Architecture Induces Structural Invariant Manifolds of Neural Network Training Dynamics

Jiajie Zhao[1] ZJJ0216@SJTU.EDU.CN
Tao Luo[1,3,*] LUOTAO41@SJTU.EDU.CN
Yaoyu Zhang[1,2,*] ZHYY.SJTU@SJTU.EDU.CN

[1] *School of Mathematical Sciences, Institute of Natural Sciences and MOE-LSC, Shanghai Jiao Tong University, Shanghai, 200240, China.*

[2] *School of Artificial Intelligence, Shanghai Jiao Tong University, Shanghai, 200240, China*

[3] *CMA-Shanghai, Shanghai Jiao Tong University, Shanghai, 200240, China.*

[*] *Corresponding authors.*

Abstract

While architecture is recognized as key to the performance of deep neural networks, its precise effect on training dynamics has been unclear due to the confounding influence of data and loss functions. This paper proposed an analytic framework based on the geometric control theory to characterize the dynamical properties intrinsic to a model's parameterization. We prove that the Structural Invariant Manifolds (SIMs) of an analytic model $F(\boldsymbol{\theta})(\boldsymbol{x})$—submanifolds that confine gradient flow trajectories independent of data and loss—are unions of orbits of the vector field family $\{\nabla_{\boldsymbol{\theta}} F(\cdot)(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathbb{R}^d\}$. We then prove that a model's symmetry, e.g., permutation symmetry for neural networks, induces SIMs. Applying this, we characterize the hierarchy of symmetry-induced SIMs in fully-connected networks, where dynamics exhibit neuron condensation and equivalence to reduced-width networks. For two-layer networks, we prove all SIMs are symmetry-induced, closing the gap between known symmetries and all possible invariants. Overall, by establishing the framework for analyzing SIMs induced by architecture, our work paves the way for a deeper analysis of neural network training dynamics and generalization in the near future.

Keywords: neural network architecture; training dynamics; geometric control theory; structural invariant manifold

AMS Subject Classification: 68T07, 34H05, 93C10, 93B03, 93B27

1. Introduction

Neural networks serve as the core engine of modern AI applications. The architecture of a network—that is, its specific scheme for parameterizing functions—is widely recognized as the primary factor influencing its training behavior and ultimate generalization performance on a given task (Krizhevsky et al., 2012; He et al., 2016; Vaswani et al., 2017). Nevertheless, the nonlinear nature of these architectures gives rise to highly nonlinear training dynamics, making the analysis of these dynamics and the precise consequences of architectural choices a persistently challenging problem (E et al., 2006).

In recent years, several theoretical developments in deep learning have shed light on this problem. One line of research focuses on a key phenomenon in nonlinear training dynamics known as condensation (Luo et al., 2021; Xu et al., 2025) (also referred to as quantization (Maennel et al., 2018), weight clustering (Brutzkus and Globerson, 2019), or

alignment (Min et al., 2024)). This widely observed process describes how neurons within a layer tend to align with one another during training. The study of condensation illuminates how neural networks adaptively extract features from data and reveals an implicit bias towards simpler functions that can be expressed by narrower networks (Xu et al., 2025). Crucially, condensation results from the nonlinear network architecture and is absent in any linear models. In addition, a series of works have revealed that the permutation symmetry of neural network architectures profoundly impacts both training dynamics and the loss landscape's critical point distribution (Fukumizu et al., 2019; Liu, 2024; Simsek et al., 2021; Zhang et al., 2021). Regarding the dynamics, it has been shown that permutation-invariant subspaces are also invariant under the training dynamics (Simsek et al., 2021; Liu, 2024). Regarding the loss landscape, the embedding principle demonstrates that a network inherits all critical points from any narrower network within its architecture (Zhang et al., 2021; Simsek et al., 2021; Fukumizu et al., 2019). Furthermore, some recent works have leveraged Lie brackets and the Frobenius theorem to systematically identify conserved quantities and the lower-dimensional invariant manifolds they induce in nonlinear models like deep linear and ReLU networks (Marcotte et al., 2023, 2025). These conservation laws, inherent to the model architecture, constrain the models' global training dynamics.

Despite these advancements, uncovering the exact impact of a nonlinear architecture on training dynamics remains challenging, primarily due to the difficulty of isolating its effect from the complications of the training data and loss function. In this paper, we address this challenge by introducing the concept of structural invariant manifold (SIM), which is defined as a submanifold of the parameter space that confines gradient flow trajectories independent of training data and loss, as the central object for our study. By employing the geometric control theory, in particular the Hermann–Nagano Theorem (Nagano, 1966), we uncover the dynamical effect of architecture as follows: **Architecture partitions the parameter space into nonintersecting orbits.** These orbits and their unions give rise to all SIMs of the gradient flow. Note that all models possess a trivial SIM, i.e., the entire parameter space $\mathbb{R}^M$. Our results yield a key insight into the dichotomy between linear and nonlinear models: a generic linear model possesses only the trivial SIM. Consequently, the existence of non-trivial SIMs, which often have much lower dimensions than the full parameter space, is a hallmark of how a nonlinear architecture fundamentally shapes training dynamics. We remark that our framework, grounded in geometric control theory, offers a unified mathematical foundation for the analysis of architecture-induced invariant structures. This approach bridges the gap between the separate treatments of invariant structures resulting from symmetry or conservation law in the literature (Simsek et al., 2021; Liu, 2024; Marcotte et al., 2023, 2025).

In general, uncovering the orbits of complex nonlinear models like neural networks is technically difficult. In this work, we identify a general family of architectural properties—namely, invariant maps and their induced symmetry groups and infinitesimal symmetries—that can conveniently reveal a series of SIMs. By determining all such symmetry groups and infinitesimal symmetries for general deep neural networks, we uncover a large family of SIMs with a hierarchical structure: non-trivial SIMs exist with dimensions ranging from low to high, where each lower-dimensional SIM exhibits neuron condensation and is functionally equivalent to a reduced-width network. While obtaining all SIMs for neural networks remains a general challenge, we take a step forward by proving that, for generic

two-layer networks, all SIMs are indeed symmetry-induced. The properties and analytical techniques we develop for studying these SIMs are broadly applicable and extend to other nonlinear models, such as matrix factorization (Bai et al., 2024; Koren et al., 2009), and to other architectures, including Convolutional Neural Networks (Krizhevsky et al., 2012; Zhang et al., 2025a) and Transformers (Vaswani et al., 2017; Chen and Luo, 2025).

Overall, this paper establishes an analytic framework for identifying SIMs induced by model architecture, thereby elucidating how specific architectural designs inherently constrain gradient flow training dynamics globally. It is important to note, however, that the realized training dynamics and ultimate generalization performance are also profoundly influenced by other factors, including the target function, training data, initialization, and loss function. An important and promising direction for future research is to study how these elements interact with these architecture-induced geometric structures to determine the actual trajectory of training and generalization performance.

The logical organization of our main results is illustrated in the flowchart in Figure 1. In this diagram, red blocks highlight the main theorems of each section. We now proceed to introduce each main section. Section 2 introduces the preliminaries including the Hermann–Nagano Theorem (Theorem 2.1, Corollary 2.1) in geometric control theory. Section 3 then defines Structural Invariant Manifolds (SIMs), proving they are orbit unions of $\mathcal{F} = \left\{\nabla_{\boldsymbol{\theta}} F(\cdot)(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathbb{R}^d\right\}$ (Theorem 3.1). This implies structural invariant sets are closed under set operations (Proposition 3.2) and that linear models possess only trivial SIMs (Proposition 3.3). Section 4 connects symmetry of model architecture to its SIMs, proving that SIMs can be induced by invariant maps (Theorem 4.1). Results on the orthogonal symmetry group (Lemma 4.1) and infinitesimal symmetry (Proposition 4.1) then lead to a characterization of symmetry-induced SIMs in deep neural networks (Theorem 4.2). Section 5 analyzes two-layer neural networks. By applying neuron independence (Lemma 5.1) and a perturbation lemma (Lemma 5.2), we establish properties for the rank of the Lie closure (Corollary 5.1–5.3). These results, combined with connectivity (Corollary 5.4, based on Proposition 5.1 and 5.2), allow us to prove a main find: for generic two-layer networks, all SIMs are symmetry-induced (Theorem 5.1). Section 6 provides a conclusion of our paper. In Appendix A, we provide the definitions and concepts from geometric control theory pertinent to the content of this paper.

2. Preliminary

2.1 Problem setting

We define a **parametric model** as a map $F : \mathbb{R}^M \to C(\mathbb{R}^d, \mathbb{R})$, where M and d are positive integers, and $C(\mathbb{R}^d, \mathbb{R})$ denotes the set of continuous functions from $\mathbb{R}^d$ to $\mathbb{R}$. Given a parameter $\boldsymbol{\theta} \in \mathbb{R}^M$, the output function $F(\boldsymbol{\theta})$ is a function from $\mathbb{R}^d$ to $\mathbb{R}$. The value of this function for an input $\boldsymbol{x} \in \mathbb{R}^d$ is denoted by $F(\boldsymbol{\theta})(\boldsymbol{x})$. Given a parametric model F, a dataset $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ and an analytic loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, we can define the empirical loss function as $L(\boldsymbol{\theta}) = \sum_{i=1}^n \ell(F(\boldsymbol{\theta})(\boldsymbol{x}_i), y_i)$. We analyze in this work the gradient flow given by $\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}t} = -\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$. By chain rule, we have

$$\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}t} = -\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) = -\sum_{i=1}^{n} \nabla \ell(F(\boldsymbol{\theta})(\boldsymbol{x}_i), y_i) \nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta})(\boldsymbol{x}_i). \tag{1}$$

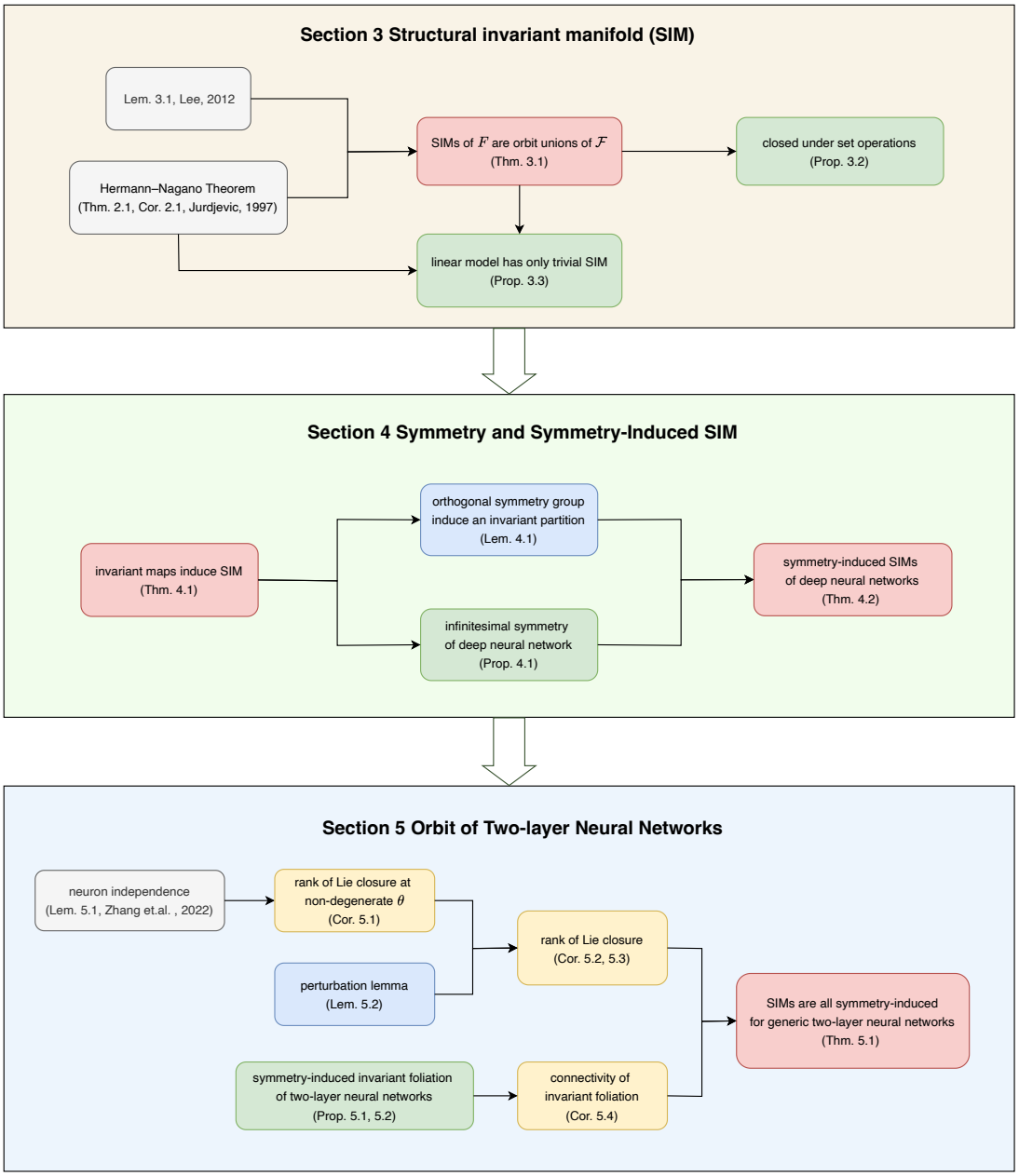

Figure 1: Flowchart of the paper's logical structure, illustrating the progression from Section 3, Section 4, to Section 5. Grey blocks represent foundational results from prior work. Red blocks denote the paper's main theorems. Green blocks represent propositions, blue blocks represent lemmas, and yellow blocks represent corollaries.

Throughout the paper, $\nabla \ell$ denotes the gradient of ℓ with respect to its first variable. The function $L(\boldsymbol{\theta})$ depends on the dataset S and the loss function ℓ, but we omit these dependencies from the notation for simplicity. All regularity assumptions in this paper are analytic. For example, we only consider analytic parametric model defined in Definition 2.1.

Definition 2.1 (analytic parametric model). A parametric model $F(\boldsymbol{\theta})(\boldsymbol{x})$, where $\boldsymbol{\theta} \in \mathbb{R}^M$ and $\boldsymbol{x} \in \mathbb{R}^d$, is called an **analytic parametric model** if F, considered as a function of $\boldsymbol{\theta}$ and $\boldsymbol{x}$, is a real-valued analytic function.

A major objective of this paper is to discuss invariant manifolds of Eq. (1) that are independent of loss function ℓ and dataset S. The definition of an invariant manifold is provided in Definition 2.2.

Definition 2.2 (vector field induced invariant set (manifold)[1]). Suppose M is a positive integer and $\mathcal{M}$ is a subset of $\mathbb{R}^M$. Let X be an analytic vector field on $\mathbb{R}^M$, and let $\boldsymbol{\theta}(t)$ denote the solution to the Cauchy problem $\dot{\boldsymbol{\theta}} = X(\boldsymbol{\theta}), \boldsymbol{\theta}(0) = \boldsymbol{\theta}_0$. We say that $\mathcal{M}$ is an **invariant set** (with respect to the vector field X) if for every $\boldsymbol{\theta}_0 \in \mathcal{M}$, the solution $\boldsymbol{\theta}(t)$ remains in $\mathcal{M}$ for all t in its maximal interval of existence. We also say $\mathcal{M}$ **is invariant under** X. Moreover, if $\mathcal{M}$ is an immersed submanifold of $\mathbb{R}^M$, we say $\mathcal{M}$ is an **invariant manifold**.

In this paper, the primary parametric models we consider are the neural networks defined in Definitions 2.3 and 2.4.

Definition 2.3 (multi-layer fully-connected neural network). Consider the neural network $F(\boldsymbol{\theta})(\boldsymbol{x})$ defined inductively by

$$\boldsymbol{a}^{(0)} = \boldsymbol{x}, \quad \boldsymbol{a}^{(l)} = \sigma\left(\boldsymbol{W}^{(l)}\boldsymbol{a}^{(l-1)} + \boldsymbol{b}^{(l)}\right), \quad l = 1, 2, \ldots, L,$$

where $F(\boldsymbol{\theta})(\boldsymbol{x}) = \boldsymbol{a}^{(L)}$ and the parameters are $\boldsymbol{\theta} = \left(\boldsymbol{W}^{(l)}, \boldsymbol{b}^{(l)}\right)_{l=1}^{L}$. Here, each $\boldsymbol{W}^{(l)}$ is an $n_l \times n_{l-1}$ matrix, and $\boldsymbol{b}^{(l)}$, $\boldsymbol{a}^{(l)}$ are vectors in $\mathbb{R}^{n_l}$. The activation function $\sigma : \mathbb{R} \to \mathbb{R}$ is assumed to be real-analytic and acts entrywise on vectors. Technically when writing $\boldsymbol{\theta} = \left(\boldsymbol{W}^{(l)}, \boldsymbol{b}^{(l)}\right)_{l=1}^{L}$ the matrix $\boldsymbol{W}^{(l)}$ should be flattened to a vector, but we omit it for simplicity. In this paper we consider scalar output, i.e. $n_L = 1$.

Definition 2.4 (two-layer neural network). The network is represented as $F(\boldsymbol{\theta})(\boldsymbol{x}) = \sum_{i=1}^{m} a_i \sigma(\boldsymbol{w}_i^{\intercal}\boldsymbol{x})$, where $\boldsymbol{x} \in \mathbb{R}^d, a_i \in \mathbb{R}, \boldsymbol{w}_i \in \mathbb{R}^d$, $\boldsymbol{\theta} = (a_i, \boldsymbol{w}_i)_{i=1}^{m} \in \mathbb{R}^{(d+1)m}$, and m is the width of the network. The function $\sigma : \mathbb{R} \to \mathbb{R}$ is the activation function, which is assumed to be a non-polynomial, real analytic function.

Numerous symmetries exist in neural networks. We will introduce the concept here and discuss it in detail in Section 4.

Definition 2.5 (symmetry group). Let F be an analytic parametric model, and let G be a group (or semigroup) acting on the parameter space of F. If the action of any element $g \in G$ leaves the output of F invariant for all inputs, i.e., $F(g(\boldsymbol{\theta}))(\mathbf{x}) = F(\boldsymbol{\theta})(\mathbf{x}), \forall g \in G, \forall \boldsymbol{\theta} \in \mathbb{R}^M, \forall \mathbf{x} \in \mathbb{R}^d$, then G, together with its action, is called a **symmetry group (or semigroup)** of F. Furthermore, if every action of G is an orthogonal linear transformation, then G is called an **orthogonal symmetry group**.

1. Please see Appendix A for the definition of analytic vector field.

2.2 Orbit

To analyze the dynamics of $\boldsymbol{\theta}(t)$, we now introduce concepts from geometric control theory. A detailed introduction of geometric control theory is provided in Appendix A.

Definition 2.6 (orbit, page 33 of Jurdjevic (1997)). [2] Let $\mathcal{F}$ be a family of analytic vector fields on an analytic manifold $\mathcal{M}$. Let $G = G(\mathcal{F})$ be the group (pseudogroup) of diffeomorphisms (local diffeomorphisms) generated by $\{e^{tX} \mid t \in \mathbb{R}, X \in \mathcal{F}\}$ under composition. For any $\boldsymbol{\theta} \in \mathcal{M}$, we define the **orbit** of $\mathcal{F}$ through $\boldsymbol{\theta}$ as $\{g(\boldsymbol{\theta}) \mid g \in G\}$, which we denote by $O_{\mathcal{F}}(\boldsymbol{\theta})$.

Given a family of analytic vector fields, each of its orbits forms an analytic immersed submanifold. The dimension of an orbit is determined by the Lie closure of the vector field family, as stated in Definition 2.7 and Theorem 2.1.

In Jurdjevic (1997), Theorem 2.1 is stated under an analytic regularity assumption, whereas Corollary 2.1 is presented as a theorem under smooth regularity. Given that this paper operates within analytic regularity, we introduce an analytic version of Corollary 2.1. This version is a direct consequence of Theorem 2.1, and we therefore designate it as a corollary.

Definition 2.7 (Lie closure). Let $\mathcal{M}$ be an analytic manifold and $\mathcal{F}$ be a family of analytic vector fields on $\mathcal{M}$. We use $\mathrm{Lie}(\mathcal{F})$ to denote the Lie algebra of analytic vector fields generated by $\mathcal{F}$. For any point $\boldsymbol{\theta} \in \mathcal{M}$, $\mathrm{Lie}_{\boldsymbol{\theta}}(\mathcal{F})$ is defined to be the set of all tangent vectors $V(\boldsymbol{\theta})$ with V in $\mathrm{Lie}(\mathcal{F})$. We call $\mathrm{Lie}_{\boldsymbol{\theta}}(\mathcal{F})$ the **Lie closure** of $\mathcal{F}$ at $\boldsymbol{\theta}$.

Theorem 2.1. (Hermann–Nagano Theorem, Theorem 6 in Section 2 of Jurdjevic (1997)) Let $\mathcal{M}$ be an analytic manifold, and $\mathcal{F}$ a family of analytic vector fields on $\mathcal{M}$. Then:

(i) Each orbit of $\mathcal{F}$ is an (immersed) analytic submanifold of $\mathcal{M}$.

(ii) If $\mathcal{N}$ is an orbit of $\mathcal{F}$, then the tangent space of $\mathcal{N}$ at $\boldsymbol{\theta}$ is given by $\mathrm{Lie}_{\boldsymbol{\theta}}(\mathcal{F})$. In particular, the dimension of $\mathrm{Lie}_{\boldsymbol{\theta}}(\mathcal{F})$ is constant as $\boldsymbol{\theta}$ varies over $\mathcal{N}$.

Corollary 2.1. (Theorem 3 in Section 2 of Jurdjevic (1997)) Let $\mathcal{M}$ be an analytic manifold and $\mathcal{F}$ be a family of analytic vector fields on $\mathcal{M}$. Suppose that $\mathcal{F}$ is such that $\mathrm{Lie}_{\boldsymbol{\theta}_0}(\mathcal{F}) = T_{\boldsymbol{\theta}_0}\mathcal{M}$ for some $\boldsymbol{\theta}_0$ in $\mathcal{M}$. Then the orbit of $\mathcal{F}$ through $\boldsymbol{\theta}_0$ is open. If, in addition, $\mathrm{Lie}_{\boldsymbol{\theta}}(\mathcal{F}) = T_{\boldsymbol{\theta}}\mathcal{M}$ for each $\boldsymbol{\theta}$ in $\mathcal{M}$, and if $\mathcal{M}$ is connected, then there is only one orbit of $\mathcal{F}$ equal to $\mathcal{M}$.

3. Structural Invariant Manifold (SIM) and its framework

3.1 SIM as a key tool for the recovery puzzle

Considering the simplest setup where a parametric model $F(\boldsymbol{\theta})(\boldsymbol{x})$ is used to recover a target function $f^* \in \{F(\boldsymbol{\theta})(\cdot) \mid \boldsymbol{\theta} \in \mathbb{R}^M\}$ from n training samples $\{(\boldsymbol{x}_i, f^*(\boldsymbol{x}_i))\}_{i=1}^n$. The gradient flow training dynamics is written as

$$\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}t} = -\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) = -\sum_{i=1}^{n} \nabla \ell(F(\boldsymbol{\theta})(\boldsymbol{x}_i), f^*(\boldsymbol{x}_i)) \nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta})(\boldsymbol{x}_i). \tag{2}$$

2. Please see Appendix A for the definition of e^{tX}, analytic manifold, analytic vector field and pseudogroup.

A fundamental question in machine learning, known as the recovery problem, is to identify the conditions that allow above dynamics to successfully find the target function f^*. If $F(\boldsymbol{\theta})(\boldsymbol{x})$ is a linear model in $\boldsymbol{\theta}$ with linearly independent basis and a proper loss $\ell(\cdot,\cdot)$, then it is well-known that f^* can be recovered generically from $n \geq M$ samples (Zhang et al., 2025a). However, when we change $F(\boldsymbol{\theta})(\boldsymbol{x})$ to a nonlinear model, our understanding becomes extremely limited.

