_1,j} : j = 1, 2, \cdots, d_2\} = \{y_{i_2,j} : j = 1, 2, \cdots, d_2\}. \tag{137}$$

After a permutation, we may assume that $x_{1,1} = y_{1,1}$ and therefore $x_{i,1} = y_{i,1}$ and $x_{1,j} = y_{1,j}$ for all i, j.

Suppose now, $x_{I,J} \neq y_{I,J}$, then there must exist $i \neq I$ such that $x_{I,J} = y_{i,J}$ and $j \neq J$ such that $x_{I,J} = y_{I,j}$. Therefore, $\boldsymbol{y}$ is not in general position. Similarly, we can deduce that $\boldsymbol{x}$ is not in general position, contradicting the assumption $\overline{s}(\boldsymbol{x}, \boldsymbol{y}) > 0$.

For transversal transitivity, it suffices to show for example that

$$Q_{\mathsf{a}} = \{x_{1,1} > x_{1,2} \cdots, > x_{d_1,d_2}\} \tag{138}$$

and

$$Q_{\mathsf{b}} = \{x_{1,2} > x_{1,1} \cdots, > x_{d_1,d_2}\}. \tag{139}$$

are $\mathcal{F}_{s,2D,1}$-connected. Clearly, we can choose $\boldsymbol{z} \in \partial Q_{\mathsf{a}} \cap \partial Q_{\mathsf{b}}$ such that $\sum_{j=1}^{d_2} z_{1,j} \neq \sum_{j=1}^{d_2} z_{2,j}$. Therefore, we can show the transversal transitivity. ■

Appendix D. Temporal Discretization

Our main result applies to continuous-time dynamics, which corresponds to a continuum idealization of practical deep neural networks (E, 2017; Haber and Ruthotto, 2017; Li et al., 2018). It is thus natural to consider the implications of our results for the discrete case in applications. In this section, we discuss how temporal discretization of continuous-time flows can inherit approximation properties.

In the following, we consider for simplicity only single step integrators for discretizing the continuous flow. This includes but is not limited to the important example of forward Euler discretization, which corresponds to the ResNet family of architectures (He et al., 2016). The numerical integrator is abstracted as follows: for a given $\boldsymbol{f} \in \mathcal{F}$, for each time interval of size t we define the numerical scheme as a mapping $\hat{\boldsymbol{\phi}}(\boldsymbol{f}, t)(\cdot) : \mathbb{R}^n \to \mathbb{R}^n$, serving as an approximation of the flow map of the continuous dynamics $\boldsymbol{\phi}(\boldsymbol{f}, t)(\cdot)$. For a time horizon T, the discrete flow with respect to the partition $\Delta : 0 = t_0 < t_1 < \cdots < t_s = T$ is given as

$$\hat{\boldsymbol{\phi}}(\boldsymbol{f}, \Delta) := \hat{\boldsymbol{\phi}}(\boldsymbol{f}, t_s - t_{s-1}) \circ \hat{\boldsymbol{\phi}}(\boldsymbol{f}, t_{s-1} - t_{s-2}) \cdots \hat{\boldsymbol{\phi}}(\boldsymbol{f}, t_1 - t_0) \tag{140}$$

Definition 41 (Convergence of Numerical Scheme) *We say the numerical scheme is convergent if*

1. $\hat{\boldsymbol{\phi}}(\boldsymbol{f}, t) \to \mathrm{Id}$ *uniformly in any compact set* $K \subset \mathbb{R}^n$, *as* $t \to 0$, *where* Id *is the identity mapping.*

2. *For any* T,
$$\hat{\boldsymbol{\phi}}(\boldsymbol{f}, \Delta) \to \boldsymbol{\phi}(\boldsymbol{f}, T) \tag{141}$$
uniformly in any compact set $K \subset \mathbb{R}^n$, *as* $|\Delta| := \max(t_i - t_{i-1}) \to 0$.

