A Unified Framework to Enforce, Discover, and Promote Symmetry in Machine Learning

Samuel E. Otto[*]   S.OTTO@CORNELL.EDU
AI Institute in Dynamic Systems
University of Washington
Seattle, WA 98195-4322, USA

Nicholas Zolman   NZOLMAN@UW.EDU
AI Institute in Dynamic Systems
University of Washington
Seattle, WA 98195-4322, USA

J. Nathan Kutz   KUTZ@UW.EDU
AI Institute in Dynamic Systems
University of Washington
Seattle, WA 98195-4322, USA

Steven L. Brunton   SBRUNTON@UW.EDU
AI Institute in Dynamic Systems
University of Washington
Seattle, WA 98195-4322, USA

Abstract

Symmetry is present throughout nature and continues to play an increasingly central role in machine learning. In this paper, we provide a unifying theoretical and methodological framework for incorporating Lie group symmetry into machine learning models in three ways: **1. enforcing** known symmetry when training a model; **2. discovering** unknown symmetries of a given model or data set; and **3. promoting** symmetry during training by learning a model that breaks symmetries within a user-specified candidate group only when the data provide sufficient evidence. We show that these tasks can be cast within a common mathematical framework whose central object is the Lie derivative. We extend and unify several existing results by showing that enforcing and discovering symmetry are linear-algebraic tasks that are dual under the bilinear pairing induced by the Lie derivative. We also propose a novel way to promote symmetry by introducing a class of convex regularizers, built from the Lie derivative with a nuclear-norm relaxation, that penalizes symmetry breaking during training. We explain how these ideas can be applied to a wide range of machine learning models including basis-function regression, dynamical-systems discovery, neural networks, and neural operators acting on fields.

Keywords: symmetry, machine learning, Lie groups, manifolds, invariance, equivariance, neural networks, deep learning

---

[*]. Current affiliation: Sibley School of Mechanical and Aerospace Engineering, Cornell University, Ithaca, NY, USA

Contents

1. Introduction

Symmetry is present throughout nature, and as David Gross (Gross, 1996) argues, the discovery of fundamental symmetries has played an increasingly central role in physics since the beginning of the 20th century. He asserts that

> "Einstein's great advance in 1905 was to put symmetry first, to regard the symmetry principle as the primary feature of nature that constrains the allowable dynamical laws."

According to Einstein's special theory of relativity, physical laws—including those of electromagnetism and quantum mechanics—are Poincaré-invariant, meaning that after the corresponding transformations (actions of the Poincaré group), these laws can be applied in any non-accelerating reference frame, anywhere in the universe, at all times. Specifically, these transformations form a ten-parameter group consisting of four space-time translations, three spatial rotations, and three velocity shifts called *boosts*. For small boosts of velocity, these transformations become the Galilean symmetries of classical mechanics. Similarly, the theorems of Euclidean geometry are invariant under arbitrary translations, rotations, and reflections, which together constitute the Euclidean group. In fluid mechanics, the conformal (angle-preserving) symmetry of Laplace's equation reduces the study of idealized flows in complicated geometries to canonical flows in simple domains. In dynamical systems, the celebrated theorem of Noether (1918) establishes a correspondence between symmetries and conservation laws, an idea which has become a central pillar of mechanics (Abraham and Marsden, 2008). These examples illustrate the diversity of symmetry groups and their physical applications. More importantly, they illustrate how *symmetric models and theories automatically extrapolate beyond the available data in an explainable way.*

In machine learning, models that exploit symmetry can be trained with less data and use fewer parameters compared to their asymmetric counterparts. Early examples include augmenting data with known transformations (see Shorten and Khoshgoftaar, 2019; Van Dyk and Meng, 2001) or using convolutional neural networks (CNNs) to achieve translation equivariance for image processing tasks (see Fukushima, 1980; LeCun et al., 1989). More recently, equivariant neural networks respecting Euclidean symmetries have achieved state-of-the-art performance for predicting potentials in molecular dynamics (Batzner et al., 2022). As with physical laws, symmetries and invariances allow machine learning models to extrapolate beyond the training data, and achieve high performance with fewer modeling parameters.

However, many problems are only weakly symmetric. Gravity, friction, and other external forces can break some or all of the Poincaré or Galilean symmetries. Interactions between particles can be viewed as breaking symmetries possessed by noninteracting particles. Written characters have translation and scaling symmetries, but not rotation (cf. 6 vs 9; d vs p; N vs Z) or reflection (cf. b vs d; b vs p). One of the main contributions of this work is to propose a method of enforcing a new principle of parsimony in machine learning. This *maximal-symmetry parsimony principle* states that a model should break a symmetry within a set of physically reasonable transformations (such as Poincaré, Galilean, Euclidean, or conformal symmetry) only when there is sufficient evidence in the data.

In this paper, we provide a unifying theoretical and methodological framework for incorporating symmetry into machine learning models in the following three ways:

Task 1. **Enforce.** Train a model with known symmetry.

Task 2. **Discover.** Identify the symmetries of a given model or data set.

Task 3. **Promote.** Train a model with as many symmetries as possible (from among candidates), breaking symmetries only when the data provide sufficient evidence.

While these tasks have been studied separately to varying extents, we show how they can be situated within a common mathematical framework whose central object is the Lie derivative arising from fiber-linear Lie group actions on vector bundles. As a special case, the Lie derivative recovers the linear constraints derived by Finzi et al. (2021) for weights in equivariant multilayer perceptrons. In full generality, *we show that known symmetries can be enforced as linear constraints derived from Lie derivatives for a large class of problems* including learning vector fields, tensor fields, and equivariant integral operators acting on such fields. For example, the kernels of *steerable* CNNs developed by Weiler et al. (2018); Weiler and Cesa (2019) are constructed to automatically satisfy these constraints for the groups SO(3) (rotations in three dimensions) and SE(2) (rotations and translations in two dimensions). We show how analogous steerable networks for other groups, such as subgroups of SE(n), can be constructed by numerically enforcing linear constraints derived from the Lie derivative on the integral kernels defining each layer. Symmetries, conservation laws, and symplectic structure can also be enforced when learning dynamical systems via linear constraints on the vector field. These constraints arise from the Lie derivative and can be incorporated into neural-network architectures and basis-function regression models such as Sparse Identification of Nonlinear Dynamics (SINDy) (Brunton et al., 2016).

Moskalev et al. (2022) identifies the connected subgroup of symmetries of a trained neural network by computing the nullspace of a linear operator. Likewise, Kaiser et al. (2018, 2021) recover the symmetries and conservation laws of a dynamical system by computing the nullspace of a different linear operator. *We observe that these operators—and others for which the nullspaces encode the symmetries of more general models—can be derived directly from the Lie derivative in a manner that is dual to the operators used to enforce symmetry.* Specifically, the nullspaces of the operators we construct reveal the largest connected subgroups of symmetries for large classes of models. This extends work by Gruver et al. (2022) using the Lie derivative to test whether a trained neural network is equivariant with respect to a given one-parameter group, e.g., rotation of images. Generalizing the ideas in Cahill et al. (2023) to nonlinear actions on manifolds, we also show that the symmetries of point clouds approximating submanifolds could be recovered by computing the nullspaces of associated linear operators. This could enable unsupervised mining of data for hidden symmetries.

The idea of relaxed symmetry was introduced by Wang et al. (2022), along with architecture-specific symmetry-promoting regularization functions involving sums or integrals over the candidate group. The Augerino method (Benton et al., 2020) uses regularization to promote equivariance with respect to a larger collection of candidate transformations. Promoting physical constraints via the loss is also core to Physics-Informed Neural Networks (PINNs; Raissi et al., 2019). *Our approach to the third task (promoting symmetry) is to introduce a unified and broadly applicable class of convex regularization functions based on the Lie derivative to penalize symmetry breaking during training of machine learning models.* As we describe above, the Lie derivative yields an operator whose nullspace corresponds to the symmetries of a given model. Hence, the lower the rank of this operator, the more

symmetric the model is. The nuclear norm has been used extensively as a convex relaxation of the rank with favorable theoretical properties for compressed sensing and low-rank matrix recovery (Recht et al., 2010; Gross, 2011), as well as in robust PCA (Candès et al., 2011; Bouwmans et al., 2018). Penalizing the nuclear norm of our symmetry-encoding operator yields a convex regularization function that can be added to the loss when training machine learning models, including basis-function regression and neural networks. Likewise, we show how nuclear norm penalties can be used to promote conservation laws and Hamiltonian structure with respect to candidate symplectic structures when fitting dynamical systems to data. This approach lets us train a model and promote data-consistent symmetries simultaneously.

2. Executive Summary

This paper provides a linear-algebraic framework to enforce, discover, and promote symmetry of machine learning models. To illustrate, consider a model defined by a function $F : \mathbb{R}^m \to \mathbb{R}^n$. By a symmetry, we mean an invertible transformation θ and an invertible linear transformation T so that

$$F(\theta(x)) = TF(x) \qquad \text{for all } x \in \mathbb{R}^m. \tag{1}$$

Examples to keep in mind are rotations and translations. If (θ_a, T_a) is a symmetry, then so is its inverse $(\theta_a^{-1}, T_a^{-1})$, and if (θ_b, T_b) is another symmetry, then so is the composition $(\theta_b \circ \theta_a, T_b T_a)$. We work with sets of transformations $\{(\theta_g, T_g)\}_{g \in G}$, called groups, which have an identity element and are closed under composition and inversion. Specifically, we consider Lie group actions on manifolds.

2.1 Enforcing Symmetry

Given a group of transformations, the symmetry condition (1) imposes linear constraints on F that can be enforced during the fitting process. However, there is one constraint per transformation, making direct enforcement impractical for continuous Lie groups such as rotations or translations. For smoothly parameterized connected groups it suffices to enforce a finite collection of linear constraints $\mathcal{L}_{\xi_i} F = 0$ where ξ_i are elements in the *Lie algebra* and $\mathcal{L}_\xi$ is the *Lie derivative* defined by

$$\mathcal{L}_\xi F(x) = \left(\frac{\partial F(x)}{\partial x} \left. \frac{\partial \theta_g(x)}{\partial g} \right|_{g=\text{Id}} - \left. \frac{\partial T_g F(x)}{\partial g} \right|_{g=\text{Id}} \right) \xi. \tag{2}$$

This operation is linear with respect to ξ and with respect to F.

2.2 Discovering Symmetry

The symmetries of a given model F form a subgroup that we seek to identify within a given group of candidates. For continuous Lie groups of transformations, the connected component of the subgroup containing the identity is revealed by the nullspace of the linear map

$$L_F : \xi \mapsto \mathcal{L}_\xi F. \tag{3}$$

5

More generally, the symmetries of a smooth surface in $\mathbb{R}^n$ can be determined from data sampled from this surface by computing the nullspace of a positive semidefinite operator. When the surface is the graph of the function F in $\mathbb{R}^m \times \mathbb{R}^n$, this operator is $L_F^* L_F$ with L_F^* being an adjoint operator.

2.3 Promoting Symmetry

Here, we seek to learn a model F that both fits data and possesses as many symmetries as possible from a given candidate group of transformations. Since the nullspace of the operator L_F defined in (3) corresponds with the symmetries of F, we seek to minimize the rank of L_F during training. To do this, we regularize optimization problems for F using a convex relaxation of the rank given by the nuclear norm (sum of singular values)

$$\|L_F\|_* = \sum_{i=1}^{\dim G} \sigma_i(L_F).$$

This is convex with respect to F because $F \mapsto L_F$ is linear and the nuclear norm is convex.

3. Related Work

3.1 Enforcing Symmetry

Data augmentation, as reviewed by Shorten and Khoshgoftaar (2019); Van Dyk and Meng (2001), is one of the simplest ways to incorporate known symmetries into machine learning models. Usually this entails training a neural network architecture on training data to which known transformations have been applied. The theoretical foundations of these methods are explored by Chen et al. (2020). Data augmentation has also been used by Benton et al. (2020) to construct equivariant neural networks by averaging over transformed network inputs and outputs. Finally, Brandstetter et al. (2022) applied data augmentation strategies with known Lie point symmetries to improve neural PDE solvers.

Symmetry can also be enforced directly on the machine learning architecture. For example, Convolutional Neural Networks (CNNs), introduced by Fukushima (1980) and popularized by LeCun et al. (1989), achieve translational equivariance by employing convolutional filters with trainable kernels in each layer. CNNs have been generalized to provide equivariance with respect to symmetry groups other than translation. Group-Equivariant CNNs (G-CNNs) (Cohen and Welling, 2016) provide equivariance with respect to arbitrary discrete groups generated by translations, reflections, and rotations. Rotational equivariance can be enforced on three-dimensional scalar, vector, or tensor fields using the 3D Steerable CNNs developed by Weiler et al. (2018). Spherical CNNs (Cohen et al., 2018; Esteves et al., 2018) allow for rotation-equivariant maps to be learned for fields (such as projected images of 3D objects) on spheres. Essentially any group-equivariant linear map (defining a layer of an equivariant neural network) acting on fields can be described by group convolution (Kondor and Trivedi, 2018; Cohen et al., 2019), with the spaces of appropriate convolution kernels characterized by Cohen et al. (2019). Finzi et al. (2020) provides a practical way to construct convolutional layers that are equivariant with respect to arbitrary Lie groups and for general data types. For dynamical systems, Marsden and Ratiu (1999); Rowley et al. (2003); Abraham and

Marsden (2008) describe techniques for symmetry reduction of the original problem to a quotient space where the known symmetry group has been factored out. Related approaches have been used by Peitz et al. (2023); Steyert (2022) to approximate Koopman operators for symmetric dynamical systems (see Koopman, 1931; Mezić, 2005; Mauroy et al., 2020; Otto and Rowley, 2021; Brunton et al., 2022).

A general method for constructing equivariant neural networks was introduced by Finzi et al. (2021), based on enforcing linear constraints on layer weights. For graph neural networks, Maron et al. (2018) characterizes the subspaces of linear layers satisfying permutation equivariance. Similarly, Ahmadi and Khadir (2020, 2023) show that discrete symmetries and other types of side information including interpolation at finite sets of points, nonnegativity, monotonicity, invariance of sets, and gradient or Hamiltonian structure can be enforced via linear or convex constraints in learning problems for dynamical systems. Our work builds on Finzi et al. (2021), Weiler et al. (2018), Cohen et al. (2019), and Ahmadi and Khadir (2020, 2023) by showing that equivariance can be enforced in a systematic and unified way via linear constraints for large classes of functions and neural networks. Concurrent work by Yang et al. (2024) shows how to enforce known or discovered Lie group symmetries on latent dynamics using hard linear constraints or soft penalties.

3.2 Discovering Symmetry

Early work by Rao and Ruderman (1999); Miao and Rao (2007) used nonlinear optimization to learn infinitesimal generators describing transformations between images. Later, Cahill et al. (2023) recognized that linear-algebraic methods could be used to uncover the generators of continuous linear symmetries of arbitrary point clouds in Euclidean space. Similarly, Kaiser et al. (2018) and Moskalev et al. (2022) show how conserved quantities of dynamical systems and invariances of trained neural networks can be revealed by computing the nullspaces of associated linear operators. We connect these linear-algebraic methods to the Lie derivative, and provide generalizations to nonlinear group actions on manifolds. The Lie derivative has been used by Gruver et al. (2022) to quantify the extent to which a trained network is equivariant with respect to a given one-parameter subgroup of transformations. Our results show how the Lie derivative can reveal the entire connected subgroup of symmetries of a trained model via symmetric eigendecomposition.

More sophisticated nonlinear optimization techniques use Generative Adversarial Networks (GANs) to learn the transformations that leave a data distribution unchanged. These methods include SymmetryGAN developed by Desai et al. (2022) and LieGAN developed by Yang et al. (2023b). In contrast, our methods for detecting symmetry are entirely linear-algebraic.

While symmetries may exist in data, their representation may be difficult to describe. Yang et al. (2023a) introduced an extension LieGAN, called Latent LieGAN (LaLieGAN), to find linear representations of symmetries in a latent space. Recently this has been applied to dynamics discovery (Yang et al., 2024). Likewise, Liu and Tegmark (2022) discover hidden symmetries by optimizing nonlinear transformations into spaces where candidate symmetries hold. Similar to our approach for promoting symmetry, they use a cost function to measure whether a given symmetry holds. In contrast, our regularization functions enable subgroups of candidate symmetry groups to be identified.

3.3 Promoting Symmetry

Biasing a network towards increased symmetry can be accomplished through methods such as symmetry regularization. Analogous to the physics-informed loss developed in PINNs by Raissi et al. (2019), which penalizes a solution for violating known dynamics, one can penalize symmetry violation for a known group; for example, Akhound-Sadegh et al. (2024) extends the PINN framework to penalize deviations from known Lie point symmetries of a PDE. More generally, however, one can consider a *candidate* group of symmetries and promote as much symmetry as possible while remaining *consistent with the available data.* Wang et al. (2022) discusses these approaches, along with architecture-specific methods, including regularization functions involving summations or integrals over the candidate group of symmetries. While our regularization functions resemble these for discrete groups, we use a different approach for continuous Lie groups. By leveraging the Lie algebra, our regularization functions eliminate the need to numerically integrate over the group—a task that is already prohibitive for the 10-dimensional noncompact group of Galilean symmetries in classical mechanics.

Automated data augmentation techniques introduced by Cubuk et al. (2019); Hataya et al. (2020); Benton et al. (2020) are another class of methods that arguably promote symmetry. These techniques optimize the distribution of transformations applied to augment the data during training. For example, Augerino is an elegant method developed by Benton et al. (2020) which averages an arbitrary network's output over the augmentation distribution and relies on regularization to prevent the distribution of transformations from becoming concentrated near the identity. In essence, the regularization biases the averaged network towards increased symmetry.

In contrast, our regularization functions promote symmetry at the architectural level for the original network. This eliminates the need to perform averaging, which grows more costly for larger collections of symmetries. While a distribution over symmetries can be useful for learning interesting partial symmetries (e.g., 6 stays 6 for small rotations, before turning into 9), as is done by Benton et al. (2020) and Romero and Lohit (2022), it is unclear how sampling from a continuous distribution would identify lower-dimensional subgroups (measure-zero). On the other hand, a linear-algebraic approach provides a natural way to identify and promote symmetries in lower-dimensional connected subgroups.

3.4 Additional Approaches and Applications

There are several other approaches that incorporate various aspects of enforcing, discovering, and promoting symmetries. For example, Baddoo et al. (2023) developed algorithms to enforce and promote known symmetries in dynamic mode decomposition, through manifold constrained learning and regularization, respectively. They also show that discovering unknown symmetries is a dual problem to enforcing symmetry. Exploiting symmetry has also been a central theme in the reduced-order modeling of fluids for decades (Holmes et al., 2012). As machine learning methods have become widely used to construct these models (Brunton et al., 2020), the themes of enforcing and discovering symmetries in these models are increasingly relevant. Known fluid symmetries have been enforced in SINDy for fluid systems (Loiseau and Brunton, 2018) through linear equality constraints; this approach was generalized to enforce more complex constraints (Champion et al., 2020).

Unknown symmetries were similarly uncovered for electroconvective flows (Guan et al., 2021). Symmetry breaking is also important in many turbulent flows (Callaham et al., 2022).

4. Elementary Theory of Lie Group Actions

We provide background and notation required to understand the main results of this paper in a less abstract but still highly useful setting of Lie groups acting on vector spaces. In Section 5 we use this theory to study the symmetries of continuously differentiable functions between vector spaces. Such functions form the basic building blocks of many machine learning models such as basis-function regression models, the layers of multilayer perceptrons, and the kernels of integral operators acting on spatial fields such as images. We emphasize that this is not the most general setting for our results, but we provide this section and simpler versions of our main theorems in Section 5 in order to make the presentation more accessible. We develop our main results in the more general setting of fiber-linear Lie group actions on sections of vector bundles in Section 11.

4.1 Lie Groups and Subgroups

Lie groups are ubiquitous in science and engineering. Some familiar examples include the general linear group $\mathrm{GL}(n)$ consisting of all real, invertible, $n \times n$ matrices; the orthogonal group

$$\mathrm{O}(n) = \left\{ Q \in \mathbb{R}^{n \times n} \ : \ Q^T Q = I \right\}; \tag{4}$$

and the special Euclidean group

$$\mathrm{SE}(n) = \left\{ \begin{bmatrix} Q & b \\ 0 & 1 \end{bmatrix} \ : \ Q \in \mathbb{R}^{n \times n}, \ b \in \mathbb{R}^n, \ Q^T Q = I, \ \det(Q) = 1 \right\}, \tag{5}$$

representing rotations and translations in real n-dimensional space, $\mathbb{R}^n$, embedded in $\mathbb{R}^{n+1}$ via $x \mapsto (x, 1)$. Observe that the sets $\mathrm{GL}(n)$, $\mathrm{O}(n)$, and $\mathrm{SE}(n)$ contain the identity matrix and are closed under matrix multiplication and inversion, making them into (noncommutative) groups. They are also smooth manifolds, which makes them *Lie groups*. In general, a Lie group is a smooth manifold that is simultaneously an algebraic group whose composition and inversion operations are smooth maps. The identity element is usually denoted e (German: Einheitselement), which in a matrix Lie group is the identity matrix $e = I$. This section summarizes some basic results that can be found in references such as (Abraham et al., 1988; Lee, 2013; Varadarajan, 1984; Hall, 2015).