A particularly mysterious phenomenon is that nonlinear models like neural networks can recover certain targets even under severe overparameterization $n \ll M$ (Zhao et al., 2024; Zhang et al., 2025a,b, 2023). This phenomenon sparks the following recovery puzzle:

How neural networks recover targets under overparameterization?

Note that this puzzle is a specialization of the widely acknowledged generalization puzzle in deep learning theory—why overparameterized neural networks often generalize well (Breiman, 2018; Zhang et al., 2017). Yet, we argue that the recovery puzzle well serves as the cornerstone of the generalization puzzle: (i) the notion of recovery resolves the ambiguity in the notion of "generalize well"; (ii) understanding the conditions for recovery is often the first and the key step towards understanding generalization as in the cases of linear regression, signal processing (Shannon, 1948; Luke, 1999) (e.g., Nyquist-Shannon sampling theorem (Shannon, 1948)), and compressed sensing (Candès et al., 2006; Donoho, 2006; Candès and Wakin, 2008).

In recent years, progress has been made by solving some weaker versions of the recovery puzzle. Zhang et al. (2025a) proves a recovery guarantee for neural networks under overparameterization in the sense of local linear recovery, i.e., recovering targets in the tangent space of some optimal point in the target set $F^{-1}(f^*)$. Zhang et al. (2023) makes a step further to prove a recovery guarantee in the sense of local recovery, i.e., recovering targets in the neighbourhood of the target set $F^{-1}(f^*)$. Despite the progress, how one can leverage these local recovery guarantees to a global one remains an extremely difficult problem. Particularly, we lack means to globally back-trace the gradient flow dynamics from the vicinity of the target set to see if there exists a generic initialization that reliably access $F^{-1}(f^*)$ for $n < M$ training samples.

Motivated by recent results that demonstrate existence of lower dimensional ($< M$) invariant subspaces independent to training data and loss induced by the permutation symmetry of neural networks architecture (Simsek et al., 2021; Liu, 2024), we realize that these architecture-induced invariant manifolds could serve as the key tool for the global tracing of gradient flow dynamics. For the convenience of study, we first provide a formal definition as follows.

Definition 3.1 (structural invariant manifold (SIM)). Let $F(\boldsymbol{\theta})(\boldsymbol{x}), \boldsymbol{\theta} \in \mathbb{R}^M, \boldsymbol{x} \in \mathbb{R}^d$ be an analytic parametric model. For a subset $\mathcal{M} \subset \mathbb{R}^M$, we say $\mathcal{M}$ is a **structural invariant set** if it is invariant under $-\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$ in Eq. (1) for any real analytic loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ and dataset S. Moreover, if $\mathcal{M}$ is an immersed submanifold of $\mathbb{R}^M$, we say $\mathcal{M}$ is a **structural invariant manifold.**[3]

The concept of SIM is introduced to capture the intrinsic dynamical consequence of a model's architecture, independent of any particular dataset or loss function. Example 3.1

3. By convention, the empty set is neither a structural invariant set nor a structural invariant manifold.

illustrates a nontrivial SIM that arises in a simple nonlinear model. As we will show in the next section, this manifold emerges as a consequence of the model's permutation symmetry.

Example 3.1 (SIMs of two-neuron exponential neural network). Consider a two-neuron neural network with exponential activation and a one-dimensional input x. The model is given by:

$$F(\boldsymbol{\theta})(x) = a_1 e^{w_1 x} + a_2 e^{w_2 x},$$

where $\boldsymbol{\theta} = (a_1, w_1, a_2, w_2) \in \mathbb{R}^4, x \in \mathbb{R}$. It is easy to verify that $\mathcal{M} = \{(a_1, w_1, a_2, w_2) \in \mathbb{R}^4 \mid a_1 = a_2, w_1 = w_2\}$ is a SIM. Later in Theorem 5.1 we will see that all SIMs of F are $\mathbb{R}^4$, $\mathcal{M}$ and $\mathbb{R}^4 \setminus \mathcal{M}$.

The justification for the utility of lower dimensional SIMs in global trajectory tracing is as follows:
(i) **Enabling recovery under overparameterization:** a $d < M$ dimensional SIM enables us to study the gradient flow on this confined lower dimensional manifold. If the target set $F^{-1}(f^*)$ intersecting with certain $d < M$ dimensional SIM, then target recovery by intuition is possible with $d \leq n < M$ training samples.
(ii) **Enabling strong complexity control:** a $d < M$ dimensional SIM makes it possible for the model to keep the complexity (marked by the effective degrees of freedom) $\leq d$ for an arbitrarily long time: (1) on the SIM the complexity is constrained for infinite time; (2) the closer some $\boldsymbol{\theta}(t)$ is to the SIM, the longer afterwards the output complexity is upper bounded approximately by d.

SIMs emerge directly from a model's architecture, providing the key utilities for analyzing global dynamics described above. Therefore, this work makes an effort to establish a theoretical foundation—using geometric control theory—for the systematic identification of all SIMs in analytic parametric models with a focus on the neural network architecture.

3.2 SIMs as orbit unions of $\mathcal{F}$

A central challenge in the study of SIMs for Eq. (1) lies in isolating the influence of model architecture from the confounding effects of training data and loss function. In this work, we propose a relaxation of the dynamics that resembles a geometric control problem, revealing a connection between the SIMs of the model F and the orbits of the induced vector fields $\mathcal{F}$, defined as follows.

Definition 3.2 (induced vector fields). Let $F(\boldsymbol{\theta})(\boldsymbol{x})$ be an analytic parametric model with $\boldsymbol{\theta} \in \mathbb{R}^M$ and $\boldsymbol{x} \in \mathbb{R}^d$. Define the family of vector fields

$$\mathcal{F} = \left\{ \nabla_{\boldsymbol{\theta}} F(\cdot)(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathbb{R}^d \right\}.$$

$\mathcal{F}$ is called the **induced vector fields of the model**.

Our relaxation of the dynamics in Eq. (1) proceeds as follows. For each parameter vector $\boldsymbol{\theta} \in \mathbb{R}^M$, we observe that the gradient $-\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$ lies within the span of the model's gradients, i.e.,

$$-\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) \in \operatorname{span}\left(\{\nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta})(\boldsymbol{x}_i)\}_{i=1}^n\right) \subseteq \operatorname{span}\left(\{\nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta})(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathbb{R}^d\}\right).$$

This observation implies that the gradient flow trajectories are encapsulated in the orbits of $\{\nabla_{\boldsymbol{\theta}} F(\cdot)(\boldsymbol{x}_i)\}_{i=1}^n$, hence in the orbits of $\mathcal{F}$, which is determined solely by the model architecture.

With a detailed theoretical derivation below, we arrive at the first key result of our work in Theorem 3.1, which ensures that the orbits of $\mathcal{F}$ and their unions give rise to all SIMs. This theorem serves as the foundation for all our later analysis as it translates the seemingly complicated task of identifying all SIMs into a clean one: computing the orbits of $\mathcal{F}$. Building on this result, we can further explore several key questions: What SIMs arise under different neural network architectures? How do these manifolds emerge?

The proof of Theorem 3.1 relies on Lemma 3.1 as an auxiliary result. We therefore begin by stating and proving Lemma 3.1, and then proceed to the proof of Theorem 3.1.

Lemma 3.1 (Problem 9-2 of Lee (2012)). Let X be a smooth vector field on $\mathbb{R}^M$, and consider the Cauchy problem

$$\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}t} = X(\boldsymbol{\theta}), \quad \boldsymbol{\theta}(0) = \boldsymbol{\theta}_0, \tag{3}$$

with solution denoted by $\boldsymbol{\theta}(t)$. Suppose $\mathcal{M} \subset \mathbb{R}^M$ is an immersed submanifold such that $X(\boldsymbol{\theta}) \in T_{\boldsymbol{\theta}}\mathcal{M}$ for all $\boldsymbol{\theta} \in \mathcal{M}$. Then for any initial condition $\boldsymbol{\theta}_0 \in \mathcal{M}$, there exists $\delta > 0$ such that $\boldsymbol{\theta}(t) \in \mathcal{M}$ for all $|t| < \delta$. Moreover, if $\mathcal{M}$ is closed in $\mathbb{R}^M$, then $\boldsymbol{\theta}(t) \in \mathcal{M}$ for all t in the maximal interval of existence.

Proof. We provide its proof for completeness. We prove local invariance first, then extend to global invariance under the closedness assumption.

Local invariance. Let $\boldsymbol{\theta}_0 \in \mathcal{M}$, and let $\boldsymbol{\theta}(t)$ be the solution to Eq. (3). Pick any $t_0 \in \mathbb{R}$ such that $\boldsymbol{\theta}(t_0) \in \mathcal{M}$. Since $\mathcal{M}$ is an immersed submanifold, there exists a local parameterization $\psi : U \subset \mathbb{R}^k \to \mathbb{R}^M$ such that $\psi(U) \subset \mathcal{M}$, $\boldsymbol{\theta}(t_0) \in \psi(U)$, and $D\psi(u)$ has full column rank for all $u \in U$. Let $\boldsymbol{u}_0 := \psi^{-1}(\boldsymbol{\theta}(t_0)) \in U$, and consider the ODE in $\mathbb{R}^k$:

$$\frac{\mathrm{d}\boldsymbol{u}}{\mathrm{d}t} = (D\psi^\intercal D\psi)^{-1} D\psi^\intercal X(\psi(\boldsymbol{u}(t))), \quad \boldsymbol{u}(t_0) = \boldsymbol{u}_0.$$

This defines a smooth vector field in $\mathbb{R}^k$, hence admits a unique solution $\boldsymbol{u}(t)$ near t_0. Define $\tilde{\boldsymbol{\theta}}(t) := \psi(\boldsymbol{u}(t))$. By the chain rule,

$$\frac{\mathrm{d}\tilde{\boldsymbol{\theta}}}{\mathrm{d}t} = D\psi(\boldsymbol{u}(t)) \cdot \frac{\mathrm{d}\boldsymbol{u}}{\mathrm{d}t} = D\psi(D\psi^\intercal D\psi)^{-1} D\psi^\intercal X(\psi(\boldsymbol{u}(t))).$$

Since $X(\psi(\boldsymbol{u}(t))) \in T_{\psi(\boldsymbol{u}(t))}\mathcal{M}$, $X(\psi(\boldsymbol{u}(t))$ is in the image of $D\psi$. Thus, $\frac{\mathrm{d}\tilde{\boldsymbol{\theta}}}{\mathrm{d}t} = X(\tilde{\boldsymbol{\theta}}(t))$. Therefore, $\tilde{\boldsymbol{\theta}}(t)$ satisfies the same ODE as $\boldsymbol{\theta}(t)$ and coincides with it at $t = t_0$. By uniqueness, $\boldsymbol{\theta}(t) = \tilde{\boldsymbol{\theta}}(t) \in \mathcal{M}$ near t_0. Taking $t_0 = 0$, we obtain $\delta > 0$ such that $\boldsymbol{\theta}(t) \in \mathcal{M}$ for all $|t| < \delta$.

Global invariance (if $\mathcal{M}$ is closed). Pick arbitrary T in the maximal interval of existence. Without loss of generality we assume $T > 0$. Define $A = \{t \in [0, T] \mid \boldsymbol{\theta}(t) \in \mathcal{M}\}$. By the local invariance, A is open in $[0, T]$. Since $\boldsymbol{\theta}(t)$ is continuous and $\mathcal{M}$ is closed, A is also closed in $[0, T]$. Since $0 \in A$, $A = [0, T]$ by connectedness. Therefore $\boldsymbol{\theta}(T) \in \mathcal{M}$. Since T was arbitrary within the maximal interval of existence, it follows that $\boldsymbol{\theta}(t) \in \mathcal{M}$ for all t in the maximal interval of existence. □

Theorem 3.1 (SIMs of F are orbit unions of $\mathcal{F}$). Let $F(\boldsymbol{\theta})(\boldsymbol{x}), \boldsymbol{\theta} \in \mathbb{R}^M, \boldsymbol{x} \in \mathbb{R}^d$ be an analytic parametric model. Let $\mathcal{F} = \left\{\nabla_{\boldsymbol{\theta}} F(\cdot)(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathbb{R}^d\right\}$. Let $\mathcal{M} \neq \emptyset$ be a subset (or immersed submanifold) of $\mathbb{R}^M$. Then $\mathcal{M}$ is a structural invariant set (or SIM) if and only if $\mathcal{M}$ is invariant under every vector field in $\mathcal{F}$, equivalently, $\mathcal{M}$ is union of orbits of $\mathcal{F}$.

Proof. Let $\mathcal{M} \subseteq \mathbb{R}^M$. Consider the following statements: **(i)** $\mathcal{M}$ is a structural invariant set. **(ii)** $\mathcal{M}$ is invariant under every vector field in $\mathcal{F}$. **(iii)** $\mathcal{M}$ is a union of orbits of $\mathcal{F}$. We will show that statements **(i)(ii)(iii)** are equivalent.

(i) $\implies$ **(ii)**: Assume $\mathcal{M}$ is invariant under the gradient flow $-\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$ for any real analytic loss function and dataset. Let $\boldsymbol{x} \in \mathbb{R}^d$ be arbitrary. Consider the loss function $\ell(s,t) = -s$ and the dataset $S = \{(\boldsymbol{x}, y)\}$ for some $y \in \mathbb{R}$. Then $L(\boldsymbol{\theta}) = -F(\boldsymbol{\theta})(\boldsymbol{x})$, and hence $-\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta})(\boldsymbol{x})$. So $\nabla_{\boldsymbol{\theta}} L(\cdot)$ is a vector field in $\mathcal{F}$. Since $\boldsymbol{x} \in \mathbb{R}^d$ is arbitrary, $\mathcal{M}$ is invariant under every vector field in $\mathcal{F}$.

(ii) $\implies$ **(iii)**: Assume that $\mathcal{M} \subset \mathbb{R}^M$ is invariant under every vector field in $\mathcal{F}$. Since this invariance is preserved under the composition of flows, it follows that $\mathcal{M}$ is also invariant under every (local) diffeomorphism in the (pseudo) group generated by $\mathcal{F}$. Consequently, $O_{\mathcal{F}}(\boldsymbol{\theta}) \subset \mathcal{M}$ for any $\boldsymbol{\theta} \in \mathcal{M}$. Also, $\mathcal{M} \neq \emptyset$. Therefore $\mathcal{M}$ is a union of orbits of $\mathcal{F}$.

(iii) $\implies$ **(i)**: We begin by presenting a lemma along with its proof.

Lemma 3.2. Let $\mathcal{F}$ be an arbitrary family of analytic vector fields on $\mathbb{R}^M$. Assume $X(\boldsymbol{\theta})$ is an analytic vector field that satisfies the condition $X(\boldsymbol{\theta}) \in \mathrm{Lie}_{\boldsymbol{\theta}}(\mathcal{F}), \forall \boldsymbol{\theta} \in \mathbb{R}^M$. Then each orbit of $\mathcal{F}$ is invariant under $X(\boldsymbol{\theta})$.

Proof for Lemma 3.2: Let $\mathcal{M}$ be an orbit of $\mathcal{F}$. Fix any $\boldsymbol{\theta}_0 \in \mathcal{M}$, and let $\boldsymbol{\theta}(t), t \in I$ be the solution to the Cauchy problem $\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}t} = X(\boldsymbol{\theta}), \boldsymbol{\theta}(0) = \boldsymbol{\theta}_0$, where I is the maximal interval of existence. We now prove that $\boldsymbol{\theta}(t) \in \mathcal{M}, \forall t \in I$. Suppose for contradiction that $\{t \in [0, +\infty) \cap I \mid \boldsymbol{\theta}(t) \notin \mathcal{M}\} \neq \emptyset$. Let $t_1 = \inf\{t \in [0, +\infty) \cap I \mid \boldsymbol{\theta}(t) \notin \mathcal{M}\}$. By Theorem 2.1, $\mathcal{M}$ is an immersed submanifold, and its tangent space is given by $T_{\boldsymbol{\theta}}\mathcal{M} = \mathrm{Lie}_{\boldsymbol{\theta}}(\mathcal{F})$. Since $X(\boldsymbol{\theta}) \in \mathrm{Lie}_{\boldsymbol{\theta}}(\mathcal{F})$ for any $\boldsymbol{\theta} \in \mathbb{R}^M$, we have $X(\boldsymbol{\theta}) \in T_{\boldsymbol{\theta}}\mathcal{M}, \forall \boldsymbol{\theta} \in \mathcal{M}$. By Lemma 3.1, there exists $\delta > 0$ such that $\boldsymbol{\theta}(t) \in \mathcal{M}$ for all $t \in [0, \delta)$, which implies $t_1 > 0$.

Applying Lemma 3.1 again at $\boldsymbol{\theta}(t_1)$, we obtain a $\delta_1 \in (0, t_1)$ such that $\boldsymbol{\theta}(t) \in O_{\mathcal{F}}(\boldsymbol{\theta}(t_1)), \forall t \in [t_1 - \delta_1, t_1 + \delta_1]$. Since $\boldsymbol{\theta}(t_1 - \delta_1) \in \mathcal{M} \cap O_{\mathcal{F}}(\boldsymbol{\theta}(t_1))$, $\mathcal{M} = O_{\mathcal{F}}(\boldsymbol{\theta}(t_1))$. Therefore for all $0 < t \leq t_1 + \delta_1$, $\boldsymbol{\theta}(t) \in \mathcal{M}$, which contradicts that t_1 is the infimum. Hence, the set $\{t \in [0, +\infty) \cap I \mid \boldsymbol{\theta}(t) \notin \mathcal{M}\}$ is empty. Similarly, we have $\{t \in (-\infty, 0] \cap I \mid \boldsymbol{\theta}(t) \notin \mathcal{M}\} = \emptyset$. So $\boldsymbol{\theta}(t) \in \mathcal{M}$ for all $t \in I$. Therefore $\mathcal{M}$ is invariant under $X(\boldsymbol{\theta})$. □

We now return to the proof of **(iii)** $\implies$ **(i)**. Assume that $\mathcal{M}$ is a union of orbits of $\mathcal{F} = \left\{\nabla_{\boldsymbol{\theta}} F(\cdot)(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathbb{R}^d\right\}$. Consider an arbitrary dataset S and loss function ℓ. For any $\boldsymbol{\theta} \in \mathbb{R}^M$, the vector $-\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$ in Eq. (1) is a linear combination of vectors in $\mathcal{F}|_{\boldsymbol{\theta}}$, where $\mathcal{F}|_{\boldsymbol{\theta}}$ is the evaluation of $\mathcal{F}$ at $\boldsymbol{\theta}$. Thus, for any $\boldsymbol{\theta} \in \mathbb{R}^M$, we have $-\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) \in \mathrm{span}(\mathcal{F}|_{\boldsymbol{\theta}}) \subset \mathrm{Lie}_{\boldsymbol{\theta}}(\mathcal{F})$. By Lemma 3.2, each orbit of $\mathcal{F}$ is invariant under the vector field $-\nabla_{\boldsymbol{\theta}} L(\cdot)$. Hence, each orbit of $\mathcal{F}$ is a SIM. Since $\mathcal{M}$ is a union of orbits of $\mathcal{F}$, it follows readily from the definition that $\mathcal{M}$ is a SIM.

Thus the three statements are equivalent. Furthermore, if $\mathcal{M}$ is assumed to be an immersed submanifold of $\mathbb{R}^M$, statement **(i)** may be replaced by the assertion that $\mathcal{M}$ is a SIM.

□

In defining SIM, we require it to be invariant to the gradient flow under any loss function ℓ and dataset S. However, Proposition 3.1 shows that, under mild assumptions on the loss function ℓ_0, an immersed submanifold is a SIM if and only if it is invariant to the gradient flow under this loss function ℓ_0 and any dataset S. Intuitively, data-independent invariance is strong enough to induce structural invariance.

Proposition 3.1. Let $F(\boldsymbol{\theta})(\boldsymbol{x})$ be an analytic parametric model with $\boldsymbol{\theta} \in \mathbb{R}^M$, and let $\mathcal{M}$ be an immersed submanifold of $\mathbb{R}^M$. Suppose the loss function $\ell_0(s,t)$ is real analytic and satisfies: $\forall s \in \mathbb{R}, \exists t \in \mathbb{R}$ such that $\nabla \ell_0(s,t) \neq 0$. Then $\mathcal{M}$ is a SIM if and only if $\mathcal{M}$ is invariant under $-\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$ in Eq. (1) for this loss function ℓ_0 and any dataset S.

Proof. By definition of SIM, one direction is trivial. To prove the other direction, assume that $\mathcal{M}$ is invariant under $-\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$ in Eq. (1) for any dataset S and the loss function ℓ_0.

Let $\boldsymbol{x}_0 \in \mathbb{R}^d$ be arbitrary, and let $S_0 = \{(\boldsymbol{x}_0, y)\}$ for some $y \in \mathbb{R}$. Denote $X(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta})(\boldsymbol{x}_0)$. Then under the dataset S_0 and the loss function ℓ_0, we have $-\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) = -\nabla l_0(F(\boldsymbol{\theta})(\boldsymbol{x}_0), y) X(\boldsymbol{\theta})$. By our initial assumption, $\mathcal{M}$ is invariant under $-\nabla l_0(F(\boldsymbol{\theta})(\boldsymbol{x}_0), y) X(\boldsymbol{\theta})$ for any $y \in \mathbb{R}$. Define $\mathcal{F}' = \{-\nabla l_0(F(\boldsymbol{\theta})(\boldsymbol{x}_0), y) X(\boldsymbol{\theta}) \mid y \in \mathbb{R}\}$. Since $\mathcal{M}$ is invariant under any vector field in $\mathcal{F}'$, it follows that $\mathcal{M}$ is invariant under compositions of flows generated by vector fields in $\mathcal{F}'$. So $O_{\mathcal{F}'}(\boldsymbol{\theta}) \subset \mathcal{M}, \forall \boldsymbol{\theta} \in \mathcal{M}$. Now, fix any $\boldsymbol{\theta}_0 \in \mathcal{M}$. Then $O_{\mathcal{F}'}(\boldsymbol{\theta}_0) \subset \mathcal{M}$. Let $\boldsymbol{\theta}(t), t \in I$ denote the solution of the Cauchy problem $\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}t} = X(\boldsymbol{\theta}), \boldsymbol{\theta}(0) = \boldsymbol{\theta}_0$, where I is the maximal interval of existence. We now prove that $\boldsymbol{\theta}(t) \in \mathcal{M}, \forall t \in I$. Since $O_{\mathcal{F}'}(\boldsymbol{\theta}_0) \subset \mathcal{M}$, it is sufficient to prove that $\boldsymbol{\theta}(t) \in O_{\mathcal{F}'}(\boldsymbol{\theta}_0), \forall t \in I$. By assumption of ℓ_0, for any $\boldsymbol{\theta} \in \mathbb{R}^M$, there exists $y_0 \in \mathbb{R}$ such that $\nabla \ell(F(\boldsymbol{\theta})(\boldsymbol{x}_0), y_0)) \neq 0$. Therefore, for any $\boldsymbol{\theta} \in \mathbb{R}^M$, we have $X(\boldsymbol{\theta}) \in \mathcal{F}'|_{\boldsymbol{\theta}} \subset \mathrm{Lie}_{\boldsymbol{\theta}}(\mathcal{F}')$. Applying Lemma 3.2, it follows that $O_{\mathcal{F}'}(\boldsymbol{\theta}_0)$ is invariant under $X(\boldsymbol{\theta})$. So $\boldsymbol{\theta}(t) \in O_{\mathcal{F}'}(\boldsymbol{\theta}_0) \subset \mathcal{M}, \forall t \in I$. Thus, $\mathcal{M}$ is invariant under $X(\boldsymbol{\theta})$. Since our choice of $\boldsymbol{x}_0$ is arbitrary, $\mathcal{M}$ is invariant under any vector field in $\mathcal{F} = \left\{\nabla_{\boldsymbol{\theta}} F(\cdot)(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathbb{R}^d\right\}$. It follows from Theorem 3.1 that $\mathcal{M}$ is a SIM. □

Structural invariant sets are closed under set operations, as shown in Proposition 3.2.