3. *There exists a continuous function* $g : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$ *such that* $|\hat{\boldsymbol{\phi}}(\boldsymbol{f}, \Delta)(\boldsymbol{x})| \leq g(\boldsymbol{x}, T)$.

As an example, we consider the forward Euler scheme, where

$$\hat{\boldsymbol{\phi}}(\boldsymbol{f}, t)(\boldsymbol{x}) := \boldsymbol{x} + t\boldsymbol{f}(\boldsymbol{x}). \tag{142}$$

Standard numerical analysis shows that the forward Euler scheme is convergent under fairly general conditions on $\boldsymbol{f}$, e.g. Lipschitz continuity, see Atkinson (2008).

The main result on numerical discretization below shows that the composition of convergent discretizations can be used to approximate flow maps.

Theorem 42 (Temporal Discretization of Compositional Flow Map) *Consider a target mapping* $\boldsymbol{\varphi}$ *in the form of a composition of* l *flow maps* $\boldsymbol{\varphi} = \boldsymbol{\phi}(\boldsymbol{f}^l, T_l - T_{l-1}) \circ \cdots \circ \boldsymbol{\phi}(\boldsymbol{f}^1, T_1 - T_0)$, *shortened as*

$$\boldsymbol{\varphi} = \boldsymbol{\phi}_L \circ \cdots \circ \boldsymbol{\phi}_1. \tag{143}$$

Suppose that for each l, $\hat{\boldsymbol{\phi}}(\boldsymbol{f}^l, \cdot)$ *is a convergent numerical integrator, as defined in Definition 41.*

Consider the following function with respect to a partition $\Delta : 0 = t_0 < t_1 < \cdots < t_s = T$

$$\hat{\boldsymbol{\varphi}}(\cdot, \Delta) = \hat{\boldsymbol{\phi}}(\boldsymbol{f}^{l_s}, t_s - t_{s-1}) \circ \cdots \circ \hat{\boldsymbol{\phi}}(\boldsymbol{f}^{l_1}, t_1 - t_0), \tag{144}$$

where $l_s = \min\{l : T_l > s\}$. *Then,* $\hat{\boldsymbol{\varphi}}(\cdot, \Delta) \to \boldsymbol{\varphi}(\cdot)$ *uniformly on compact sets as* $|\Delta| \to 0$.

As a consequence of this result, we can conclude that for any $\mathcal{H}_{\text{ode}}$ possessing invariant UAP, any discrete architecture corresponding to a convergent discretization inherits the UAP.

Corollary 43 *Suppose for a given control family $\mathcal{F}$ and a terminal family $\mathcal{G}$ that*

$$\mathcal{H}_{ode} := \{\boldsymbol{g} \circ \boldsymbol{\varphi} : \boldsymbol{g} \in \mathcal{G}, \boldsymbol{\varphi} \in \mathcal{A}_{\mathcal{F}}\} \tag{145}$$

possesses UAP. Moreover, suppose for each $\boldsymbol{f} \in \mathcal{F}$, the numerical scheme $\hat{\boldsymbol{\phi}}_{\boldsymbol{f}}$ is convergent. We define the discretized attainable set

$$\hat{\mathcal{A}}_{\mathcal{F}} := \{\hat{\boldsymbol{\phi}}(\boldsymbol{f}^l, t_l - t_{l-1}) \circ \cdots \hat{\boldsymbol{\phi}}(\boldsymbol{f}^1, t_1 - t_0) \ : \ l \geq 1, 0 = t_0 < t_1 < \cdots < t_l, \boldsymbol{f}^l \in \mathcal{F}\}. \tag{146}$$

Then, the discretized hypothesis space

$$\hat{\mathcal{H}} := \{\boldsymbol{g} \circ \hat{\boldsymbol{\varphi}} : \boldsymbol{g} \in \mathcal{G}, \hat{\boldsymbol{\varphi}} \in \hat{\mathcal{A}}_{\mathcal{F}}\}, \tag{147}$$

corresponding to the deep neural network architecture $\boldsymbol{x}^{l+1} = \hat{\boldsymbol{\phi}}(\boldsymbol{f}^l, t_{l+1} - t_l)(\boldsymbol{x}^l)$ possesses the UAP.