A key property of a Lie group is that much of its structure is captured by an associated vector space called its *Lie algebra*. This allows global nonlinear questions about the group— such as which elements leave a function unchanged—to be answered using linear algebra. If G is a Lie group, its Lie algebra, commonly denoted $\mathrm{Lie}(G)$ or $\mathfrak{g}$, is the vector space consisting of all smooth vector fields on G that remain invariant when pushed forward by left translation $L_g : h \mapsto g \cdot h$. Translating back and forth from the identity element, the Lie algebra can be identified with the tangent space $\mathrm{Lie}(G) \cong T_e G$. For example, the Lie algebra

of the orthogonal group $O(n)$ consists of all skew-symmetric matrices, and is denoted

$$\mathfrak{o}(n) = \left\{ S \in \mathbb{R}^{n \times n} \; : \; S + S^T = 0 \right\}.$$

A key fact is that the Lie algebra of G is closed under the *Lie bracket* of vector fields[1]

$$[\xi, \eta] = \xi\eta - \eta\xi \in \mathrm{Lie}(G).$$

For matrix Lie groups, the Lie bracket coincides with the usual matrix commutator $\xi, \eta \in T_I G$.

The key tool relating global properties of a Lie group back to its Lie algebra is the exponential map $\exp : \mathrm{Lie}(G) \to G$. A vector field $\xi \in \mathrm{Lie}(G)$ has a unique integral curve $\gamma : (-\infty, \infty) \to G$ passing through the identity $\gamma(0) = e$ and satisfying $\gamma'(t) = \xi|_{\gamma(t)}$. The exponential map defined by

$$\exp(\xi) := \gamma(1)$$

reproduces the entire integral curve via $\exp(t\xi) = \gamma(t)$. An exponential curve is illustrated in Figure 1. For a matrix Lie group and $\xi \in T_I G$, the exponential map is the matrix exponential

$$\exp(\xi) = e^\xi = \sum_{k=0}^{\infty} \frac{1}{k!} \xi^k.$$

Many of the basic properties of the exponential map, such as $\exp((s+t)\xi) = \exp(s\xi) \cdot \exp(t\xi)$, $\exp(\xi)^{-1} = \exp(-\xi)$, and $d\exp(0) = \mathrm{Id}_{T_e G}$ are summarized by Lee (2013) in Proposition 20.8. The exponential map also provides a diffeomorphism of an open neighborhood of the origin 0 in $\mathrm{Lie}(G)$ and an open neighborhood of the identity element e in G.

The connected component of G containing the identity element is called the *identity component* of the Lie group and is denoted G_0. Any element in this component can then be expressed as a finite product of exponentials, that is

$$G_0 = \left\{ \exp(\xi_1) \cdots \exp(\xi_N) \; : \; \xi_1, \ldots, \xi_N \in \mathrm{Lie}(G), \; N = 1, 2, 3, \ldots \right\}.$$

Moreover, the identity component is a normal subgroup of G and all of the other connected components of G are diffeomorphic cosets of G_0, as we illustrate in Figure 1. For example, the special Euclidean group $SE(n)$ is connected, and thus equal to its identity component. On the other hand, the orthogonal group $O(n)$ is compact and has two components consisting of orthogonal matrices Q whose determinants are 1 and -1. The identity component of the orthogonal group is called the special orthogonal group and is denoted $SO(n)$. It is a general fact that when a Lie group such as $SO(n)$ is connected and compact, it is equal to the image of the exponential map without the need to consider products of exponentials (see Tao, 2011 and Appendix C.1 of Lezcano-Casado and Martınez-Rubio, 2019).

A subgroup H of a Lie group G is called a *Lie subgroup* when H is an immersed submanifold of G and the group operations are smooth when restricted to H. An immersed submanifold does not necessarily inherit its topology as a subset of G, but rather H has

---

1. In $\mathbb{R}^n$ a vector field $V = (V^1, \ldots, V^n)$ is equivalent to a directional derivative operator $V^1 \frac{\partial}{\partial x^1} + \cdots + V^n \frac{\partial}{\partial x^n}$. A vector field on a smooth manifold is defined as an analogous linear operator acting on the space of smooth functions. The Lie bracket is defined as the commutator of these operators.
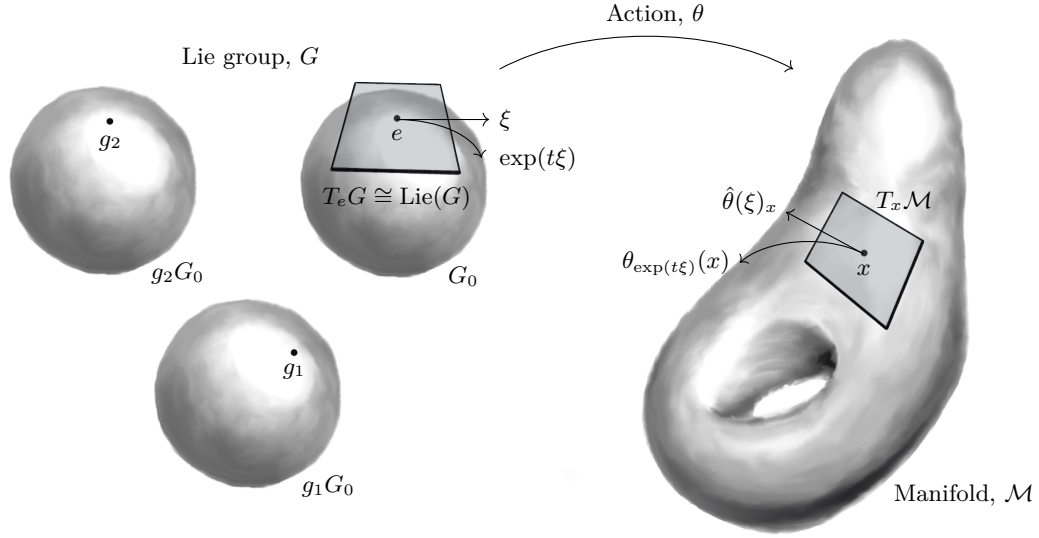
Figure 1: A Lie group G and its action θ on a manifold $\mathcal{M}$. The Lie group G consists of three connected components with G_0 being the one that contains the identity element e. Each non-identity component of G is a coset $g_i G_0$ formed by translating the identity component by an arbitrary element g_i in the component. The Lie algebra $\mathrm{Lie}(G)$ is identified with the tangent space $T_e G$ and an exponential curve $\exp(t\xi)$ generated by an element $\xi \in \mathrm{Lie}(G)$ is shown. The infinitesimal generator $\hat{\theta}(\xi)$ is the vector field on $\mathcal{M}$ whose flow corresponds with the action $\theta_{\exp(t\xi)}$ of group elements along $\exp(t\xi)$.

a topology and smooth structure that makes the inclusion $\iota_H : H \hookrightarrow G$ smooth and its derivative injective. The tangent space to a Lie subgroup $H \subset G$ at the identity, defined as $T_e H = \mathrm{Range}(\mathrm{d}\,\iota_H(I)) \subset \mathrm{Lie}(G)$, is closed under the Lie bracket and thus forms a *Lie subalgebra* of $\mathrm{Lie}(G)$, denoted $\mathrm{Lie}(H)$. Conversely, any subalgebra $\mathfrak{h} \subset \mathrm{Lie}(G)$, that is, any subspace closed under the Lie bracket, corresponds to a unique connected Lie subgroup $H \subset G_0$ satisfying $\mathrm{Lie}(H) = \mathfrak{h}$ (Theorem 19.26; Lee, 2013). Later on, we will use this fact to identify the connected subgroups of symmetries of machine learning models based on infinitesimal criteria. Another useful result is the *closed subgroup theorem* (Theorem 20.12; Lee, 2013). It says that if $H \subset G$ is a closed subset and is closed under the group operations of G, then H is automatically an embedded Lie subgroup of G. Interestingly, while a Lie subgroup $H \subset GL(n)$ need not be a closed subset, it turns out that H can always be embedded as a closed subgroup in a larger $GL(n')$, $n' \geq n$, thanks to Theorem 9 in Gotô (1950).

4.2 Group Representations, Actions, and Infinitesimal Generators

A Lie group homomorphism is a smooth map $\Phi : G_1 \to G_2$ between Lie groups that respects the group product, that is,

$$\Phi(g_1 g_2) = \Phi(g_1) \cdot \Phi(g_2).$$

The tangent map $\phi := \mathrm{d}\,\Phi(e) : T_e G_1 \cong \mathfrak{g}_1 \to T_e G_2 \cong \mathfrak{g}_2$ is a Lie algebra homomorphism, meaning that it is a linear map respecting the Lie bracket:

$$\phi([\xi_1, \xi_2]) = [\phi(\xi_1), \phi(\xi_2)].$$

Moreover, the Lie group homomorphism and its induced Lie algebra homomorphism are related by the exponential maps on G_1 and G_2 via the identity

$$\Phi\big(\exp(\xi)\big) = \exp\big(\phi(\xi)\big). \tag{6}$$

Another fundamental result is that any Lie algebra homomorphism $\mathrm{Lie}(G_1) \to \mathrm{Lie}(G_2)$ corresponds to a unique Lie group homomorphism $G_1 \to G_2$ when G_1 is simply connected. When G_2 is the general linear group on a vector space, then the Lie group and Lie algebra homomorphisms are called Lie group and Lie algebra *representations*.

A Lie group G can act on a vector space $\mathcal{V}$ via a representation $\Phi : G \to \mathrm{GL}(\mathcal{V})$ by

$$\theta : (x, g) \mapsto \Phi(g^{-1})x, \tag{7}$$

with $x \in \mathcal{V}$ and $g \in G$. More generally, a nonlinear right action of a Lie group G on a manifold $\mathcal{M}$ is any smooth map $\theta : \mathcal{M} \times G \to \mathcal{M}$ satisfying

$$\theta(\theta(x, g_1), g_2) = \theta(x, g_1 g_2) \qquad \text{and} \qquad \theta(x, e) = x$$

for every $x \in \mathcal{M}$ and $g_1, g_2 \in G$. Figure 1 depicts the action of a Lie group on a manifold. We make frequent use of the maps $\theta_g = \theta(\cdot, g)$, which have smooth inverses $\theta_{g^{-1}}$, and the *orbit maps* $\theta^{(x)} = \theta(x, \cdot)$. For example, using a representation $\Phi : \mathrm{SE}(3) \to \mathrm{GL}(\mathbb{R}^7)$, the position q and velocity v of a particle in $\mathbb{R}^3$ can be rotated and translated via the action

$$\theta\left(\begin{bmatrix} Q & b \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} q \\ v \\ 1 \end{bmatrix}\right) = \Phi\left(\begin{bmatrix} Q^T & -Q^T b \\ 0 & 1 \end{bmatrix}\right)\begin{bmatrix} q \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} Q^T & 0 & -Q^T b \\ 0 & Q^T & 0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} q \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} Q^T(q - b) \\ Q^T v \\ 1 \end{bmatrix}. \tag{8}$$

The positions and velocities of n particles arranged as a vector $(q_1, \ldots, q_n, v_1, \ldots, v_n, 1)$ can be simultaneously rotated and translated via an analogous representation $\Phi : \mathrm{SE}(3) \to \mathrm{GL}(\mathbb{R}^{6n+1})$.

We will make use of the fact that a Lie group action is almost completely characterized by a linear map called its *infinitesimal generator*. The generator $\hat{\theta}$ assigns to each element $\xi \in \mathrm{Lie}(G)$ in the Lie algebra, a smooth vector field $\hat{\theta}(\xi)$ defined at $x \in \mathcal{M}$ by

$$\hat{\theta}(\xi)_x = \frac{\mathrm{d}}{\mathrm{d}\,t}\left.\theta_{\exp(t\xi)}(x)\right|_{t=0} = \mathrm{d}\,\theta^{(x)}(e)\xi.$$

The infinitesimal generator and its relation to the group action are illustrated in Figure 1. For the linear action in (7), the infinitesimal generator is the linear vector field given by the matrix-vector product $\hat{\theta}(\xi)_x = -\phi(\xi)x$. Crucially, the generator's flow is identical to the

group action along the exponential curve $\exp(t\xi)$, i.e.,

$$\mathrm{Fl}^t_{\hat{\theta}(\xi)}(x) = \theta_{\exp(t\xi)}(x).$$

For the linear right action in (7), this is easily verified by differentiation, applying (6), and the fact that solutions of smooth ordinary differential equations are unique. For a nonlinear right action this follows from Lemma 20.14 and Proposition 9.13 in Lee (2013).

Remark 1 *In contrast to a right action $\theta : \mathcal{M} \times G \to \mathcal{M}$, a left action $\theta^L : G \times \mathcal{M} \to \mathcal{M}$ satisfies $\theta^L(g_2, \theta^L(g_1, x)) = \theta^L(g_2 g_1, x)$. While our main results work for left actions too, e.g., $\theta^L(g, x) = \Phi(g)x$, we fix right actions throughout because the infinitesimal generator is a Lie algebra homomorphism, i.e.,*

$$\hat{\theta}([\xi, \eta]) = [\hat{\theta}(\xi), \hat{\theta}(\eta)],$$

whereas this holds with a sign change for left actions. Every left action can be converted into an equivalent right action defined by $\theta^R(x, g) = \theta^L(g^{-1}, x)$, and vice versa.

5. Fundamental Operators for Studying Symmetry

Here we introduce our main theoretical results for studying symmetries of machine learning models by focusing on a concrete and useful special case. The basic building blocks of the machine learning models we consider here are continuously differentiable functions $F : \mathcal{V} \to \mathcal{W}$ between finite-dimensional vector spaces. The spaces of functions $\mathcal{V} \to \mathcal{W}$ with continuous derivatives up to order $k \in \mathbb{N} \cup \{\infty\}$ are denoted $C^k(\mathcal{V}; \mathcal{W})$, with addition and scalar multiplication defined point-wise. These functions could be layers of a multilayer neural network, integral kernels applied to spatio-temporal fields, or simply linear combinations of user-specified basis functions in a regression task as in Brunton et al. (2016). General versions of our results for sections of vector bundles are developed later in Section 11. Our main results show that two families of fundamental linear operators encode the symmetries of these functions. The fundamental operators allow us to enforce, discover, and promote symmetry in machine learning models as we describe in Sections 6, 7, and 8.

Letting G be a Lie group, we consider a right action $\Theta : \mathcal{V} \times \mathcal{W} \times G \to \mathcal{V} \times \mathcal{W}$,

$$\Theta(x, y, g) = (\theta(x, g),\ T(x, g)y), \tag{9}$$

where θ is a general (perhaps nonlinear) right G-action on $\mathcal{V}$ and $T : \mathcal{V} \times G \to \mathrm{GL}(\mathcal{W})$ is a smooth function satisfying

$$T(x, g_1 g_2) = T(\theta(x, g_1), g_2)T(x, g_1) \qquad \text{and} \qquad T(x, e) = I$$

for all $x \in \mathcal{V}$ and every $g_1, g_2 \in G$. The infinitesimal generator of Θ with $\xi \in \mathrm{Lie}(G)$ is

$$\hat{\Theta}(\xi)_{(x,y)} = (\hat{\theta}(\xi)_x,\ \hat{T}(\xi)_x y), \quad \text{where} \quad \hat{T}(\xi)_x = \frac{\mathrm{d}}{\mathrm{d}\,t} T(x, \exp(t\xi))\Big|_{t=0} = \mathrm{d}\,T(x, e)(0, \xi).$$
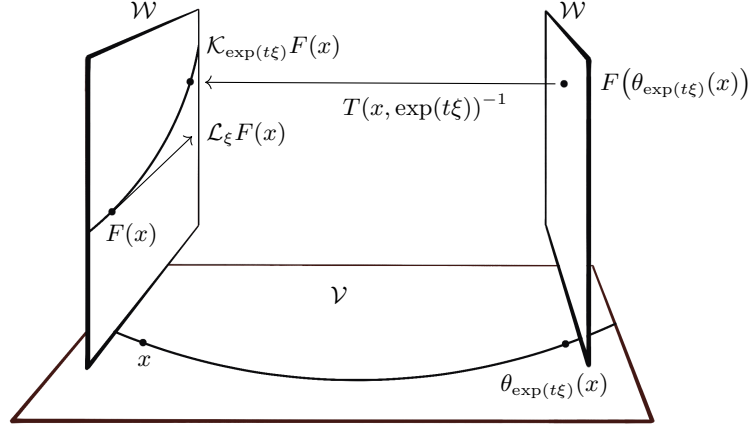
13

Figure 2: The fundamental operators for functions between vector spaces and linear Lie group actions defined by representations. The finite transformation operators $\mathcal{K}_g$ act on the function $F : \mathcal{V} \to \mathcal{W}$ by composing it with the transformation θ_g and then applying $T(x, g)^{-1}$ to the values in $\mathcal{W}$. The function is g-equivariant when this process does not alter the function. The Lie derivative $\mathcal{L}_\xi$ is formed by differentiating $t \mapsto \mathcal{K}_{\exp(t\xi)}$ at $t = 0$. Geometrically, $\mathcal{L}_\xi F(x)$ is the vector in $\mathcal{W}$ tangent to the curve $t \mapsto \mathcal{K}_{\exp(t\xi)} F(x)$ in $\mathcal{W}$ passing through $F(x)$ at $t = 0$.

For example, using representations $\Phi : G \to \mathrm{GL}(\mathcal{V})$, $\Psi : G \to \mathrm{GL}(\mathcal{W})$, the action can be defined by

$$\theta(x, g) = \Phi(g^{-1})x \qquad \text{and} \qquad T(x, g)y = \Psi(g^{-1})y. \tag{10}$$

With ϕ and ψ being the corresponding Lie algebra representations, the generator has components

$$\hat{\theta}(\xi)_x = -\phi(\xi)x \qquad \text{and} \qquad \hat{T}(\xi)_x y = -\psi(\xi)y. \tag{11}$$

The definition of equivariance, the symmetry group of a function, and the first family of fundamental operators are introduced by the following:

Definition 2 *We say that F is **equivariant** with respect to a group element $g \in G$ if*

$$\boxed{(\mathcal{K}_g F)(x) := T(x, g)^{-1} F(\theta_g(x)) = F(x)} \tag{12}$$

for every $x \in \mathcal{V}$. These elements form a subgroup of G denoted $\mathrm{Sym}_G(F)$.

Note that when the action is defined by representations (10), then (12) becomes

$$(\mathcal{K}_g F)(x) := \Psi(g) F(\Phi(g)^{-1}x) = F(x).$$

The transformation operators $\mathcal{K}_g$ are linear maps sending functions in $C^k(\mathcal{V}; \mathcal{W})$ to functions in $C^k(\mathcal{V}; \mathcal{W})$. These fundamental operators form a group with composition $\mathcal{K}_g \mathcal{K}_h = \mathcal{K}_{gh}$ and inversion $\mathcal{K}_g^{-1} = \mathcal{K}_{g^{-1}}$. Hence, $g \mapsto \mathcal{K}_g$ is an infinite-dimensional representation of G in $C^k(\mathcal{V}; \mathcal{W})$ for any k. These operators are useful for studying discrete symmetries of functions.

However, for a continuous group G it is impractical to work directly with the uncountable family $\{\mathcal{K}_g\}_{g \in G}$.

The second family of fundamental operators are the key objects we use to study continuous symmetries of functions. These are the Lie derivatives $\mathcal{L}_\xi : C^1(\mathcal{V}; \mathcal{W}) \to C^0(\mathcal{V}; \mathcal{W})$ defined along each $\xi \in \mathrm{Lie}(G)$ by

$$\boxed{(\mathcal{L}_\xi F)(x) = \left.\frac{\mathrm{d}}{\mathrm{d}\,t}\right|_{t=0} (\mathcal{K}_{\exp(t\xi)} F)(x) = \frac{\partial F(x)}{\partial x} \hat{\theta}(\xi)_x - \hat{T}(\xi)_x F(x).} \tag{13}$$

Note that when the action is defined using representations (10), then (13) becomes

$$(\mathcal{L}_\xi F)(x) = \psi(\xi) F(x) - \frac{\partial F(x)}{\partial x} \phi(\xi) x. \tag{14}$$

Evident from (13) is the fact that the Lie derivative is linear with respect to both ξ and F, and sends functions in C^{k+1} to functions in C^k for every $k \geq 0$. The geometric construction of the fundamental operators $\mathcal{K}_g$ and $\mathcal{L}_\xi$ are depicted in Figure 2. It turns out (see Proposition 24) that $\xi \mapsto \mathcal{L}_\xi$ is the Lie algebra representation corresponding to $g \mapsto \mathcal{K}_g$ on $C^\infty(\mathcal{V}; \mathcal{W})$, meaning that on this space we have the handy relations

$$\frac{\mathrm{d}}{\mathrm{d}\,t} \mathcal{K}_{\exp(t\xi)} = \mathcal{L}_\xi \mathcal{K}_{\exp(t\xi)} = \mathcal{K}_{\exp(t\xi)} \mathcal{L}_\xi \qquad \text{and} \qquad \mathcal{L}_{[\xi,\eta]} = \mathcal{L}_\xi \mathcal{L}_\eta - \mathcal{L}_\eta \mathcal{L}_\xi.$$

The results stated below are special cases of more general results developed later in Section 11.