Proposition 3.2 (structural invariant sets are closed under set operations). Let $F(\boldsymbol{\theta})(\boldsymbol{x})$ be an analytic parametric model with parameter $\boldsymbol{\theta} \in \mathbb{R}^M$, and let $\mathcal{E}$ denote the collection of all structural invariant sets of the model F, augmented by the empty set. Then the following properties hold:

(i) If $\mathcal{M} \in \mathcal{E}$, then its complement $\mathbb{R}^M \setminus \mathcal{M} \in \mathcal{E}$.

(ii) If $\{\mathcal{M}_i\}_{i \in I} \subseteq \mathcal{E}$ is any collection of structural invariant sets indexed by I, then both the intersection $\bigcap_{i \in I} \mathcal{M}_I$ and the union $\bigcup_{i \in I} \mathcal{M}_i$ belong to $\mathcal{E}$.

Proof. By definition, $\mathbb{R}^M = \bigcup_{j \in J} O_j$, where J is an index set, $O_j, j \in J$ are all orbits of $\mathcal{F} = \{\nabla_{\boldsymbol{\theta}} F(\cdot)(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathbb{R}^d\}$. Besides, $O_s \cap O_t = \emptyset$ if $s, t \in J$ and $s \neq t$. By Theorem 3.1, elements of $\mathcal{E}$ is of the form $\mathcal{M} = \bigcup_{j \in J'} O_j$, where $J' \subset J$ is an arbitrary index set (J' may be empty).

(i) Complement: Let $\mathcal{M} = \bigcup_{j \in J'} O_j$ be an arbitrary structural invariant set, where $J' \subset J$ is an index set. Then $\mathbb{R}^M \setminus \mathcal{M} = \bigcup_{j \in J \setminus J'} O_j$. Therefore $\mathbb{R}^M \setminus \mathcal{M} \in \mathcal{E}$.

(ii) Union and Intersection: Let $\mathcal{M}_i = \bigcup_{j \in J_i} O_j$ be arbitrary structural invariant sets for $i \in I$, where I is an arbitrary index set, and $J_i \subset J$ for all $i \in I$. Then it is straightforward

to verify $\bigcup_{i\in I}\mathcal{M}_i = \bigcup_{j\in\cup_{i\in I}J_i} O_j$, and $\bigcap_{i\in I}\mathcal{M}_i = \bigcup_{j\in\cap_{i\in I}J_i} O_j$. Therefore both $\bigcup_{i\in I}\mathcal{M}_i$ and $\bigcap_{i\in I}\mathcal{M}_i$ are in $\mathcal{E}$. □

As demonstrated in Proposition 3.3, linear models possess only trivial SIM $\mathbb{R}^M$, highlighting that the presence of nontrivial SIMs is a distinctive characteristic of nonlinear systems.

Proposition 3.3 (linear model has only trivial SIM). Let $\{\psi_1(\boldsymbol{x}),\ldots,\psi_M(\boldsymbol{x})\}$ be a set of linearly independent analytic functions defined on $\mathbb{R}^d$. Consider the linear model $F(\boldsymbol{\theta})(\boldsymbol{x}) = \sum_{i=1}^M \theta_i\psi_i(\boldsymbol{x})$, where $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_M) \in \mathbb{R}^M, \boldsymbol{x} \in \mathbb{R}^d$. Then F has only the trivial SIM, $\mathbb{R}^M$.

Proof. By Theorem 3.1, a SIM is a union of orbits of the family $\mathcal{F} = \{(\psi_1(\boldsymbol{x}),\ldots,\psi_M(\boldsymbol{x})) \mid \boldsymbol{x}\in\mathbb{R}^d\}$. Therefore, it suffices to show that $\mathcal{F}$ has a single orbit equal to $\mathbb{R}^M$.

Fix any $\boldsymbol{\theta}_0 \in \mathbb{R}^M$. Let $\mathcal{F}|_{\boldsymbol{\theta}_0}$ denote the evaluation of $\mathcal{F}$ at some $\boldsymbol{\theta}_0$. Suppose, for contradiction that $\mathrm{span}(\mathcal{F}|_{\boldsymbol{\theta}_0}) \subsetneq \mathbb{R}^M$. Then there exists a nonzero vector $\boldsymbol{c} = (c_1,\ldots,c_M) \in \mathbb{R}^M$ such that $\boldsymbol{c}$ is orthogonal to $\mathrm{span}(\mathcal{F}|_{\boldsymbol{\theta}_0})$, i.e., $\sum_{i=1}^M c_i\psi_i(\boldsymbol{x}) = 0, \forall \boldsymbol{x}\in\mathbb{R}^d$. This contradicts the assumption that $\{\psi_1(\boldsymbol{x}),\ldots,\psi_M(\boldsymbol{x})\}$ is a linearly independent set of functions. Therefore, $\mathrm{span}(\mathcal{F}|_{\boldsymbol{\theta}}) = \mathbb{R}^M$ for all $\boldsymbol{\theta}\in\mathbb{R}^M$. Since $\mathrm{span}(\mathcal{F}|_{\boldsymbol{\theta}}) \subset \mathrm{Lie}_{\boldsymbol{\theta}}(\mathcal{F}) \subset \mathbb{R}^M$, it follows that $\mathrm{Lie}_{\boldsymbol{\theta}}(\mathcal{F}) = \mathbb{R}^M$ for all $\boldsymbol{\theta}\in\mathbb{R}^M$. By Corollary 2.1, this implies that $\mathcal{F}$ has a single orbit equal to $\mathbb{R}^M$. □

4. Symmetry and Symmetry-Induced SIM

Based on Theorem 3.1, the problem of identifying all SIMs of an analytic parametric model F reduces to computing all orbits of $\mathcal{F}$. This task is particularly challenging, especially for deep neural networks (DNNs) with three or more layers. In the following, we prove a very general and useful mechanism for SIM generation, i.e., the invariant map, which induces a large subset of all SIMs.

4.1 Symmetry-induced SIM

We begin by examining how invariant maps give rise to SIMs. In Definition 4.1, we formally define two types of invariant maps: infinitesimal invariant maps and global invariant maps, which we collectively refer to as **invariant maps**. It is worth noting that every global invariant map is also an infinitesimal invariant map, but not vice versa.

The notion of infinitesimal invariant maps is introduced for the following reasons. First, the symmetries that give rise to SIMs often do not require global invariance; instead, it is sufficient for the invariance to hold in a local neighborhood of the manifold, or even merely at the level of tangent. This motivates the generalization from global to infinitesimal invariance. Second, in the context of neural networks, one encounters invariant maps that are globally defined but possess only tangent invariance (Proposition 4.1). Such maps are still capable of inducing SIMs despite exhibiting only this weaker form of invariance.

Definition 4.1 (infinitesimal and global invariant map). Let $F(\boldsymbol{\theta})(\boldsymbol{x})$ be an analytic parametric model with $\boldsymbol{\theta}\in\mathbb{R}^M$ and $\boldsymbol{x}\in\mathbb{R}^d$. For an analytic map $g:\mathbb{R}^M\to\mathbb{R}^M$, we say g is an **infinitesimal invariant map** if $\mathcal{M} := \{\boldsymbol{\theta}' \mid g(\boldsymbol{\theta}') = \boldsymbol{\theta}'\} \neq \emptyset$, and for any $\boldsymbol{\theta}\in\mathcal{M}$,

any $\boldsymbol{x} \in \mathbb{R}^d$, $D_{\boldsymbol{\theta}}\left(F(g(\boldsymbol{\theta}))(\boldsymbol{x})\right) = D_{\boldsymbol{\theta}}\left(F(\boldsymbol{\theta})(\boldsymbol{x})\right)$. Here $D_{\boldsymbol{\theta}}$ denotes the Jacobian matrix. Moreover, if $F(g(\boldsymbol{\theta}))(\boldsymbol{x}) = F(\boldsymbol{\theta})(\boldsymbol{x}), \forall \boldsymbol{\theta} \in \mathbb{R}^M, \boldsymbol{x} \in \mathbb{R}^d$, we say g is a **global invariant map**.

Theorem 4.1 provides a set of general conditions under which the fixed-point set of invariant maps forms a SIM. In contrast, Example 4.1 shows that invariant maps—even globally invariant ones—do not necessarily induce SIMs without the other conditions. Theorem 4.1 subsumes prior results such as the O-mirror symmetry in Liu (2024) and the symmetric loss in Simsek et al. (2021), up to a subtle distinction: those works consider symmetries of the empirical loss function $L(\boldsymbol{\theta})$, whereas we focus on symmetries of the parametric model. Ignoring this difference, our result can be viewed as a generalization of these earlier cases. Moreover, Theorem 4.1 applies not only to linear but also to nonlinear invariant maps, as illustrated in Example 4.2.

In addition to the symmetries considered in Theorem 4.1, continuous symmetries, as discussed in Liu et al. (2024), represent another class capable of inducing SIMs. These continuous symmetries typically manifest in homogeneous networks and matrix factorization models. However, the scope of this paper is intentionally focused on the discrete symmetries detailed in Theorem 4.1 as they are generally shared by all neural networks.

Theorem 4.1 (invariant maps induced SIM). Let $F(\boldsymbol{\theta})(\boldsymbol{x})$ be an analytic parametric model with $\boldsymbol{\theta} \in \mathbb{R}^M$ and $\boldsymbol{x} \in \mathbb{R}^d$. Let $\{g_i\}_{i\in I}$ be family of invariant maps of F. Define $\mathcal{M} = \{\boldsymbol{\theta} \mid g_i(\boldsymbol{\theta}) = \boldsymbol{\theta}, \forall i \in I\}$. Assume $\mathcal{M}$ is an immersed submanifold of $\mathbb{R}^M$ with its tangent space satisfying $T_{\boldsymbol{\theta}}\mathcal{M} = \bigcap_{i\in I} \ker(Dg_i^\intercal(\boldsymbol{\theta}) - \mathrm{id}_M), \forall \boldsymbol{\theta} \in \mathcal{M}$. Then $\mathcal{M}$ is a SIM.[4]

Proof. By Theorem 3.1, it suffices to prove that for any $\boldsymbol{x}_0 \in \mathbb{R}^d, \boldsymbol{\theta}_0 \in \mathcal{M}$, the solution $\boldsymbol{\theta}(t)$ to the Cauchy problem $\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}t} = \nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta}(t))(\boldsymbol{x}_0), \boldsymbol{\theta}(0) = \boldsymbol{\theta}_0$ remains in $\mathcal{M}$ for all t in its maximal interval of existence. Fix any $\boldsymbol{x}_0 \in \mathbb{R}^d$, and define the vector field $X(\boldsymbol{\theta}) := \nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta})(\boldsymbol{x}_0)$. To apply Lemma 3.1, we now show that $X(\boldsymbol{\theta}) \in T_{\boldsymbol{\theta}}\mathcal{M}, \forall \boldsymbol{\theta} \in \mathcal{M}$.

Since a global invariant map is always an infinitesimal invariant map, without loss of generality we assume g_i is an infinitesimal invariant map for each $i \in I$. Since g_i is an infinitesimal invariant map, $Dg_i(\boldsymbol{\theta})^\intercal \nabla_{\boldsymbol{\theta}} F(g_i(\boldsymbol{\theta}))(\boldsymbol{x}_0) = \nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta})(\boldsymbol{x}_0), \forall \boldsymbol{\theta} \in \mathcal{M}$. When $\boldsymbol{\theta} \in \mathcal{M}$, we have $g_i(\boldsymbol{\theta}) = \boldsymbol{\theta}$, so the equation becomes $Dg_i(\boldsymbol{\theta})^\intercal \nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta})(\boldsymbol{x}_0) = \nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta})(\boldsymbol{x}_0)$, implying $\nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta})(\boldsymbol{x}_0) \in \ker(Dg_i(\boldsymbol{\theta})^\intercal - \mathrm{id}_M)$. Since this holds for all $i \in I$, $X(\boldsymbol{\theta}) \in \bigcap_{i\in I} \ker(Dg_i(\boldsymbol{\theta})^\intercal - \mathrm{id}_M)$. By assumption, $\bigcap_{i\in I} \ker(Dg_i(\boldsymbol{\theta})^\intercal - \mathrm{id}_M) = T_{\boldsymbol{\theta}}\mathcal{M}$. Therefore $X(\boldsymbol{\theta}) \in T_{\boldsymbol{\theta}}\mathcal{M}, \forall \boldsymbol{\theta} \in \mathcal{M}$.

Moreover, each fixed-point set $\{\boldsymbol{\theta} \mid g_i(\boldsymbol{\theta}) = \boldsymbol{\theta}\}$ is closed (since g_i is continuous), so $\mathcal{M}$, being their intersection, is closed in $\mathbb{R}^M$. Since $X(\boldsymbol{\theta}) \in T_{\boldsymbol{\theta}}\mathcal{M}$ and $\mathcal{M}$ is a closed immersed submanifold, it follows from Lemma 3.1 that $\boldsymbol{\theta}(t) \in \mathcal{M}$ for all t in the maximal interval of existence. So $\mathcal{M}$ is a SIM. □

Remark 4.1. The assumption $\mathcal{M}$ is an immersed submanifold and $T_{\boldsymbol{\theta}}\mathcal{M} = \bigcap_{i\in I} \ker(Dg_i^\intercal(\boldsymbol{\theta}) - \mathrm{id}_M), \forall \boldsymbol{\theta} \in \mathcal{M}$ often holds under the following two conditions:
(i) $\mathcal{M}$ is an immersed submanifold with $T_{\boldsymbol{\theta}}\mathcal{M} = \bigcap_{i\in I} \ker(Dg_i(\boldsymbol{\theta}) - \mathrm{id}_M)$, which is automatically satisfied when all g_i are linear maps.

4. Dg_i is the jacobian matrix of g_i, and $Dg_i^\intercal$ is its transpose. id_M is the $M \times M$ identity matrix.

(ii) $\bigcap_{i\in I}\ker(Dg_i^\intercal(\boldsymbol{\theta})-\mathrm{id}_M)=\bigcap_{i\in I}\ker(Dg_i(\boldsymbol{\theta})-\mathrm{id}_M)$ for all $\boldsymbol{\theta}\in\mathcal{M}$, which is automatically satisfied when $Dg_i(\boldsymbol{\theta})$ is a linear normal operator for all $i\in I,\boldsymbol{\theta}\in\mathcal{M}$.

As a direct corollary, the assumption $\mathcal{M}$ is an immersed submanifold and $T_{\boldsymbol{\theta}}\mathcal{M}=\bigcap_{i\in I}\ker(Dg_i^\intercal(\boldsymbol{\theta})-\mathrm{id}_M),\forall\boldsymbol{\theta}\in\mathcal{M}$ holds if all g_i are linear normal operators. Besides, the regularity assumptions on $F(\boldsymbol{\theta})(\boldsymbol{x})$ and $g_i(\boldsymbol{\theta})$ can be weakened to C^1, and the domain of $\boldsymbol{\theta}$ can be taken to be any open set $U\subset\mathbb{R}^M$.

Example 4.1 (global invariant map may not induce SIM). Let $\boldsymbol{\theta}=(\theta_1,\theta_2)\in\mathbb{R}^2$, and define $F(\boldsymbol{\theta})(x)=(\theta_1-\theta_2)x$ for all $\boldsymbol{\theta}\in\mathbb{R}^2$, $x\in\mathbb{R}$. Consider the map $g(\boldsymbol{\theta})=(\theta_1+\theta_2,2\theta_2)$. Then g is a global invariant map of F. However, the fixed-point set of g, given by $\mathcal{M}=\{\boldsymbol{\theta}\mid\theta_2=0\}$, is not a SIM, since it is not invariant under all vector fields in $\mathcal{F}=\{(x,-x)\mid x\in\mathbb{R}\}$.

Example 4.2 (nonlinear symmetry). Let $F(\boldsymbol{\theta})(x)=(\sqrt{\theta_1^2+\theta_2^2}+\frac{1}{\sqrt{\theta_1^2+\theta_2^2}}+x)^2$, where $\boldsymbol{\theta}=(\theta_1,\theta_2)\in\mathbb{R}^2\setminus\{0\}$ and $x\in\mathbb{R}$. Define the map $g:\mathbb{R}^2\setminus\{0\}\to\mathbb{R}^2\setminus\{0\}$ by $g(\theta_1,\theta_2)=(\frac{\theta_1}{\theta_1^2+\theta_2^2},\frac{\theta_2}{\theta_1^2+\theta_2^2})$. Then g is a global invariant map of F, and its fixed-point set is $\mathcal{M}=\{(\theta_1,\theta_2)\in\mathbb{R}^2\mid\theta_1^2+\theta_2^2=1\}$. One can verify that the assumptions of Theorem 4.1 are satisfied. Therefore, $\mathcal{M}$ is a SIM.

Invariant maps can form a semigroup under composition. In neural networks, this semigroup is typically an **orthogonal symmetry group**, as defined in Definition 2.5. If a collection of SIMs forms a disjoint partition of the parameter space, we refer to this collection as an **invariant partition**. As shown in Lemma 4.1, a finite orthogonal symmetry group can induce such an invariant partition, where each leaf corresponds to a set of parameters sharing the same stabilizer subgroup.

In the context of invariant partitions, a natural partial order can be defined based on partition coarseness: given two partitions $\mathcal{P}_1$ and $\mathcal{P}_2$ of a set, we say that $\mathcal{P}_1$ is finer than $\mathcal{P}_2$ if every block of $\mathcal{P}_1$ is contained within some block of $\mathcal{P}_2$. Given an analytic model, the collection of orbits of $\mathcal{F}$ naturally forms an invariant partition. By Theorem 3.1, any SIM is a union of such orbits. It follows that the orbit partition is the finest invariant partition. Consequently, any invariant partition induced by an orthogonal symmetry group provides an upper bound for the orbit partition under this ordering.

Lemma 4.1 (invariant partition induced by an orthogonal symmetry group). Let $F(\boldsymbol{\theta})(\boldsymbol{x})$ with $\boldsymbol{\theta}\in\mathbb{R}^M$ be an analytic model, and let G be an orthogonal symmetry group of finite elements. For each $\boldsymbol{\theta}\in\mathbb{R}^M$, define its stabilizer subgroup as

$$S(\boldsymbol{\theta}):=\{g\in G\mid g(\boldsymbol{\theta})=\boldsymbol{\theta}\}.$$

Define an equivalence relation on the parameter space by $\boldsymbol{\theta}_1\sim\boldsymbol{\theta}_2\iff S(\boldsymbol{\theta}_1)=S(\boldsymbol{\theta}_2)$. Denote by $[\boldsymbol{\theta}]$ the equivalence class containing $\boldsymbol{\theta}$. Then the collection $\{[\boldsymbol{\theta}]\mid\boldsymbol{\theta}\in\mathbb{R}^M\}$ is an invariant foliation.

Proof. For any $g\in G$, its fixed-point set $\mathcal{M}_g:=\{\boldsymbol{\theta}\in\mathbb{R}^M\mid g(\boldsymbol{\theta})=\boldsymbol{\theta}\}$ is a SIM by Theorem 4.1 and Remark 4.1. A straightforward verification confirms that the defined relation satisfies the properties of an equivalence relation. Thus, it suffices to prove that for any $\boldsymbol{\theta}\in\mathbb{R}^M$, $[\boldsymbol{\theta}]$ is a SIM. Fix any $\boldsymbol{\theta}\in\mathbb{R}^M$. By definition, $[\boldsymbol{\theta}]$ consists of all parameters

whose stabilizer is exactly $S(\boldsymbol{\theta})$. This allows $[\boldsymbol{\theta}]$ to be expressed in terms of set operations on the family $\{\mathcal{M}_g \mid g \in G\}$ as follows:

$$[\boldsymbol{\theta}] = \left(\bigcap_{h \in S(\boldsymbol{\theta})} \mathcal{M}_h \right) \cap \left(\bigcap_{g \in G \setminus S(\boldsymbol{\theta})} (\mathbb{R}^M \setminus \mathcal{M}_g) \right).$$

Since the family of structural invariant sets is closed under finite intersection and complement (Proposition 3.2), $[\boldsymbol{\theta}]$ is a structural invariant set.

Moreover, since G has finite elements, the set $[\boldsymbol{\theta}]$ can be viewed as the linear space $\bigcap_{h \in S(\boldsymbol{\theta})} \mathcal{M}_h$ with finitely many linear subspaces $\left(\bigcap_{h \in S(\boldsymbol{\theta})} \mathcal{M}_h\right) \cap \mathcal{M}_g$ (for $g \notin S(\boldsymbol{\theta})$) removed. Consequently, $[\boldsymbol{\theta}]$ is relatively open in the linear subspace $\bigcap_{h \in S(\boldsymbol{\theta})} \mathcal{M}_h$, and hence is an immersed submanifold of $\mathbb{R}^M$. Therefore, $[\boldsymbol{\theta}]$ is a SIM.

□

4.2 Symmetries of neural networks

Symmetries are prevalent in deep neural networks, as detailed in Proposition 4.1 and Theorem 4.2 below. These symmetries generally originate from two principal sources. First, the indistinguishability of neurons within a given layer gives rise to the **permutation symmetry group**, denoted G_{per} (Theorem 4.2). Second, the symmetry of the activation function, $\sigma(x)$, constitute another source. Specifically, if $\sigma(x)$ possesses definite parity (i.e., is an odd or even function), a **reflection symmetry group**, G_{sign} or G'_{sign}, emerges as an orthogonal symmetry group (Theorem 4.2). Given that global invariant maps and orthogonal maps are closed under composition, the permutation and reflection group can generate more complex orthogonal symmetry groups under composition. Furthermore, local symmetries can also be identified. If $\sigma(0) = 0$ or $\sigma'(0) = 0$, the activation function $\sigma(x)$ exhibits infinitesimal odd or even behavior in the vicinity of the origin. This local property results in the actions of elements in G_{sign} or G'_{sign} manifesting as infinitesimal symmetry maps (Proposition 4.1).

To present Proposition 4.1 and Theorem 4.2, we first introduce Definition 4.2. The group $S_2^p \rtimes S_p$ in Definition 4.2 is also known as the hyper-octahedral group (Young, 1928).

Definition 4.2. Consider the multi-layer neural network F from Definition 2.3, with layer widths $n_0, \ldots, n_L$. For any positive integer p, let S_2^p denote the group of $p \times p$ diagonal sign matrices (entries in $\{\pm 1\}$), and let S_p denote the group of $p \times p$ permutation matrices. Define the semidirect product $S_2^p \rtimes S_p$ with the group operation given by

$$(\boldsymbol{\Lambda}_1, \boldsymbol{P}_1)(\boldsymbol{\Lambda}_2, \boldsymbol{P}_2) = \left(\boldsymbol{\Lambda}_1 \boldsymbol{P}_1 \boldsymbol{\Lambda}_2 \boldsymbol{P}_1^\top,\, \boldsymbol{P}_1 \boldsymbol{P}_2\right),$$

where $\boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2 \in S_2^p$ and $\boldsymbol{P}_1, \boldsymbol{P}_2 \in S_p$. One can readily verify that this structure satisfies the axioms of a semidirect product. We define the following groups and describe their action on the parameter space of F:

(i) Define the group $G_{\text{per}} = S_{n_1} \times \cdots \times S_{n_{L-1}}$, where $\times$ denotes the direct sum. For any $(\boldsymbol{P}^{(1)}, \ldots, \boldsymbol{P}^{(L-1)}) \in G_{\text{per}}$, define its action on the parameter space as

$$(\boldsymbol{P}^{(1)}, \ldots, \boldsymbol{P}^{(L-1)}) : \left(\boldsymbol{W}^{(l)}, \boldsymbol{b}^{(l)}\right)_{l=1}^{L} \mapsto \left(\boldsymbol{P}^{(l)} \boldsymbol{W}^{(l)} \boldsymbol{P}^{(l-1)\intercal}, \boldsymbol{P}^{(l)} \boldsymbol{b}^{(l)}\right)_{l=1}^{L},$$

with the conventions $\boldsymbol{P}^{(0)} = \mathrm{id}_{n_0}$ and $\boldsymbol{P}^{(L)} = \mathrm{id}_{n_L}$.