In addition, if $\mathcal{G}$ is G invariant, $\mathcal{F}$ is G equivariant, and $\hat{\boldsymbol{\phi}}_{\boldsymbol{f}}$ is G equivariant for each $\boldsymbol{f} \in \mathcal{F}$, then $\hat{\mathcal{H}}$ possesses the G UAP.

Remark 44 *The first part of the above corollary extends the approximations results in Li et al. (2023) for continuous dynamics to all convergent discretizations. The second part covers symmetry considerations.*

Proof We prove an equivalent result, which does not require $t_s = T$, and the convergence holds as $|\Delta| \to 0$ and $|t_s - T| \to 0$. For convenience, we partition the product in (144) into L products $\boldsymbol{\pi}_L \circ \cdots \circ \boldsymbol{\pi}_1$, where $\boldsymbol{\pi}_i$ is the composition of $\hat{\boldsymbol{\phi}}(\boldsymbol{f}^{l_j}, t_j - t_{j-1})$ such that $l_j = i$.

Without loss of generality, we assume that $K = [-\kappa, \kappa]^n$. We begin by induction on L. We first prove the base case. By assumption, it holds that $t_s < T$, then by definition, for ε there exists δ, such that

$$|\hat{\boldsymbol{\phi}}(\boldsymbol{f}, T - t_s) \circ \hat{\boldsymbol{\phi}}(\boldsymbol{f}, \Delta) - \boldsymbol{\phi}(\boldsymbol{f}, T)| \leq \varepsilon \tag{148}$$

as $|\Delta| \leq \delta$ and $|T - t_s| \leq \delta$. A direct calculation yields

$$|\hat{\boldsymbol{\phi}}(\boldsymbol{f}, \Delta)\boldsymbol{x} - \boldsymbol{\phi}(\boldsymbol{f}, T)\boldsymbol{x}| \leq \varepsilon + |(\hat{\boldsymbol{\phi}}(\boldsymbol{f}, T - t_s) - \mathrm{Id}) \circ \hat{\boldsymbol{\phi}}(\boldsymbol{x}, \Delta)|. \tag{149}$$

The result then follows from the definition of convergence.

Suppose that for $\varepsilon > 0$, there exists a sufficiently small δ, if $\max(t_{i+1} - t_i) \leq \delta$. Then,

$$|\boldsymbol{\pi}_{L-1} \circ \cdots \circ \boldsymbol{\pi}_1(\boldsymbol{x}) - \boldsymbol{\phi}_{L-1} \circ \cdots \circ \boldsymbol{\phi}_1(\boldsymbol{x})| \leq \varepsilon. \tag{150}$$

The key idea is to perform a telescope decomposition, i.e., for $\boldsymbol{x} \in K$, the estimation

$$|(\boldsymbol{\pi}_L \circ \boldsymbol{\Pi}_L - \boldsymbol{\phi}_L \circ \boldsymbol{\Phi}_L)\boldsymbol{x}| \leq |(\boldsymbol{\pi}_L \circ \boldsymbol{\Pi}_L - \boldsymbol{\phi}_L \circ \boldsymbol{\Pi}_L)\boldsymbol{x}| + |(\boldsymbol{\phi}_L \circ \boldsymbol{\Pi}_L - \boldsymbol{\phi}_L \circ \boldsymbol{\Phi}_L)\boldsymbol{x}| \tag{151}$$

where

$$\mathbf{\Pi}_L = \boldsymbol{\pi}_{L-1} \circ \boldsymbol{\pi}_{L_2} \circ \cdots \circ \boldsymbol{\pi}_1, \tag{152}$$

and

$$\mathbf{\Phi}_L = \boldsymbol{\phi}_{L-1} \circ \boldsymbol{\phi}_{L-2} \circ \cdots \circ \boldsymbol{\phi}_1. \tag{153}$$