Our first main result provides necessary and sufficient conditions for a continuously differentiable function $F : \mathcal{V} \to \mathcal{W}$ to be equivariant with respect to the Lie group actions on $\mathcal{V}$ and $\mathcal{W}$. This generalizes the constraints derived by Finzi et al. (2021) for the linear layers of equivariant multilayer perceptrons by allowing for nonlinear group actions and nonlinear functions.

Theorem 3 *Let $\{\xi_i\}_{i=1}^q$ generate (via linear combinations and Lie brackets) the Lie algebra $\mathrm{Lie}(G)$ and let $\{g_j\}_{j=1}^{n_G-1}$ contain one element from each non-identity component of G. Then $F \in C^1(\mathcal{V}; \mathcal{W})$ is G-equivariant if and only if*

$$\mathcal{L}_{\xi_i} F = 0 \qquad \text{and} \qquad \mathcal{K}_{g_j} F - F = 0 \tag{15}$$

for every $i = 1, \ldots, q$ and every $j = 1, \ldots, n_G - 1$. This is a special case of Theorem 26.

Since the fundamental operators $\mathcal{L}_\xi$ and $\mathcal{K}_g$ are linear, Theorem 3 provides linear constraints for a continuously differentiable function F to be G-equivariant.

Our second main result shows that the continuous symmetries of a given continuously differentiable function $F : \mathcal{V} \to \mathcal{W}$ are encoded by its Lie derivatives.

Theorem 4 *Given $F \in C^1(\mathcal{V}; \mathcal{W})$, the symmetry group $\mathrm{Sym}_G(F)$ is a closed, embedded Lie subgroup of G with Lie subalgebra*

$$\mathfrak{sym}_G(F) = \{\xi \in \mathrm{Lie}(G) \ : \ \mathcal{L}_\xi F = 0\}. \tag{16}$$

This is a special case of Theorem 25.

This result completely characterizes the identity component of the symmetry group $\mathrm{Sym}_G(F)$ because the connected Lie subgroups of G are in one-to-one correspondence with Lie subalgebras of $\mathrm{Lie}(G)$ (Theorem 19.26; Lee, 2013). Moreover, the Lie subalgebra of symmetries of a C^1 function F can be identified via linear algebra because $\mathfrak{sym}_g(F)$ is the nullspace of the linear operator $L_F : \mathrm{Lie}(G) \to C^0(\mathcal{V}; \mathcal{W})$ defined by

$$L_F : \xi \mapsto \mathcal{L}_\xi F. \tag{17}$$

Discretization methods suitable for linear-algebraic computations with the fundamental operators will be discussed in Section 9. The key point is that when the functions F lie in a finite-dimensional subspace $\mathcal{F} \subset C^1(\mathcal{V}; \mathcal{W})$, the ranges of the restricted Lie derivatives $\{\mathcal{L}_\xi|_{\mathcal{F}}\}_{\xi \in \mathrm{Lie}(G)}$, hence the ranges of $\{L_F\}_{F \in \mathcal{F}}$, are contained in a corresponding finite-dimensional subspace $\mathcal{F}' \subset C^0(\mathcal{V}; \mathcal{W})$ on which inner products can be defined using sampling or quadrature.

The preceding two theorems already show the duality between enforcing and discovering continuous symmetries with respect to the Lie derivative, viewed as a bilinear form $(\xi, F) \mapsto \mathcal{L}_\xi F$. To discover symmetries, we seek generators $\xi \in \mathrm{Lie}(G)$ satisfying $\mathcal{L}_\xi F = 0$ for a known function F. On the other hand, to enforce a connected group of symmetries, we seek functions F satisfying $\mathcal{L}_{\xi_i} F = 0$ with known generators $\xi_1, \ldots, \xi_q$ of $\mathrm{Lie}(G)$.

6. Enforcing Symmetry with Linear Constraints

Methods to enforce symmetry in neural networks and other machine learning models have been studied extensively, as we reviewed briefly in Section 3.1. A unifying theme in these techniques has been the use of linear constraints to enforce symmetry (Finzi et al., 2021; Loiseau and Brunton, 2018; Weiler et al., 2018; Cohen et al., 2019; Ahmadi and Khadir, 2020). The purpose of this section is to show how several of these methods can be understood in terms of the fundamental operators and linear constraints provided by Theorem 3.

6.1 Multilayer Perceptrons

Finzi et al. (2021) studied enforcing symmetry in multilayer perceptrons. They provide a practical method based on enforcing linear constraints on the weights defining each layer of a neural network. The network uses specialized nonlinearities that are automatically equivariant, meaning that the constraints need only be enforced on the linear component of each layer. We show that the constraints derived by Finzi et al. (2021) are the same as those given by Theorem 3.

Specifically, each linear layer $F^{(l)} : \mathcal{V}_{l-1} \to \mathcal{V}_l$, for $l = 1, \ldots, L$, is defined by

$$F^{(l)}(x) = W^{(l)}x + b^{(l)},$$

where $W^{(l)}$ are weight matrices and $b^{(l)}$ are bias vectors. Defining group representations $\Phi_l : G \to \mathrm{GL}(\mathcal{V}_l)$ for each layer, yields fundamental operators given by

$$\mathcal{K}_g F^{(l)}(x) = \Phi_l(g)W^{(l)}\Phi_{l-1}(g)^{-1}x + \Phi_l(g)b^{(l)}$$
$$\mathcal{L}_\xi F^{(l)}(x) = (\phi_l(\xi)W^{(l)} - W^{(l)}\phi_{l-1}(\xi))x + \phi_l(\xi)b^{(l)}.$$

Let $\{\xi_i\}_{i=1}^q$ generate $\mathrm{Lie}(G)$ and let $\{g_j\}_{j=1}^{n_G-1}$ consist of an element from each non-identity component of G. Using the fundamental operators and Theorem 3, it follows that the layer $F^{(l)}$ is G-equivariant if and only if the weights and biases satisfy

$$\phi_l(\xi_i)W^{(l)} = W^{(l)}\phi_{l-1}(\xi_i), \quad \text{and} \quad \Phi_l(g_j)W^{(l)} = W^{(l)}\Phi_{l-1}(g_j),$$

$$\phi_l(\xi_i)b^{(l)} = 0, \quad \text{and} \quad \Phi_l(g_j)b^{(l)} = b^{(l)}$$

for every $i = 1, \ldots, q$ and $j = 1, \ldots, n_G - 1$. These are the same as those derived in Finzi et al. (2021). The equivariant linear layers are then combined with specialized equivariant nonlinearities $\sigma^{(l)} : \mathcal{V}_l \to \mathcal{V}_l$ to produce an equivariant network

$$F = \sigma^{(L)} \circ F^{(L)} \circ \cdots \circ \sigma^{(1)} \circ F^{(1)} : \mathcal{V}_0 \to \mathcal{V}_L.$$

The composition of equivariant functions is equivariant, as one can easily check using Definition 2.

6.2 Neural Operators Acting on Fields

Enforcing symmetry in neural networks acting on spatial fields has been studied extensively by Weiler et al. (2018); Cohen et al. (2018); Esteves et al. (2018); Kondor and Trivedi (2018); Cohen et al. (2019) among others. Many of these techniques use integral operators to define equivariant linear layers, which are coupled with equivariant nonlinearities, such as the gated nonlinearities proposed by Weiler et al. (2018). Networks built by composing integral operators with nonlinearities constitute a large class of *neural operators* described by Kovachki et al. (2023); Goswami et al. (2023); Boullé and Townsend (2023). The key task is to identify appropriate bases for equivariant kernels. For certain groups, such as the Special Euclidean group $G = \mathrm{SE}(3)$, bases can be constructed explicitly using spherical harmonics, as in Weiler et al. (2018). We show that equivariance with respect to arbitrary group actions can be enforced via linear constraints on the integral kernels derived using the fundamental operators introduced in Section 5. Appropriate bases of kernel functions can then be constructed numerically by computing an appropriate nullspace, as is done by Finzi et al. (2021) for multilayer perceptrons.

For the sake of simplicity we consider integral operators acting on vector-valued functions $F : \mathbb{R}^m \to \mathcal{V}$, where $\mathcal{V}$ is a finite-dimensional vector space. Later on in Section 11.4 we study higher-order integral operators acting on sections of vector bundles. If $\mathcal{W}$ is another finite-dimensional vector space, an integral operator acting on F to produce a new function $\mathbb{R}^n \to \mathcal{W}$ is defined by

$$\mathcal{T}_K F(x) = \int_{\mathbb{R}^m} K(x, y)F(y)\,\mathrm{d}\,y, \tag{18}$$

where the *kernel* function K provides a linear map $K(x, y) : \mathcal{V} \to \mathcal{W}$ at each $(x, y) \in \mathbb{R}^n \times \mathbb{R}^m$. In other words, the kernel is a function on $\mathbb{R}^n \times \mathbb{R}^m$ taking values in the tensor product space $\mathcal{W} \otimes \mathcal{V}^*$, where $\mathcal{V}^*$ denotes the algebraic dual of $\mathcal{V}$. Many of the neural operator architectures described by Kovachki et al. (2023); Goswami et al. (2023); Boullé and Townsend (2023) are constructed by composing layers defined by integral operators (18) with nonlinear activation functions, usually acting pointwise. The kernel functions K are optimized during training of the neural operator.

17

With group actions defined by representations on $\mathbb{R}^m, \mathbb{R}^n, \mathcal{V}, \mathcal{W}$, functions $F : \mathbb{R}^m \to \mathcal{V}$ transform according to:

$$\mathcal{K}_g^{(\mathbb{R}^m, \mathcal{V})} F(x) = \Phi_{\mathcal{V}}(g) F(\Phi_{\mathbb{R}^m}(g)^{-1} x)$$

for $g \in G$. Likewise, functions $\mathbb{R}^n \to \mathcal{W}$ transform via an analogous operator $\mathcal{K}_g^{(\mathbb{R}^n, \mathcal{W})}$.

Definition 5 *The integral operator $\mathcal{T}_K$ in (18) is **equivariant** with respect to $g \in G$ when*

$$\mathcal{K}_g^{(\mathbb{R}^n, \mathcal{W})} \circ \mathcal{T}_K \circ \mathcal{K}_{g^{-1}}^{(\mathbb{R}^m, \mathcal{V})} = \mathcal{T}_K.$$

The elements g satisfying this equation form a subgroup of G denoted $\mathrm{Sym}_G(\mathcal{T}_K)$.

By changing variables in the integral, the operator on the left is given by

$$\mathcal{K}_g^{(\mathbb{R}^n, \mathcal{W})} \circ \mathcal{T}_K \circ \mathcal{K}_{g^{-1}}^{(\mathbb{R}^m, \mathcal{V})} F(x) = \int_{\mathbb{R}^m} \mathcal{K}_g K(x, y) F(y) \, \mathrm{d}\, y,$$

where

$$\mathcal{K}_g K(x, y) = \Phi_{\mathcal{W}}(g) K\big(\Phi_{\mathbb{R}^n}(g)^{-1} x, \Phi_{\mathbb{R}^m}(g)^{-1} y\big) \Phi_{\mathcal{V}}(g)^{-1} \det\big[\Phi_{\mathbb{R}^m}(g)^{-1}\big]. \qquad (19)$$

The following result provides equivariance conditions in terms of the kernel, generalizing Lemma 1 in Weiler et al. (2018) beyond a specific action of SE(3).

Proposition 6 *Let K be continuous and suppose that $\mathcal{T}_K$ acts on a function space containing all smooth, compactly supported fields. Then*

$$\mathrm{Sym}_G(\mathcal{T}_K) = \{g \in G \; : \; \mathcal{K}_g K = K\}. \qquad (20)$$

We give a proof in Appendix A

The Lie derivative of a continuously differentiable kernel function is given by

$$\begin{aligned} \mathcal{L}_\xi K(x, y) = \phi_{\mathcal{W}}(\xi) K(x, y) - K(x, y) \phi_{\mathcal{V}}(\xi) - K(x, y) \mathrm{Tr}[\phi_{\mathbb{R}^m}(\xi)] \\ - \frac{\partial K(x, y)}{\partial x} \phi_{\mathbb{R}^n}(\xi) x - \frac{\partial K(x, y)}{\partial y} \phi_{\mathbb{R}^m}(\xi) y \end{aligned} \qquad (21)$$

The operators $\mathcal{K}_g$ and $\mathcal{L}_\xi$ are the fundamental operators from Section 5 because the transformation law for the kernel can be written as

$$\mathcal{K}_g K = \Phi_{\mathcal{W} \otimes \mathcal{V}^*}(g) K \circ \Phi_{\mathbb{R}^m \times \mathbb{R}^m}(g)^{-1},$$

where

$$\Phi_{\mathbb{R}^n \times \mathbb{R}^m}(g) : (x, y) \mapsto \left(\Phi_{\mathbb{R}^n}(g) x, \Phi_{\mathbb{R}^m}(g) y\right),$$
$$\Phi_{\mathcal{W} \otimes \mathcal{V}^*}(g) : A \mapsto \Phi_{\mathcal{W}}(g) A \Phi_{\mathcal{V}}(g)^{-1} \det\big[\Phi_{\mathbb{R}^m}(g)^{-1}\big]$$

are representations of G in $\mathbb{R}^m \times \mathbb{R}^m$ and $\mathcal{W} \otimes \mathcal{V}^*$.

As an immediate consequence of Theorem 3, we have the following corollary establishing linear constraints for the kernel to produce an equivariant integral operator.

Corollary 7 *Let $\{\xi_i\}_{i=1}^{q}$ generate the Lie algebra $\mathrm{Lie}(G)$ and let $\{g_j\}_{j=1}^{n_G-1}$ contain one element from each non-identity component of G. Under the same hypotheses as Proposition 6 and assuming K is continuously differentiable, the integral operator $\mathcal{T}_K$ in (18) is G-equivariant in the sense of Definition 5 if and only if*

$$\mathcal{L}_{\xi_i} K = 0 \qquad and \qquad \mathcal{K}_{g_j} K - K = 0$$

for every $i = 1, \ldots, q$ and every $j = 1, \ldots, n_G - 1$.

These linear constraint equations must be satisfied to enforced equivariance with respect to a known symmetry G in the machine learning process. By discretizing the operators $\mathcal{K}_g$ and $\mathcal{L}_\xi$, as discussed later in Section 9, one can solve these constraints numerically to construct a basis of kernel functions for equivariant integral operators.

It follows immediately from Theorem 4 that the Lie derivative of the kernel encodes the continuous symmetries of a given integral operator.

Corollary 8 *Under the same hypotheses as Proposition 6 and assuming K is continuously differentiable, it follows that $\mathrm{Sym}_G(\mathcal{T}_K)$ is a closed, embedded Lie subgroup of G with Lie subalgebra*

$$\mathfrak{sym}_G(\mathcal{T}_K) = \{\xi \in \mathrm{Lie}(G) \ : \ \mathcal{L}_\xi K = 0\}.$$

This result will be useful for methods that promote symmetry of the integral operator, as we describe later in Section 8.

7. Discovering Symmetry by Computing Nullspaces

In this section, we show that in a wide range of settings the continuous symmetries of a manifold, point cloud, or map can be recovered by computing the nullspace of a linear operator. For functions, this is already covered by Theorem 4, which allows us to compute the connected subgroup of symmetries by identifying its Lie subalgebra

$$\mathfrak{sym}_G(F) = \mathrm{Null}(L_F)$$

where $L_F : \xi \mapsto \mathcal{L}_\xi$ is the linear operator defined by (17). Hence, if a machine learning model F has a symmetry group $\mathrm{Sym}_G(F)$, then its Lie algebra is equal to the nullspace of L_F.

This section explains how this is actually a special case of a more general result allowing us to reveal the symmetries of submanifolds via the nullspace of a closely related operator. We begin with the more general case where we study the symmetries of a submanifold of Euclidean space, and we explain how to recover symmetries from point clouds approximating submanifolds. The Lie derivative described in Section 5 is then recovered when the submanifold is the graph of a function. We also briefly describe how the fundamental operators from Section 5 can be used to recover symmetries and conservation laws of dynamical systems.

7.1 Symmetries of Submanifolds

We begin by studying the symmetries of smooth submanifolds $\mathcal{M}$ of Euclidean space $\mathbb{R}^d$ using an approach similar to Cahill et al. (2023). However, we use a different operator that generalizes more naturally to nonlinear group actions on arbitrary manifolds (see Section 12)
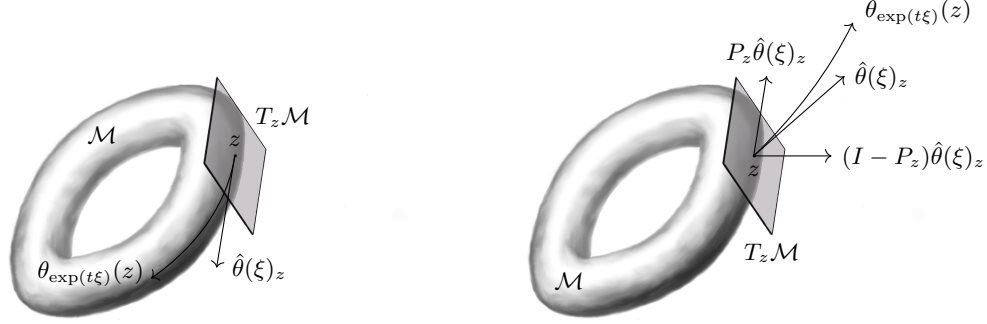
Figure 3: Tangency of infinitesimal generators reveals symmetries of submanifolds. The infinitesimal generator $\hat{\theta}(\xi)$ is everywhere tangent to the submanifold $\mathcal{M}$ if and only if the curves $t \mapsto \theta_{\exp(t\xi)}(z)$, with $z \in \mathcal{M}$, lie in $\mathcal{M}$ for all t. The Lie algebra elements ξ satisfying this tangency condition form the Lie subalgebra of symmetries of $\mathcal{M}$. To test for tangency of the infinitesimal generator we use a family of projections P_z onto the tangent spaces $T_z\mathcal{M}$, for every $z \in \mathcal{M}$. Here, $(I - P_z)\hat{\theta}(\xi)_z$ extracts the component of the infinitesimal generator not lying tangent to $\mathcal{M}$. Hence, ξ generates a symmetry of $\mathcal{M}$ if and only if $(I - P_z)\hat{\theta}(\xi)_z = 0$ for all $z \in \mathcal{M}$.

and recovers the Lie derivative (see Section 7.2). With a right action $\theta : \mathbb{R}^d \times G \to \mathbb{R}^d$ of a Lie group, we define invariance of a submanifold as follows:

Definition 9 *A submanifold* $\mathcal{M} \subset \mathbb{R}^d$ *is **invariant** with respect to an element* $g \in G$ *if*

$$\theta_g(z) \in \mathcal{M}$$

for every $z \in \mathcal{M}$. *These elements form a subgroup of* G *denoted* $\mathrm{Sym}_G(\mathcal{M})$.

The subgroup of symmetries of a submanifold is characterized by the following theorem.

Theorem 10 *Let* $\mathcal{M}$ *be a smooth, closed, embedded submanifold of* $\mathbb{R}^d$. *Then* $\mathrm{Sym}_G(\mathcal{M})$ *is a closed, embedded Lie subgroup of* G *whose Lie subalgebra is*

$$\mathfrak{sym}_G(\mathcal{M}) = \{\xi \in \mathrm{Lie}(G) \ : \ \hat{\theta}(\xi)_z \in T_z\mathcal{M} \quad \forall z \in \mathcal{M}\}.$$

This is a special case of Theorem 35.

The meaning of this result and its practical use for detecting symmetry are illustrated in Figure 3.

To reveal the connected component of $\mathrm{Sym}_G(\mathcal{M})$, we let $P_z : \mathbb{R}^d \to \mathbb{R}^d$ be a family of linear projections onto $T_z\mathcal{M} \subset \mathbb{R}^d$. These are assumed to vary continuously with respect to $z \in \mathcal{M}$. Then, under the assumptions of the theorem above, $\mathfrak{sym}_G(\mathcal{M})$ is the nullspace of

the symmetric, positive-semidefinite operator $S_{\mathcal{M}} : \mathrm{Lie}(G) \to \mathrm{Lie}(G)$ defined by

$$\langle \eta, \ S_{\mathcal{M}} \xi \rangle_{\mathrm{Lie}(G)} = \int_{\mathcal{M}} \hat{\theta}(\eta)_z^T (I - P_z)^T (I - P_z) \hat{\theta}(\xi)_z \ \mathrm{d}\,\mu(z) \tag{22}$$

for every $\eta, \xi \in \mathrm{Lie}(G)$. The choice of inner product on $\mathrm{Lie}(G)$ fixes the adjoint used later. We see in Figure 3 that $(I - P_z)\hat{\theta}(\xi)_z$ measures the component of the infinitesimal generator not tangent to the submanifold at z. Here, μ is any strictly positive measure on $\mathcal{M}$ that makes all of these integrals finite. The above formula is useful for computing the matrix of $S_{\mathcal{M}}$ in an orthonormal basis for $\mathrm{Lie}(G)$.