(ii) Define the group $G_{\text{sign}} = S_2^{n_1} \times \cdots \times S_2^{n_{L-1}}$. For any $(\boldsymbol{\Lambda}^{(1)}, \ldots, \boldsymbol{\Lambda}^{(L-1)}) \in G_{\text{sign}}$, define its action on the parameter space as

$$(\boldsymbol{\Lambda}^{(1)}, \ldots, \boldsymbol{\Lambda}^{(L-1)}) : \left(\boldsymbol{W}^{(l)}, \boldsymbol{b}^{(l)}\right)_{l=1}^{L} \mapsto \left(\boldsymbol{\Lambda}^{(l)} \boldsymbol{W}^{(l)} \boldsymbol{\Lambda}^{(l-1)}, \boldsymbol{\Lambda}^{(l)} \boldsymbol{b}^{(l)}\right)_{l=1}^{L},$$

with the conventions $\boldsymbol{\Lambda}^{(0)} = \mathrm{id}_{n_0}$ and $\boldsymbol{\Lambda}^{(L)} = \mathrm{id}_{n_L}$.

(iii) Define the group $G'_{\text{sign}} = S_2^{n_1} \times \cdots \times S_2^{n_L}$. For any $(\boldsymbol{\Lambda}^{(1)}, \ldots, \boldsymbol{\Lambda}^{(L)}) \in G'_{\text{sign}}$, define its action on the parameter space as

$$(\boldsymbol{\Lambda}^{(1)}, \ldots, \boldsymbol{\Lambda}^{(L)}) : \left(\boldsymbol{W}^{(l)}, \boldsymbol{b}^{(l)}\right)_{l=1}^{L} \mapsto \left(\boldsymbol{\Lambda}^{(l)} \boldsymbol{W}^{(l)}, \boldsymbol{\Lambda}^{(l)} \boldsymbol{b}^{(l)}\right)_{l=1}^{L}.$$

(iv) Define the group $G_{\text{combine}} = (S_2^{n_1} \rtimes S_{n_1}) \times \cdots \times (S_2^{n_{L-1}} \rtimes S_{n_{L-1}})$. For $g = ((\boldsymbol{\Lambda}^{(1)}, \boldsymbol{P}^{(1)}), \ldots, (\boldsymbol{\Lambda}^{(L-1)}, \boldsymbol{P}^{(L-1)})) \in G_{\text{combine}}$, define its action on parameter space as

$$g : \ \left(\boldsymbol{W}^{(l)}, \boldsymbol{b}^{(l)}\right)_{l=1}^{L} \mapsto \left(\boldsymbol{\Lambda}^{(l)} \boldsymbol{P}^{(l)} \boldsymbol{W}^{(l)} \boldsymbol{P}^{(l-1)\intercal} \boldsymbol{\Lambda}^{(l-1)}, \ \boldsymbol{\Lambda}^{(l)} \boldsymbol{P}^{(l)} \boldsymbol{b}^{(l)}\right)_{l=1}^{L},$$

with the conventions $\boldsymbol{P}^{(0)} = \boldsymbol{\Lambda}^{(0)} = \mathrm{id}_{n_0}, \boldsymbol{P}^{(L)} = \boldsymbol{\Lambda}^{(L)} = \mathrm{id}_{n_L}$.

Proposition 4.1 (infinitesimal symmetry of deep neural networks). Consider the multi-layer neural network from Definition 2.3, with layer widths $n_0, \ldots, n_L$. Let G_{sign} and G'_{sign} be the groups defined in Definition 4.2. Then the following statements hold:

(i) If $\sigma(0) = 0$, then the action of any element in G_{sign} is an infinitesimal invariant map.

(ii) If $\sigma'(0) = 0$, then the action of any element in G'_{sign} is an infinitesimal invariant map.

Proof. We use backpropagation to derive the gradients. For $l = 1, \ldots, L$, define $\boldsymbol{z}^{(l)} = \boldsymbol{W}^{(l)} \boldsymbol{a}^{(l-1)} + \boldsymbol{b}^{(l)} \in \mathbb{R}^{n_l}$ and $\boldsymbol{\delta}^{(l)} = \frac{\partial F}{\partial \boldsymbol{z}^{(l)}} \in \mathbb{R}^{n_l}$. Then $\boldsymbol{\delta}^{(L)} = \frac{\partial F}{\partial \boldsymbol{z}^{(L)}} = \sigma'(\boldsymbol{z}^{(L)})$, and $\boldsymbol{\delta}^{(l)} = \mathrm{diag}(\sigma'(\boldsymbol{z}^{(l)}))(\boldsymbol{W}^{(l+1)})^\intercal \boldsymbol{\delta}^{(l+1)}$ for $l = L-1, \ldots, 1$. The partial derivatives for $\boldsymbol{W}^{(l)}$ and $\boldsymbol{b}^{(l)}$ are given by $\frac{\partial F}{\partial \boldsymbol{W}^{(l)}} = \boldsymbol{\delta}^{(l)} (\boldsymbol{a}^{(l-1)})^\intercal \in \mathbb{R}^{n_l \times n_{l-1}}$ and $\frac{\partial F}{\partial \boldsymbol{b}^{(l)}} = \boldsymbol{\delta}^{(l)} \in \mathbb{R}^{n_l}$.

(i) Assume $\sigma(0) = 0$. Let $\boldsymbol{\Lambda} = (\boldsymbol{\Lambda}^{(1)}, \ldots, \boldsymbol{\Lambda}^{(L-1)})$ be an arbitrary element in G_{sign}. Let $\mathcal{M}$ be the set of fixed points of $\boldsymbol{\Lambda}$. Since $\boldsymbol{0} \in \mathcal{M}$, $\mathcal{M} \neq \emptyset$. Pick any $\boldsymbol{\theta} = \left(\boldsymbol{W}^{(l)}, \boldsymbol{b}^{(l)}\right)_{l=1}^{L} \in \mathcal{M}$ and any $\boldsymbol{x} \in \mathbb{R}^d$. For simplicity, we write $F(\boldsymbol{\theta})(\boldsymbol{x})$ as F. Since the action of $\boldsymbol{\Lambda}$ is linear and orthogonal, by Remark 4.1, to show that $\boldsymbol{\Lambda}$ is an infinitesimal invariant map, it suffices to prove $\frac{\partial F}{\partial \boldsymbol{W}^{(l)}} = \boldsymbol{\Lambda}^{(l)} \frac{\partial F}{\partial \boldsymbol{W}^{(l)}} \boldsymbol{\Lambda}^{(l-1)}$ and $\frac{\partial F}{\partial \boldsymbol{b}^{(l)}} = \boldsymbol{\Lambda}^{(l)} \frac{\partial F}{\partial \boldsymbol{b}^{(l)}}$ for $l = 1, \ldots, L$.

Since $\boldsymbol{\theta} \in \mathcal{M}$, $\boldsymbol{\Lambda}^{(l)} \boldsymbol{W}^{(l)} \boldsymbol{\Lambda}^{(l-1)} = \boldsymbol{W}^{(l)}$ and $\boldsymbol{\Lambda}^{(l)} \boldsymbol{b}^{(l)} = \boldsymbol{b}^{(l)}$ for $l = 1, \ldots, L$. For $l = 1, \ldots, L$, define $I_l = \{i \in \{1, \ldots, n_l\} \mid \boldsymbol{\Lambda}_{ii}^{(l)} = -1\}$ (note I_0 and I_L are empty sets as $\boldsymbol{\Lambda}^{(0)} = \boldsymbol{I}, \boldsymbol{\Lambda}^{(L)} = \boldsymbol{I}$). Since $\boldsymbol{\Lambda}^{(1)} \boldsymbol{W}^{(1)} \boldsymbol{\Lambda}^{(0)} = \boldsymbol{W}^{(1)}$, the j-th row $\boldsymbol{W}_j^{(1)} = \boldsymbol{0}$ for all $j \in I_1$. Similarly, $\boldsymbol{b}_j^{(1)} = 0$ for all $j \in I_1$. Thus $\boldsymbol{z}_j^{(1)} = 0$ for all $j \in I_1$. Since $\sigma(0) = 0$, $\boldsymbol{a}_j^{(1)} = 0$ for all $j \in I_1$.

Next, we prove by induction that $\boldsymbol{a}_j^{(l)} = 0$ for all $l \in \{1, \ldots, L\}$ and $j \in I_l$. Assume that $\boldsymbol{a}_j^{(l)} = 0, \forall j \in I_l$ holds for some $l \in \{1, \ldots, L-1\}$. Since $\boldsymbol{\Lambda}^{(l+1)} \boldsymbol{W}^{(l+1)} \boldsymbol{\Lambda}^{(l)} = \boldsymbol{W}^{(l+1)}$, we know $\boldsymbol{W}_{ij}^{(l+1)} = 0$ if $i \in I_{l+1}$ and $j \notin I_l$, or if $i \notin I_{l+1}$ and $j \in I_l$. Since $\boldsymbol{\Lambda}^{(l+1)} \boldsymbol{b}^{(l+1)} = \boldsymbol{b}^{(l+1)}$,

we have $\boldsymbol{b}_i^{(l+1)} = 0, \forall i \in I_{l+1}$. Consider $\boldsymbol{z}_i^{(l+1)} = \sum_{j=1}^{n_l} \boldsymbol{W}_{ij}^{(l+1)} \boldsymbol{a}_j^{(l)} + \boldsymbol{b}_i^{(l+1)}$ for any $i \in I_{l+1}$. If $j \notin I_l$, then $\boldsymbol{W}_{ij}^{(l+1)} = 0$. If $j \in I_l$, then $\boldsymbol{a}_j^{(l)} = 0$ by the induction hypothesis. In both cases, $\boldsymbol{W}_{ij}^{(l+1)} \boldsymbol{a}_j^{(l)} = 0$. As $\boldsymbol{b}_i^{(l+1)} = 0$, we have $\boldsymbol{z}_i^{(l+1)} = 0, \forall i \in I_{l+1}$. Since $\boldsymbol{a}^{(l+1)} = \sigma(\boldsymbol{z}^{(l+1)})$ and $\sigma(0) = 0$, $\boldsymbol{a}_i^{(l+1)} = 0, \forall i \in I_{l+1}$. By mathematical induction, $\boldsymbol{a}_j^{(l)} = 0$ for all $l \in \{1, \ldots, L\}, j \in I_l$.

Next, we prove by backward induction that $\boldsymbol{\delta}_i^{(l)} = 0$ for all $l \in \{1, \ldots, L\}, i \in I_l$. When $l = L$, $I_L = \emptyset$, so the statement holds vacuously. Assume $\boldsymbol{\delta}_i^{(l)} = 0, \forall i \in I_l$ for some $l \in \{2, \ldots, L\}$. We have $\boldsymbol{\delta}_i^{(l-1)} = \sigma'(\boldsymbol{z}_i^{(l-1)}) \sum_{j=1}^{n_l} \boldsymbol{W}_{ji}^{(l)} \boldsymbol{\delta}_j^{(l)}$. For any $i \in I_{l-1}$, if $j \in I_l$, then $\boldsymbol{\delta}_j^{(l)} = 0$ by the induction hypothesis. If $j \notin I_l$, then $\boldsymbol{W}_{ji}^{(l)} = 0$ by the fixed point condition $\boldsymbol{\Lambda}^{(l)} \boldsymbol{W}^{(l)} \boldsymbol{\Lambda}^{(l-1)} = \boldsymbol{W}^{(l)}$, as $i \in I_{l-1}$. Thus, the sum is zero, and $\boldsymbol{\delta}_i^{(l-1)} = 0, \forall i \in I_{l-1}$. By induction, $\boldsymbol{\delta}_i^{(l)} = 0, \forall l \in \{1, \ldots, L\}, i \in I_l$.

For any $l = 1, \ldots, L$, $\frac{\partial F}{\partial \boldsymbol{W}^{(l)}} = \boldsymbol{\delta}^{(l)} (\boldsymbol{a}^{(l-1)})^\intercal$. Since $\boldsymbol{\delta}_i^{(l)} = 0, \forall i \in I_l$, the i-th row of $\frac{\partial F}{\partial \boldsymbol{W}^{(l)}}$ is zero. Since $\boldsymbol{a}_j^{(l-1)} = 0, \forall j \in I_{l-1}$, the j-th column is zero. This implies $\frac{\partial F}{\partial \boldsymbol{W}^{(l)}} = \boldsymbol{\Lambda}^{(l)} \frac{\partial F}{\partial \boldsymbol{W}^{(l)}} \boldsymbol{\Lambda}^{(l-1)}$. For any $l = 1, \ldots, L$, $\frac{\partial F}{\partial \boldsymbol{b}^{(l)}} = \boldsymbol{\delta}^{(l)}$. Since $\boldsymbol{\delta}_i^{(l)} = 0, \forall i \in I_l$, this implies $\frac{\partial F}{\partial \boldsymbol{b}^{(l)}} = \boldsymbol{\Lambda}^{(l)} \frac{\partial F}{\partial \boldsymbol{b}^{(l)}}$. Therefore, $\boldsymbol{\Lambda}$ is an infinitesimal invariant map.

(ii) Assume $\sigma'(0) = 0$. Let $\boldsymbol{\Lambda} = (\boldsymbol{\Lambda}^{(1)}, \ldots, \boldsymbol{\Lambda}^{(L)})$ be an arbitrary element in G'_{sign}. Let $\mathcal{M}$ be the set of fixed points of $\boldsymbol{\Lambda}$. Since $\mathbf{0} \in \mathcal{M}$, $\mathcal{M} \neq \emptyset$. Pick any $\boldsymbol{\theta} = \left(\boldsymbol{W}^{(l)}, \boldsymbol{b}^{(l)}\right)_{l=1}^{L} \in \mathcal{M}$. Similar to (1), to show $\boldsymbol{\Lambda}$ is an infinitesimal invariant map, we only need to prove $\frac{\partial F}{\partial \boldsymbol{W}^{(l)}} = \boldsymbol{\Lambda}^{(l)} \frac{\partial F}{\partial \boldsymbol{W}^{(l)}}$ and $\frac{\partial F}{\partial \boldsymbol{b}^{(l)}} = \boldsymbol{\Lambda}^{(l)} \frac{\partial F}{\partial \boldsymbol{b}^{(l)}}$ for $l = 1, \ldots, L$.

For $l = 1, \ldots, L$, define $I_l = \{i \in \{1, \ldots, n_l\} \mid \boldsymbol{\Lambda}_{ii}^{(l)} = -1\}$. Since $\boldsymbol{\theta} \in \mathcal{M}$, $\boldsymbol{\Lambda}^{(l)} \boldsymbol{W}^{(l)} = \boldsymbol{W}^{(l)}$ and $\boldsymbol{\Lambda}^{(l)} \boldsymbol{b}^{(l)} = \boldsymbol{b}^{(l)}$ for $l = 1, \ldots, L$. This implies that for any $l = 1, \ldots, L$ and $j \in I_l$, the j-th row $\boldsymbol{W}_j^{(l)} = \mathbf{0}$ and $\boldsymbol{b}_j^{(l)} = 0$. Thus, $\boldsymbol{z}_j^{(l)} = (\boldsymbol{W}^{(l)} \boldsymbol{a}^{(l-1)})_j + \boldsymbol{b}_j^{(l)} = \mathbf{0} \cdot \boldsymbol{a}^{(l-1)} + 0 = 0$ for all $j \in I_l$. For $l < L$, $\boldsymbol{\delta}^{(l)} = \operatorname{diag}(\sigma'(\boldsymbol{z}^{(l)}))(\boldsymbol{W}^{(l+1)})^\intercal \boldsymbol{\delta}^{(l+1)}$. Since $\boldsymbol{z}_j^{(l)} = 0$ for $j \in I_l$, the j-th diagonal entry $\sigma'(\boldsymbol{z}_j^{(l)}) = \sigma'(0) = 0$, which implies $\boldsymbol{\delta}_j^{(l)} = 0$. Thus, for all $l = 1, \ldots, L$, we have $\boldsymbol{\delta}_j^{(l)} = 0$ for all $j \in I_l$. When $l = L$, since $I_L = \emptyset$, $\boldsymbol{\delta}_j^{(l)} = 0, \forall j \in I_l$ holds.

Since $\frac{\partial F}{\partial \boldsymbol{W}^{(l)}} = \boldsymbol{\delta}^{(l)} (\boldsymbol{a}^{(l-1)})^\intercal$ and $\frac{\partial F}{\partial \boldsymbol{b}^{(l)}} = \boldsymbol{\delta}^{(l)}$, the j-th row of both $\frac{\partial F}{\partial \boldsymbol{W}^{(l)}}$ and $\frac{\partial F}{\partial \boldsymbol{b}^{(l)}}$ is zero for all $j \in I_l$. This directly implies $\frac{\partial F}{\partial \boldsymbol{W}^{(l)}} = \boldsymbol{\Lambda}^{(l)} \frac{\partial F}{\partial \boldsymbol{W}^{(l)}}$ and $\frac{\partial F}{\partial \boldsymbol{b}^{(l)}} = \boldsymbol{\Lambda}^{(l)} \frac{\partial F}{\partial \boldsymbol{b}^{(l)}}$ for all $l = 1, \ldots, L$. Thus, $\boldsymbol{\Lambda}$ is an infinitesimal invariant map. □

Theorem 4.2 (symmetry-induced SIMs of deep neural networks). Consider the multi-layer neural network from Definition 2.3, with layer widths $n_0, \ldots, n_L$. Consider the groups and actions as defined in Definition 4.2. Then the following statements hold:

(i) G_{per} is an orthogonal symmetry group and thus induces an invariant partition.

(ii) If $\sigma(x)$ is an odd function, then G_{sign} is an orthogonal symmetry group. Moreover, G_{sign} and G_{per} generate a new orthogonal symmetry group under map composition, which equals G_{combine}. Therefore, G_{combine} induces an invariant partition. [5]

5. The case in which σ is even is analogous; we omit it for brevity.

(iii) Assume $\sigma(0) = 0$. Let I_0 and I_L to be empty sets. For any choice of subsets $I_l \subset \{1, \dots, n_l\}$ for each $l \in \{1, \dots, L-1\}$, define

$$\mathcal{M} = \left\{ \left(\boldsymbol{W}^{(l)}, \boldsymbol{b}^{(l)}\right)_{l=1}^{L} \;\middle|\; \begin{array}{l} \boldsymbol{W}_{ij}^{(l)} = 0, \quad \forall l \in \{1, \dots, L\},\ (i,j) \in I_l \times I_{l-1}^c \cup I_l^c \times I_{l-1}, \\ \boldsymbol{b}_i^{(l)} = 0, \quad \forall l \in \{1, \dots, L\},\ i \in I_l, \\ \boldsymbol{W}_{ij}^{(l)} = c_{ij}^{(l)}, \quad \forall l \in \{1, \dots, L\},\ (i,j) \in I_l \times I_{l-1}. \end{array} \right\},$$

where each $c_{ij}^{(l)}$ is an arbitrary real number, $\boldsymbol{W}_{ij}^{(l)}$ represents the matrix entry of $\boldsymbol{W}^{(l)}$ at position (i, j), and $I_l^c = \{1, \dots, n_l\} \setminus I_l$. Then $\mathcal{M}$ is a SIM.

(iv) If $\sigma'(0) = 0$, then for each $l \in \{1, \dots, L\}$ and $j \in \{1, \dots, n_l\}$, the set

$$\mathcal{M}_{l,j} = \left\{ \left(\boldsymbol{W}^{(k)}, \boldsymbol{b}^{(k)}\right)_{k=1}^{L} \mid \boldsymbol{W}_j^{(l)} = \boldsymbol{0} \text{ and } \boldsymbol{b}_j^{(l)} = 0 \right\}$$

is a SIM. Here, $\boldsymbol{W}_j^{(l)}$ denotes the j-th row of $\boldsymbol{W}^{(l)}$.

Proof. (i) Let $\boldsymbol{P} = (\boldsymbol{P}^{(1)}, \dots, \boldsymbol{P}^{(L-1)})$ and $\boldsymbol{P}' = (\boldsymbol{P}'^{(1)}, \dots, \boldsymbol{P}'^{(L-1)})$ be elements of G_{per}. Then the composition of their actions is given by

$$\boldsymbol{P}' \circ \boldsymbol{P} : \left(\boldsymbol{W}^{(l)}, \boldsymbol{b}^{(l)}\right)_{l=1}^{L} \mapsto \left(\boldsymbol{P}'^{(l)} \boldsymbol{P}^{(l)} \boldsymbol{W}^{(l)} (\boldsymbol{P}'^{(l-1)} \boldsymbol{P}^{(l-1)})^{\intercal}, \boldsymbol{P}'^{(l)} \boldsymbol{P}^{(l)} \boldsymbol{b}^{(l)}\right)_{l=1}^{L},$$

which is exactly the action of $\boldsymbol{P}'\boldsymbol{P}$. Therefore, it is a group action. Recall $\boldsymbol{a}^{(l)} = \sigma(\boldsymbol{W}^{(l)}\boldsymbol{a}^{(l-1)} + \boldsymbol{b}^{(l)})$. We claim that the action of $\boldsymbol{P}$ changes $\boldsymbol{a}^{(l)}$ to $\boldsymbol{P}^{(l)}\boldsymbol{a}^{(l)}$ for $l = 0, 1, \dots, L$. For $l = 0$, $\boldsymbol{a}^{(0)} = \boldsymbol{x}$. Since $\boldsymbol{P}^{(0)} = \text{id}_{\text{n}_0}$, the claim holds when $l = 0$. Suppose this holds for some $l \in \{0, 1, \dots, L-1\}$. Then $\boldsymbol{a}^{(l+1)}$ is changed to $\sigma(\boldsymbol{P}^{(l+1)}\boldsymbol{W}^{(l+1)}\boldsymbol{P}^{(l)\intercal}\boldsymbol{P}^{(l)}\boldsymbol{a}^{(l)} + \boldsymbol{P}^{(l+1)}\boldsymbol{b}^{(l+1)}) = \sigma(\boldsymbol{P}^{(l+1)}(\boldsymbol{W}^{(l+1)}\boldsymbol{a}^{(l)} + \boldsymbol{b}^{(l+1)})) = \boldsymbol{P}^{(l+1)}\sigma(\boldsymbol{W}^{(l+1)}\boldsymbol{a}^{(l)} + \boldsymbol{b}^{(l+1)}) = \boldsymbol{P}^{(l+1)}\boldsymbol{a}^{(l+1)}$. By mathematical induction, the action of $\boldsymbol{P}$ changes $\boldsymbol{a}^{(l)}$ to $\boldsymbol{P}^{(l)}\boldsymbol{a}^{(l)}$. Since $\boldsymbol{P}^{(L)} = \text{id}_{n_L}$, the action of $\boldsymbol{P}$ does not change the output of the model. Moreover, the permutation of coordinates is linear and does not change the norm of a vector. Therefore G_{per} is an orthogonal symmetry group.

(ii) Similar to the proof of (i), one can verify that the any action of element in G_{sign} is a linear orthogonal operator that does not change the output of the model. So G_{sign} is an orthogonal symmetry group. The check that G_{combine} is the group generated by G_{sign} and G_{per} is straightforward, and we omit the details. Because both global invariant maps and orthogonal maps are closed under composition, G_{combine} is also an orthogonal symmetry group.