By induction, we have $|\Pi_L \boldsymbol{x} - \Phi_L \boldsymbol{x}| \leq \varepsilon < \kappa$. Consider

$$\tilde{K} := \{\boldsymbol{x} : \operatorname{dist}(\boldsymbol{x}, \Phi_L(K)) \leq \varepsilon\}. \tag{154}$$

Then, by the definition of convergence on $\tilde{K}$, for sufficiently small δ, if $\max(t_{i+1} - t_i) \leq \delta$, we have

$$|(\pi_L \circ \Pi_L - \phi_L \circ \Pi_L)\boldsymbol{x}| \leq \varepsilon. \tag{155}$$

For the second term, just using the Lipschitz continuity of Φ_L, then

$$|(\phi_L \circ \Pi_L - \phi_L \circ \Phi_L)\boldsymbol{x}| \leq \operatorname{Lip} \Phi_L \varepsilon. \tag{156}$$

Combining both we conclude the result.

■

Bibliography

V. I. Arnold. *Ordinary differential equations*. 1973.

Kendall E Atkinson. *An introduction to numerical analysis*. John wiley & sons, 2008.

Aharon Azulay and Yair Weiss. Why do deep convolutional networks generalize so poorly to small image transformations? *arXiv preprint arXiv:1805.12177*, 2018.

Chenglong Bao, Qianxiao Li, Cheng Tai, Lei Wu, and Xueshuang Xiang. Approximation analysis of convolutional neural networks. *Submitted.*, 2019.

Benjamin Bloem-Reddy and Yee Whye Teh. Probabilistic Symmetries and Invariant Neural Networks. page 61, 2020.

Alexander Bogatskiy, Brandon Anderson, Jan Offermann, Marwah Roussi, David Miller, and Risi Kondor. Lorentz group equivariant neural network for particle physics. In *International Conference on Machine Learning*, pages 992–1002. PMLR, 2020.

Helmut Bolcskei, Philipp Grohs, Gitta Kutyniok, and Philipp Petersen. Optimal approximation with sparsely connected deep neural networks. *SIAM Journal on Mathematics of Data Science*, 1(1):8–45, 2019.

Keith T Butler, Daniel W Davies, Hugh Cartwright, Olexandr Isayev, and Aron Walsh. Machine learning for molecular and materials science. *Nature*, 559(7715):547–555, 2018.

R. Qi Charles, Hao Su, Mo Kaichun, and Leonidas J. Guibas. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 77–85, Honolulu, HI, July 2017. IEEE. ISBN 978-1-5386-0457-1. doi: 10.1109/CVPR.2017.16.

Chi Chen, Weike Ye, Yunxing Zuo, Chen Zheng, and Shyue Ping Ong. Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals. *Chemistry of Materials*, 31(9):3564–3572, May 2019. ISSN 0897-4756. doi: 10.1021/acs.chemmater.9b01294.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 6572–6583, 2018.

Jingpu Cheng, Qianxiao Li, Ting Lin, and Zuowei Shen. Interpolation, Approximation and Controllability of Deep Neural Networks, 2023. URL `http://arxiv.org/abs/2309.06015`.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2990–2999, New York, New York, USA, 20–22 Jun 2016. PMLR. URL `https://proceedings.mlr.press/v48/cohenc16.html`.

Christa Cuchiero, Martin Larsson, and Josef Teichmann. Deep neural networks, generic universal interpolation, and controlled odes. *SIAM Journal on Mathematics of Data Science*, 2(3):901–919, 2020.

Weinan E. A Proposal on Machine Learning via Dynamical Systems. *Communications in Mathematics and Statistics*, 5(1):1–11, 2017. ISSN 2194-6701.