Alternatively, when the dimension of G is large, one can compute the nullspace using a Krylov algorithm such as the one described in Finzi et al. (2021). Such algorithms rely solely on queries of $S_{\mathcal{M}}$ acting on vectors $\xi \in \mathrm{Lie}(G)$. When $\theta_g(z) = \Phi(g^{-1})z$ and $\hat{\theta}(\xi)_z = -\phi(\xi)z$ are given by a Lie group representation (see Section 4.2), then the operator defined in (22) is given explicitly by

$$S_{\mathcal{M}} \xi = \int_{\mathcal{M}} \mathrm{d}\,\Phi(e)^* \left[(I - P_z)^T (I - P_z) \phi(\xi) z z^T \right] \ \mathrm{d}\,\mu(z),$$

where $\mathrm{d}\,\Phi(e)^* : \mathbb{R}^{d \times d} \to \mathrm{Lie}(G)$ is the adjoint of $\mathrm{d}\,\Phi(e) : \mathrm{Lie}(G) \to \mathbb{R}^{d \times d}$. If $G \subset \mathbb{R}^{d \times d}$ is a matrix Lie group and Φ is the identity representation, then $\mathrm{d}\,\Phi(e)$ is the injection $\mathrm{Lie}(G) \hookrightarrow \mathbb{R}^{d \times d}$. When $\mathrm{Lie}(G) \subset \mathbb{R}^{d \times d}$ inherits its inner product from $\mathbb{R}^{d \times d}$, $\mathrm{d}\,\Phi(e)^*$ is the orthogonal projection of $\mathbb{R}^{d \times d}$ onto $\mathrm{Lie}(G)$. For example, if Φ is the identity representation of $\mathrm{SE}(d)$ in $\mathbb{R}^{d+1}$ with the inner product on $\mathfrak{se}(d) \subset \mathbb{R}^{(d+1) \times (d+1)}$ given by the usual inner product of matrices $\langle M_1, M_2 \rangle = \mathrm{Tr}(M_1^T M_2)$, then it can be readily verified that

$$\mathrm{d}\,\Phi(e)^* \left(\begin{bmatrix} A & b \\ c^T & a \end{bmatrix} \right) = \begin{bmatrix} \frac{1}{2}(A - A^T) & b \\ 0 & 0 \end{bmatrix}, \qquad A \in \mathbb{R}^{d \times d}, \ b \in \mathbb{R}^d, \ c \in \mathbb{R}^d, \ a \in \mathbb{R}.$$

In practice, one can use sample points z_i on the manifold to obtain a Monte-Carlo estimate of $S_{\mathcal{M}}$ with approximate projections P_{z_i} computed using local principal component analysis (PCA), as described in Cahill et al. (2023). More accurate estimates of the tangent spaces can be obtained using the methods in Berry and Giannakis (2020). Assuming the P_{z_i} are accurate, the following proposition shows that the correct Lie subalgebra of symmetries is revealed using finitely many sample points z_i. However, this result does not tell us how many samples to use, or even when to stop sampling.

Proposition 11 *Let μ be a strictly positive probability measure on a smooth manifold $\mathcal{M}$ such that $\langle \xi, S_{\mathcal{M}} \xi \rangle < \infty$ for every $\xi \in \mathrm{Lie}(G)$. Let z_i be drawn independently from the distribution μ and let $S_m : \mathrm{Lie}(G) \to \mathrm{Lie}(G)$ be defined by*

$$\langle \eta, \ S_m \xi \rangle_{\mathrm{Lie}(G)} = \frac{1}{m} \sum_{i=1}^{m} \hat{\theta}(\eta)_{z_i}^T (I - P_{z_i})^T (I - P_{z_i}) \hat{\theta}(\xi)_{z_i}. \tag{23}$$

Then there is almost surely an integer M_0 such that for every $m \geq M_0$ we have $\mathrm{Null}(S_m) = \mathfrak{sym}_G(\mathcal{M})$. We provide a proof in Appendix B.

7.2 Symmetries of Functions as Symmetries of Submanifolds

The method described for studying symmetries of submanifolds can be applied to smooth maps between finite-dimensional vector spaces by identifying the map $F : \mathbb{R}^m \to \mathbb{R}^n$ with its graph

$$\mathrm{gr}(F) = \{(x, F(x)) \in \mathbb{R}^m \times \mathbb{R}^n \ : \ x \in \mathbb{R}^m\}.$$

The graph is a smooth, closed, embedded submanifold of the space $\mathbb{R}^m \times \mathbb{R}^n$ by Proposition 5.7 in Lee (2013). We show that this approach recovers the Lie derivative and the result in Theorem 4.

We consider the action of G on $\mathbb{R}^m \times \mathbb{R}^n$ defined by (9) with $\mathcal{V} = \mathbb{R}^m$ and $\mathcal{W} = \mathbb{R}^n$. Defining the smoothly-varying family of projections

$$P_{(x,F(x))} = \begin{bmatrix} I & 0 \\ \mathrm{d}\,F(x) & 0 \end{bmatrix} \tag{24}$$

onto $T_{(x,F(x))}\,\mathrm{gr}(F)$, it is easy to check that

$$\begin{bmatrix} 0 \\ \mathcal{L}_\xi F(x) \end{bmatrix} = - \underbrace{\left(\begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} - \begin{bmatrix} I & 0 \\ \mathrm{d}\,F(x) & 0 \end{bmatrix} \right)}_{I - P_{(x,F(x))}} \underbrace{\begin{bmatrix} \hat{\theta}(\xi)_x \\ \hat{T}(\xi)_x F(x) \end{bmatrix}}_{\hat{\Theta}(\xi)_{(x,F(x))}}. \tag{25}$$

We note that this is a special case of Theorem 39 describing the Lie derivative in terms of a projection onto the tangent space of a function's graph. The resulting operator $S_{\mathrm{gr}(F)}$ in (22) is given by

$$\left\langle \eta, S_{\mathrm{gr}(F)}\xi \right\rangle_{\mathrm{Lie}(G)} = \int_{\mathbb{R}^m} (\mathcal{L}_\eta F(x))^T \mathcal{L}_\xi F(x) \ \mathrm{d}\,\mu(x),$$

for $\eta, \xi \in \mathrm{Lie}(G)$ and an appropriate positive measure μ on $\mathbb{R}^m$ that makes the integrals finite. Therefore, Theorem 4 is recovered from the result on symmetries of submanifolds in Theorem 10.

Related quantities have been used to study the symmetries of trained neural networks, with F being the network and its derivatives computed via back-propagation. The quantity $\left\langle \xi, S_{\mathrm{gr}(F)}\xi \right\rangle_{\mathrm{Lie}(G)} = \|\mathcal{L}_\xi F\|_{L^2(\mu)}$ was used by Gruver et al. (2022) to construct the Local Equivariant Error or (LEE), measuring the extent to which a trained neural network F fails to respect symmetries in the one-parameter group $\{\exp(t\xi)\}_{t \in \mathbb{R}}$. The nullspace of $\xi \mapsto \mathcal{L}_\xi F$ in the special case where $T(x, g) = I$ acts trivially was used by Moskalev et al. (2022) to identify the connected subgroup with respect to which a given network is invariant.

By viewing a function as a submanifold, we propose a simple, data-driven technique for estimating the Lie derivative and subgroup of symmetries of the function. To approximate $\mathcal{L}_\xi F$, $S_{\mathrm{gr}(F)}$, and $\mathfrak{sym}_G(F)$ using input–output pairs $(x_i, y_i = F(x_i))$, it suffices to approximate the projection in (24) using these data. To do this, we could obtain $(n+m) \times m$ matrices U_i with columns spanning $T_{(x_i,y_i)}\,\mathrm{gr}(F)$ by applying local PCA to the data $z_i = (x_i, y_i)$, or by pruning the frames computed in Berry and Giannakis (2020). With $E = \begin{bmatrix} I_{m \times m} & 0_{m \times n} \end{bmatrix}$,

the projection in (24) is given by

$$P_{z_i} = U_i(EU_i)^{-1}E$$

because any projection is uniquely determined by its range and nullspace (see Section 5.9 of Meyer, 2000). Evaluating the generator $\hat{\Theta}(\xi)_{(x_i, y_i)}$ and applying the projection as in (25) gives a potentially simple way to approximate $(\mathcal{L}_\xi F)(z_i)$, $S_{\text{gr}(F)}$, and $\mathfrak{sym}_G(F)$ using the input–output pairs. However, many such pairs would be needed since the tangent space to the graph of F at x_i is well-approximated by local PCA only when there are at least m neighboring samples sufficiently close to x_i. Even more samples are likely needed when they are noisy. The convergence properties of the spectral methods in Berry and Giannakis (2020) are better, but they still require enough samples to estimate integrals over $\text{gr}(F)$ using quadrature or Monte Carlo.

7.3 Symmetries and Conservation Laws of Dynamical Systems

Here, we consider the case when $F : \mathbb{R}^n \to \mathbb{R}^n$ is a smooth function defining a dynamical system

$$\frac{\mathrm{d}}{\mathrm{d}t} x(t) = F(x(t)) \tag{26}$$

with state variables $x(t) \in \mathbb{R}^n$. The solution of this equation is described by the flow map $\text{Fl} : (t, x(\tau)) \mapsto x(\tau + t)$, which is defined on a maximal connected open set D containing $0 \times \mathbb{R}^n$. In many cases we write $\text{Fl}^t(\cdot) = \text{Fl}(t, \cdot)$. Given a right action $\theta : \mathbb{R}^n \times G \to \mathbb{R}^n$, equivariance for the dynamical system is defined as follows:

Definition 12 *The dynamical system in* (26) *is* ***equivariant*** *with respect to a group element* $g \in G$ *if the flow map satisfies*

$$\mathcal{K}_g \text{Fl}^t(x) := \theta_{g^{-1}}\left(\text{Fl}^t\left(\theta_g(x)\right)\right) = \text{Fl}^t(x)$$

for every $(t, x) \in D$.

Differentiating at $t = 0$ shows that equivariance of the dynamical system implies that F is equivariant in the sense of Definition 2 with $T(x, g) = \mathrm{d}\,\theta_g(x)$. The converse is also true thanks to Corollary 9.14 in Lee (2013), meaning that equivariance for the dynamical system is equivalent to equivariance of F. Therefore, we can study equivariance of the dynamical system in (26) by directly applying the tools developed in Section 5 to the function F. Thanks to Theorem 4, identifying the connected subgroup of symmetries for the dynamical system is a simple matter of computing the nullspace of the linear map $\xi \mapsto \mathcal{L}_\xi F$, that is

$$\mathfrak{sym}_G(F) = \{\xi \in \text{Lie}(G) \; : \; \mathcal{L}_\xi F = 0\}. \tag{16}$$

In this case $\hat{T}(\xi)_x = \frac{\partial}{\partial x}\hat{\theta}(\xi)_x$, meaning the Lie derivative is given by

$$\mathcal{L}_\xi F(x) = \frac{\partial F(x)}{\partial x}\hat{\theta}(\xi)_x - \frac{\partial \hat{\theta}(\xi)_x}{\partial x}F(x) = [\hat{\theta}(\xi), F]_x,$$

where $[\hat{\theta}(\xi), F]_x$ is the Lie bracket of the vector fields $\hat{\theta}(\xi)$ and F evaluated at x. Symmetries can also be enforced as linear constraints on F described by Theorem 3. This was done by Ahmadi and Khadir (2020) for polynomial dynamical systems with discrete symmetries. Later on in Section 11.1 we show that analogous results apply to dynamical systems on manifolds.

A conserved quantity for the system in (26) is defined as follows:

Definition 13 *A scalar valued quantity $f : \mathbb{R}^n \to \mathbb{R}$ is said to be **conserved** when*

$$\mathcal{K}_t f(x) := f(\mathrm{Fl}^t(x)) = f(x) \qquad \forall (t, x) \in D.$$

In this setting, the composition operators $\mathcal{K}_t$ are often referred to as Koopman operators (see Koopman, 1931; Mezić, 2005; Mauroy et al., 2020; Otto and Rowley, 2021; Brunton et al., 2022). It is easy to see that a smooth function f is conserved if and only if

$$\mathcal{L}_F f := \left. \frac{\mathrm{d}}{\mathrm{d}\,t} \right|_{t=0} \mathcal{K}_t f = \frac{\partial f}{\partial x} F = 0.$$

This relation is used by Kaiser et al. (2018, 2021) to identify conserved quantities by computing the nullspace of $\mathcal{L}_F$ restricted to finite-dimensional spaces of candidate functions. When the flow is defined for all $t \in \mathbb{R}$, the operators $\mathcal{K}_t$ and $\mathcal{L}_F$ are the fundamental operators from Section 5 for the right action of the Lie group $G = (\mathbb{R}, +)$ defined by $\theta(x, t) = \mathrm{Fl}^t(x)$ and $T(x, t) = I$.

Remark 14 *For Hamiltonian dynamical systems Noether's theorem establishes an equivalence between the symmetries of the Hamiltonian and conserved quantities of the system. We study Hamiltonian systems later in Section 11.3.*

8. Promoting Symmetry with Convex Penalties

In this section, we show how to design custom convex regularization functions to promote symmetries within a given candidate group during training of a machine learning model. This allows us to train a model with as many symmetries as possible from among the candidates, while breaking candidate symmetries only when the data provide sufficient evidence. We study both discrete and continuous groups of candidate symmetries. We quantify symmetry breaking within a candidate group using the fundamental operators described in Section 5. For discrete groups we use the transformation operators $\{\mathcal{K}_g\}_{g \in G}$ and for continuous groups we use the Lie derivatives $\{\mathcal{L}_\xi\}_{\xi \in \mathrm{Lie}(G)}$. In the continuous case we penalize a convex relaxation of the codimension of the subgroup of symmetries given by a nuclear norm (Schatten 1-norm) of the operator $\xi \mapsto \mathcal{L}_\xi F$ defined by (17); minimizing this codimension via the proxy nuclear norm will promote the largest nullspace possible, and hence the largest admissible symmetry group. Once these regularization functions are developed abstractly in Sections 8.1 and 8.2, we explain how to apply the approach to basis-function regression (Section 8.3) and neural networks (Section 8.4). We evaluate these methods on numerical examples in Section 10.

As in Section 5, the basic building blocks of the machine learning models we consider are continuously differentiable (C^1) functions $F : \mathcal{V} \to \mathcal{W}$ between finite-dimensional vector

spaces. While we consider this restricted setting here, our results readily generalize to sections of vector bundles, as we describe later in Section 11. These functions could be layers of a multilayer perceptron, integral kernels to be applied to spatio-temporal fields, or simply linear combinations of user-specified basis functions in a regression task. We consider parametric models where F is constrained to lie in a given finite-dimensional subspace $\mathcal{F} \subset C^1(\mathcal{V}; \mathcal{W})$ of continuously differentiable functions. Working within a finite-dimensional subspace of functions will allow us to discretize the fundamental operators in Section 9.

We consider the same setting as Section 5, i.e., candidate symmetries are described by a Lie group G acting via (9) on the domain and codomain of functions $F \in \mathcal{F}$. Equivariance in this setting is described by Definition 2. When fitting the function F to data, our regularization functions penalize the size of $G \setminus \mathrm{Sym}_G(F)$. For reasons that will become clear, we use different penalties corresponding to different notions of "size" when G is a discrete group versus when G is continuous. The main result describing the continuous symmetries of F is Theorem 4.

8.1 Discrete Symmetries

When the group G has finitely many elements, one can measure the size of $G \setminus \mathrm{Sym}_G(F)$ simply by counting its elements:

$$R_{G,0}(F) = |G \setminus \mathrm{Sym}_G(F)|.$$

However, this penalty is impractical for optimization owing to its discrete values and nonconvexity. Letting $\| \cdot \|$ be any norm on the space $\mathcal{F}'' = \mathrm{span}\{\mathcal{K}_g F \ : \ g \in G, \ F \in \mathcal{F}\}$ yields a convex relaxation of the above penalty given by

$$R_{G,1}(F) = \sum_{g \in G} \|\mathcal{K}_g F - F\|. \tag{27}$$

This is a convex function on $\mathcal{F}$ because $\mathcal{K}_g$ is a linear operator and vector space norms are convex. For example, if $\boldsymbol{c} = (c_1, \ldots, c_N)$ are the coefficients of F in a basis for $\mathcal{F}''$ and $\boldsymbol{K}_g$ is the matrix of $\mathcal{K}_g$ in this basis, then the Euclidean norm can be used to define

$$R_{G,1}(F) = \sum_{g \in G} \|\boldsymbol{K}_g \boldsymbol{c} - \boldsymbol{c}\|_2.$$

This is directly analogous to the group sparsity penalty proposed in Yuan and Lin (2006).

8.2 Continuous Symmetries

We now consider the case where G is a Lie group with positive dimension and Lie algebra $\mathfrak{g}$. Here we use the dimension of $\mathrm{Sym}_G(F)$ to measure the symmetry of F, seeking to penalize the complementary dimension or *codimension*, given by

$$R_{\mathfrak{g},0}(F) = \mathrm{codim}(\mathrm{Sym}_G(F)) = \dim(G) - \dim(\mathrm{Sym}_G(F)).$$

We take this approach in the continuous case because it is no longer possible to simply count the number of broken symmetries. While it is possible in principle to replace the sum in (27) by an integral of $\|\mathcal{K}_g F - F\|$ over $g \in G$, the numerical quadrature required to approximate it becomes prohibitive for higher-dimensional candidate groups. This difficulty is exacerbated by the fact that the integrand is not smooth. The space $\mathcal{F}''$ can also become infinite-dimensional when G has positive dimension, making it challenging to compute the norm $\| \cdot \|$.

In contrast, it is much easier to measure the "size" of a continuous symmetry group using its dimension because this can be computed with linear algebra. Specifically, the dimension of $\mathrm{Sym}_G(F)$ is equal to that of its Lie algebra. Thanks to Theorem 4, this is the nullspace of a linear operator $L_F : \mathrm{Lie}(G) \to C^0(\mathcal{V}; \mathcal{W})$ defined by

$$L_F : \xi \mapsto \mathcal{L}_\xi F, \tag{17}$$

where $\mathcal{L}_\xi$ is the Lie derivative in (13). By the rank and nullity theorem, the codimension of $\mathrm{Sym}_G(F)$ is equal to the rank of this operator:

$$R_{\mathfrak{g},0}(F) = \mathrm{codim}(\mathfrak{sym}_G(F)) = \mathrm{rank}(L_F).$$

Penalizing the rank of an operator is impractical for optimization owing to its discrete values and nonconvexity. A commonly used convex relaxation of the rank is provided by the Schatten 1-norm, also known as the *nuclear norm*, given by

$$R_{\mathfrak{g},*}(F) = \|L_F\|_* = \sum_{i=1}^{\dim(G)} \sigma_i(L_F). \tag{28}$$

Here $\sigma_i(L_F)$ denotes the ith singular value of L_F with respect to inner products on $\mathrm{Lie}(G)$ and $\mathcal{F}' = \mathrm{span}\{\mathcal{L}_\xi F : \xi \in \mathrm{Lie}(G), F \in \mathcal{F}\}$. This space is finite-dimensional, being spanned by $\{\mathcal{L}_{\xi_i} F_j\}_{i,j}$ where ξ_i and F_j are basis elements for $\mathrm{Lie}(G)$ and $\mathcal{F}$. This enables computations with discrete inner products on $\mathcal{F}'$, as we describe in Section 9. For certain rank minimization problems, penalizing the nuclear norm is guaranteed to recover the true minimum rank solution (Candès and Recht, 2009; Recht et al., 2010; Gross, 2011). Our numerical results in Section 10.1 suggest that nuclear norm minimization can also recover symmetric functions using (28).

The proposed regularization function (28) is convex on $\mathcal{F}$ because $F \mapsto L_F$ is linear and the nuclear norm is convex. For example, if $(c_1, \dots, c_N)$ are the coefficients of F in a basis $\{F_1, \dots, F_N\}$ for $\mathcal{F}$ and $\boldsymbol{L}_{F_i}$ are the matrices of L_{F_i} in orthonormal bases for $\mathrm{Lie}(G)$ and $\mathcal{F}'$, then

$$R_{\mathfrak{g},*}(F) = \|c_1 \boldsymbol{L}_{F_1} + \cdots + c_N \boldsymbol{L}_{F_N}\|_*.$$

With $\{\xi_1, \dots, \xi_{\dim(G)}\}$ and $\{u_1, \dots, u_{N'}\}$ being the orthonormal bases for $\mathrm{Lie}(G)$ and $\mathcal{F}'$, one can compute and store the rank-3 tensor $[\boldsymbol{L}_{F_i}]_{j,k} = \langle u_j, \mathcal{L}_{\xi_k} F_i \rangle_{\mathcal{F}'}$. Practical methods for constructing and computing with inner products on $\mathcal{F}'$ will be discussed in Section 9.