(iii) In the proof of the first statement of Proposition 4.1, one sees that the set

$$\mathcal{M}' = \left\{ \left(\boldsymbol{W}^{(l)}, \boldsymbol{b}^{(l)}\right)_{l=1}^{L} \;\middle|\; \begin{array}{l} \boldsymbol{W}_{ij}^{(l)} = 0, \quad \forall l \in \{1, \dots, L\},\ (i,j) \in I_l \times I_{l-1}^c \cup I_l^c \times I_{l-1}, \\ \boldsymbol{b}_i^{(l)} = 0, \quad \forall l \in \{1, \dots, L\},\ i \in I_l. \end{array} \right\}$$

is the fixed point of $\boldsymbol{\Lambda} = (\boldsymbol{\Lambda}^{(1)}, \dots, \boldsymbol{\Lambda}^{(L-1)})$, where $\boldsymbol{\Lambda}^{(l)}$ is the diagonal matrix with entries equal to -1 for row indices in I_l and $+1$ otherwise. By Proposition 4.1, $\boldsymbol{\Lambda}$ is an infinitesimal invariant map. Since $\boldsymbol{\Lambda}$ is linear and orthogonal, by Theorem 4.1 and Remark 4.1, $\mathcal{M}'$ is a

SIM. By the proof of the first statement of Proposition 4.1, the i-th row and j-th column of $\frac{\partial F}{\partial \boldsymbol{W}^{(l)}}$ are zero for any $i \in I_l, j \in I_{l-1}$. Therefore, for any $i \in I_l, j \in I_{l-1}$, $\boldsymbol{W}_{ij}^{(l)}$ remains constant during training. Therefore $\mathcal{M}$ is a SIM.

(iv) Fix any $l \in \{1, \dots, L\}$ and $j \in \{1, \dots, n_l\}$. Define $\boldsymbol{\Lambda}^{(l)}$ to be the $n_l \times n_l$ diagonal matrix such that the diagonal satisfies $\boldsymbol{\Lambda}_{jj}^{(l)} = -1$ and $\boldsymbol{\Lambda}_{ii}^{(l)} = 1, i \neq j$. Define $\boldsymbol{\Lambda} = (\mathrm{id}_{n_1}, \dots, \mathrm{id}_{n_{l-1}}, \boldsymbol{\Lambda}^{(l)}, \mathrm{id}_{n_{l+1}}, \dots, \mathrm{id}_{n_L})$. By Proposition 4.1, $\boldsymbol{\Lambda}$ is an infinitesimal invariant map. It is easy to see that $\boldsymbol{\Lambda}$ is an orthogonal linear map. By Theorem 4.1 and Remark 4.1, the fixed point of $\boldsymbol{\Lambda}$ is a SIM. Since the fixed point of $\boldsymbol{\Lambda}$ is $\mathcal{M}_{l,j}$, $\mathcal{M}_{l,j}$ is a SIM. □

Remark 4.2. The symmetries in Theorem 4.2 can be extended to other architectures, such as ResNet, Convolutional Neural Networks, and Transformers.

5. Orbits of Two-layer Neural Networks

This section considers the two-layer neural network from Definition 2.4. The following two propositions characterize the invariant partitions induced by the permutation symmetry group G_{per} and the combined symmetry groups G_{combine} in Definition 4.2.

Proposition 5.1 (invariant partition induced by the permutation symmetry group). Consider the two-layer neural network of width m and the group G_{per} in Definition 4.2. By Theorem 4.2, G_{per} induces an invariant partition. Let $\mathfrak{P}_m$ denote the set of all partitions of $\{1, \dots, m\}$. For each partition $\mathcal{P} = \{B_1, \dots, B_s\} \in \mathfrak{P}_m$, define

$$\mathcal{M}_{\mathcal{P}} := \left\{ (a_i, \boldsymbol{w}_i)_{i=1}^m \in \mathbb{R}^{(d+1)m} \;\middle|\; \begin{array}{ll} (a_i, \boldsymbol{w}_i) = (a_j, \boldsymbol{w}_j), & \forall p \in \{1, \dots, s\}, \forall i, j \in B_p \\ (a_i, \boldsymbol{w}_i) \neq (a_j, \boldsymbol{w}_j), & \forall p, p' \in \{1, \dots, s\}, p \neq p', \forall i \in B_p, j \in B_{p'} \end{array} \right\}.$$

Then the collection $\{\mathcal{M}_{\mathcal{P}} \mid \mathcal{P} \in \mathfrak{P}_m\}$ equals the invariant partition induced by G_{per}.

Proof. For two-layer neural networks of width m, G_{per} is simply the group of $m \times m$ permutation matrices, denoted by S_m . Since it is isomorphic to the symmetric group of degree m, we sightly abuse the notation by denoting S_m to be the symmetric group of degree m in the proof. The action of S_m on the parameter space is given by:

$$\pi : (a_i, \boldsymbol{w}_i)_{i=1}^m \mapsto (a_{\pi^{-1}(i)}, \boldsymbol{w}_{\pi^{-1}(i)})_{i=1}^m, \forall \pi \in S_m.$$

For any parameter vector $\boldsymbol{\theta} = (a_i, \boldsymbol{w}_i)_{i=1}^m \in \mathbb{R}^{(d+1)m}$, we define an equivalence relation $\sim$ on the set of indices $\{1, \dots, m\}$ such that $i \sim j$ if and only if $(a_i, \boldsymbol{w}_i) = (a_j, \boldsymbol{w}_j)$. This relation induces a partition of $\{1, \dots, m\}$, which we denote by $\mathcal{P}_{\boldsymbol{\theta}} = \{B_1, \dots, B_s\}$.

Fix any $\boldsymbol{\theta} = (a_i, \boldsymbol{w}_i)_{i=1}^m \in \mathbb{R}^{(d+1)m}$. Let $S(\boldsymbol{\theta})$ be the stabilizer subgroup of $\boldsymbol{\theta}$ in S_m. An element $\pi \in S_m$ belongs to $S(\boldsymbol{\theta})$ if and only if $\pi(\boldsymbol{\theta}) = \boldsymbol{\theta}$, i.e. $(a_{\pi^{-1}(i)}, \boldsymbol{w}_{\pi^{-1}(i)}) = (a_i, \boldsymbol{w}_i), \forall i \in \{1, \dots, m\}$. By definition of $\mathcal{P}_{\boldsymbol{\theta}}$, this holds if and only if for each i, the indices i and $\pi^{-1}(i)$ belong to the same block in the partition $\mathcal{P}_{\boldsymbol{\theta}}$. This is true if and only if π permutes the indices within each block B_l for $l = 1, \dots, s$. Consequently,

$$S(\boldsymbol{\theta}) = S_{|B_1|} \times S_{|B_2|} \times \cdots \times S_{|B_s|}, \tag{4}$$

where $S_{|B_l|}$ is the symmetric group on the set B_l. One can readily verify that
(i): If $\boldsymbol{\theta} \in \mathcal{M}_{\mathcal{P}_{\boldsymbol{\theta}_0}}$, then $\mathcal{P}_{\boldsymbol{\theta}} = \mathcal{P}_{\boldsymbol{\theta}_0}$. Thus, $S(\boldsymbol{\theta}) = S(\boldsymbol{\theta}_0)$.
(ii): If $\boldsymbol{\theta} \notin \mathcal{M}_{\mathcal{P}_{\boldsymbol{\theta}_0}}$, then $\mathcal{P}_{\boldsymbol{\theta}} \neq \mathcal{P}_{\boldsymbol{\theta}_0}$. Thus, $S(\boldsymbol{\theta}) \neq S(\boldsymbol{\theta}_0)$.
Then by definition of the equivalence class, we have $[\boldsymbol{\theta}_0] = \mathcal{M}_{\mathcal{P}_{\boldsymbol{\theta}_0}}$. Since $\boldsymbol{\theta}_0$ is arbitrary, $\{\mathcal{M}_{\mathcal{P}} \mid \mathcal{P} \in \mathfrak{P}_m\}$ is the invariant partition induced by G_{per}.

□

Proposition 5.2 (invariant partition induced by the combined symmetry group). Consider the two-layer neural network of width m with odd activation function. Consider the group G_{combine} in Definition 4.2. By Theorem 4.2, G_{combine} induces an invariant partition. Let $\mathfrak{P}_m$ denote the set of all partitions of $\{1, \ldots, m\}$. For each partition $\mathcal{P} = \{B_1, \ldots, B_s\} \in \mathfrak{P}_m$ (B_1 can be empty) and each $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_m) \in \{-1, 1\}^m$, define

$$\mathcal{M}_{\mathcal{P},\boldsymbol{\gamma}} := \left\{ (a_i, \boldsymbol{w}_i)_{i=1}^m \in \mathbb{R}^{(d+1)m} \,\middle|\, \begin{array}{ll} (a_i, \boldsymbol{w}_i) = \boldsymbol{0}, & \forall i \in B_1, \\ \gamma_i(a_i, \boldsymbol{w}_i) = \gamma_j(a_j, \boldsymbol{w}_j), & \forall p \in \{2, \ldots, s\}, \forall i, j \in B_p, \\ (a_i, \boldsymbol{w}_i) \neq \pm(a_j, \boldsymbol{w}_j), & \forall p, p' \in \{1, \ldots, s\}, p \neq p', \forall i \in B_p, j \in B_{p'} \end{array} \right\}.$$

Then the collection $\{\mathcal{M}_{\mathcal{P},\boldsymbol{\gamma}} \mid \mathcal{P} \in \mathfrak{P}_m,\ \boldsymbol{\gamma} \in \{-1, 1\}^m\}$ equals the invariant partition induced by G_{combine}.

Proof. For two-layer neural networks, G_{combine} is simply $S_2^m \rtimes S_m$ defined in Definition 4.2. With a slight abuse of notation, S_m and S_2^m are denoted to be the symmetric group of order m and $\{-1, 1\}^m$ (m products of the group $\{-1, 1\}$), respectively. Then an element of G is a pair $(\boldsymbol{\delta}, \pi)$ where $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_m) \in \{-1, 1\}^m$ and $\pi \in S_m$. The action of the pair $(\boldsymbol{\delta}, \pi)$ is given by:

$$(\boldsymbol{\delta}, \pi) : (a_i, \boldsymbol{w}_i)_{i=1}^m \mapsto \left(\delta_i a_{\pi^{-1}(i)}, \delta_i \boldsymbol{w}_{\pi^{-1}(i)}\right)_{i=1}^m.$$

First, we note that for any $\boldsymbol{\theta} \in \mathbb{R}^{(d+1)m}$, we can find a partition $\mathcal{P} \in \mathfrak{P}_m$ and a sign vector $\boldsymbol{\gamma} \in \{-1, 1\}^m$ such that $\boldsymbol{\theta} \in \mathcal{M}_{\mathcal{P},\boldsymbol{\gamma}}$. Thus, the collection of sets $\{\mathcal{M}_{\mathcal{P},\boldsymbol{\gamma}}\}$ covers the entire parameter space. Therefore, to prove that this collection is the invariant partition of $S_2^m \rtimes S_m$, we only need to show that for any $\mathcal{P} \in \mathfrak{P}_m, \boldsymbol{\gamma} \in \{-1, 1\}^m$, any $\boldsymbol{\theta}_0 \in \mathcal{M}_{\mathcal{P},\boldsymbol{\gamma}}$, the identity $[\boldsymbol{\theta}_0] = \mathcal{M}_{\mathcal{P},\boldsymbol{\gamma}}$ holds. Now fix any $\mathcal{P} = \{B_1, \ldots, B_s\} \in \mathfrak{P}_m$, any $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_m) \in \{-1, 1\}^m$, and any $\boldsymbol{\theta}_0 \in \mathcal{M}_{\mathcal{P},\boldsymbol{\gamma}}$.
Step 1: Prove $\mathcal{M}_{\mathcal{P},\boldsymbol{\gamma}} \subset [\boldsymbol{\theta}_0]$. Let $\boldsymbol{\theta} = (a_i, \boldsymbol{w}_i)_{i=1}^m$ be an arbitrary element in $\mathcal{M}_{\mathcal{P},\boldsymbol{\gamma}}$. Denote $G = S_2^m \rtimes S_m$. An element $(\boldsymbol{\delta}, \pi) \in G$ is in $S(\boldsymbol{\theta})$ if and only if $(\boldsymbol{\delta}, \pi) \cdot \boldsymbol{\theta} = \boldsymbol{\theta}$, which means:

$$(a_i, \boldsymbol{w}_i) = \delta_i(a_{\pi^{-1}(i)}, \boldsymbol{w}_{\pi^{-1}(i)}) \quad \text{for all } i \in \{1, \ldots, m\}. \tag{5}$$

If $(\boldsymbol{\delta}, \pi) \in S(\boldsymbol{\theta})$, then $(\boldsymbol{\delta}, \pi)$ must satisfy the following conditions:

(i) From the third condition of $\mathcal{M}_{\mathcal{P},\boldsymbol{\gamma}}$, we have $(a_i, \boldsymbol{w}_i) \neq \pm(a_j, \boldsymbol{w}_j)$ if i and j are in different blocks of the partition $\mathcal{P}$. Equation (5) can only hold if for every block $B_l \in \mathcal{P}$, the permutation π maps B_l to itself, i.e., $\pi(B_l) = B_l$.

(ii) For any $i \in B_1$, we have $(a_i, \boldsymbol{w}_i) = \boldsymbol{0}$. Since $\pi(B_1) = B_1$, $\pi^{-1}(i)$ is also in B_1, so $(a_{\pi^{-1}(i)}, \boldsymbol{w}_{\pi^{-1}(i)}) = \boldsymbol{0}$. The condition becomes $\boldsymbol{0} = \delta_i \boldsymbol{0}$, which holds for any $\delta_i \in \{-1, 1\}$.

(iii) For any $i \in B_l$ with $l \in \{2, \ldots, s\}$, we have $(a_i, \boldsymbol{w}_i) \neq \boldsymbol{0}$. From the second condition of $\mathcal{M}_{\mathcal{P},\boldsymbol{\gamma}}$, we know that $\gamma_i(a_i, \boldsymbol{w}_i) = \gamma_j(a_j, \boldsymbol{w}_j)$ for any $i, j \in B_l$. Applying this to the

pair $i, \pi^{-1}(i) \in B_l$, we get $\gamma_i(a_i, \boldsymbol{w}_i) = \gamma_{\pi^{-1}(i)}(a_{\pi^{-1}(i)}, \boldsymbol{w}_{\pi^{-1}(i)})$. Substituting this into the stabilizer condition Eq. (5), we find:

$$\gamma_i(a_i, \boldsymbol{w}_i) = \gamma_{\pi^{-1}(i)} \left(\frac{1}{\delta_i}(a_i, \boldsymbol{w}_i) \right) \implies \delta_i = \frac{\gamma_{\pi^{-1}(i)}}{\gamma_i}.$$

Since $\gamma_k \in \{-1, 1\}$, this is equivalent to $\delta_i = \gamma_i \gamma_{\pi^{-1}(i)}$.

Denote H to be the set of all pairs $(\boldsymbol{\delta}, \pi) \in G$ such that:

(i) $\pi(B_l) = B_l$ for all $l \in \{1, \ldots, s\}$.

(ii) $\delta_i \in \{-1, 1\}$ is arbitrary for $i \in B_1$.

(iii) $\delta_i = \gamma_i \gamma_{\pi^{-1}(i)}$ for all $i \in B_l$ where $l \in \{2, \ldots, s\}$.

By previous argument, $S(\boldsymbol{\theta}) \subset H$. Conversely, it is easy to verify that Eq. (5) holds whenever $(\boldsymbol{\delta}, \pi) \in H$. Therefore $H \subset S(\boldsymbol{\theta})$. By both inclusions, $H = S(\boldsymbol{\theta})$. Since H depends only on the partition $\mathcal{P}$ and the sign vector $\boldsymbol{\gamma}$, all elements of $\mathcal{M}_{\mathcal{P},\boldsymbol{\gamma}}$ have the same stabilizer subgroup, and thus $\mathcal{M}_{\mathcal{P},\boldsymbol{\gamma}} \subset [\boldsymbol{\theta}_0]$.

Step 2: Prove $[\boldsymbol{\theta}_0] \subset \mathcal{M}_{\mathcal{P},\boldsymbol{\gamma}}$. Let $\boldsymbol{\theta}' = (a_i', \boldsymbol{w}_i')_{i=1}^m$ be an arbitrary element in $[\boldsymbol{\theta}_0]$. We will show that $\boldsymbol{\theta}' \in \mathcal{M}_{\mathcal{P},\boldsymbol{\gamma}}$. Since $S(\boldsymbol{\theta}') = S(\boldsymbol{\theta}_0)$, the following conditions hold:

(i) For any $i \in B_1$, the element $(\boldsymbol{\delta}, \mathrm{id})$ where $\delta_i = -1$ and $\delta_j = 1$ for $j \neq i$ is in $S(\boldsymbol{\theta}_0)$. Since $S(\boldsymbol{\theta}') = S(\boldsymbol{\theta}_0)$, $(\boldsymbol{\delta}, \mathrm{id}) \in S(\boldsymbol{\theta}')$. Therefore $(\boldsymbol{\delta}, \mathrm{id}) \cdot \boldsymbol{\theta}' = \boldsymbol{\theta}'$, which implies $(a_i', \boldsymbol{w}_i') = -(a_i', \boldsymbol{w}_i')$, so $(a_i', \boldsymbol{w}_i') = \boldsymbol{0}$. This holds for all $i \in B_1$.

(ii) For any $l \in \{2, \ldots, s\}$ and any $i, j \in B_l$, let π_{ij} be the transposition of i and j. The element $(\boldsymbol{\delta}, \pi_{ij})$ with $\delta_k = \gamma_k \gamma_{\pi_{ij}^{-1}(k)}$ is in $S(\boldsymbol{\theta}_0)$. Applying it to $\boldsymbol{\theta}'$ at index i gives $(a_i', \boldsymbol{w}_i') = \delta_i(a_j', \boldsymbol{w}_j') = (\gamma_i \gamma_j)(a_j', \boldsymbol{w}_j')$, which implies $\gamma_i(a_i', \boldsymbol{w}_i') = \gamma_j(a_j', \boldsymbol{w}_j')$.

(iii) If $\boldsymbol{\theta}'$ violated the third condition, i.e., if $(a_i', \boldsymbol{w}_i') = \pm(a_j', \boldsymbol{w}_j')$ for some $i \in B_l, j \in B_{l'}$ with $l \neq l'$, then $S(\boldsymbol{\theta}')$ would contain elements $(\boldsymbol{\delta}, \pi)$ where $\pi(i) = j$. Such elements are not in $S(\boldsymbol{\theta}_0)$, contradicting $S(\boldsymbol{\theta}_0) = S(\boldsymbol{\theta}')$.

Thus, $\boldsymbol{\theta}'$ must satisfy all three conditions defining $\mathcal{M}_{\mathcal{P},\boldsymbol{\gamma}}$. So $\boldsymbol{\theta}' \in \mathcal{M}_{\mathcal{P},\boldsymbol{\gamma}}$. Thus $[\boldsymbol{\theta}_0] \subset \mathcal{M}_{\mathcal{P},\boldsymbol{\gamma}}$. □

As stated in Lemma 4.1, the orthogonal symmetry group induces an invariant partition that gives an "upper bound" for orbits. This leads to a natural question: **Are the two invariant partitions equal for two-layer neural networks?** Our strategy for addressing this question proceeds in three steps:

(i) Develop a neuron independence result to calculate the rank of the Lie closure at non-degenerate $\boldsymbol{\theta}$ (Lemma 5.1, Corollary 5.1).

(ii) Establish a perturbation lemma to transform the degenerate cases into non-degenerate whenever possible (Lemma 5.2). This lemma allows us to calculate the rank of the Lie closure at all $\boldsymbol{\theta}$ (Corollary 5.2, 5.3).

(iii) Analyze the connectivity of the leaves of invariant partitions (Corollary 5.4).

These preparatory results, in conjunction with Theorem 2.1, establish that all SIMs are symmetry-induced for generic two-layer neural networks (Theorem 5.1).

5.1 Rank of Lie closure

We begin by presenting a lemma that establishes the foundation for the rank analysis carried out in this subsection.

Lemma 5.1 (neuron independence (Zhang et al., 2025a)). Let $\sigma : \mathbb{R} \to \mathbb{R}$ be any analytic function such that $\sigma^{(n_j)}(0) \neq 0$ for an infinite sequence of distinct indices $\{n_j\}_{j=1}^{\infty}$. Given $d \in \mathbb{N}$ and m distinct weights $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_m \in \mathbb{R}^d \backslash \{\mathbf{0}\}$, such that $\boldsymbol{w}_k \neq \pm \boldsymbol{w}_j$ for all $1 \leqslant k < j \leqslant m$. Then $\{\sigma\left(\boldsymbol{w}_i^{\intercal} \boldsymbol{x}\right), \sigma'\left(\boldsymbol{w}_i^{\intercal} \boldsymbol{x}\right) x_1, \ldots, \sigma'\left(\boldsymbol{w}_i^{\intercal} \boldsymbol{x}\right) x_d\}_{i=1}^{m}$ is a linearly independent function set.

The linear independence established in Lemma 5.1 holds only when parameters satisfy certain conditions (e.g., $\boldsymbol{w}_i = \pm \boldsymbol{w}_j$). We formally define parameters that do not meet these conditions or have some $a_i = 0$ as **degenerate** in Definition 5.1. The definition of degeneracy is designed to ensure the Lie closure of $\mathcal{F}$ at non-degenerate $\boldsymbol{\theta}$ has full rank, as stated in Corollary 5.1.

Definition 5.1 (degenerate and non-degenerate). Consider the two-layer neural network. For $\boldsymbol{\theta} = (a_i, \boldsymbol{w}_i)_{i=1}^m$, if $\boldsymbol{\theta}$ satisfies (i): $a_k \neq 0, \boldsymbol{w}_k \neq \mathbf{0}$ for all $k \in \{1, \ldots, m\}$, and (ii): $\boldsymbol{w}_i \neq \pm \boldsymbol{w}_j$ for any $i, j \in \{1, \ldots, m\}$ and $i \neq j$, then $\boldsymbol{\theta}$ is said to be **non-degenerate**. Otherwise $\boldsymbol{\theta}$ is said to be **degenerate**.

Corollary 5.1. Consider the two-layer neural network, and suppose that $\boldsymbol{\theta} \in \mathbb{R}^M$ is non-degenerate. Then $\dim(\mathrm{Lie}_{\boldsymbol{\theta}}(\mathcal{F})) = (d+1)m$.

Proof. Let $\boldsymbol{\theta} = (a_i, \boldsymbol{w}_i)_{i=1}^m$ be non-degenerate. By calculation, we have

$$\nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta})(\boldsymbol{x}) = \left(\sigma(\boldsymbol{w}_i^{\intercal}\boldsymbol{x}), a_i\sigma'(\boldsymbol{w}_i^{\intercal}\boldsymbol{x})x_1, \ldots, a_i\sigma'(\boldsymbol{w}_i^{\intercal}\boldsymbol{x})x_d\right)_{i=1}^m.$$

To show that $\dim(\mathrm{Lie}_{\boldsymbol{\theta}}(\mathcal{F})) = (d+1)m$, we will prove that the set of vectors $U := \{\nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta})(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathbb{R}^d\}$ spans the entire parameter space $\mathbb{R}^{(d+1)m}$.