Lawrence C Evans. An introduction to mathematical optimal control theory version 0.2. *Lecture notes available at http://math. berkeley. edu/˜ evans/control. course. pdf*, 1983.

Marc Finzi, Samuel Stanton, Pavel Izmailov, and Andrew Gordon Wilson. Generalizing convolutional neural networks for equivariance to lie groups on arbitrary continuous data. In *International Conference on Machine Learning*, pages 3165–3176. PMLR, 2020.

Maximilien Germain, Mathieu Laurière, Huyên Pham, and Xavier Warin. DeepSets and their derivative networks for solving symmetric PDEs. (arXiv:2103.00838), January 2022. doi: 10.48550/arXiv.2103.00838.

Garrett B Goh, Nathan O Hodas, and Abhinav Vishnu. Deep learning for computational chemistry. *Journal of computational chemistry*, 38(16):1291–1307, 2017.

Eldad Haber and Lars Ruthotto. Stable architectures for deep neural networks. *Inverse Problems*, 34(1):14004, 2017. ISSN 0266-5611.

Sydney R Hall, Frank H Allen, and I David Brown. The crystallographic information file (cif): a new standard archive file for crystallography. *Acta Crystallographica Section A: Foundations of Crystallography*, 47(6):655–685, 1991.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016. ISBN 9781467388504. doi: 10.1109/CVPR.2016.90.

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

Qianxiao Li, Long Chen, Cheng Tai, and Weinan E. Maximum Principle Based Algorithms for Deep Learning. *Journal of Machine Learning Research*, 18(1):1–29, 2018. ISSN 1532-4435.

Qianxiao Li, Ting Lin, and Zuowei Shen. Deep learning via dynamical systems: An approximation perspective. *Journal of the European Mathematical Society*, 25(5):1671–1709, 2023. doi: 10.4171/JEMS/1221.

Zhong Li, Jiequn Han, Weinan E, and Qianxiao Li. On the curse of memory in recurrent neural networks: Approximation and optimization analysis. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=8Sqhl-nF50`.

Zhong Li, Jiequn Han, Weinan E, and Qianxiao Li. Approximation and Optimization Theory for Linear Continuous-Time Recurrent Neural Networks. *Journal of Machine Learning Research*, 23(42):1–85, 2022. ISSN 1533-7928. Publisher: Microtome Publishing.

Jianfeng Lu, Zuowei Shen, Haizhao Yang, and Shijun Zhang. Deep network approximation for smooth functions. *SIAM Journal on Mathematical Analysis*, 53(5):5465–5506, 2021.

Haggai Maron, Ethan Fetaya, Nimrod Segol, and Yaron Lipman. On the universality of invariant networks. In *International Conference on Machine Learning*, pages 4363–4371. PMLR, 2019.

Frank Noé, Alexandre Tkatchenko, Klaus-Robert Müller, and Cecilia Clementi. Machine learning for molecular simulation. *Annual Review of Physical Chemistry*, 71:361–390, 2020.

Kenta Oono and Taiji Suzuki. Approximation and non-parametric estimation of resnet-type convolutional neural networks. In *International Conference on Machine Learning*, pages 4922–4931. PMLR, 2019.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

Kristin Persson. Materials data on nacl (sg:225) by materials project, 11 2014a.

Kristin Persson. Materials data on caco3 (sg:167) by materials project, 7 2014b.

Philipp Petersen and Felix Voigtlaender. Equivalence of approximation by convolutional neural networks and fully-connected networks. *Proceedings of the American Mathematical Society*, 148(4):1567–1581, 2020.

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*, pages 652–660, 2017.

Siamak Ravanbakhsh. Universal equivariant multilayer perceptrons. In *International Conference on Machine Learning*, pages 7996–8006. PMLR, 2020.