8.3 Promoting Symmetry in Basis-Function Regression

To demonstrate how the symmetry-promoting regularization functions proposed above can be used in practice, consider a regression problem for a function $F : \mathbb{R}^m \to \mathbb{R}^n$. It is common to parameterize this problem by expressing $F(x) = W\mathcal{D}(x)$ in a dictionary $\mathcal{D} : \mathbb{R}^m \to \mathbb{R}^N$ consisting of user-defined smooth functions with a matrix of weights $W \in \mathbb{R}^{n \times N}$ to be fit during the training process. For example, the sparse identification of nonlinear dynamics (SINDy) algorithm (Brunton et al., 2016) belongs to this type of learning, among other machine learning algorithms (Brunton and Kutz, 2022). The fundamental operators (Section 5) for this class of functions are given by

$$\mathcal{K}_g F(x) = T(x, g)^{-1} W\mathcal{D}(\theta_g(x)),$$

$$\mathcal{L}_\xi F(x) = W\frac{\partial \mathcal{D}(x)}{\partial x}\hat{\theta}(\xi)_x - \hat{T}(\xi)_x W\mathcal{D}(x).$$

These can be used directly in (27) or (28) to construct a symmetry-promoting regularizer $R_{G,1}(W\mathcal{D})$ or $R_{\mathfrak{g},*}(W\mathcal{D})$ that is convex with respect to the weight matrix W. We can use one of these regularizers R to promote symmetry in a least-squares fit with training data consisting of input–output pairs $\{(x_j, y_j)\}_{j=1}^M$ by solving the convex optimization problem

$$\underset{W \in \mathbb{R}^{n \times N}}{\text{minimize}} \quad \frac{1}{M}\sum_{j=1}^M \|y_j - W\mathcal{D}(x_j)\|^2 + \gamma R(W\mathcal{D}). \tag{29}$$

Here, $\gamma \geq 0$ is a parameter controlling the strength of the regularization that can be determined using cross-validation. We demonstrate this approach in Section 10.2 by using it to identify a symmetric linear dynamical system in a setting that requires extrapolation from limited data.

Remark 15 *The solutions $F = W\mathcal{D}$ of (29) do not depend on how the dictionary functions are normalized due to the fact that the function being minimized can be written entirely in terms of F and the data (x_j, y_j). This is in contrast to other types of regularized regression problems that penalize the weights W directly, and therefore depend on how the functions in $\mathcal{D}$ are normalized.*

8.4 Promoting Symmetry in Neural Networks

In this section, we describe a convex regularizing penalty to promote G-equivariance in feed-forward neural networks

$$F = F^{(L)} \circ \cdots \circ F^{(1)}$$

composed of layers $F^{(l)} : \mathcal{V}_{l-1} \to \mathcal{V}_l$ acted upon by group representations $\Phi_l : G \to \mathrm{GL}(\mathcal{V}_l)$. We penalize the breaking of G-symmetries as a means to regularize the neural network and to learn which subgroup of symmetries are compatible with the data and which are not. Since the composition is g-equivariant if every layer is g-equivariant, the main idea is to measure the symmetries shared by all of the layers. Specifically, we aim to maximize the

"size" of the subgroup

$$
\bigcap_{l=1}^{L} \mathrm{Sym}_G\left(F^{(l)}\right) = \{g \in G \; : \; \mathcal{K}_g F^{(l)} = F^{(l)}, \; l = 1, \ldots, L\} \subset \mathrm{Sym}_G(F), \tag{30}
$$

where the notion of "size" we adopt depends on whether G is discrete or continuous. The same ideas can be applied to neural operators acting on fields with layers defined by integral operators as described in Section 6.2. In this case we consider symmetries shared by all of the integral kernels.

As in Section 8.1, when G is a discrete group with finitely many elements, a convex relaxation of the cardinality of $G \setminus \bigcap_{l=1}^{L} \mathrm{Sym}_G(F^{(l)})$ is

$$
R_{G,1}(F^{(1)}, \ldots, F^{(l)}) = \sum_{g \in G} \sqrt{\sum_{l=1}^{L} \left\| \mathcal{K}_g F^{(l)} - F^{(l)} \right\|^2}. \tag{31}
$$

Again, this is analogous to the group-lasso penalty developed in Yuan and Lin (2006).

When G is a Lie group with nonzero dimension, we follow the approach in Section 8.2 using the following observation:

Proposition 16 *The subgroup in* (30) *is closed and embedded in G; its Lie subalgebra is*

$$
\bigcap_{l=1}^{L} \mathfrak{sym}_G\left(F^{(l)}\right) = \left\{ \xi \in \mathrm{Lie}(G) \; : \; \mathcal{L}_\xi F^{(l)} = 0, \; l = 1, \ldots, L \right\}. \tag{32}
$$

We provide a proof in Appendix A.

The Lie subalgebra in the proposition is equal to the nullspace of the linear operator $L_{F^{(1)},\ldots,F^{(L)}} : \mathrm{Lie}(G) \to \bigoplus_{l=1}^{L} C^\infty(\mathcal{V}_{l-1}; \mathcal{V}_l)$ defined by

$$
L_{F^{(1)},\ldots,F^{(L)}} : \xi \mapsto \left(\mathcal{L}_\xi F^{(1)}, \ldots, \mathcal{L}_\xi F^{(L)} \right). \tag{33}
$$

By the rank and nullity theorem, minimizing the rank of this operator is equivalent to maximizing the dimension of the subgroup of symmetries shared by all of the layers in the network. As in Section 8.2, a convex relaxation of the rank is provided by the nuclear norm

$$
R_{\mathfrak{g},*}(F^{(1)}, \ldots, F^{(l)}) = \left\| L_{F^{(1)},\ldots,F^{(L)}} \right\|_* = \left\| \begin{bmatrix} \boldsymbol{L}_{F^{(1)}} \\ \vdots \\ \boldsymbol{L}_{F^{(L)}} \end{bmatrix} \right\|_*, \tag{34}
$$

where $\boldsymbol{L}_{F^{(l)}}$ are the matrices of $L_{F^{(l)}}$ in orthonormal bases for $\mathrm{Lie}(G)$ and the associated spaces $\mathcal{F}'_l$.

We introduce the Equivariance-Promoting Neural Network (EPNN) architecture shown in Figure 4 to promote Lie group symmetries via the convex regularization (34). This draws from the approach of Finzi et al. (2021), where trainable G-equivariant linear layers are composed with G-equivariant nonlinearities. The EPNN is built from Equivariance-Promoting (EP)

28
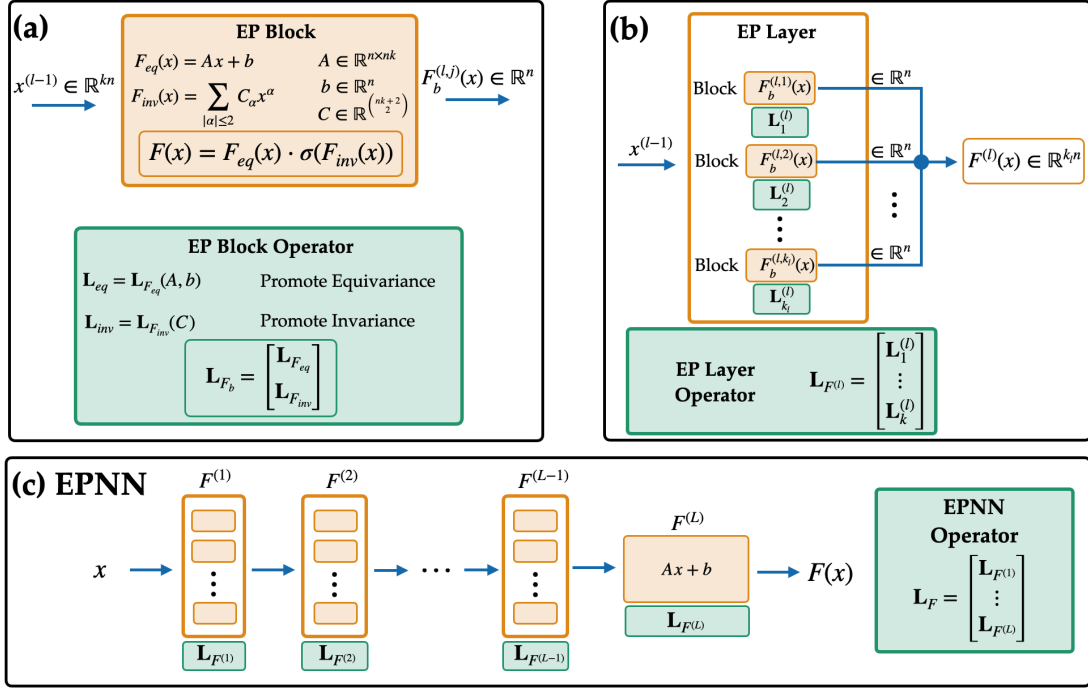
Figure 4: The Equivariance-Promoting Neural Network (EPNN) is a composition of Equivariance-Promoting (EP) layers built from *EP blocks*. An EP block is parameterized by weights A, b, C defining a vector-valued affine function F_{eq} and a scalar-valued quadratic function F_{inv}. The symmetries shared by these components are captured by the nullspace of the *EP block operator*. These are assembled into operators capturing shared symmetries of components within each layer, and then into an EPNN operator for the whole network.

blocks,

$$F_{\text{b}}(x) = \underbrace{(Ax + b)}_{F_{\text{eq}}(x)} \cdot \sigma\left(\underbrace{\sum_{|\alpha| \leq 2} C_\alpha x^\alpha}_{F_{\text{inv}}(x)}\right), \tag{35}$$

where α is a multi-index; A, b, and $C = (C_\alpha)_{|\alpha| \leq 2}$ are the trainable weights of the block; and σ is a nonlinear activation function such as the sigmoid $\sigma(z) = (1 + e^{-z})^{-1}$. Using the sigmoid ensures the gradient of the EP block is bounded. An EP block is g-equivariant with respect to linear group actions when F_{eq} is g-equivariant and F_{inv} is g-invariant. The Lie subalgebra of symmetries shared by F_{eq} and F_{inv} forms the nullspace of the discretized EP block operator $\boldsymbol{L}_{F_{\text{b}}}$ (see Figure 4). The matrix entries of this operator are linear functions of the trainable parameters.

An *EP layer* is built from EP blocks $F_{\text{b}}^{(l,j)} : \mathcal{V}_{l-1} \to \tilde{\mathcal{V}}_l$ by concatenation

$$F^{(l)}(x) = \left(F_{\text{b}}^{(l,1)}(x), \ \ldots, \ F_{\text{b}}^{(l,k_l)}(x)\right), \tag{36}$$

producing an element in $\mathcal{V}_l = \oplus_{j=1}^{k_l} \tilde{\mathcal{V}}_l$. An EPNN is then formed by composing $L-1$ of these layers, with a final affine EP output layer. The G-actions on the hidden layers $l = 1, \ldots, L-1$ are defined by direct sum representations

$$\Phi_l(g) = \begin{bmatrix} \tilde{\Phi}_l(g) & & \\ & \ddots & \\ & & \tilde{\Phi}_l(g) \end{bmatrix}, \qquad \text{where} \qquad \tilde{\Phi}_l : G \to \mathrm{GL}(\tilde{\mathcal{V}}_l). \qquad (37)$$

The key observation is that if the components $F_{\mathrm{eq}}^{(l,j)}$ and $F_{\mathrm{inv}}^{(l,j)}$ of each EP block are respectively g-equivariant and g-invariant, i.e.,

$$\tilde{\Phi}_l(g) F_{\mathrm{eq}}^{(l,j)}\big(\Phi_{l-1}(g)^{-1} x\big) = F_{\mathrm{eq}}^{(l,j)}(x), \qquad \text{and} \qquad F_{\mathrm{inv}}^{(l,j)}\big(\Phi_{l-1}(g)^{-1} x\big) = F_{\mathrm{inv}}^{(l,j)}(x),$$

and the affine EP output layer is g-equivariant, then the entire EPNN is g-equivariant. The Lie subalgebra of symmetries shared by all components in an EPNN is given by the nullspace of the discretized EPNN operator $\boldsymbol{L}_F$ (see Figure 4). We obtain a convex symmetry-promoting penalty (34) by taking the nuclear norm of the EPNN operator, whose matrix entries are linear functions of the trainable weights. Of course, it is also possible to enforce G-equivariance in the EPNN through a collection of linear constraints on the trainable weights using an approach analogous to Finzi et al. (2021) (see Section 6). We evaluate EPNNs on a simple example in Section 10.3.

9. Discretizing the Operators

This section describes how to construct matrices for the operators $\mathcal{L}_\xi$ and L_F for continuously differentiable functions F in a user-specified finite-dimensional subspace $\mathcal{F} \subset C^1(\mathcal{V}; \mathcal{W})$. By choosing bases for the finite-dimensional vector spaces $\mathcal{V}$ and $\mathcal{W}$, it suffices without loss of generality to consider the case in which $\mathcal{V} = \mathbb{R}^m$ and $\mathcal{W} = \mathbb{R}^n$. We assume that $\mathrm{Lie}(G)$ and $\mathcal{F}$ are endowed with inner products and that $\{\xi_1, \ldots \xi_{\dim(G)}\}$ and $\{F_1, \ldots F_{\dim(\mathcal{F})}\}$ are orthonormal bases for these spaces, respectively. The key task is to endow the finite-dimensional subspace

$$\mathcal{F}' = \mathrm{span}\left\{\mathcal{L}_\xi F \ : \ \xi \in \mathrm{Lie}(G), \ F \in \mathcal{F}\right\} \subset C^0(\mathbb{R}^m; \mathbb{R}^n)$$

with a convenient inner product. Once this is done, an orthonormal basis $\{u_1, \ldots, u_N\}$ for $\mathcal{F}'$ can be constructed by applying a Gram-Schmidt process to the functions $\mathcal{L}_{\xi_i} F_j$, which span $\mathcal{F}'$. Matrices for $\mathcal{L}_\xi$ and L_F are then easily obtained by computing

$$\big[\boldsymbol{\mathcal{L}}_\xi\big]_{i,j} = \big\langle u_i, \ \mathcal{L}_\xi F_j \big\rangle_{\mathcal{F}'}, \qquad \big[\boldsymbol{L}_F\big]_{i,k} = \big\langle u_i, \ \mathcal{L}_{\xi_k} F \big\rangle_{\mathcal{F}'}.$$

The issue at hand is to choose the inner product on $\mathcal{F}'$ in a way that makes computing these matrices easy. A natural choice is to equip $\mathcal{F}'$ with an $L^2(\mathbb{R}^m, \mu; \mathbb{R}^n)$ inner product where μ is a positive measure on $\mathbb{R}^m$ (such as a Gaussian distribution) for which the L^2 norms of functions in $\mathcal{F}'$ are finite. The problem is that it is usually challenging or inconvenient to

compute the required integrals

$$\langle \mathcal{L}_{\xi_i} F_j, \ \mathcal{L}_{\xi_k} F_l \rangle_{L^2(\mu)} = \int_{\mathbb{R}^m} (\mathcal{L}_{\xi_i} F_j)(x)^T (\mathcal{L}_{\xi_k} F_l)(x) \, d\mu(x) \tag{38}$$

analytically. In this section, we discuss inner products that are easy to compute in practice.

9.1 Numerical Quadrature and Monte Carlo

When (38) cannot be computed analytically, numerical quadrature or Monte Carlo can be used. In both cases the integral is approximated by a weighted sum, yielding a semi-inner product

$$\langle f, \ g \rangle_{L^2(\mu_M)} = \frac{1}{M} \sum_{i=1}^{M} w_i f(x_i)^T g(x_i) \tag{39}$$

that converges to $\langle f, \ g \rangle_{L^2(\mu)}$ as $M \to \infty$. The following proposition means that we do not have to pass to the limit $M \to \infty$ in order to obtain a valid inner product defined by (39) on $\mathcal{F}'$.

Proposition 17 *Suppose that $\mathcal{F}'$ is finite-dimensional and $\langle f, \ g \rangle_{L^2(\mu_M)} \to \langle f, \ g \rangle_{L^2(\mu)}$ as $M \to \infty$ for every $f, g \in \mathcal{F}'$. Then there is an M_0 such that (39) is an inner product on $\mathcal{F}'$ for every $M \geq M_0$. We give a proof in Appendix B.*

In Monte Carlo approximation, the samples x_i are drawn independently from a distribution ν with the assumption that both μ and ν are σ-finite and μ is absolutely continuous with respect to ν. The weights are given by the Radon-Nikodym derivative $w_i = \frac{d\mu}{d\nu}(x_i)$. Then for every $f, g \in L^2(\mu)$ the approximate integral converges $\langle f, \ g \rangle_{L^2(\mu_M)} \to \langle f, \ g \rangle_{L^2(\mu)}$ as $M \to \infty$, almost surely, thanks to the strong law of large numbers (Theorem 7.7; Koralov and Sinai, 2012). By the proposition, there is almost surely a finite M_0 such that (39) is an inner product on $\mathcal{F}'$ for every $M \geq M_0$.

9.2 Subspaces of Polynomials

Here we consider the special case when $\mathcal{F}$ is a finite-dimensional subspace consisting of polynomial functions $\mathbb{R}^m \to \mathbb{R}^n$ and the action (9) is defined by representations as in (10). Examining (14), it is evident that $\mathcal{L}_\xi F$ is also a polynomial function $\mathbb{R}^m \to \mathbb{R}^n$ with degree not greater than that of $F \in \mathcal{F}$. Thus, $\mathcal{F}'$ is also a space of polynomial functions with degree not exceeding the maximum degree in $\mathcal{F}$. Since a polynomial that vanishes on an open set must be identically zero, we can take the integrals defining the inner product in (39) over a cube, such as $[0,1]^m \subset \mathbb{R}^m$. This is convenient because polynomial integrals over the cube can be calculated analytically.

We can also use the sample-based inner product in (39) with randomly chosen points x_i and positive weights w_i. The following proposition tells us exactly how many sample points we need.

Proposition 18 *Let $\mathcal{F}'$ be a space of real polynomial functions $\mathbb{R}^m \to \mathbb{R}^n$ and let $\pi_i : \mathbb{R}^n \to \mathbb{R}$ be the ith coordinate projection $\pi(c_1, \ldots, c_n) = c_i$. Let*

$$M \geq M_0 := \max_{1 \leq i \leq n} \dim(\pi_i(\mathcal{F}')) \tag{40}$$

and let $w_1, \ldots, w_M > 0$ be positive weights. Then for almost every set of points $(x_1, \ldots, x_M) \in (\mathbb{R}^m)^M$ with respect to Lebesgue measure, (39) is an inner product on $\mathcal{F}'$. We give a proof in Appendix B.

This means that we can draw $M \geq M_0$ sample points independently from any absolutely continuous measure (such as a Gaussian distribution or the uniform distribution on a cube), and with probability one, (39) will be an inner product on $\mathcal{F}'$. When $\mathcal{F}$ consists of polynomials with degree at most d, then it is sufficient to take

$$M \geq \left| \left\{ (p_0, \ldots, p_m) \in \mathbb{N}_0^{m+1} \ : \ p_0 + \cdots + p_m = d \right\} \right| = \binom{d+m}{m}.$$

10. Numerical Experiments using New Symmetry-Promoting Penalties

We perform several numerical experiments to evaluate our proposed approach for promoting Lie group symmetry using nuclear-norm regularization. All of our code was written in Python and is available at `https://github.com/nzolman/unified-symmetry-ml`.

10.1 Sample Complexity to Recover Symmetric Polynomials

Can promoting symmetry help us learn an unknown symmetric function using fewer samples? To begin answering this question, we perform numerical experiments studying the amount of sampled data needed to recover structured polynomial functions on $\mathbb{R}^n$ of the form

$$F_{\mathrm{rad}}(x) = \varphi_{\mathrm{rad}}\big(\|x - c_1\|_2^2, \ \ldots, \ \|x - c_r\|_2^2\big) \quad \text{and} \tag{41}$$

$$F_{\mathrm{lin}}(x) = \varphi_{\mathrm{lin}}\big(u_1^T x, \ \ldots, \ u_r^T x\big). \tag{42}$$

These possess various rotation and translation invariances when $r < n$, as characterized in detail below by Proposition 19 and its corollaries.