We proceed by contradiction. Assume that $\mathrm{span}(U)$ is a proper subspace of $\mathbb{R}^{(d+1)m}$. Then there must exist a non-zero constant vector $\boldsymbol{c} = (c_1, \mathbf{v}_1^{\intercal}, \ldots, c_m, \mathbf{v}_m^{\intercal})^{\intercal} \in \mathbb{R}^{(d+1)m}$, where $c_i \in \mathbb{R}$ and $\mathbf{v}_i = (v_{i,1}, \ldots, v_{i,d}) \in \mathbb{R}^d$, that is orthogonal to every vector in U. This orthogonality condition, $\boldsymbol{c}^{\intercal}\nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta})(\boldsymbol{x}) = 0$ for all $\boldsymbol{x} \in \mathbb{R}^d$, expands to:

$$\sum_{i=1}^{m}\left(c_i\sigma(\boldsymbol{w}_i^{\intercal}\boldsymbol{x}) + \sum_{j=1}^{d}(a_i v_{i,j})\sigma'(\boldsymbol{w}_i^{\intercal}\boldsymbol{x})x_j\right) = 0.$$

This equation is a linear combination of the functions in the set $\{\sigma\left(\boldsymbol{w}_i^{\intercal} \boldsymbol{x}\right), \sigma'\left(\boldsymbol{w}_i^{\intercal} \boldsymbol{x}\right) x_1, \ldots, \sigma'\left(\boldsymbol{w}_i^{\intercal} \boldsymbol{x}\right) x_d\}_{i=1}^{m}$ that is identically zero for all $\boldsymbol{x} \in \mathbb{R}^d$.

Since $\boldsymbol{\theta}$ is non-degenerate, Lemma 5.1 guarantees that this set of functions is linearly independent. Consequently, all coefficients of the linear combination must be zero. This implies:

(i) $c_i = 0$ for all $i = 1, \ldots, m$.

(ii) $a_i v_{i,j} = 0$ for all $i = 1, \ldots, m$ and $j = 1, \ldots, d$.

The non-degeneracy of $\boldsymbol{\theta}$ ensures that $a_i \neq 0$ for all $i = 1, \ldots, m$. From the second point, we must have $v_{i,j} = 0$ for all $i = 1, \ldots, m, j = 1, \ldots, d$, which means $\mathbf{v}_i = \mathbf{0}$ for all $i = 1, \ldots, m$. This implies that the entire vector $\boldsymbol{c}$ is the zero vector, which contradicts our assumption that $\boldsymbol{c}$ was non-zero.

Therefore, the initial assumption must be false, and $\text{span}(U) = \mathbb{R}^{(d+1)m}$. Since $\text{span}(U) \subset \text{Lie}_{\boldsymbol{\theta}}(\mathcal{F}) \subset \mathbb{R}^{(d+1)m}$, $\dim(\text{Lie}_{\boldsymbol{\theta}}(\mathcal{F})) = (d+1)m$. □

Remark 5.1. Theorem 4.2 implies that the degenerate case of $\boldsymbol{\theta}$ can give rise to SIMs. For completeness, we provide its proof here.

(i) For any $i, j \in \{1, \ldots, m\}$, the set $\{(a_k, \boldsymbol{w}_k)_{k=1}^m \in \mathbb{R}^{(d+1)m} \mid (a_i, \boldsymbol{w}_i) = (a_j, \boldsymbol{w}_j)\}$ is a SIM.

(ii) If $\sigma(x)$ is an odd function, then for any $i, j \in \{1, \ldots, m\}$, the set $\{(a_k, \boldsymbol{w}_k)_{k=1}^m \in \mathbb{R}^{(d+1)m} \mid (a_i, \boldsymbol{w}_i) = -(a_j, \boldsymbol{w}_j)\}$ is a SIM.

(iii) If $\sigma(x)$ is an even function, then for any $i, j \in \{1, \ldots, m\}$, the set $\{(a_k, \boldsymbol{w}_k)_{k=1}^m \in \mathbb{R}^{(d+1)m} \mid (a_i, \boldsymbol{w}_i) = (a_j, -\boldsymbol{w}_j)\}$ is a SIM.

(iv) If $\sigma(0) = 0$, then for any $i \in \{1, \ldots, m\}$, the set $\{(a_k, \boldsymbol{w}_k)_{k=1}^m \in \mathbb{R}^{(d+1)m} \mid (a_i, \boldsymbol{w}_i) = \mathbf{0}\}$ is a SIM.

(v) If $\sigma'(0) = 0$, then for any $i \in \{1, \ldots, m\}$, the set $\{(a_k, \boldsymbol{w}_k)_{k=1}^m \in \mathbb{R}^{(d+1)m} \mid \boldsymbol{w}_i = \mathbf{0}\}$ is a SIM.

Proof. We prove the five statements item by item. The procedure is the same for each: we denote the set in question as a manifold $\mathcal{M}$ and then verify that it is a SIM. Since each of these five sets is a linear subspace of $\mathbb{R}^{(d+1)m}$, by Lemma 3.1, we only need to show that for any parameter $\boldsymbol{\theta} \in \mathcal{M}$ and any input $\boldsymbol{x} \in \mathbb{R}^d$, the gradient $\nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta})(\boldsymbol{x})$ also lies in the tangent space of $\mathcal{M}$, which for a linear subspace means $\nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta})(\boldsymbol{x}) \in \mathcal{M}$. For any $\boldsymbol{\theta} = (a_k, \boldsymbol{w}_k)_{k=1}^m \in \mathcal{M}$ and $\boldsymbol{x} \in \mathbb{R}^d$, the i-th component of the gradient (corresponding to the parameters $(a_i, \boldsymbol{w}_i)$ of neuron i) is given by:

$$\nabla_i F(\boldsymbol{\theta})(\boldsymbol{x}) = (\sigma(\boldsymbol{w}_i^\intercal \boldsymbol{x}), a_i \sigma'(\boldsymbol{w}_i^\intercal \boldsymbol{x}) \boldsymbol{x}^\intercal).$$

We now analyze each item:

(i) Define $\mathcal{M} = \{(a_k, \boldsymbol{w}_k)_{k=1}^m \mid (a_i, \boldsymbol{w}_i) = (a_j, \boldsymbol{w}_j)\}$. If $\boldsymbol{\theta} \in \mathcal{M}$, we have $(a_i, \boldsymbol{w}_i) = (a_j, \boldsymbol{w}_j)$. This directly implies that their corresponding gradient components are equal:

$$\nabla_i F(\boldsymbol{\theta})(\boldsymbol{x}) = (\sigma(\boldsymbol{w}_i^\intercal \boldsymbol{x}), a_i \sigma'(\boldsymbol{w}_i^\intercal \boldsymbol{x}) \boldsymbol{x}^\intercal) = (\sigma(\boldsymbol{w}_j^\intercal \boldsymbol{x}), a_j \sigma'(\boldsymbol{w}_j^\intercal \boldsymbol{x}) \boldsymbol{x}^\intercal) = \nabla_j F(\boldsymbol{\theta})(\boldsymbol{x}).$$

Thus, $\nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta})(\boldsymbol{x}) \in \mathcal{M}$.

(ii) Assume $\sigma(x)$ is an odd function, and define $\mathcal{M} = \{(a_k, \boldsymbol{w}_k)_{k=1}^m \mid (a_i, \boldsymbol{w}_i) = -(a_j, \boldsymbol{w}_j)\}$. We use the property that the derivative of an odd function is an even function, i.e., $\sigma'(-z) = \sigma'(z)$. If $\boldsymbol{\theta} \in \mathcal{M}$, we have $a_i = -a_j$ and $\boldsymbol{w}_i = -\boldsymbol{w}_j$. The i-th component of the gradient is:

$$\begin{aligned}
\nabla_i F(\boldsymbol{\theta})(\boldsymbol{x}) &= (\sigma(\boldsymbol{w}_i^\intercal \boldsymbol{x}), a_i \sigma'(\boldsymbol{w}_i^\intercal \boldsymbol{x}) \boldsymbol{x}^\intercal) \\
&= (\sigma(-\boldsymbol{w}_j^\intercal \boldsymbol{x}), -a_j \sigma'(-\boldsymbol{w}_j^\intercal \boldsymbol{x}) \boldsymbol{x}^\intercal) \\
&= (-\sigma(\boldsymbol{w}_j^\intercal \boldsymbol{x}), -a_j \sigma'(\boldsymbol{w}_j^\intercal \boldsymbol{x}) \boldsymbol{x}^\intercal) && \text{(since } \sigma \text{ is odd, } \sigma' \text{ is even)} \\
&= -(\sigma(\boldsymbol{w}_j^\intercal \boldsymbol{x}), a_j \sigma'(\boldsymbol{w}_j^\intercal \boldsymbol{x}) \boldsymbol{x}^\intercal) = -\nabla_j F(\boldsymbol{\theta})(\boldsymbol{x}).
\end{aligned}$$

Thus, $\nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta})(\boldsymbol{x}) \in \mathcal{M}$.

(iii) Assume $\sigma(x)$ is an even function, and define $\mathcal{M} = \{(a_k, \boldsymbol{w}_k)_{k=1}^m \mid (a_i, \boldsymbol{w}_i) = (a_j, -\boldsymbol{w}_j)\}$. If $\boldsymbol{\theta} \in \mathcal{M}$, we have $a_i = a_j$ and $\boldsymbol{w}_i = -\boldsymbol{w}_j$. Let $\nabla_k F = (\nabla_{a_k} F, \nabla_{\boldsymbol{w}_k} F)$. Then we have:

$$\begin{aligned}
\nabla_i F(\boldsymbol{\theta})(\boldsymbol{x}) &= (\sigma(\boldsymbol{w}_i^\intercal \boldsymbol{x}), a_i \sigma'(\boldsymbol{w}_i^\intercal \boldsymbol{x}) \boldsymbol{x}^\intercal) \\
&= (\sigma(-\boldsymbol{w}_j^\intercal \boldsymbol{x}), a_j \sigma'(-\boldsymbol{w}_j^\intercal \boldsymbol{x}) \boldsymbol{x}^\intercal) \\
&= (\sigma(\boldsymbol{w}_j^\intercal \boldsymbol{x}), a_j (-\sigma'(\boldsymbol{w}_j^\intercal \boldsymbol{x})) \boldsymbol{x}^\intercal) \qquad \text{(since } \sigma \text{ is even, } \sigma' \text{ is odd)} \\
&= (\nabla_{a_j} F, -\nabla_{\boldsymbol{w}_j} F).
\end{aligned}$$

Thus, $\nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta})(\boldsymbol{x}) \in \mathcal{M}$.

(iv) Assume $\sigma(0) = 0$, and define $\mathcal{M} = \{(a_k, \boldsymbol{w}_k)_{k=1}^m \mid (a_i, \boldsymbol{w}_i) = \boldsymbol{0}\}$. If $\boldsymbol{\theta} \in \mathcal{M}$, we have $(a_i, \boldsymbol{w}_i) = (0, \boldsymbol{0})$. The i-th component of the gradient is:

$$\nabla_i F(\boldsymbol{\theta})(\boldsymbol{x}) = (\sigma(\boldsymbol{0}^\intercal \boldsymbol{x}), 0 \cdot \sigma'(\boldsymbol{0}^\intercal \boldsymbol{x}) \boldsymbol{x}^\intercal) = (\sigma(0), \boldsymbol{0}).$$

Given the condition $\sigma(0) = 0$, this becomes $(0, \boldsymbol{0})$. Thus, $\nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta})(\boldsymbol{x}) \in \mathcal{M}$.

(v) Assume $\sigma'(0) = 0$, and define $\mathcal{M} = \{(a_k, \boldsymbol{w}_k)_{k=1}^m \mid \boldsymbol{w}_i = \boldsymbol{0}\}$. If $\boldsymbol{\theta} \in \mathcal{M}$, we have $\boldsymbol{w}_i = \boldsymbol{0}$. We only need to examine the component of the gradient corresponding to $\boldsymbol{w}_i$:

$$\nabla_{\boldsymbol{w}_i} F(\boldsymbol{\theta})(\boldsymbol{x}) = a_i \sigma'(\boldsymbol{w}_i^\intercal \boldsymbol{x}) \boldsymbol{x}^\intercal = a_i \sigma'(\boldsymbol{0}^\intercal \boldsymbol{x}) \boldsymbol{x}^\intercal = a_i \sigma'(0) \boldsymbol{x}^\intercal.$$

Given the condition $\sigma'(0) = 0$, the expression becomes $a_i \cdot 0 \cdot \boldsymbol{x}^\intercal = \boldsymbol{0}$. This satisfies the condition for the gradient vector to be in $\mathcal{M}$.

□

To deal with the degenerate case, we introduce Lemma 5.2, which allows us to perturb those degenerate $\boldsymbol{\theta}$ to non-degenerate whenever possible.

Lemma 5.2 (perturbation lemma). Consider the two-layer neural network. For $\boldsymbol{\theta}^* = (a_i^*, \boldsymbol{w}_i^*)_{i=1}^m$, the following statement holds:

(i) Assume $i \in \{1, \ldots, m\}$ and $a_i^* = 0$. Then

$$\exists \delta > 0, B_\delta(\boldsymbol{\theta}^*) \cap O_{\mathcal{F}}(\boldsymbol{\theta}^*) \subset \{(a_i, \boldsymbol{w}_i)_{i=1}^m \mid a_i = 0\} \Longleftrightarrow \boldsymbol{w}_i^* = \boldsymbol{0}, \sigma(0) = 0.$$

(ii) Assume $i \in \{1, \ldots, m\}$ and $\boldsymbol{w}_i^* = \boldsymbol{0}, a_i^* \neq 0$. Then

$$\exists \delta > 0, B_\delta(\boldsymbol{\theta}^*) \cap O_{\mathcal{F}}(\boldsymbol{\theta}^*) \subset \{(a_i, \boldsymbol{w}_i)_{i=1}^m \mid \boldsymbol{w}_i = \boldsymbol{0}\} \Longleftrightarrow \sigma'(0) = 0.$$

(iii) Assume $i, j \in \{1, \ldots, m\}$ and $\boldsymbol{w}_i^* = \boldsymbol{w}_j^* \neq \boldsymbol{0}$. Then

$$\exists \delta > 0, B_\delta(\boldsymbol{\theta}^*) \cap O_{\mathcal{F}}(\boldsymbol{\theta}^*) \subset \{(a_i, \boldsymbol{w}_i)_{i=1}^m \mid \boldsymbol{w}_i = \boldsymbol{w}_j\} \Longleftrightarrow a_i^* = a_j^*.$$

(iv) Assume $i, j \in \{1, \ldots, m\}$ and $\boldsymbol{w}_i^* = -\boldsymbol{w}_j^* \neq \boldsymbol{0}$. Then

$$\begin{aligned}
&\exists \delta > 0, B_\delta(\boldsymbol{\theta}^*) \cap O_{\mathcal{F}}(\boldsymbol{\theta}^*) \subset \{(a_i, \boldsymbol{w}_i)_{i=1}^m \mid \boldsymbol{w}_i = -\boldsymbol{w}_j\} \\
&\Longleftrightarrow a_i^* = -a_j^*, \sigma(x) \text{ is odd} \quad \text{or} \quad a_i^* = a_j^*, \sigma(x) \text{ is even.}
\end{aligned}$$

Here, $B_\delta(\boldsymbol{\theta}^*)$ denotes the open δ-ball around $\boldsymbol{\theta}^*$.

Proof. In each case, the goal is to characterize when the intersection $B_\delta(\boldsymbol{\theta}^*) \cap O_{\mathcal{F}}(\boldsymbol{\theta}^*)$ is contained within a specific linear subspace $\mathcal{M}$. Since $O_{\mathcal{F}}(\boldsymbol{\theta}^*)$ is immersed submanifold of $\mathbb{R}^M$, it follows that for all $\boldsymbol{\theta} \in B_\delta(\boldsymbol{\theta}^*) \cap O_{\mathcal{F}}(\boldsymbol{\theta}^*)$, the tangent space $T_{\boldsymbol{\theta}}(B_\delta(\boldsymbol{\theta}^*) \cap O_{\mathcal{F}}(\boldsymbol{\theta}^*)) = T_{\boldsymbol{\theta}} O_{\mathcal{F}}(\boldsymbol{\theta}^*)$ is contained in $T_{\boldsymbol{\theta}}\mathcal{M} = \mathcal{M}$. By Theorem 2.1, $T_{\boldsymbol{\theta}} O_{\mathcal{F}}(\boldsymbol{\theta}^*) = \mathrm{Lie}_{\boldsymbol{\theta}}(\mathcal{F})$, hence $\mathrm{Lie}_{\boldsymbol{\theta}}(\mathcal{F}) \subset \mathcal{M}$ for all $\boldsymbol{\theta} \in B_\delta(\boldsymbol{\theta}^*) \cap O_{\mathcal{F}}(\boldsymbol{\theta}^*)$. Since $\nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta})(\boldsymbol{x}) \in \mathrm{Lie}_{\boldsymbol{\theta}}(\mathcal{F}), \forall \boldsymbol{\theta} \in \mathbb{R}^M, \boldsymbol{x} \in \mathbb{R}^d$, we have $\nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta})(\boldsymbol{x}) \in \mathcal{M}, \forall \boldsymbol{\theta} \in B_\delta(\boldsymbol{\theta}^*) \cap O_{\mathcal{F}}(\boldsymbol{\theta}^*), \boldsymbol{x} \in \mathbb{R}^d$. Particularly $\nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta}^*)(\boldsymbol{x}) \in \mathcal{M}, \forall \boldsymbol{x} \in \mathbb{R}^d$. We analyze this condition for the four statements to be proved. For simplicity we use $\boldsymbol{\theta}$ to denote $(a_i, \boldsymbol{w}_i)_{i=1}^m \in \mathbb{R}^{(d+1)m}$.

(i) $\Longrightarrow$: Assume $O_{\mathcal{F}}(\boldsymbol{\theta}^*) \subset \{\boldsymbol{\theta} \mid a_i = 0\}$. The condition $\nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta}^*)(\boldsymbol{x}) \in \{\boldsymbol{\theta} \mid a_i = 0\}, \forall \boldsymbol{x} \in \mathbb{R}^d$ indicates that $\sigma(\boldsymbol{w}_i^{*\intercal} \boldsymbol{x}) = 0$ for all $\boldsymbol{x} \in \mathbb{R}^d$. Since σ is a real, non-polynomial, analytic function, this can only hold if $\boldsymbol{w}^* = \boldsymbol{0}$ and $\sigma(0) = 0$.

$\Longleftarrow$: Assume $\boldsymbol{w}_i^* = \boldsymbol{0}$ and $\sigma(0) = 0$. By Remark 5.1, when $\sigma(0) = 0$, $\{\boldsymbol{\theta} \mid (a_i, \boldsymbol{w}_i) = \boldsymbol{0}\}$ is a SIM. Therefore $O_{\mathcal{F}}(\boldsymbol{\theta}^*) \subset \{\boldsymbol{\theta} \mid (a_i, \boldsymbol{w}_i) = \boldsymbol{0}\} \subset \{\boldsymbol{\theta} \mid a_i = 0\}$.

(ii) $\Longrightarrow$: Assume $O_{\mathcal{F}}(\boldsymbol{\theta}^*) \subset \{\boldsymbol{\theta} \mid \boldsymbol{w}_i = \boldsymbol{0}\}$. This requires $\dot{\boldsymbol{w}}_i = a_i \sigma'(\boldsymbol{w}_i^\intercal \boldsymbol{x}) \boldsymbol{x}$ to be zero at $\boldsymbol{\theta}^*$ for all $\boldsymbol{x} \in \mathbb{R}^d$. So $a_i^* \sigma'(\boldsymbol{w}_i^{*\intercal} \boldsymbol{x}) \boldsymbol{x} = \boldsymbol{0}$ for all $\boldsymbol{x} \in \mathbb{R}^d$. Given $\boldsymbol{w}_i^* = \boldsymbol{0}$, the equation becomes $a_i^* \sigma'(0) \boldsymbol{x} = \boldsymbol{0}$. Since $a_i^* \neq 0$ and this must hold for all $\boldsymbol{x}$, it follows that $\sigma'(0) = 0$.

$\Longleftarrow$: Assume $\boldsymbol{w}_i^* = \boldsymbol{0}$ and $\sigma'(0) = 0$. By Remark 5.1, when $\sigma'(0) = 0$, $\{\boldsymbol{\theta} \mid \boldsymbol{w}_i = \boldsymbol{0}\}$ is a SIM. Therefore $O_{\mathcal{F}}(\boldsymbol{\theta}^*) \subset \{\boldsymbol{\theta} \mid \boldsymbol{w}_i = \boldsymbol{0}\}$.

(iii) $\Longrightarrow$ Assume $O_{\mathcal{F}}(\boldsymbol{\theta}^*) \subset \{\boldsymbol{\theta} \mid \boldsymbol{w}_i = \boldsymbol{w}_j\}$. Then $\dot{\boldsymbol{w}}_i = \dot{\boldsymbol{w}}_j$ at $\boldsymbol{\theta}^*$ for all $\boldsymbol{x} \in \mathbb{R}^d$. This means $a_i^* \sigma'(\boldsymbol{w}_i^{*\intercal} \boldsymbol{x}) \boldsymbol{x} = a_j^* \sigma'(\boldsymbol{w}_j^{*\intercal} \boldsymbol{x}) \boldsymbol{x}, \forall \boldsymbol{x} \in \mathbb{R}^d$. As $\boldsymbol{w}_i^* = \boldsymbol{w}_j^*$, this simplifies to $(a_i^* - a_j^*) \sigma'(\boldsymbol{w}_i^{*\intercal} \boldsymbol{x}) \boldsymbol{x} = \boldsymbol{0}, \forall \boldsymbol{x} \in \mathbb{R}^d$. Since $\boldsymbol{w}_i^* \neq \boldsymbol{0}$ and σ is not a zero function, the function $\sigma'(\boldsymbol{w}_i^{*\intercal} \boldsymbol{x}) \boldsymbol{x}$ is not identically zero. Thus, we must have $a_i^* = a_j^*$.

$\Longleftarrow$: Assume $a_i^* = a_j^*$ and $\boldsymbol{w}_i^* = \boldsymbol{w}_j^*$. By the permutation symmetry introduced in Remark 5.1, the set $\{\boldsymbol{\theta} \mid a_i = a_j, \boldsymbol{w}_i = \boldsymbol{w}_j\}$ is a SIM. Therefore $O_{\mathcal{F}}(\boldsymbol{\theta}^*) \subset \{\boldsymbol{\theta} \mid a_i = a_j, \boldsymbol{w}_i = \boldsymbol{w}_j\} \subset \{\boldsymbol{\theta} \mid \boldsymbol{w}_i = \boldsymbol{w}_j\}$.

(iv) $\Longrightarrow$: Assume $O_{\mathcal{F}}(\boldsymbol{\theta}^*) \subset \{\boldsymbol{\theta} \mid \boldsymbol{w}_i = -\boldsymbol{w}_j\}$. Given $\boldsymbol{w}_i^* = -\boldsymbol{w}_j^*$, it requires $\dot{\boldsymbol{w}}_i = -\dot{\boldsymbol{w}}_j$ at $\boldsymbol{\theta}^*$. This implies $a_i^* \sigma'(\boldsymbol{w}_i^{*\intercal} \boldsymbol{x}) \boldsymbol{x} = -a_j^* \sigma'(\boldsymbol{w}_j^{*\intercal} \boldsymbol{x}) \boldsymbol{x} = -a_j^* \sigma'(-\boldsymbol{w}_i^{*\intercal} \boldsymbol{x}) \boldsymbol{x}$. Since this must hold for all $\boldsymbol{x} \in \mathbb{R}^d$ and $\boldsymbol{w}_i^* \neq \boldsymbol{0}$, it follows that $a_i^* \sigma'(t) + a_j^* \sigma'(-t) = 0$ for all $t \in \mathbb{R}$. Integrating with respect to t yields $a_i^* \sigma(t) - a_j^* \sigma(-t) = C$ for some constant C. Replacing t with $-t$ gives $a_i^* \sigma(-t) - a_j^* \sigma(t) = C$. Equating the two expressions gives $(a_i^* + a_j^*)(\sigma(t) - \sigma(-t)) = 0$.

This implies two cases. First, if $\sigma(t) = \sigma(-t)$ for all t (i.e., σ is an even function), then σ' is odd. The condition $a_i^* \sigma'(t) + a_j^* \sigma'(-t) = 0$ becomes $(a_i^* - a_j^*) \sigma'(t) = 0$. As σ' is not identically zero, we must have $a_i^* = a_j^*$.