Siamak Ravanbakhsh, Jeff Schneider, and Barnabas Poczos. Equivariance through parameter-sharing. In *International Conference on Machine Learning*, pages 2892–2901. PMLR, 2017.

Zekun Ren, Juhwan Noh, Siyu Tian, Felipe Oviedo, Guangzong Xing, Qiaohao Liang, Armin Aberle, Yi Liu, Qianxiao Li, Senthilnath Jayavelu, et al. Inverse design of crystals using generalized invertible crystallographic representation. *arXiv preprint arXiv:2005.07609*, 2020.

Domènec Ruiz-Balet and Enrique Zuazua. Neural ode control for classification, approximation and transport. *arXiv preprint arXiv:2104.05278*, 2021.

Akiyoshi Sannai, Yuuki Takai, and Matthieu Cordonnier. Universal approximations of permutation invariant/equivariant functions by deep neural networks. *arXiv preprint arXiv:1903.01939*, 2019.

Zuowei Shen, Haizhao Yang, and Shijun Zhang. Deep network approximation characterized by number of neurons. *arXiv preprint arXiv:1906.05497*, 2019.

Zuowei Shen, Haizhao Yang, and Shijun Zhang. Deep network with approximation error being reciprocal of width to power of square root of depth. *Neural Computation*, 33 (4):1005–1036, 2021a.

Zuowei Shen, Haizhao Yang, and Shijun Zhang. Neural network approximation: Three hidden layers are enough. *Neural Networks*, 141:160–173, 2021b.

Zuowei Shen, Haizhao Yang, and Shijun Zhang. Optimal approximation rate of relu networks in terms of width and depth. *Journal de Mathématiques Pures et Appliquées*, 157:101–135, 2022.

Paulo Tabuada and Bahman Gharesifard. Universal approximation power of deep residual neural networks through the lens of control. *IEEE Transactions on Automatic Control*, 2022.

Siyu Isaac Parker Tian, Ting Lin, Tonio Buonassisi, and Qianxiao Li. Material property prediction using permutation invariant dynamical systems. *In preparation*, 2022.

Oriol Vinyals, Samy Bengio, and Manjunath Kudlur. Order Matters: Sequence to sequence for sets. November 2015. doi: 10.48550/arXiv.1511.06391.

Edward Wagstaff, Fabian B Fuchs, Martin Engelcke, Michael A Osborne, and Ingmar Posner. Universal approximation of functions on sets. *Journal of Machine Learning Research*, 23(151):1–56, 2022.

Maurice Weiler, Mario Geiger, Max Welling, Wouter Boomsma, and Taco Cohen. 3d steerable cnns: Learning rotationally equivariant features in volumetric data. *arXiv preprint arXiv:1807.02547*, 2018.

Tian Xie and Jeffrey C. Grossman. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Physical Review Letters*, 120(14):145301, April 2018. ISSN 0031-9007, 1079-7114. doi: 10.1103/PhysRevLett.120.145301.

Yunfei Yang, Zhen Li, and Yang Wang. Approximation in shift-invariant spaces with deep relu neural networks. *Neural Networks*, 153:269–281, 2022.

Dmitry Yarotsky. Optimal approximation of continuous functions by very deep relu networks. In *Conference on Learning Theory*, pages 639–649. PMLR, 2018.

Dmitry Yarotsky. Universal approximations of invariant maps by neural networks. *Constructive Approximation*, 55(1):407–474, 2022.

Julia M Yeomans. *Statistical mechanics of phase transitions*. Clarendon Press, 1992.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper/2017/file/f22e4747da1aa27e363d86d40ff442fe-Paper.pdf`.

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

Ding-Xuan Zhou. Universality of deep convolutional neural networks. *Applied and computational harmonic analysis*, 48(2):787–794, 2020.

Aaron Zweig and Joan Bruna. A functional perspective on learning symmetric functions with neural networks. In *International Conference on Machine Learning*, pages 13023–13032. PMLR, 2021.