We aim to recover the unknown function F_* within the space $\mathcal{P}_d(\mathbb{R}^n)$ of polynomial functions on $\mathbb{R}^n$ with degrees up to $d = \deg F_*$ based on the values $y_j = F_*(x_j)$ at sample points $x_1, \ldots, x_N$. Our approximation $\hat{F}$ of F_* is computed by solving the convex optimization problem

$$\underset{F \in \mathcal{P}_d(\mathbb{R}^n)}{\text{minimize}} \ R_{\mathfrak{g},*}(F) = \|L_F\|_* \qquad \text{s.t.} \qquad F(x_j) = y_j, \quad j = 1, \ldots, N, \tag{43}$$

where G is a candidate Lie group of symmetries acting on $\mathbb{R}^n$. This was done using the CVXPY software package developed by Diamond and Boyd (2016); Agrawal et al. (2018). The nuclear norm in (43) was defined with respect to inner products on the corresponding Lie algebras given by $\langle \xi, \eta \rangle_{\mathrm{Lie}(G)} = \mathrm{Tr}(\xi^T \eta)$. As described in Section 9, the inner product on the space $\mathcal{F}'$ containing the ranges of every L_F was defined by (39) with unit weights $w_i = 1$ and $M = \dim \mathcal{P}_d(\mathbb{R}^n)$ points drawn uniformly from the cube $[-1, 1]^n$. When using the group of translations $G = (\mathbb{R}^n, +)$, the Lie derivative always lowers the degree by 1, so we only use $M = \dim \mathcal{P}_{d-1}(\mathbb{R}^n)$ points in this case. Note that the discretization points are not the same as the sample points x_j in (43).

The dimensions of the resulting discretized operators are summarized in Table 1. One of the largest experiments we performed is shown in Figure 6 with $\deg(\phi_{\text{lin}}) = 2$, $\dim(x) = 30$, and $G = \text{SE}(30)$. We computed and stored the entries of the $496 \times 496 \times 465$ tensor $[\boldsymbol{L}_{F_i}]_{j,k}$, which were accessed by CVXPY. This problem took about 40 minutes to solve using an Intel Xeon Gold 6230 CPU. On the other hand, the higher-dimensional problem with $\dim(x) = 50$ shown in Figure 7 was solved in less than 1 minute using the translation group $G = (\mathbb{R}^{50}, +)$. Here, the tensor had dimensions $1326 \times 51 \times 50$, meaning the problem had more optimization variables (1326), but involved nuclear norms of much smaller 51×50 matrices.

Group	i-dim. $= \dim \mathcal{F} = \dim \mathcal{P}_d(\mathbb{R}^n)$	j-dim. $= \dim \mathcal{F}' = M$	k-dim. $= \dim G$
$\text{SE}(n)$	$\binom{n+d}{d}$	$\binom{n+d}{d}$	$\frac{n(n+1)}{2}$
$(\mathbb{R}^n, +)$	$\binom{n+d}{d}$	$\binom{n+d-1}{d-1}$	n

Table 1: Dimensions of discretized rank-3 tensors $[\boldsymbol{L}_{F_i}]_{j,k}$ for symmetric polynomial recovery.

To study the sample complexity for (43) to recover functions in the form of F_{rad} and F_{lin}, we perform multiple experiments using random realizations of these functions sampled at random points x_j. In each experiment, the vectors c_i were drawn uniformly from the cube $[-1, 1]^n$ and the vectors u_i were formed from the columns of a random $n \times r$ orthonormal matrix (specifically, the left singular vectors of an $n \times r$ matrix with standard Gaussian entries). The coefficients of φ_{rad} and φ_{lin} in a basis of monomials up to a specified degree were sampled uniformly from the interval $[0, 1]$. This yielded random polynomial functions F_{rad} and F_{lin} with degrees $\deg F_{\text{rad}} = 2 \deg \varphi_{\text{rad}}$ and $\deg F_{\text{lin}} = \deg \varphi_{\text{lin}}$.

The sample points x_j for each experiment were drawn uniformly from the cube $[-1, 1]^n$. A total of $\dim \mathcal{P}_d(\mathbb{R}^n)$ with $d = \deg F_*$ sample points were drawn, which is sufficient to recover the function almost surely regardless of regularization. For each experiment, we determine the smallest N_* so that recovery is achieved by (43) with $\hat{F} = F_*$ using the sample points $x_1, \ldots, x_N$ for every $N \geq N_*$. To be precise, successful recovery is declared when all coefficients describing $\hat{F}$ and F_* in the monomial basis for $\mathcal{P}_d(\mathbb{R}^n)$ agree to a tolerance of 5×10^{-3} times the magnitude of the largest coefficient of F_*. The range of values for N_* across 10 such random experiments provides an estimate of the sample complexity. In Figures 5, 6, and 7 we plot the range of values for N_* as shaded regions with the average values displayed as a solid lines.

In Figure 5, we use the special Euclidean group $G = SE(n)$ as a candidate group to recover functions of the form F_{rad} with the degree of φ_{rad} specified. The number of radial features $r \leq \deg \varphi_{\text{rad}}$ is selected in accordance with Corollary 20 in order to ensure that $\mathfrak{sym}_G(F_{\text{rad}}) = \mathfrak{g}_{\text{rad}}$ has the known form and dimension stated in Proposition 19. The symmetry-promoting regularization significantly reduces the number of samples needed to recover F_{rad} compared to the number of samples needed to solve the linear system specifying this function within the space of polynomials with the same or lesser degree. As the number of radial features r increases, so does the sample complexity to recover F_{rad}. This is likely due to the decreased dimension of $\mathfrak{sym}_G(F_{\text{rad}})$.

In Figures 6 and 7, we use $G = SE(n)$ and the group of translations $G = (\mathbb{R}^n, +)$ as candidate symmetry groups to recover function of the form F_{lin} with the degree of φ_{lin}
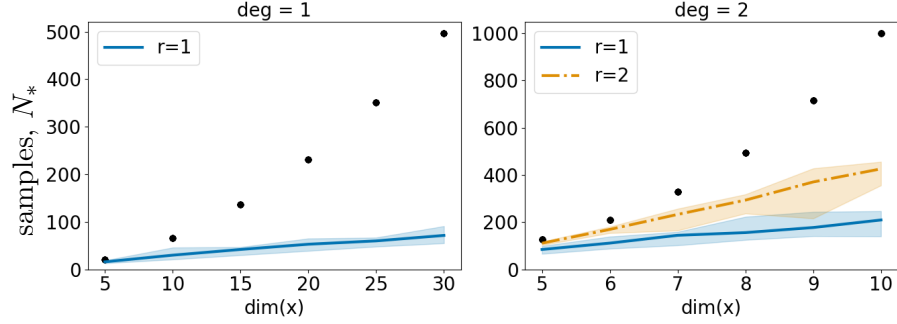
Figure 5: Sample complexity (shaded region: min–max across 10 trials; line: mean) to recover polynomial functions F_{rad} of r radial features, i.e., (41) with polynomial φ_{rad} of the specified degree, by solving (43) using the special Euclidean group $G = SE(n)$. Black dots indicate the number of dictionary functions, $\dim \mathcal{P}_d(\mathbb{R}^n)$, thus the number of samples needed to recover F_{rad} without regularization.
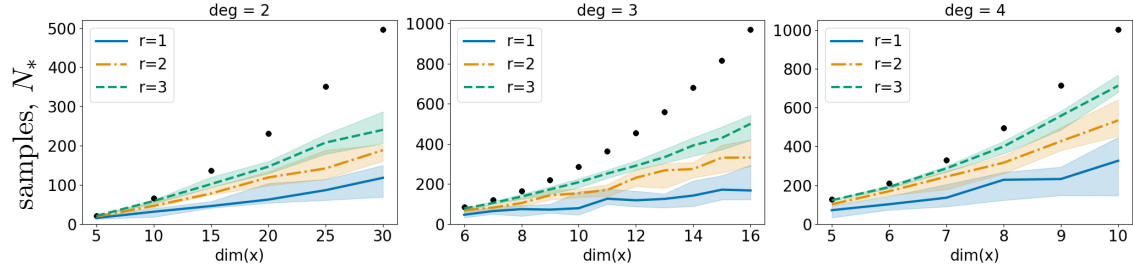
Figure 6: Sample complexity (shaded region: min–max across 10 trials; line: mean) to recover polynomial functions F_{lin} of r linear features, i.e., (42) with polynomial φ_{lin} of the specified degree, by solving (43) using the special Euclidean group $G = SE(n)$. Black dots indicate the number of dictionary functions, $\dim \mathcal{P}_d(\mathbb{R}^n)$, thus the number of samples needed to recover F_{lin} without regularization.
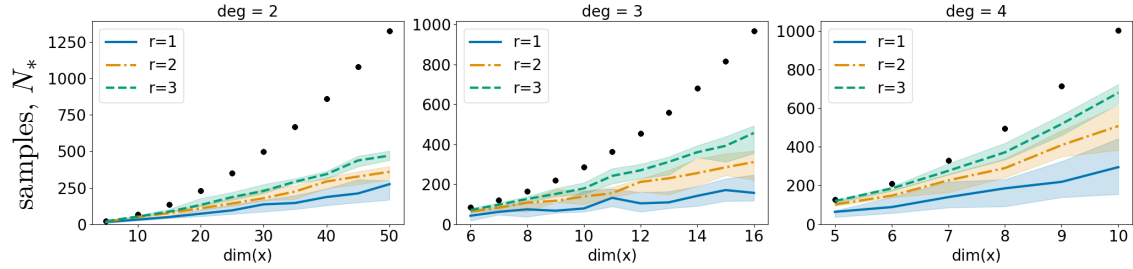
Figure 7: Analogue of Fig. 6 using the group of translations $G = (\mathbb{R}^n, +)$.

specified. Obviously, F_{lin} has an $(n-r)$-dimensional subgroup of translation symmetries orthogonal to span$\{u_1, \ldots, u_r\}$. By Corollary 21, choosing deg $\varphi_{\text{lin}} \geq 2$ is sufficient to ensure that $\mathfrak{sym}_{\text{SE}(n)}(F_{\text{lin}}) = \mathfrak{g}_{\text{lin}}$ has the known form and dimension stated in Proposition 19. The results in Figures 6 and 7 show that the symmetry-promoting regularization reduces the sample complexity to recover F_{lin}. Moreover, fewer samples are needed when F_{lin} depends on fewer linear features, as might be expected because the dimension of $\mathfrak{sym}_G(F_{\text{rad}})$ increases as r decreases.

Proposition 19 *Let $r \leq n$ and suppose that $\{c_k - c_1\}_{k=2}^r$ and $\{u_k\}_{k=1}^r$ are sets of linearly-independent vectors in $\mathbb{R}^n$. Then, $\mathfrak{sym}_{\text{SE}(n)}(F_{rad})$ contains the $\frac{1}{2}(n-r)(n-r+1)$-dimensional subalgebra*

$$\mathfrak{g}_{rad} = \left\{ \begin{bmatrix} S & v \\ 0 & 0 \end{bmatrix} \;:\; S^T = -S \text{ and } Sc_1 = \cdots = Sc_r = -v \right\} \tag{44}$$

and $\mathfrak{sym}_{\text{SE}(n)}(F_{lin})$ contains the $\frac{1}{2}(n-r)(n-r+1)$-dimensional subalgebra

$$\mathfrak{g}_{lin} = \left\{ \begin{bmatrix} S & v \\ 0 & 0 \end{bmatrix} \;:\; S^T = -S, \; Su_1 = \cdots = Su_r = 0, \text{ and } u_1^T v = \cdots = u_r^T v = 0 \right\}. \tag{45}$$

Either every polynomial φ_{rad} with degree $\leq d$ gives $\mathfrak{sym}_{\text{SE}(n)}(F_{rad}) \neq \mathfrak{g}_{rad}$, or else the set of polynomials φ_{rad} with degree $\leq d$ satisfying $\mathfrak{sym}_{\text{SE}(n)}(F_{rad}) \neq \mathfrak{g}_{rad}$ is a set of measure zero. Likewise, for φ_{lin}, F_{lin}, and $\mathfrak{g}_{lin}$. See Appendix C for a proof.

Corollary 20 *With the same hypotheses as Proposition 19, let $d \geq r$. The set of polynomials φ_{rad} with degree $\leq d$ satisfying $\mathfrak{sym}_{\text{SE}(n)}(F_{rad}) \neq \mathfrak{g}_{rad}$ is a set of measure zero. A proof is given in Appendix C.*

Corollary 21 *With the same hypotheses as Proposition 19, let $d \geq 2$. The set of polynomials φ_{lin} with degree $\leq d$ satisfying $\mathfrak{sym}_{\text{SE}(n)}(F_{lin}) \neq \mathfrak{g}_{lin}$ is a set of measure zero. A proof is given in Appendix C.*

10.2 Extrapolatory Recovery of a Multi-Spring-Mass System

We demonstrate how symmetry-promoting nuclear-norm regularization can identify equivariant dynamics that correctly extrapolate from a restricted data set. We consider a system of $n_p = 5$ particles with identical masses $m_i = 1$, whose positions and momenta are vectors $\boldsymbol{q}^i = (q_x^i, q_y^i, q_z^i)$ and $\boldsymbol{p}^i = (p_x^i, p_y^i, p_z^i)$ in $\mathbb{R}^3$. However, the training data are confined to the x–z plane, as explained below. The particles fall under the influence of gravitational acceleration $\boldsymbol{g} = (0, 0, -1)$ and are coupled together by springs with random stiffness $K_{i,j} = \sum_{k=1}^{n_p} (1 - \delta_{i,j}) B_{i,k} B_{j,k}$, where $B_{i,j}$ are drawn uniformly from the interval $[0.1, 1]$. This yields a linear dynamical system

$$\frac{\mathrm{d}}{\mathrm{d}\, t} \begin{bmatrix} \boldsymbol{q}^i \\ \boldsymbol{p}^i \end{bmatrix} = \sum_{j=1}^{n_p} \underbrace{\begin{bmatrix} \boldsymbol{0} & \delta_{i,j} m_i^{-1} \boldsymbol{I} \\ \left(K_{i,j} - \delta_{i,j} \sum_{k=1}^{n_p} K_{i,k} \right) \boldsymbol{I} & \boldsymbol{0} \end{bmatrix}}_{\boldsymbol{A}_{i,j}} \begin{bmatrix} \boldsymbol{q}^j \\ \boldsymbol{p}^j \end{bmatrix} + \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{g} \end{bmatrix}, \tag{46}$$
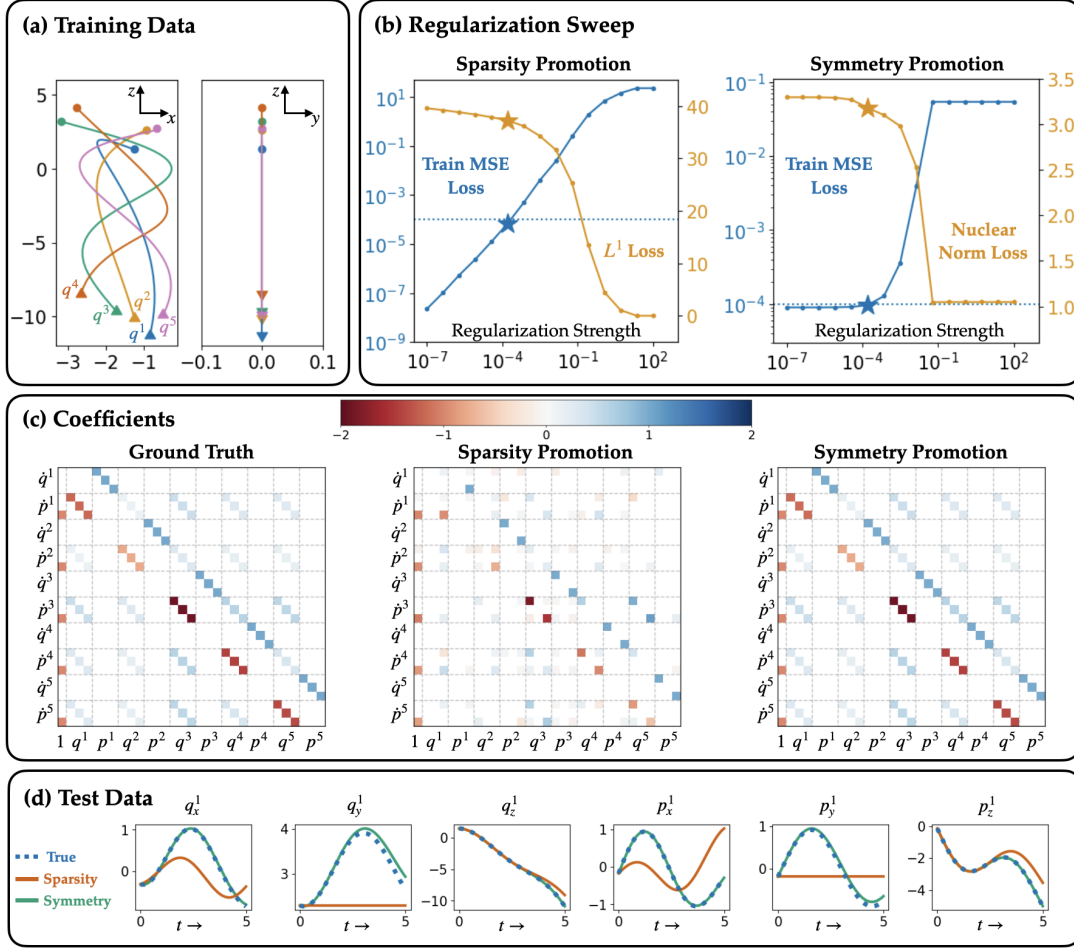
Figure 8: The multi-spring-mass system is learned from training data confined to the x–z plane using SE(3)-symmetry-promoting regularized regression. Results are compared to Sparse Identification of Nonlinear Dynamics (SINDy) using sparsity-promoting ℓ^1 regularization.

expressible in the form (26) with $\dot{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{x}$ using state variable $\boldsymbol{x} = (\boldsymbol{q}^1, \boldsymbol{p}^1, \ldots, \boldsymbol{q}^{n_p}, \boldsymbol{p}^{n_p}, 1)$. The entries of the matrix $\boldsymbol{A}$ are shown in the leftmost plot of Figure 8(c).

When there is no gravity, the system is SE(3)-equivariant under the action (8) applied simultaneously to all particles. Gravity breaks the rotational symmetries about all but the z-axis, leaving the system equivariant to translations and rotations of the entire system of particles about the z-axis, corresponding to the subgroup of SE(3) given by

$$\mathrm{Sym}_{\mathrm{SE}(3)}(\boldsymbol{A}) = \left\{ \begin{bmatrix} \cos(\theta) & -\sin(\theta) & 0 & b_x \\ \sin(\theta) & \cos(\theta) & 0 & b_y \\ 0 & 0 & 1 & b_z \\ 0 & 0 & 0 & 1 \end{bmatrix} : \theta, b_x, b_y, b_z \in \mathbb{R} \right\}. \tag{47}$$

36

We use nuclear norm–based symmetry-promotion (28) within the group $G = \mathrm{SE}(3)$ to learn the dynamics (46) and its symmetries (47) from trajectories that only evolve in the x–z plane. The training data consists of $M = 100$ state-time-derivative pairs $(\boldsymbol{x}_j, \dot{\boldsymbol{x}}_j)$ collected along two trajectories, each with 50 samples spaced uniformly at time intervals $\Delta t = 0.1$. One of the trajectories is shown in Figure 8(a). These were generated by choosing normally-distributed initial conditions

$$\boldsymbol{q}^i \sim \mathcal{N}\left(\boldsymbol{q}_0, \ \mathrm{diag}\begin{bmatrix} 1 & 0 & 1 \end{bmatrix}\right), \quad \boldsymbol{p}^i \sim \mathcal{N}\left(\boldsymbol{p}_0, \ \mathrm{diag}\begin{bmatrix} 0.01 & 0 & 0.01 \end{bmatrix}\right), \quad i = 1, \ldots, n_p,$$

$$\boldsymbol{q}_0 \sim \mathcal{N}(\boldsymbol{0}, \ \mathrm{diag}\begin{bmatrix} 1 & 0 & 1 \end{bmatrix}), \qquad \boldsymbol{p}_0 \sim \mathcal{N}(\boldsymbol{0}, \ \mathrm{diag}\begin{bmatrix} 0.01 & 0 & 0.01 \end{bmatrix}),$$

with the stated means and covariance so that every particle has zero y-position and zero y-momentum.

We employ the basis-function regression framework described in Section 8.3 with loss function (29) and linear dictionary $\mathcal{D}$ to fit the system matrix $\boldsymbol{A}$ to the training data. This approach is compared to Sparse Identification of Nonlinear Dynamics (SINDy) employing sparsity-promoting ℓ^1 regularization (see Brunton et al., 2016; Tibshirani, 1996) on the entries of $\boldsymbol{A}$. The components of the training loss are plotted against the regularization strength γ in Figure 8(b). The chosen regularization strengths (indicated by stars) were the maximum values for which the training mean square error (MSE) remained less than 10^{-4}.