Second, if σ is not an even function, then we must have $a_i^* + a_j^* = 0$. Choose $0 < \delta' < \delta$ such that for all $\boldsymbol{\theta}' = (a_i', \boldsymbol{w}_i')_{i=1}^m \in B_{\delta'}(\boldsymbol{\theta}^*)$, we have $\boldsymbol{w}_i' \neq \boldsymbol{0}$ and $\boldsymbol{w}_j' \neq \boldsymbol{0}$. For any $\boldsymbol{\theta}' \in B_{\delta'}(\boldsymbol{\theta}^*) \cap O_{\mathcal{F}}(\boldsymbol{\theta}^*)$, there exists δ'' such that $B_{\delta''}(\boldsymbol{\theta}') \subset B_\delta(\boldsymbol{\theta}^*)$. Thus, $B_{\delta''}(\boldsymbol{\theta}') \cap O_{\mathcal{F}}(\boldsymbol{\theta}') \subset B_\delta(\boldsymbol{\theta}^*) \cap O_{\mathcal{F}}(\boldsymbol{\theta}') = B_\delta(\boldsymbol{\theta}^*) \cap O_{\mathcal{F}}(\boldsymbol{\theta}^*) \subset \{\boldsymbol{\theta} \mid \boldsymbol{w}_i = -\boldsymbol{w}_j\}$. Since σ is not an even function, the same derivation implies $a_i' + a_j' = 0$. Therefore, $B_{\delta'}(\boldsymbol{\theta}^*) \cap O_{\mathcal{F}}(\boldsymbol{\theta}^*) \subset \{\boldsymbol{\theta} \mid a_i = -a_j\}$. This implies $\dot{a_i} + \dot{a_j} = 0$ at $\boldsymbol{\theta}^*$ for any input $\boldsymbol{x} \in \mathbb{R}^d$. Thus, $\sigma(\boldsymbol{w}_i^{*\intercal} \boldsymbol{x}) + \sigma(-\boldsymbol{w}_i^{*\intercal} \boldsymbol{x}) = 0$ for all $\boldsymbol{x} \in \mathbb{R}^d$. Since $\boldsymbol{w}_i^* \neq \boldsymbol{0}$, σ must be an odd function.

$\Longleftarrow$: Assume σ is odd and $a_i^* = -a_j^*$, $\boldsymbol{w}_i^* = -\boldsymbol{w}_j^*$. By Remark 5.1 the submanifold $\{\boldsymbol{\theta} \mid a_i = -a_j, \boldsymbol{w}_i = -\boldsymbol{w}_j\}$ is a SIM. Therefore $O_{\mathcal{F}}(\boldsymbol{\theta}^*) \subset \{\boldsymbol{\theta} \mid a_i = -a_j, \boldsymbol{w}_i = -\boldsymbol{w}_j\} \subset \{\boldsymbol{\theta} \mid \boldsymbol{w}_i = -\boldsymbol{w}_j\}$. The case when σ is even and $a_i^* = a_j^*$, $\boldsymbol{w}_i^* = -\boldsymbol{w}_j^*$ is similar.

□

Remark 5.2. Lemma 5.2 establishes that the $\boldsymbol{\theta}^*$ can be perturbed arbitrarily slightly outside the specified subset. For example, in the first case, if $\boldsymbol{w}_i^* \neq \mathbf{0}$ or $\sigma(0) \neq 0$, then for any $\epsilon > 0$, there exists a perturbed parameter $\boldsymbol{\theta}' \in O_{\mathcal{F}}(\boldsymbol{\theta}^*)$ such that $\|\boldsymbol{\theta}' - \boldsymbol{\theta}^*\|_2 < \epsilon$ and $\boldsymbol{\theta}' \notin \{(a_i, \boldsymbol{w}_i)_{i=1}^m \mid a_i = 0\}$. This is why we refer to it as the perturbation lemma.

As observed in Remark 5.1 and Lemma 5.2, there are numerous cases to consider for the activation function $\sigma(x)$, such as whether $\sigma(x)$ is odd or even, whether $\sigma(0) = 0$, and whether $\sigma'(0) = 0$. To manage this complexity, we focus on two representative types of activation: generic activation and generic odd activation.

Definition 5.2 (generic activation and generic odd activation). A real analytic function $\sigma : \mathbb{R} \to \mathbb{R}$ is called a **generic activation** if it satisfies the following conditions: (i): $\sigma(x)$ is not a polynomial; (ii): $\sigma(x)$ is neither an odd function nor an even function; (iii): $\sigma(0) \neq 0$ and $\sigma'(0) \neq 0$.

Similarly, a real analytic function $\sigma : \mathbb{R} \to \mathbb{R}$ is called a **generic odd activation** if it satisfies the following conditions: (i): $\sigma(x)$ is not a polynomial; (ii): $\sigma(x)$ is an odd function; (iii): $\sigma'(0) \neq 0$.

When the activation function is either a generic activation or a generic odd activation, any degenerate parameter $\boldsymbol{\theta}$ with trivial stabilizer subgroup $S(\boldsymbol{\theta})$ can always be perturbed to a non-degenerate one. Once such a perturbation is made, Theorem 2.1 implies that the Lie closure at $\boldsymbol{\theta}$ has rank equal to $(d+1)m$. This idea is illustrated in Corollary 5.2.

Corollary 5.2. Consider the two-layer neural network. For $\boldsymbol{\theta} = (a_i, \boldsymbol{w}_i)_{i=1}^m \in \mathbb{R}^{(d+1)m}$, the following holds:

(i) Assume $\sigma(x)$ is a generic activation, and $(a_i, \boldsymbol{w}_i) \neq (a_j, \boldsymbol{w}_j)$ for any $i, j \in \{1, \ldots, m\}$ with $i \neq j$. Then there exists $\boldsymbol{\theta}' \in O_{\mathcal{F}}(\boldsymbol{\theta})$ such that $\boldsymbol{\theta}'$ is non-degenerate. Thus, $\dim(\mathrm{Lie}_{\boldsymbol{\theta}}(\mathcal{F})) = (d+1)m$.

(ii) Assume $\sigma(x)$ is a generic odd activation, and $(a_i, \boldsymbol{w}_i) \neq \pm(a_j, \boldsymbol{w}_j)$ for any $i, j \in \{1, \ldots, m\}$. Then there exists $\boldsymbol{\theta}' \in O_{\mathcal{F}}(\boldsymbol{\theta})$ such that $\boldsymbol{\theta}'$ is non-degenerate. Thus, $\dim(\mathrm{Lie}_{\boldsymbol{\theta}}(\mathcal{F})) = (d+1)m$.

Proof. (i) Assume that $\sigma(x)$ is generic activation, and assume $(a_i, \boldsymbol{w}_i) \neq (a_j, \boldsymbol{w}_j)$ for any $i, j \in \{1, \ldots, m\}, i \neq j$. We now prove that there exists $\boldsymbol{\theta}' \in O_{\mathcal{F}}(\boldsymbol{\theta})$ such that $\boldsymbol{\theta}'$ is non-degenerate. Since $\sigma(0) \neq 0$, by the first statement of Lemma 5.2, there exists $\boldsymbol{\theta}_1 = (a_i^1, \boldsymbol{w}_i^1)_{i=1}^m \in O_{\mathcal{F}}(\boldsymbol{\theta})$ such that $a_i^1 \neq 0$ for all $i \in \{1, \ldots, m\}$. Moreover, $\boldsymbol{\theta}_1$ can be arbitrary close to $\boldsymbol{\theta}$ such that $(a_i^1, \boldsymbol{w}_i^1) \neq (a_j^1, \boldsymbol{w}_j^1)$ for any $i, j \in \{1, \ldots, m\}, i \neq j$. Since $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}$ are on the same orbit, we regard them as equivalent. In sense of this equivalence, without loss of generality we can assume that $a_i \neq 0$ for all $i \in \{1, \ldots, m\}$. By the second statement of Lemma 5.2, without loss of generality we can assume that $\boldsymbol{w}_i \neq \mathbf{0}$ for all $i \in \{1, \ldots, m\}$. If there exists $i \neq j$ such that $\boldsymbol{w}_i = \boldsymbol{w}_j$, then $a_i \neq a_j$. By the third statement of Lemma 5.2, without loss of generality we can assume that $\boldsymbol{w}_i \neq \boldsymbol{w}_j$ for all $i, j \in \{1, \ldots, m\}$ and $i \neq j$. By the fourth statement of Lemma 5.2 we can assume that $\boldsymbol{w}_i \neq -\boldsymbol{w}_j$ for all $i \neq j$. Therefore there exists non-degenerate $\boldsymbol{\theta}'$ on the orbit of $\boldsymbol{\theta}$.

(ii) Assume that $\sigma(x)$ is generic odd activation, and for any $i, j \in \{1, \ldots, m\}$, $(a_i, \boldsymbol{w}_i) \neq \pm(a_j, \boldsymbol{w}_j)$. Therefore for each $i \in \{1, \ldots, m\}$, either $a_i \neq 0$ or $\boldsymbol{w}_i \neq \boldsymbol{0}$. If $a_i \neq 0$, by the second statement of Lemma 5.2, without loss of generality we can assume that $\boldsymbol{w}_i \neq \boldsymbol{0}$. If $\boldsymbol{w}_i \neq \boldsymbol{0}$, by the first statement of Lemma 5.2, without loss of generality we can assume that $a_i \neq 0$. In both cases we can assume that $a_i \neq 0, \boldsymbol{w}_i \neq \boldsymbol{0}$ for all $i \in \{1, \ldots, m\}$. If there exists $i \neq j$ such that $\boldsymbol{w}_i = \boldsymbol{w}_j$, then $a_i \neq a_j$. By the third statement of Lemma 5.2, without loss of generality we can assume that $\boldsymbol{w}_i \neq \boldsymbol{w}_j$ for all $i, j \in \{1, \ldots, m\}$ and $i \neq j$. Similarly, if there exists $i \neq j$ such that $\boldsymbol{w}_i = -\boldsymbol{w}_j$, then $a_i \neq -a_j$. By the fourth statement of Lemma 5.2, without loss of generality we can assume that $\boldsymbol{w}_i \neq -\boldsymbol{w}_j$ for all $i, j \in \{1, \ldots, m\}$ and $i \neq j$. Therefore there exists a non-degenerate $\boldsymbol{\theta}'$ on the orbit of $\boldsymbol{\theta}$.

In both cases there exists a non-degenerate $\boldsymbol{\theta}' \in O_{\mathcal{F}}(\boldsymbol{\theta})$. By Corollary 5.1, $\dim(\mathrm{Lie}_{\boldsymbol{\theta}'}(\mathcal{F})) = (d+1)m$. By Theorem 2.1, $\dim(\mathrm{Lie}_{\boldsymbol{\theta}}(\mathcal{F})) = \dim(\mathrm{Lie}_{\boldsymbol{\theta}'}(\mathcal{F})) = (d+1)m$.

□

The result of Corollary 5.2 can be extended to any parameter $\boldsymbol{\theta}$ lying in each leaf of the invariant partition induced by the orthogonal symmetry group, as shown in Corollary 5.3.

Corollary 5.3 (rank of Lie closure). Consider the two-layer neural network. Assume that $\sigma(x)$ is generic activation or generic odd activation. Then for any $\boldsymbol{\theta} \in \mathbb{R}^{(d+1)m}$, $\dim(\mathrm{Lie}_{\boldsymbol{\theta}}(\mathcal{F}))$ is equal to the dimension of $[\boldsymbol{\theta}]$.[6]

Proof. We claim that, for two-layer neural networks, the calculation of Lie algebra is neuron-wise. We begin with the following definitions. Define the network of width k as $F_k(\boldsymbol{\theta})(\boldsymbol{x}) = \sum_{i=1}^{k} a_i \sigma(\boldsymbol{w}_i^\intercal \boldsymbol{x}), \boldsymbol{x} \in \mathbb{R}^d, \boldsymbol{\theta} = (a_i, \boldsymbol{w}_i) \in \mathbb{R}^{(d+1)k}$, and define $\mathcal{F}_k = \{\nabla_{\boldsymbol{\theta}} F_k(\cdot)(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathbb{R}^d\}$. Denote $\mathrm{Lie}(\mathcal{F}_k)$ to be the Lie algebra generated by $\mathcal{F}_k$. Define $\mathcal{F}_k' = \{(X(a_i, \boldsymbol{w}_i))_{i=1}^k \mid X \in \mathrm{Lie}(\mathcal{F}_1)\}$, which is a family of vector fields on $\mathbb{R}^{(d+1)k}$. The specific model considered is $F(\boldsymbol{\theta})(\boldsymbol{x}) = \sum_{i=1}^{m} a_i \sigma(\boldsymbol{w}_i^\intercal \boldsymbol{x}), \boldsymbol{x} \in \mathbb{R}^d, \boldsymbol{\theta} = (a_i, \boldsymbol{w}_i) \in \mathbb{R}^{(d+1)m}$ for fixed $m \in \mathbb{N}^+$. For notional simplicity, we omit the subscript m, using $F, \mathrm{Lie}(\mathcal{F}), \mathcal{F}'$ to denote $F_m, \mathrm{Lie}(\mathcal{F}_m), \mathcal{F}_m'$, respectively.

For any positive integer k, and any $X_1, \ldots, X_k \in \mathcal{F}_1$, define $Y_j = (X_j(a_i, \boldsymbol{w}_i))_{i=1}^m$ for $j = 1, \ldots, k$. Define the nested Lie brackets $X = [X_1, [X_2, [\cdots [X_{k-1}, X_k] \cdots]]]$ and $Y = [Y_1, [Y_2, [\cdots [Y_{k-1}, Y_k] \cdots]]]$. It is straightforward to verify by induction that $Y = (X(a_i, \boldsymbol{w}_i))_{i=1}^m$.

Let $\mathfrak{g}^{(k)}$ and $\mathfrak{g}_1^{(k)}$ denote the k-th terms in the lower central series of $\mathcal{F}$ and $\mathcal{F}_1$, respectively. The lower central series $\mathfrak{g}^{(1)}, \mathfrak{g}^{(2)}, \ldots$ of a family of vector fields $\tilde{\mathcal{F}}$ is defined recursively by $\mathfrak{g}^{(1)} = \tilde{\mathcal{F}}$ and $\mathfrak{g}^{(k+1)} = [\mathfrak{g}, \mathfrak{g}^{(k)}]$, where the bracket denotes the Lie bracket. From the identity $Y = (X(a_i, \boldsymbol{w}_i))_{i=1}^m$, and the fact that $\mathcal{F} = \{(X(a_i, \boldsymbol{w}_i))_{i=1}^m \mid X \in \mathcal{F}_1\}$, it follows that $\mathfrak{g}^{(k)} = \{(X(a_i, \boldsymbol{w}_i))_{i=1}^m \mid X \in \mathfrak{g}_1^{(k)}\}$ for all $k \in \mathbb{N}^+$. Consequently, we conclude that $\mathcal{F}' = \mathrm{Lie}(\mathcal{F})$.

(i) Assume the activation function to be a generic activation. Fix any $\boldsymbol{\theta} = (a_i, \boldsymbol{w}_i)_{i=1}^m \in \mathbb{R}^{(d+1)m}$. By Proposition 5.1, $[\boldsymbol{\theta}] = \mathcal{M}_{\mathcal{P}}$ for some partition $\mathcal{P} = \{B_1, \ldots, B_s\}$ of $\{1, \ldots, m\}$. By definition of $\mathcal{M}_{\mathcal{P}}$, we have $\dim([\boldsymbol{\theta}]) = \dim(\mathcal{M}_{\mathcal{P}}) = (d+1)s$. Next we calculate $\dim(\mathrm{Lie}_{\boldsymbol{\theta}}(\mathcal{F}))$. Since $\mathrm{Lie}(\mathcal{F}) = \mathcal{F}'$, we have $\mathrm{Lie}_{\boldsymbol{\theta}}(\mathcal{F}) = \mathcal{F}'|_{\boldsymbol{\theta}}$. For each $p = 1, \ldots, s$, select $k_p \in B_p$. We then define $\boldsymbol{\theta}' = (a_{k_p}, \boldsymbol{w}_{k_p})_{p=1}^s$. Similarly, $\mathrm{Lie}_{\boldsymbol{\theta}'}(\mathcal{F}_{\mathrm{s}}) = \mathcal{F}_s'|_{\boldsymbol{\theta}'}$. Define a

6. For generic activation and generic odd activation, the orthogonal symmetry group is the permutation symmetry group G_{per} and the combined symmetry group G_{combine}, respectively.

linear map $P : \mathcal{F}'|_{\boldsymbol{\theta}} \to \mathbb{R}^{(d+1)s}$ by: $(a_i^*, \boldsymbol{w}_i^*)_{i=1}^m \mapsto (a_{k_p}^*, \boldsymbol{w}_{k_p}^*)_{p=1}^s, \forall (a_i^*, \boldsymbol{w}_i^*)_{i=1}^m \in \mathcal{F}'|_{\boldsymbol{\theta}}$. By definition of $\mathcal{F}'_s|_{\boldsymbol{\theta}'}$, P is a surjection onto $\mathcal{F}'_s|_{\boldsymbol{\theta}'}$. It implies that $\dim(\mathrm{Lie}_{\boldsymbol{\theta}}(\mathcal{F})) = \dim(\mathcal{F}'|_{\boldsymbol{\theta}}) \geq \dim(\mathcal{F}'_s|_{\boldsymbol{\theta}'}) = \dim(\mathrm{Lie}_{\boldsymbol{\theta}'}(\mathcal{F}_s))$. Since $\boldsymbol{\theta} \in \mathcal{M}_{\mathcal{P}}$, it follows that $(a_{k_p}, \boldsymbol{w}_{k_p}) \neq (a_{k_q}, \boldsymbol{w}_{k_q})$ for any $p, q \in \{1, \ldots, s\}$ and $p \neq q$. By Corollary 5.2, $\dim(\mathrm{Lie}_{\boldsymbol{\theta}'}(\mathcal{F}_{\mathrm{s}})) = (d+1)s$. Therefore $\dim(\mathrm{Lie}_{\boldsymbol{\theta}}(\mathcal{F})) \geq \dim(\mathrm{Lie}_{\boldsymbol{\theta}'}(\mathcal{F}_{\mathrm{s}})) = (d+1)s$. By Theorem 2.1, the tangent space of $O_{\mathcal{F}}(\boldsymbol{\theta})$ at $\boldsymbol{\theta}$ is $\mathrm{Lie}_{\boldsymbol{\theta}}(\mathcal{F})$. By Lemma 4.1, $[\boldsymbol{\theta}]$ is a SIM. So $O_{\mathcal{F}}(\boldsymbol{\theta}) \subset [\boldsymbol{\theta}]$. Take this inclusion to tangent space gives $\dim(\mathrm{Lie}_{\boldsymbol{\theta}}(\mathcal{F})) \leq (d+1)s$. Thus, $\dim(\mathrm{Lie}_{\boldsymbol{\theta}}(\mathcal{F})) = (d+1)s$.

(ii) We now consider the case where the activation function is a generic odd function and the symmetry group is the combined orthogonal group. The proof is analogous to that of (i); hence, we omit several details for brevity. By Proposition 5.2, we have $[\boldsymbol{\theta}] = \mathcal{M}_{\mathcal{P},\boldsymbol{\gamma}}$ for some $\mathcal{P} = \{B_1, \ldots, B_s\}$ and $\boldsymbol{\gamma} \in \{-1, 1\}^m$. Then dimension is therefore $\dim([\boldsymbol{\theta}]) = (d+1)(s-1)$. A similar argument establishes that $\mathrm{Lie}_{\boldsymbol{\theta}}(\mathcal{F}) \geq \mathrm{Lie}_{\boldsymbol{\theta}'}(\mathcal{F}_{s-1})$, where $\boldsymbol{\theta}' = (a_{k_p}, \boldsymbol{w}_{k_p})_{p=2}^s$ for $k_p \in B_p$ with $p = 2, \ldots, s$. Subsequently, by Corollary 5.2, it follows that $\dim(\mathrm{Lie}_{\boldsymbol{\theta}'}(\mathcal{F})) = (d+1)(s-1)$. Then some straightforward reasoning leads to $\dim(\mathrm{Lie}_{\boldsymbol{\theta}}(\mathcal{F})) = \dim([\boldsymbol{\theta}]) = (d+1)(s-1)$.

□

5.2 Orbits

Corollary 5.3 establishes that, on each leaf of the invariant partition, the rank of the Lie closure equals the dimension of the leaf itself. In this case, Theorem 2.1 indicates that determining the orbit reduces to verifying the connectivity of each leaf. In Corollary 5.4, we show that every leaf is indeed connected.

Corollary 5.4. Consider the two-layer neural network, and assume that $\sigma(x)$ is either a generic activation or a generic odd activation. Then, for all $\boldsymbol{\theta} \in \mathbb{R}^{(d+1)m}$, the equivalence class $[\boldsymbol{\theta}]$ is connected.

Proof. We begin by presenting a standard lemma from manifold geometry (Conrad).

Lemma 5.3 (derived from Theorem 1.1 in Conrad). Let $A_1, \ldots, A_n$ be linear subspaces of $\mathbb{R}^M$, each with codimension at least 2. Then the set $\mathbb{R}^M \setminus \bigcup_{i=1}^n A_i$ is connected.

We now return to the proof of the corollary. As shown in Propositions 5.1 and 5.2, for any $\boldsymbol{\theta} \in \mathbb{R}^M$, the corresponding leaf $[\boldsymbol{\theta}]$ takes the form $[\boldsymbol{\theta}] = B \setminus (\bigcup_{i=1}^s A_i)$, where $B \subset \mathbb{R}^M$ is a linear subspace, and each $A_i \subset B$ is a linear subspace of B with $\dim(A_i) \leq \dim(B) - 2$. By Lemma 5.3, such a set $[\boldsymbol{\theta}]$ is connected. □

With the necessary preliminaries established, we are now in a position to present Theorem 5.1, which serves as the principal result of this section.

Theorem 5.1 (SIMs are all symmetry-induced for generic two-layer neural networks). Consider the two-layer neural network, and assume that $\sigma(x)$ is either a generic activation or a generic odd activation. Then for any $\boldsymbol{\theta} \in \mathbb{R}^{(d+1)m}$, $[\boldsymbol{\theta}] = O_{\mathcal{F}}(\boldsymbol{\theta})$.

Proof. For any $\boldsymbol{\theta} \in \mathbb{R}^{(d+1)m}$, Corollary 5.3 implies that $\dim(\mathrm{Lie}_{\boldsymbol{\theta}}(\mathcal{F})) = \dim([\boldsymbol{\theta}])$, where $\dim([\boldsymbol{\theta}])$ denotes the dimension of $[\boldsymbol{\theta}]$. By Corollary 5.4, the set $[\boldsymbol{\theta}]$ is connected. According to Lemma 4.1, $[\boldsymbol{\theta}]$ is a SIM. So $\mathcal{F}$ can be regarded as a family of vector fields defined on $[\boldsymbol{\theta}]$. Applying Theorem 2.1, it then follows that $[\boldsymbol{\theta}] = O_{\mathcal{F}}(\boldsymbol{\theta}), \forall \boldsymbol{\theta} \in \mathbb{R}^{(d+1)m}$. □

6. Conclusions

In this work, we lay the theoretical foundation for identifying SIMs in analytic parametric models by employing geometric control theory. By uncovering the hierarchy of symmetry-induced SIMs for deep neural networks and enumerating all SIMs for the two-layer network, we unravel the profound dynamical consequence of the layer-wise neural network architecture. These SIMs display condensation behavior and underlie the remarkable potential for target recovery in overparameterized settings. Although the milestone of fully solving the recovery puzzle for neural networks has not yet been reached, these SIMs offer powerful tools for tracing global training dynamics. Building on our findings, we anticipate major breakthroughs in solving the recovery puzzle in the near future. Such advances will pave the way for a comprehensive generalization theory that clarifies how architecture design, target properties, training samples, nonlinear dynamics, and parameter tuning collectively shape the generalization of neural networks.