The system matrices learned using sparsity-promoting ℓ^1 regularization and SE(3)-symmetry-promoting regularization are shown in Figure 8(c). We observe that symmetry-promoting regularization recovers the correct matrix structure, while promoting sparsity does not. Because the training data have zero y-positions and momenta, a maximally-sparse fit will not have terms involving these components. On the other hand, promoting SE(3)-symmetry enables identification of the rotational symmetry about the z-axis, allowing the learned model to extrapolate from training data that lives only in the x–z plane. This is confirmed in Figure 8(d), where we plot the true and predicted trajectories of the first particle on a testing data set where all three components of each particle's initial position and momentum were chosen at random according to

$$\boldsymbol{q}^i \sim \mathcal{N}\left(\boldsymbol{q}_0, \ \boldsymbol{I}\right), \qquad \boldsymbol{p}^i \sim \mathcal{N}\left(\boldsymbol{p}_0, \ 0.1\boldsymbol{I}\right), \qquad i = 1, \ldots, n_p,$$

$$\boldsymbol{q}_0 \sim \mathcal{N}(\boldsymbol{0}, \ \boldsymbol{I}), \qquad \boldsymbol{p}_0 \sim \mathcal{N}(\boldsymbol{0}, \ 0.1\boldsymbol{I}).$$

10.3 Neural ODEs Model of a Double-Spring Pendulum using EPNNs

We use the double-spring pendulum example from Finzi et al. (2020) to study the nuclear-norm regularization technique for neural networks presented in Section 8.4. As illustrated in Figure 9(a), this system consists of two masses $m_1 = m_2 = 1$ connected in series to the origin by springs with stiffness $k_1 = k_2 = 1$ and length $\ell_1 = \ell_2 = 1$. The masses are subject to downward gravitational acceleration $\boldsymbol{g} = (0, 0, -1)$. The positions $\boldsymbol{q}^i = (q_x^i, q_y^i, q_z^i)$ and momenta $\boldsymbol{p}^i = (p_x^i, p_y^i, p_z^i)$ of the masses $i = 1, 2$ are described by a nonlinear Hamiltonian

dynamical system

$$\dot{\boldsymbol{q}}^T = \frac{\partial H}{\partial \boldsymbol{p}}(\boldsymbol{q}, \boldsymbol{p}), \qquad \dot{\boldsymbol{p}}^T = -\frac{\partial H}{\partial \boldsymbol{q}}(\boldsymbol{q}, \boldsymbol{p}),$$

$$H(\boldsymbol{q}, \boldsymbol{p}) = \frac{\|\boldsymbol{p}^1\|^2}{2m_1} + \frac{\|\boldsymbol{p}^2\|^2}{2m_2} + \frac{k_1}{2}\left(\|\boldsymbol{q}^1\| - \ell_1\right)^2 + \frac{k_2}{2}\left(\|\boldsymbol{q}^1 - \boldsymbol{q}^2\| - \ell_2\right)^2 - m_1 \boldsymbol{g}^T \boldsymbol{q}^1 - m_2 \boldsymbol{g}^T \boldsymbol{q}^2,$$

expressible in the form $\dot{\boldsymbol{x}} = \boldsymbol{F}(\boldsymbol{x})$ using the 12-dimensional state variable $\boldsymbol{x} = (\boldsymbol{q}^1, \boldsymbol{p}^1, \boldsymbol{q}^2, \boldsymbol{p}^2)$. In the absence of gravity, the dynamics $\boldsymbol{F}$ would be equivariant with respect to rotations about the origin, that is, with respect to the simultaneous action of the orthogonal group O(3) on the positions and momenta of both masses. Gravity breaks most of these symmetries, preserving only the subgroup O(2) of rotational symmetries about the vertical z-axis.

We compare approximations of the nonlinear function F using EPNNs (see Figure 4 and Section 8.4) with different candidate symmetry groups to a standard multilayer perceptron (MLP) with a similar number of parameters. These are trained using neural ODEs (Chen et al., 2018) on data gathered in short chunks along trajectories. In order to study the data efficiency of the resulting symmetry-promoting learning algorithms, we train the models on different amounts of data. For each training-data set size, we perform 20 random experiments, in which the training data, testing data, and initial network weights are redrawn from their respective distributions. The box plots in Figure 9 capture variance resulting from randomness across all of these factors.

10.3.1 DATA

The training data for each experiment is generated using the same procedure as Finzi et al. (2020), and consists of $M = 10$, 25, 50, 100, 250, or 500 trajectory chunks $\left(\boldsymbol{x}_j(t_j),\ \boldsymbol{x}_j(t_j + \Delta t),\ \ldots,\ \boldsymbol{x}_j(t_j + n_t \Delta t)\right)$ sampled at $n_t = 5$ consecutive intervals of length $\Delta t = 0.2s$. The jth chunk is generated by simulating the double-spring pendulum for $30s$ from normally distributed initial conditions,

$$\begin{aligned} \boldsymbol{q}_j^1(0) \sim \mathcal{N}\left((0, 0, -1.5),\ 0.04\boldsymbol{I}\right), \qquad \boldsymbol{q}_j^2(0) \sim \mathcal{N}\left((0, 0, -3),\ 0.04\boldsymbol{I}\right), \\ \boldsymbol{p}_j^1(0),\ \boldsymbol{p}_j^2(0) \sim \mathcal{N}\left(\boldsymbol{0},\ 0.16\boldsymbol{I}\right), \end{aligned} \tag{48}$$

and choosing the starting time of the chunk t_j uniformly at random from the interval $[0, 29]$. An evaluation data set of M trajectory chunks, generated in the same way, is used for early stopping. Two testing data sets are used in each experiment. The first consists of a single randomly generated $30s$ trajectory with initial conditions drawn according to (48). The networks are evaluated by forecasting the entire trajectory from its initial condition at $t = 0$. The second testing data set in each experiment consists of 500 trajectory chunks generated in the same way as the training data.

10.3.2 OPTIMIZATION

With θ being the parameters of an EPNN approximation $\boldsymbol{F}_\theta$, and $\hat{\boldsymbol{x}}_j(t; \theta)$ being the trajectory of $\dot{\hat{\boldsymbol{x}}}_j = \boldsymbol{F}_\theta(\hat{\boldsymbol{x}}_j)$ with the same initial condition as the chunk $\hat{\boldsymbol{x}}_j(t_j; \theta) = \boldsymbol{x}_j(t_j)$, we use neural
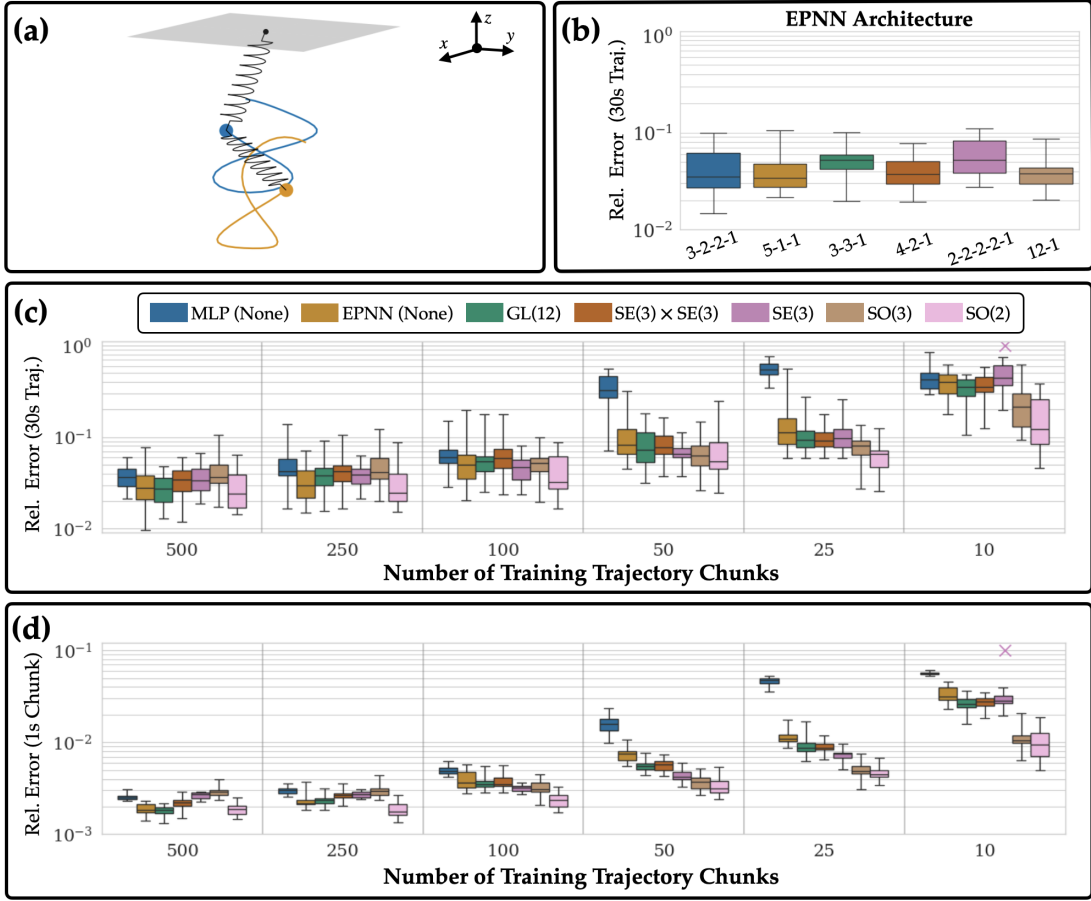
Figure 9: (a) double-spring pendulum system from Finzi et al. (2021). (b) Prediction performance on 30s trajectories using different SO(3)-EPNN architectures trained on $M = 100$ trajectory chunks. (c) Prediction performance using 3-3-1 EPNNs trained using different amounts of data and candidate groups. (d) Prediction performance on 1s trajectory chunks using the same setup as (c). The "x" represents a prediction that diverged.

ODEs and the Adam optimizer of Kingma and Ba (2014) to minimize the loss function

$$J(\theta) = \underbrace{\frac{1}{M} \sum_{j=1}^{M} \sum_{k=1}^{n_t} \left\| \boldsymbol{x}_j(t_j + k\Delta t) - \hat{\boldsymbol{x}}_j(t_j + k\Delta t; \theta) \right\|^2}_{\text{MSE}} + \underbrace{\gamma \left\| \boldsymbol{L}_{\boldsymbol{F}_\theta} \right\|_*}_{R_{\mathfrak{g}, *}(\theta)}. \qquad (49)$$

Here, $\boldsymbol{L}_{\boldsymbol{F}_\theta}$ is the discretized EPNN operator described in Section 8.4 with respect to the action of the candidate group G. The regularization strength $\gamma = 10^{-4}$ was used for $G = \mathrm{SO}(2)$, $\mathrm{SO}(3)$, and $\mathrm{SE}(3)$; $\gamma = 10^{-6}$ was used for $G = \mathrm{SE}(3) \times \mathrm{SE}(3)$; and $\gamma = 10^{-8}$ was used for $G = \mathrm{GL}(12)$. We use the default ODE solver in JAX (Bradbury et al., 2018), which is the

Dormand-Prince Runge-Kutta scheme with adaptive step size and atol = rtol = 1.4×10^{-8}. We fix a learning rate of 3×10^{-4} for all experiments and use optax (DeepMind et al., 2020) default parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$ for Adam. We use batches containing all of the training data, and we perform 10^4 optimization steps with early stopping based on the validation data set performance.

10.3.3 ARCHITECTURE

The EP blocks in our EPNNs are maps $F_{\mathrm{b}} : \mathbb{R}^{12} \to \mathbb{R}^{12}$ defined by (35) with the sigmoid activation, and the same G-action across all layers defined by a fixed representation

$$\Phi_0 = \tilde{\Phi}_1 = \cdots = \tilde{\Phi}_{L-1} = \Phi_L.$$

We compare six different candidate Lie group actions:

1. No group action,
2. GL(12) acting via matrix multiplication with the state vector,
3. SE(3) $\times$ SE(3) with each copy of SE(3) acting via (8) on the position and momentum of a corresponding mass,
4. SE(3) acting via (8) on the positions and momenta of both masses,
5. SO(3) acting to rotate the positions and momenta of both masses, and
6. SO(2) acting to rotate the positions and momenta of both masses about the z-axis.

All network weights were initialized uniformly at random in the interval $[-0.1, 0.1]$. The components $\boldsymbol{L}_{F_{\mathrm{eq}}}$ and $\boldsymbol{L}_{F_{\mathrm{inv}}}$ of each EP block operator were discretized using the Monte Carlo method presented in Section 9 with uniformly-sampled points from axis-aligned cubes centered at the origin. The cubes in the first layer have the minimum width to contain the training data, while the cubes in subsequent layers have width 2. The number of Monte Carlo samples was taken to be $4(n_{\mathrm{params}} + 1)$, where n_{params} is the number of trainable parameters in the respective component F_{eq} or F_{inv}.

We consider several architectures such as *4-2-1*, meaning there are 4 EP blocks in the first EP layer, 2 EP blocks in the second EP layer, and a final affine EP output layer. These are compared to a standard fully connected MLP with layer widths 12-42-42-42-12 and the Swish activation function from Ramachandran et al. (2017). The MLP was trained and initialized in the same way as the EPNNs, except that the nuclear norm term was not included in the loss function (49). The number of parameters for each network is reported in Table 2. All networks were trained using a single NVIDIA RTX 2080 Ti GPU. The EPNNs took roughly 10 minutes to train—about 5× longer than the MLP with identical hardware.

architecture:	3-2-2-1	5-1-1	3-3-1	4-2-1	2-2-2-2-1	12-1	12-42-42-42-12 MLP
parameters:	4585	4014	4626	4914	2474	4704	4674

Table 2: Number of trainable parameters in each network for the double-spring pendulum.

10.3.4 Results

We compare the performance of different SO(3)-EPNN architectures trained on $M = 100$ trajectory chunks in Figure 9(b). In each random experiment, we follow Finzi et al. (2021) and compute the geometric mean of the relative error for predictions $\hat{\boldsymbol{x}}(t)$ along the $30s$ testing trajectory $\boldsymbol{x}(t)$ given by

$$\text{Relative Error} = \exp\left[\frac{1}{150}\sum_{k=1}^{150}\ln\left(\frac{\|\boldsymbol{x}(k\Delta t) - \hat{\boldsymbol{x}}(k\Delta t)\|}{\|\boldsymbol{x}(k\Delta t)\| + \|\hat{\boldsymbol{x}}(k\Delta t)\|}\right)\right]. \tag{50}$$

Comparable performance is observed across a wide range of different architectures. We do not use these results to perform architecture selection, and the 3-3-1 EPNN is used as a representative architecture for the following sample-efficiency comparisons.

The sample efficiency of the MLP and 3-3-1 EPNNs with different candidate symmetry groups is compared in Figures 9(c) and 9(d) by training on different numbers of trajectory chunks. This allows us to study the degree to which larger candidate symmetry groups can still serve as useful inductive biases to improve generalization using smaller training data sets. We report results across experiments using the same error metric (50). The MLP exhibits an abrupt loss of performance when the number of training chunks is reduced from $M = 100$ to $M = 50$, while the performance of EPNNs deteriorates more gradually. Using smaller candidate symmetry groups yields performance benefits in terms of median relative error when $M = 50$ or fewer trajectory chunks are used for training. These effects are most significant for the $1\,\text{s}$ chunk-prediction task shown in Figure 9(d).

11. Generalization to Sections of Vector Bundles

The machinery for enforcing, discovering, and promoting symmetry of maps $F : \mathcal{V} \to \mathcal{W}$ between finite-dimensional vector spaces is a special case of more general machinery for sections of vector bundles presented here. Applications of this more general framework include studying the symmetries of vector fields, tensor fields, dynamical systems, and integral operators on manifolds with respect to nonlinear group actions. We rely heavily on background, definitions, and results that can be found in Lee (2013), Abraham et al. (1988), and Kolář et al. (1993).

First, we provide some background on smooth vector bundles. A rank-k smooth vector bundle E is a collection of k-dimensional vector spaces E_p, called *fibers*, organized smoothly over a base manifold $\mathcal{M}$. These fibers are organized by the *bundle projection* $\pi : E \to \mathcal{M}$, a surjective map whose preimages are the fibers $E_p = \pi^{-1}(p)$. Smoothness means that π is a smooth submersion and that E is a smooth manifold covered by smooth local trivializations

$$\psi_\alpha : \pi^{-1}(\mathcal{U}_\alpha) \subset E \to \mathcal{U}_\alpha \times \mathbb{R}^k,$$

with $\{\mathcal{U}_\alpha\}_{\alpha \in \mathcal{A}}$ being open subsets covering $\mathcal{M}$. The transition functions between local trivializations are $\mathbb{R}^k$-linear, meaning that there are smooth matrix-valued functions $\boldsymbol{T}_{\alpha,\beta} : \mathcal{U}_\alpha \cap \mathcal{U}_\beta \to \mathbb{R}^{k \times k}$ satisfying

$$\psi_\alpha \circ \psi_\beta^{-1}(p, \boldsymbol{v}) = (p, \boldsymbol{T}_{\alpha,\beta}(p)\boldsymbol{v})$$

for every $p \in \mathcal{U}_\alpha \cap \mathcal{U}_\beta$ and $\boldsymbol{v} \in \mathbb{R}^k$. The bundle with this structure is often denoted $\pi : E \to \mathcal{M}$.

A *section* of the rank-k vector bundle $\pi : E \to \mathcal{M}$ is a map $F : \mathcal{M} \to E$ satisfying $\pi \circ F = \mathrm{Id}_{\mathcal{M}}$. The space of (possibly rough) sections, denoted $\Sigma(E)$, is a vector space with addition and scalar multiplication defined pointwise in each fiber E_p. We equip $\Sigma(E)$ with the topology of pointwise convergence, making it into a locally convex space. The space of sections possessing m continuous derivatives is denoted $C^m(\mathcal{M}, E)$, with the space of merely continuous sections being $C(\mathcal{M}, E) = C^0(\mathcal{M}, E)$ and the space of smooth sections being $C^\infty(\mathcal{M}, E)$. We use $\mathfrak{X}(\mathcal{M}) = C^\infty(\mathcal{M}, T\mathcal{M})$ to denote the space of smooth vector fields on a manifold. A vector bundle and a section are depicted in Figure 10, along with the fundamental operators for a group action introduced below.

We consider a smooth *fiber-linear* right G-action $\Theta : E \times G \to E$, meaning that every $\Theta_g = \Theta(\cdot, g) : E \to E$ is a smooth vector bundle homomorphism. In other words, Θ descends under the bundle projection π to a unique smooth right G-action $\theta : \mathcal{M} \times G \to \mathcal{M}$ so that the diagram

$$
\begin{array}{ccc}
E & \xrightarrow{\ \Theta_g\ } & E \\
{\scriptstyle \pi}\downarrow & & \downarrow{\scriptstyle \pi} \\
\mathcal{M} & \xrightarrow[\ \theta_g\ ]{} & \mathcal{M}
\end{array}
$$

commutes and the restricted maps $\Theta_g|_{E_p} : E_p \to E_{\theta(p,g)}$ are linear. It is clear that (9) is such an action on the trivial vector bundle $\pi : \mathcal{V} \times \mathcal{W} \to \mathcal{V}$. We define what it means for a section to be symmetric with respect to the action Θ as follows:

Definition 22 *A section $F \in \Sigma(E)$ is **equivariant** with respect to an element $g \in G$ if*

$$
\mathcal{K}_g F := \Theta_{g^{-1}} \circ F \circ \theta_g = F. \tag{51}
$$

These transformations form a subgroup of G denoted $\mathrm{Sym}_G(F)$.

The operators $\mathcal{K}_g$ are depicted in Figure 10, which is analogous to Figure 2. Thanks to the vector bundle homomorphism properties of $\Theta_{g^{-1}}$, the operators $\mathcal{K}_g : \Sigma(E) \to \Sigma(E)$ are well-defined and linear. Moreover, they form a group under composition $\mathcal{K}_{g_1} \mathcal{K}_{g_2} = \mathcal{K}_{g_1 \cdot g_2}$, with inverses $\mathcal{K}_g^{-1} = \mathcal{K}_{g^{-1}}$.