Appendix A. Definitions

In the appendix, we present several definitions and concepts from geometric control theory that are pertinent to the content of this paper. As our analysis is conducted within the analytic category, all definitions are stated in their analytic form. The material is drawn from Jurdjevic (1997) and Ortega and Ratiu (2013).

A.1 Differential Geometry

Definition A.1 (analytic manifold, page 3 and 4 of Jurdjevic (1997))). $\mathcal{M}$ is called an n dimensional analytic manifold if $\mathcal{M}$ is a topology space such that at each point $p \in \mathcal{M}$ there exists a neighbourhood U of p and a homeomorphism ϕ from U onto an open subset of $\mathbb{R}^n$. It is assumed that n does not vary with the choice of a point p on $\mathcal{M}$. The pair (ϕ, U) is called a chart at p. Moreover:

(i) There exists a countable collection of charts $\{(\phi_i, U_i)\}_{i=1}^{\infty}$ such that $\mathcal{M} = \bigcup_{i=1}^{\infty} U_i$.

(ii) For each pair of points p_1 and p_2, there exist charts (ϕ_1, U_1) and (ϕ_2, U_2) such that $p_1 \in U_1$, $p_2 \in U_2$, and $U_1 \cap U_2 = \emptyset$. That is, points of $\mathcal{M}$ are separated by coordinate neighborhoods (i.e., $\mathcal{M}$ is Hausdorff).

(iii) For any charts (ϕ_1, U_1) and (ϕ_2, U_2) such that $U_1 \cap U_2 \neq \emptyset$, the mapping $\phi_1 \circ \phi_2^{-1}$ is analytic as a mapping from an open set in $\mathbb{R}^n$ into $\mathbb{R}^n$.

Definition A.2 (analytic vector fields, Definition 1 in Chapter 1 of Jurdjevic (1997)). Let $\mathcal{M}$ be an analytic manifold. The totality of $(p, v), p \in \mathcal{M}, v \in T_p\mathcal{M}$, is called the tangent bundle of $\mathcal{M}$ and is denoted by $T\mathcal{M}$. A vector field is a mapping $X : \mathcal{M} \to T\mathcal{M}$ such that for each $p \in \mathcal{M}$, if $\pi : T\mathcal{M} \to \mathcal{M}$ denotes the natural projection, then $\pi\big(X(p)\big) = p$. We say that X is an analytic vector field if X is an analytic map from $\mathcal{M}$ (as an analytic manifold) into $T\mathcal{M}$ (another analytic manifold).

Definition A.3 (integral curve, Definition 3 in Chapter 1 of Jurdjevic (1997)). A differential curve $p(t), t \in J$ on $\mathcal{M}$ is an integral curve of an analytic vector field X if $\frac{\mathrm{d}p}{\mathrm{d}t} = X \circ p$ for each t in J. We shall say an integral curve $p(t), t \in J$ of X is the integral curve through $p_0 \in \mathcal{M}$ if $p(0) = p_0$ and the domain $J \subset \mathbb{R}$ is maximal.

Definition A.4 (complete vector field, flow, e^{tX}, Definition 4 in Chapter 1 of Jurdjevic (1997)). A vector field X is called complete if the integral curves through each point p_0 in $\mathcal{M}$ are defined for all values of t in $\mathbb{R}$. In such a case, X is said to define a flow Φ on $\mathcal{M}$. $\Phi : \mathbb{R} \times \mathcal{M} \to \mathcal{M}$ is defined by $\Phi(t, p_0) = p(t)$, where $p(t)$ is the integral curve through p_0. For each t, define the mapping $\Phi_t(p) = \Phi(t, p)$. We shall also use e^{tX} to denote the mapping Φ_t.

Remark A.1 (page 16 of Jurdjevic (1997)). For complete vector fields, its flow Φ has following properties:

(i) $\Phi(0, p) = p$ for all $p \in \mathcal{M}$.

(ii) $\Phi(t + s, p) = \Phi(t, \Phi(s, p))$ for all (s, t) in $\mathbb{R}^2$ and all $p \in \mathcal{M}$.

(iii) $(\partial/\partial t)\Phi(t, p) = X \circ \Phi(t, p)$ for all (t, p) in $\mathbb{R} \times \mathcal{M}$.

(iv) The mapping Φ is analytic whenever X is analytic.

(v) For each t, e^{tX} is a diffeomorphism on $\mathcal{M}$.

Definition A.5 (local flow, page 17 of Jurdjevic (1997)). Let X be an analytic vector (possibly non-complete) field on an analytic manifold $\mathcal{M}$. In order to define the local flow of a vector field at p in $\mathcal{M}$, it is first necessary to define the escape times of the integral curve of X through p. The positive escape time $e^+(p)$ is defined to be the supremum of the domain of the integral curve through p. The negative escape time $e^-(p)$ is defined similarly. Let $\Delta = \{(t,p) \mid e^-(p) < t < e^+(p)\}$. Then Δ is an open subset of $\mathbb{R} \times \mathcal{M}$ and a neighborhood of $\{0\} \times \mathcal{M}$. The local flow Φ of X is defined on Δ.

Remark A.2 (page 17 of Jurdjevic (1997)). The local flow Φ satisfies the following:
(i) $\Phi(0,p) = p$ for all $p \in \Delta$.
(ii) $\Phi(t+s,p) = \Phi(t,\Phi(s,p))$ whenever each of (s,p) and $(t,\Phi(s,p))$ is contained in Δ.
(iii) $(\partial\Phi/\partial t)(t,p) = X \circ \Phi(t,p)$ for all (p,t) in Δ.
(iv) Φ is analytic whenever X is analytic.

Definition A.6 (immersed submanifold, Definition 1 in Chapter 2 of Jurdjevic (1997)). An differentiable mapping f between two differential manifolds is called an immersion if the rank of the tangent map of f at each point is equal to the dimension of the domain manifold. Then the definition of an immersed submanifold is as follows: Given two differentiable manifolds $\mathcal{M}$ and $\mathcal{N}$, if there exists an immersion $f : \mathcal{N} \to \mathcal{M}$, then $f(\mathcal{N})$ is called an immersed submanifold of $\mathcal{M}$.

A.2 Local Diffeomorphisms and Pseudogroups

The materials of A.2 are from Section 3.1 of Ortega and Ratiu (2013).

A.2.1 Local diffeomorphisms

Let $\mathcal{M}$ be an analytic manifold. The symbol $\mathrm{Diff}(\mathcal{M})$ will denote the set of diffeomorphisms of $\mathcal{M}$. The symbol $\mathrm{Diff}_L(\mathcal{M})$ will denote the set of local diffeomorphisms of $\mathcal{M}$. More explicitly, the elements of $\mathrm{Diff}_L(\mathcal{M})$ are diffeomorphisms $f : \mathrm{Dom}(f) \subset \mathcal{M} \to f(\mathrm{Dom}(f)) \subset \mathcal{M}$ of an open subset $\mathrm{Dom}(f) \subset \mathcal{M}$ onto its image $f(\mathrm{Dom}(f)) \subset \mathcal{M}$. We will denote the elements of $\mathrm{Diff}_L(\mathcal{M})$ as pairs $(f, \mathrm{Dom}(f))$. The local diffeomorphisms can be composed using the binary operation defined as

$$(f, \mathrm{Dom}(f)) \cdot (g, \mathrm{Dom}(g)) := (f \circ g, \mathrm{Dom}(f) \cap \mathrm{Dom}(g)), \tag{6}$$

for all $(f, \mathrm{Dom}(f))$, $(g, \mathrm{Dom}(g)) \in \mathrm{Diff}_L(\mathcal{M})$. It is easy to see that this operation is associative and has $(\mathrm{id}, \mathcal{M})$, the identity map of $\mathcal{M}$, as a (unique) two-sided identity element, which makes $\mathrm{Diff}_L(\mathcal{M})$ into a monoid (set with an associative operation which contains a two-sided identity element). Notice that only the elements in $\mathrm{Diff}(\mathcal{M}) \subset \mathrm{Diff}_L(\mathcal{M})$ have an inverse since, in general, for any $(f, \mathrm{Dom}(f)) \in \mathrm{Diff}_L(\mathcal{M})$, we have that

$$(f^{-1}, \mathrm{Dom}(f^{-1})) \cdot (f, \mathrm{Dom}(f)) = (\mathrm{id}|_{\mathrm{Dom}(f)}, \mathrm{Dom}(f)) \tag{7}$$

$$(f, \mathrm{Dom}(f)) \cdot (f^{-1}, \mathrm{Dom}(f^{-1})) = (\mathrm{id}|_{\mathrm{Dom}(f^{-1})}, \mathrm{Dom}(f^{-1})). \tag{8}$$

Consequently, the only way to obtain the identity element $(\mathrm{id}, \mathcal{M})$ out of the composition of f with its inverse is having $\mathrm{Dom}(f) = \mathcal{M}$. It follows from this argument that $\mathrm{Diff}(\mathcal{M})$ is the biggest subgroup contained in the monoid $\mathrm{Diff}_L(\mathcal{M})$ with respect to the composition law Eq. (6).

A.2.2 Pseudogroups

Definition A.7 (pseudogroup). For a submonoid A of $\mathrm{Diff}_L(\mathcal{M})$, if for any $f : \mathrm{Dom}(f) \to f(\mathrm{Dom}(f)) \in A$ there exists another element $f^{-1} : f(\mathrm{Dom}(f)) \to \mathrm{Dom}(f)$ also in A that satisfies the identities Eq. (7) and Eq. (8), then A is referred to as a pseudogroup of $\mathrm{Diff}_L(\mathcal{M})$.

Remark A.3. One of the important features of pseudogroups is that they have an associated orbit space. Indeed, if A is a pseudogroup we define the orbit $A \cdot p$ under A of any element $p \in \mathcal{M}$ as the set $A \cdot p := \{f(p) \mid f \in A, \text{ such that } p \in \mathrm{Dom}(f)\}$. A being a pseudogroup implies that the relation being in the same A-orbit is an equivalence relation and induces a partition of $\mathcal{M}$ into A-orbits.

A.2.3 Pseudogroups generated by arrow

A significant number of integrable pseudogroups are generated by collections of arrows (see Stefan (1974)).

Definition A.8 (arrow). An arrow is a differentiable mapping $\Phi : U \subset \mathbb{R} \times \mathcal{M} \to \mathcal{M}$ whose domain U is an open subset of $\mathbb{R} \times \mathcal{M}$ and that, additionally, satisfies:

(i) For every $t \in \mathbb{R}$, the map $\Phi_t := \Phi(t, \cdot)$ is a local diffeomorphism of $\mathcal{M}$ (possibly with empty domain).

(ii) If the point (t, p) belongs to the domain of Φ, then so does (s, p) for every $s \in [0, t]$. Moreover, $\Phi(0, p) = p$.

An example of an arrow is the flow of an analytic vector field on $\mathcal{M}$. Let E be a collection of arrows on $\mathcal{M}$. We associate E to a set $\mathcal{A}_E \subset \mathrm{Diff}_L(\mathcal{M})$ of local diffeomorphisms defined by $\mathcal{A}_E := \{\Phi_t \mid \Phi \in E, t \in \mathbb{R}\}$, which at the same time, generates a pseudogroup

$$A_E = (\mathrm{id}, \mathcal{M}) \cup \bigcup_n \{\Phi_1 \circ \cdots \circ \Phi_n \mid n \in \mathbb{N} \text{ and for all } i = 1, 2, \ldots, n, \Phi_i \in \mathcal{A}_E \text{ or } (\Phi_i)^{-1} \in \mathcal{A}_E\}.$$

A.3 Reachable Set and Orbit

Definition A.9 (integral curve of a family of vector fields, Definition 5 in Chapter 1 of Jurdjevic (1997)). Let $\mathcal{F}$ be a family of analytic vector fields on an analytic manifold $\mathcal{M}$. A continuous curve $p(t)$ in $\mathcal{M}$, defined on an interval $[0, T]$, is called an integral curve of $\mathcal{F}$ if there exist a partition $0 = t_0 < t_1 < \cdots < t_k = T$ and vector fields $X_1, \ldots, X_k$ in $\mathcal{F}$ such that the restriction of $p(t)$ to each open interval (t_{i-1}, t_i) is differentiable, and $\frac{\mathrm{d}p(t)}{\mathrm{d}t} = X_i(p(t))$ for $i = 1, \ldots, k$.

Definition A.10 (reachable set, Definition 6 in Chapter 1 of Jurdjevic (1997)). Let $\mathcal{F}$ be a family of analytic vector fields on an analytic manifold $\mathcal{M}$.

(i) For each $T \geq 0$, and each p_0 in $\mathcal{M}$, the set of points reachable from p_0 at time T, denoted by $R(p_0, T)$, is defined to be the set of the terminal points $p(T)$ of integral curves of $\mathcal{F}$ that originate at p_0.

(ii) The union of $R(p_0, T)$, for $T \geq 0$, is called the set reachable from p_0. We will denote it by $R(p_0)$.

Remark A.4 (page 28 of Jurdjevic (1997)). The reachable sets admit further geometric descriptions through the following formalism. Assuming that the elements of $\mathcal{F}$ are all complete vector fields, then each element X in $\mathcal{F}$ generates a one-parameter group of diffeomorphisms $\{e^{tX} \mid t \in \mathbb{R}^1\}$. Let $G(\mathcal{F})$ denote the subgroup of the group of diffeomorphisms in $\mathcal{M}$ generated by the union of $\{e^{tX} \mid t \in \mathbb{R}, X \in \mathcal{F}\}$. Each element Φ of $G(\mathcal{F})$ is a diffeomorphism of $\mathcal{M}$ of the form

$$\Phi = e^{t_k X_k} \circ e^{t_{k-1} X_{k-1}} \circ \cdots \circ e^{t_1 X_1}$$

for some real numbers $t_1, \ldots, t_k$ and vector fields $X_1, \ldots, X_k$ in $\mathcal{F}$. $G(\mathcal{F})$ acts on $\mathcal{M}$ in the obvious way and partitions $\mathcal{M}$ into its orbits. Then the set reachable through p_0 at time T consists of all points $\Phi(p_0)$ corresponding to elements Φ of $G(\mathcal{F})$ that can be expressed as $\Phi = e^{t_k X_k} \circ e^{t_{k-1} X_{k-1}} \circ \cdots \circ e^{t_1 X_1}$, with $t_1 \geq 0, \ldots, t_k \geq 0, t_1 + \cdots + t_k = T$, and $X_1, \ldots, X_k$ in $\mathcal{F}$. The other reachable sets have analogous descriptions. In particular, $R(p_0)$ is equal to the orbit of the semigroup $S_{\mathcal{F}}$ through p_0, with $S_{\mathcal{F}}$ equal to the semigroup of all elements Φ in $G(\mathcal{F})$ of the form $\Phi = e^{t_k X_k} \circ e^{t_{k-1} X_{k-1}} \circ \cdots \circ e^{t_1 X_1}$, with $t_1 \geq 0, \ldots, t_k \geq 0$, and $X_1, \ldots, X_k$ in $\mathcal{F}$. The orbit of $S_{\mathcal{F}}$ through p_0, written as $S_{\mathcal{F}}(p_0)$, is equal to $\{\Phi(p_0) \mid \Phi \in S_{\mathcal{F}}\}$.

When some elements of $\mathcal{F}$ are not complete, then it becomes necessary to replace the corresponding groups of diffeomorphisms by local groups, and everything else remains the same.

Definition A.11 (orbit of family of vector fields). Let $\mathcal{F}$ be a family of analytic vector fields on an analytic manifold $\mathcal{M}$. Let G denote the group (pseudogroup) of diffeomorphisms (local diffeomorphisms) generated by $\{e^{tX} \mid t \in \mathbb{R}, X \in \mathcal{F}\}$. Then the orbit of $\mathcal{F}$ through p is defined to be $\{\phi(p) \mid \phi \in G\}$.

A.4 Lie Algebra

Definition A.12 (Lie bracket, page 40 of Jurdjevic (1997)). Analytic vector fields act as derivations on the space of analytic functions. Moreover, If X denotes an analytic vector field, and h an analytic function on $\mathcal{M}$, then Xh will denote the function $p \mapsto X(p)(h)$. For any analytic vector fields X and Y, their Lie bracket $[X, Y]$ is defined by $[X, Y]h = Y(Xh) - X(Yh)$.

Remark A.5 (page 40 of Jurdjevic (1997)). If both X and Y are analytic vector fields on an analytic manifold $\mathcal{M}$, then $[X, Y]$ is an analytic vector field. Let $p \in \mathcal{M}$ and (ϕ, U) be a chart at p. The map $\phi : U \to \mathbb{R}^n$ gives a local coordinate $p \to (x_1(p), \ldots, x_n(p))$. In terms of the local coordinates, $[X, Y]$ is given by the following relations: let $X(p) = \sum_{i=1}^n a_i(x_1, \ldots, x_n)(\partial/\partial x_i)$, $Y(p) = \sum_{i=1}^n b_i(x_1, \ldots, x_n)(\partial/\partial x_i)$, and $[X, Y](p) = \sum_{i=1}^n c_i(x_1, \ldots, x_n)(\partial/\partial x_i)$. Then

$$c_i = \sum_{j=1}^{n} \left(\frac{\partial a_i}{\partial x_j} b_j - \frac{\partial b_i}{\partial x_j} a_j \right), \quad i = 1, 2, \ldots, n. \tag{9}$$

Definition A.13 (Lie algebra of analytic vector fields, page 42 of Jurdjevic (1997)). Let $\mathfrak{X}^\omega(\mathcal{M})$ denote the space of all analytic vector fields on $\mathcal{M}$. $\mathfrak{X}^\omega(\mathcal{M})$ is a real vector space under the pointwise addition of vectors

$$(\alpha X + \beta Y)(p) = \alpha X(p) + \beta Y(p) \quad \text{for all} \quad p \in \mathcal{M} \tag{10}$$

for each set of real numbers α and β and vector fields X and Y. We shall regard $\mathfrak{X}^{\omega}(\mathcal{M})$ as an algebra, with the addition given by Eq. (10) and with the product given by the Lie bracket.

Acknowledgment

We are grateful to Leyang Zhang for his valuable insights and suggestions for refining this paper[7]. This work is sponsored by the National Key R&D Program of China Grant No. 2022YFA1008200 (Y.Z., T.L.), Natural Science Foundation of China No. 1257010106 (Y.Z.), Natural Science Foundation of Shanghai No. 25ZR1402280 (Y.Z.), Shanghai Institute for Mathematics and Interdisciplinary Sciences (T.L.).

References

Zhiwei Bai, Jiajie Zhao, and Yaoyu Zhang. Connectivity shapes implicit regularization in matrix factorization models for matrix completion. *NeurIPS*, 2024.

Leo Breiman. Reflections after refereeing papers for nips. In *The Mathematics of Generalization*, pages 11–15. CRC Press, 2018.

Alon Brutzkus and Amir Globerson. Why do larger models generalize better? a theoretical perspective via the xor problem. *ICML*, 2019.

Emmanuel J Candès and Michael B Wakin. An introduction to compressive sampling. *IEEE signal processing magazine*, 25(2):21–30, 2008.

Emmanuel J Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on information theory*, 52(2):489–509, 2006.

Zheng-An Chen and Tao Luo. From condensation to rank collapse: A two-stage analysis of transformer training dynamics. *NeurIPS*, 2025.

Brian Conrad. Connectedness of hyperplane complements. `http://virtualmath1.stanford.edu/~conrad/diffgeomPage/handouts/hyperplaneconnd.pdf`.

David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4): 1289–1306, 2006.

Weinan E, Chao Ma, Lei Wu, and Stephan Wojtowytsch. Towards a mathematical understanding of neural network-based machine learning: what we know and what we don't. *CSIAM Trans. Appl. Math*, 1(4):561–615, 2006.

Kenji Fukumizu, Shoichiro Yamaguchi, Yoh-Ichi Mototake, and Mirai Tanaka. Semi-flat minima and saddle points by embedding neural networks to overparameterization. *NeurIPS*, 2019.

7. Email: leyangz_hawk@outlook.com

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

Velimir Jurdjevic. *Geometric control theory*. Cambridge university press, 1997.

Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *NeurIPS*, 2012.

John M. Lee. *Introduction to Smooth Manifolds*. Springer, New York, 2nd edition, 2012.

Ziyin Liu. Symmetry induces structure and constraint of learning. *ICML*, 2024.

Ziyin Liu, Mingze Wang, Hongchao Li, and Lei Wu. Parameter symmetry and noise equilibrium of stochastic gradient descent. *NeurIPS*, 2024.

Hans Dieter Luke. The origins of the sampling theorem. *IEEE Communications Magazine*, 37(4):106–108, 1999.

Tao Luo, Zhi-Qin John Xu, Zheng Ma, and Yaoyu Zhang. Phase diagram for two-layer relu neural networks at infinite-width limit. *Journal of Machine Learning Research*, 22(71): 1–47, 2021.

Hartmut Maennel, Olivier Bousquet, and Sylvain Gelly. Gradient descent quantizes relu network features. *arXiv preprint arXiv:1803.08367*, 2018.

Sibylle Marcotte, Rémi Gribonval, and Gabriel Peyré. Abide by the law and follow the flow: Conservation laws for gradient flows. *NeurIPS*, 2023.

Sibylle Marcotte, Gabriel Peyré, and Rémi Gribonval. Intrinsic training dynamics of deep neural networks. *arXiv preprint arXiv:2508.07370*, 2025.

Hancheng Min, Enrique Mallada, and Rene Vidal. Early neuron alignment in two-layer relu networks with small initialization. *ICLR*, 2024.

Tadashi Nagano. Linear differential systems with singularities and an application to transitive lie algebras. *Journal of the Mathematical Society of Japan*, 18(4):398–404, 1966. doi: 10.2969/jmsj/01840398.

Juan-Pablo Ortega and Tudor S Ratiu. *Momentum maps and Hamiltonian reduction*, volume 222. Springer Science & Business Media, 2013.

Claude E Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

Berfin Simsek, François Ged, Arthur Jacot, Francesco Spadaro, Clément Hongler, Wulfram Gerstner, and Johanni Brea. Geometry of the loss landscape in overparameterized neural networks: Symmetries and invariances. *ICML*, 2021.

Peter Stefan. Accessible sets, orbits, and foliations with singularities. *Proceedings of the London Mathematical Society*, 3(4):699–713, 1974.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017.

Zhi-Qin John Xu, Yaoyu Zhang, and Zhangchen Zhou. An overview of condensation phenomenon in deep learning. *arXiv preprint arXiv:2504.09484*, 2025.

Alfred Young. On quantitative substitutional analysis. *Proceedings of the London Mathematical Society*, 2(1):255–292, 1928.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *ICLR*, 2017.

Leyang Zhang, Yaoyu Zhang, and Tao Luo. Geometry and local recovery of global minima of two-layer neural networks at overparameterization. *arXiv preprint arXiv:2309.00508*, 2023.

Yaoyu Zhang, Zhongwang Zhang, Tao Luo, and Zhiqin J Xu. Embedding principle of loss landscape of deep neural networks. *NeurIPS*, 2021.

Yaoyu Zhang, Leyang Zhang, Zhongwang Zhang, and Zhiwei Bai. Local linear recovery guarantee of deep neural networks at overparameterization. *Journal of Machine Learning Research*, 26(69):1–30, 2025a. URL `http://jmlr.org/papers/v26/24-0192.html`.

Yaoyu Zhang, Zhongwang Zhang, Leyang Zhang, Zhiwei Bai, Tao Luo, and Zhi-Qin John Xu. Optimistic estimate uncovers the potential of nonlinear models. *Journal of Machine Learning*, 4(3):192–222, 2025b.

Jiajie Zhao, Zhiwei Bai, and Yaoyu Zhang. Disentangle sample size and initialization effect on perfect generalization for single-neuron target. *arXiv preprint arXiv:2405.13787*, 2024.