The *infinitesimal generator* of the group action is the linear map $\hat{\Theta} : \mathrm{Lie}(G) \to \mathfrak{X}(E)$ defined by

$$
\hat{\Theta}(\xi) = \left. \frac{\mathrm{d}}{\mathrm{d}t} \right|_{t=0} \Theta_{\exp(t\xi)}.
$$

It turns out that this vector field is Θ-related to $(0, \xi) \in \mathfrak{X}(\mathcal{M} \times G)$ (Lemma 5.13 in Kolář et al., 1993; Lemma 20.14 in Lee, 2013), meaning that the flow of $\hat{\Theta}(\xi)$ is given by

$$
\mathrm{Fl}^t_{\hat{\Theta}(\xi)} = \Theta_{\exp(t\xi)}. \tag{52}
$$

Likewise, $\theta_{\exp(t\xi)}$ is the flow of $\hat{\theta}(\xi) = \left. \frac{\mathrm{d}}{\mathrm{d}t} \right|_{t=0} \theta_{\exp(t\xi)} \in \mathfrak{X}(\mathcal{M})$, which is π-related to $\hat{\Theta}(\xi)$.
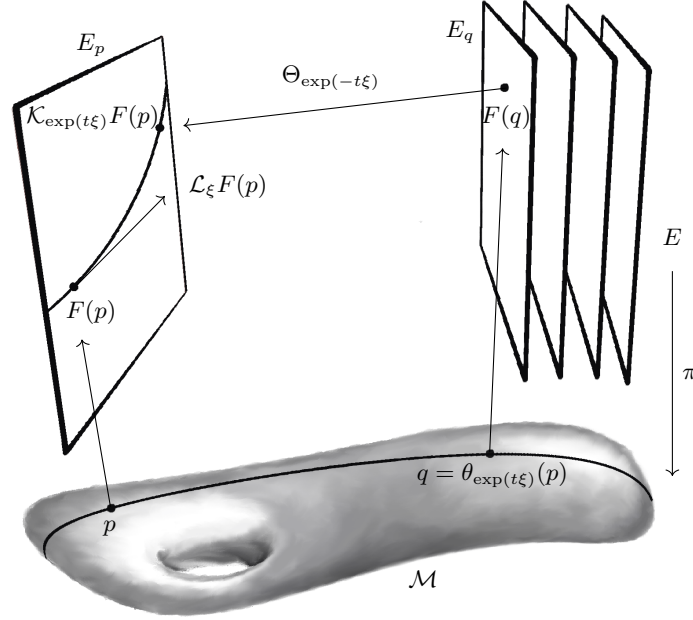
Figure 10: Fundamental operators for sections of vector bundles equipped with fiber-linear Lie group actions. The action Θ_g is linear on each fiber E_p and descends under the bundle projection $\pi : E \to \mathcal{M}$ to an action θ_g on $\mathcal{M}$. Given a section $F : \mathcal{M} \to E$, the finite transformation operators $\mathcal{K}_g$ produce a new section $\mathcal{K}_g F$ whose value at p is given by evaluating F at $q = \theta_g(p)$ and pulling the value in E_q back to E_p via the linear map $\Theta_{g^{-1}}$ on E_q. The operators $\mathcal{K}_g$ are linear thanks to linearity of $\Theta_{g^{-1}}$ on every E_q. The Lie derivative $\mathcal{L}_\xi$ is the operator on sections formed by differentiating $t \mapsto \mathcal{K}_{\exp(t\xi)}$ at $t = 0$. Geometrically, $\mathcal{L}_\xi F(p)$ is the vector in E_p lying tangent to the curve $t \mapsto \mathcal{K}_{\exp(t\xi)} F(p)$ in E_p passing through $F(p)$ at $t = 0$.

Differentiating the smooth curves $t \mapsto \mathcal{K}_{\exp(t\xi)} F(p)$ lying in E_p for each $p \in \mathcal{M}$ gives rise to the Lie derivative $\mathcal{L}_\xi : D(\mathcal{L}_\xi) \subset \Sigma(E) \to \Sigma(E)$ along $\xi \in \text{Lie}(G)$ defined by

$$\mathcal{L}_\xi F = \left. \frac{\mathrm{d}}{\mathrm{d}t} \right|_{t=0} \mathcal{K}_{\exp(t\xi)} F = \lim_{t \to 0} \frac{1}{t} \left(\Theta_{\exp(-t\xi)} \circ F \circ \theta_{\exp(t\xi)} - F \right), \tag{53}$$

where $F \in D(\mathcal{L}_\xi)$ if and only if the limit converges in $\Sigma(E)$, i.e., pointwise. Note that we implicitly identify $T_{F(p)} E_p \cong E_p$. This construction is illustrated in Figure 10. We emphasize that the Lie derivative (53) is a linear operator on sections of the vector bundle E. Moreover, it is linear with respect to ξ, as we will show in Proposition 24. This allows us to directly apply approaches described in Sections 6-8 to enforce, discover, and promote symmetry using Lie derivatives for more general sections of vector bundles. In fact, the fundamental operators described in Section 5 are versions of the operators described here for smooth

functions $F : \mathcal{V} \to \mathcal{W}$ viewed as sections $x \mapsto (x, F(x))$ of the trivial bundle $\pi : \mathcal{V} \times \mathcal{W} \to \mathcal{V}$ and acted upon by the fiber-linear right G-action (9).

Remark 23 (Lie derivatives using flows) *Thanks to* (52), *the Lie derivative defined in* (53) *only depends on the infinitesimal generator* $\hat{\Theta}(\xi) \in \mathfrak{X}(E)$, *and its flow for small time t. Hence, any vector field in* $\mathfrak{X}(E)$ *whose flow is fiber-linear, but not necessarily defined for all* $t \in \mathbb{R}$, *gives rise to an analogously-defined Lie derivative acting linearly on* $\Sigma(E)$. *These are the so-called* linear vector fields *described by Kolář et al. (1993) in Section 47.9. In fact, more general versions of the Lie derivative based on flows for maps between manifolds are described by Kolář et al. (1993) in Chapter 11. However, these generalizations are nonlinear operators, destroying the convex properties of the symmetry-promoting regularization functions in Section 8.*

Some useful properties of the fundamental operators $\mathcal{K}_g$ and $\mathcal{L}_\xi$ for studying symmetries are:

Proposition 24 *For every* $\xi, \eta \in \mathrm{Lie}(G)$, *and* $\alpha, \beta, t \in \mathbb{R}$, *we have*

$$\frac{\mathrm{d}}{\mathrm{d}\,t} \mathcal{K}_{\exp(t\xi)} F = \mathcal{L}_\xi \mathcal{K}_{\exp(t\xi)} F = \mathcal{K}_{\exp(t\xi)} \mathcal{L}_\xi F \qquad \forall F \in D(\mathcal{L}_\xi), \tag{54}$$

$$\mathcal{L}_{\alpha\xi+\beta\eta} F = (\alpha\mathcal{L}_\xi + \beta\mathcal{L}_\eta) F \qquad \forall F \in C^1(\mathcal{M}, E), \tag{55}$$

and

$$\mathcal{L}_{[\xi,\eta]} F = (\mathcal{L}_\xi \mathcal{L}_\eta - \mathcal{L}_\eta \mathcal{L}_\xi) F \qquad \forall F \in C^2(\mathcal{M}, E). \tag{56}$$

We give a proof in Appendix D.

Taken together, these results mean that $\Pi : g \mapsto \mathcal{K}_g$ and $\Pi_* : \xi \mapsto \mathcal{L}_\xi$ are (infinite-dimensional) representations of G and $\mathrm{Lie}(G)$ in $C^\infty(\mathcal{M}, E)$.

The main results of this section are the following two theorems. The first completely characterizes the identity component of $\mathrm{Sym}_G(F)$ by correspondence with its Lie subalgebra (Theorem 19.26; Lee, 2013). The second gives necessary and sufficient conditions for a section to be G-equivariant.

Theorem 25 *If* $F \in C(\mathcal{M}, E)$ *is a continuous section, then* $\mathrm{Sym}_G(F)$ *is a closed, embedded Lie subgroup of G whose Lie subalgebra is*

$$\mathfrak{sym}_G(F) = \{\xi \in \mathrm{Lie}(G) \ : \ F \in D(\mathcal{L}_\xi) \text{ and } \mathcal{L}_\xi F = 0\}. \tag{57}$$

We give a proof in Appendix D.

Theorem 26 *Suppose that G has n_G connected components with G_0 being the one that contains the identity element. Let $\xi_1, \ldots, \xi_q$ generate* $\mathrm{Lie}(G)$ *and let $g_1, \ldots, g_{n_G-1}$ be elements from each non-identity component of G. A continuous section $F \in C(\mathcal{M}, E)$ is G_0-equivariant if and only if*

$$F \in D(\mathcal{L}_{\xi_i}) \quad and \quad \mathcal{L}_{\xi_i} F = 0, \qquad i = 1, \ldots, q. \tag{58}$$

If, in addition, we have

$$\mathcal{K}_{g_j} F = F, \qquad j = 1, \dots, n_G - 1, \tag{59}$$

then F is G-equivariant. We give a proof in Appendix D.

These results allow us to promote, enforce, and discover symmetries for sections of vector bundles in fundamentally the same way we did for maps between finite-dimensional vector spaces in Sections 6, 7, and 8. In particular, symmetry can be enforced through analogous linear constraints, discovered through nullspaces of analogous operators, and promoted through analogous convex penalties based on the nuclear norm.

Remark 27 (Left actions) *Theorems 26 and 25 hold without any modification for left G-actions $\Theta^L : G \times E \to E$. This is because we can define a corresponding right G-action by $\Theta^R(p, g) = \Theta^L(g^{-1}, p)$ with associated operators related by*

$$\mathcal{K}_g^R = \mathcal{K}_{g^{-1}}^L \qquad and \qquad \mathcal{L}_\xi^R = -\mathcal{L}_\xi^L.$$

The symmetry group $\mathrm{Sym}_G(F)$ does not depend on whether it is defined by the condition $\mathcal{K}_g^R F = F$ or by $\mathcal{K}_g^L F = F$. It is slightly less natural to work with left actions because $\Pi^L : g \mapsto \mathcal{K}_g^L$ and $\Pi_*^L : \xi \mapsto \mathcal{L}_\xi^L$ are Lie group and Lie algebra anti-homomorphisms, that is,

$$\Pi^L(g_1 g_2) = \Pi^L(g_2) \Pi^L(g_1) \qquad and \qquad \Pi_*^L([\xi, \eta]) = [\Pi_*^L(\eta), \Pi_*^L(\xi)].$$

11.1 Vector Fields

Here we study the symmetries of a vector field $V \in \mathfrak{X}(\mathcal{M})$ under a right G-action $\theta : \mathcal{M} \times G \to \mathcal{M}$. This allows us to extend the discussion in Section 7.3 to dynamical systems described by vector fields on smooth manifolds and acted upon nonlinearly by arbitrary Lie groups. The tangent map of the diffeomorphism $\theta_g = \theta(\cdot, g) : \mathcal{M} \to \mathcal{M}$ transforms vector fields via the pushforward map $(\theta_g)_* : \mathfrak{X}(\mathcal{M}) \to \mathfrak{X}(\mathcal{M})$ defined by

$$((\theta_g)_* V)_{p \cdot g} = \mathrm{d}\, \theta_g(p) V_p$$

for every $p \in \mathcal{M}$.

Definition 28 *Given $g \in G$, we say that a vector field $V \in \mathfrak{X}(\mathcal{M})$ is g-**invariant** if and only if $(\theta_g)_* V = V$, that is,*

$$V_{p \cdot g} = \mathrm{d}\, \theta_g(p) V_p \qquad \forall p \in \mathcal{M}.$$

Because $(\theta_{g^{-1}})_* (\theta_g)_* = (\theta_{g^{-1}} \circ \theta_g)_* = \mathrm{Id}_{\mathfrak{X}(\mathcal{M})}$, it is clear that a vector field is g-invariant if and only if it is g^{-1}-invariant.

Recall that vector fields $V \in \mathfrak{X}(\mathcal{M})$ are smooth sections of the tangent bundle $E = T\mathcal{M}$. The right G-action θ on $\mathcal{M}$ induces a right G-action $\Theta : T\mathcal{M} \times G \to T\mathcal{M}$ on the tangent bundle defined by

$$\Theta_g(v_p) = \mathrm{d}\, \theta_g(p) v_p.$$

It is easy to see that each Θ_g is a vector bundle homomorphism descending to θ_g under the bundle projection π. The key observation that

$$\mathcal{K}_g V = \Theta_{g^{-1}} \circ V \circ \theta_g = (\theta_{g^{-1}})_* V$$

means a vector field $V \in \mathfrak{X}(\mathcal{M})$ is g-invariant if and only if it is g-equivariant as a section of $T\mathcal{M}$ with respect to the action Θ. Recall that (by Lemma 20.14 in Lee, 2013) the left-invariant vector field $\xi \in \mathrm{Lie}(G) \subset \mathfrak{X}(G)$ and its infinitesimal generator $\hat{\theta}(\xi) \in \mathfrak{X}(\mathcal{M})$ are $\theta^{(p)}$-related, where $\theta^{(p)} : g \mapsto \theta(p, g)$ is the orbit map. This means that $\theta_{\exp(t\xi)}$ is the time-t flow of $\hat{\theta}(\xi)$ by Proposition 9.13 in Lee (2013). As a result, the Lie derivative in (53) agrees with the standard Lie derivative of V along $\hat{\theta}(\xi)$ (p.228; Lee, 2013), that is,

$$\mathcal{L}_\xi V(p) = \lim_{t \to 0} \frac{1}{t} \left[\mathrm{d}\,\theta_{\exp(-t\xi)}(\theta_{\exp(t\xi)}(p)) V_{\theta_{\exp(t\xi)}(p)} - V_p \right] = [\hat{\theta}(\xi), V]_p, \qquad (60)$$

where the expression on the right is the Lie bracket of $\hat{\theta}(\xi)$ and V evaluated at $p \in \mathcal{M}$.

11.2 Tensor Fields

Symmetries of a tensor field can also be revealed using our framework. This will be useful for our study of Hamiltonian dynamics in Section 11.3 and for our study of integral operators, whose kernels can be viewed as tensor fields, in Section 11.4. For simplicity, we consider a rank-k covariant tensor field $A \in \mathfrak{T}^k(\mathcal{M})$, although our results extend to contravariant and mixed tensor fields with minimal modification. We rely on the basic definitions and machinery found in Lee (2013, Chapter 12). Under a right G-action θ on $\mathcal{M}$, the tensor field transforms via the pullback map $\theta_g^* : \mathfrak{T}^k(\mathcal{M}) \to \mathfrak{T}^k(\mathcal{M})$ defined by

$$(\theta_g^* A)_p(v_1, \ldots, v_k) = (\mathrm{d}\,\theta_g(p)^* A_{p \cdot g})(v_1, \ldots, v_k) = A_{p \cdot g}(\mathrm{d}\,\theta_g(p) v_1, \ldots, \mathrm{d}\,\theta_g(p) v_k)$$

for every $v_1, \ldots, v_k \in T_p\mathcal{M}$.

Definition 29 *Given $g \in G$, a tensor field $A \in \mathfrak{T}^k(\mathcal{M})$ is g-**invariant** if and only if $\theta_g^* A = A$, i.e,*

$$A_{p \cdot g}(\mathrm{d}\,\theta_g(p) v_1, \ldots, \mathrm{d}\,\theta_g(p) v_k) = A_p(v_1, \ldots, v_k) \qquad \forall p \in \mathcal{M}.$$

To study the invariance of tensor fields in our framework, we recall that a tensor field is a section of the tensor bundle $E = T^k T^* \mathcal{M} = \coprod_{p \in \mathcal{M}} (T_p^* \mathcal{M})^{\otimes k}$, a vector bundle over $\mathcal{M}$, where $T_p^* \mathcal{M}$ is the dual space of $T_p \mathcal{M}$. The right G-action θ on $\mathcal{M}$ induces a right G-action $\Theta : T^k T^* \mathcal{M} \times G \to T^k T^* \mathcal{M}$ defined by

$$\Theta_g(A_p) = \mathrm{d}\,\theta_{g^{-1}}(\theta_g(p))^* A_p.$$

It is clear that each Θ_g is a homomorphism of the vector bundle $T^k T^* \mathcal{M}$ descending to θ_g under the bundle projection. The key observation that

$$\mathcal{K}_g A = \Theta_{g^{-1}} \circ A \circ \theta_g = \theta_g^* A$$

means a tensor field $A \in \mathfrak{T}^k(\mathcal{M})$ is g-invariant if and only if it is g-equivariant as a section of $T^k T^* \mathcal{M}$ with respect to the action Θ. Since $\theta_{\exp(t\xi)}$ gives the time-t flow of the vector field $\hat{\theta}(\xi) \in \mathfrak{X}(\mathcal{M})$, the Lie derivative in (53) for this action agrees with the standard Lie derivative of $A \in \mathfrak{T}^k(\mathcal{M})$ along $\hat{\theta}(\xi)$ (p.321; Lee, 2013), that is

$$\boxed{(\mathcal{L}_\xi A)_p = \lim_{t \to 0} \frac{1}{t} \left[d\, \theta_{\exp(t\xi)}(p)^* A_{\theta_{\exp(t\xi)}(p)} - A_p \right] = (\mathcal{L}_{\hat{\theta}(\xi)} A)_p.} \tag{61}$$

The Lie derivative for arbitrary covariant tensor fields can be computed by applying Proposition 12.32 in Lee (2013) and its corollaries. More generally, thanks to Results 6.16-18 in Kolář et al. (1993), the Lie derivative for any tensor product of sections of natural vector bundles can be computed via the formula

$$\mathcal{L}_\xi(A_1 \otimes A_2) = (\mathcal{L}_\xi A_1) \otimes A_2 + A_1 \otimes (\mathcal{L}_\xi A_2).$$

For example, this holds when A_1, A_2 are arbitrary smooth tensor fields of mixed types. The Lie derivative of a differential form ω on $\mathcal{M}$ can be computed by Cartan's magic formula

$$\mathcal{L}_\xi \omega = \hat{\theta}(\xi) \lrcorner (d\,\omega) + d(\hat{\theta}(\xi) \lrcorner \omega), \tag{62}$$

where d is the exterior derivative.

11.3 Hamiltonian Dynamics

The dynamics of frictionless mechanical systems can be described by Hamiltonian vector fields on symplectic manifolds. Roughly speaking, these encompass systems that conserve energy, such as motion of rigid bodies and particles interacting via conservative forces. The celebrated theorem of Noether (1918) says that conserved quantities of Hamiltonian systems correspond with symmetries of the energy function (the system's Hamiltonian). In this section, we briefly illustrate how to enforce Hamiltonicity constraints on learned dynamical systems and how to promote, discover, and enforce conservation laws. A thorough treatment of classical mechanics, symplectic manifolds, and Hamiltonian systems can be found in Abraham and Marsden (2008); Marsden and Ratiu (1999). This reference also includes methods to reduce systems with known symmetries and conservation laws. The following brief introduction follows Chapter 22 of Lee (2013).

Hamiltonian systems are defined on symplectic manifolds. That is, a smooth even-dimensional manifold $\mathcal{S}$ together with a smooth, nondegenerate, closed differential 2-form ω, called the symplectic form. Nondegeneracy means that the map $\hat{\omega}_p : v \mapsto \omega_p(v, \cdot)$ is a bijective linear map of $T_p \mathcal{S}$ onto its dual $T_p^* \mathcal{S}$ for every $p \in \mathcal{S}$. Closure means that $d\,\omega = 0$, where d is the exterior derivative. Thanks to nondegeneracy, any smooth function $H \in C^\infty(\mathcal{S})$ gives rise to a smooth vector field

$$V_H = \hat{\omega}^{-1}(d\,H)$$

called the *Hamiltonian vector field* of H. A vector field $V \in \mathfrak{X}(\mathcal{S})$ is said to be Hamiltonian if $V = V_H$ for some function H, called the Hamiltonian of V. A vector field is locally Hamiltonian if it is Hamiltonian in a neighborhood of each point in $\mathcal{S}$.

The symplectic manifolds considered in classical mechanics usually consist of the cotangent bundle $\mathcal{S} = T^*\mathcal{M}$ of an m-dimensional manifold $\mathcal{M}$ describing the *configuration* of the system, e.g., the positions of particles. The cotangent bundle has a canonical symplectic form given by

$$\omega = \sum_{i=1}^{m} \mathrm{d}\, x^i \wedge \mathrm{d}\, \xi_i, \tag{63}$$

where (x^i, ξ_i) are any choice of natural coordinates on a patch of $T^*\mathcal{M}$ (see Proposition 22.11 in Lee, 2013). Here, each x^i is a generalized coordinate describing the configuration and ξ_i is its *conjugate* or *generalized momentum*. The Darboux theorem (Theorem 22.13; Lee, 2013) says that any symplectic form on a manifold can be put into the form of (63) by a choice of local coordinates. In these *Darboux coordinates*, the dynamics of a Hamiltonian system are governed by the equations

$$\frac{\mathrm{d}}{\mathrm{d}\,t}\, x^i = V_H(x^i) = \frac{\partial H}{\partial \xi_i}, \qquad \frac{\mathrm{d}}{\mathrm{d}\,t}\, \xi_i = V_H(\xi_i) = -\frac{\partial H}{\partial x^i},$$

which should be familiar to anyone who has studied undergraduate mechanics.

Enforcing local Hamiltonian structure on a vector field $V \in \mathfrak{X}(\mathcal{S})$ is equivalent to the linear constraint

$$\mathcal{L}_V \omega = 0$$

thanks to Proposition 22.17 in Lee (2013). Here $\mathcal{L}_V$ is the Lie derivative of the tensor field $\omega \in \mathfrak{T}^2(\mathcal{S})$, i.e., (61) with θ being the flow of V and its generator being the identity $\hat{\theta}(V) = V$. Note that the Lie derivative still makes sense even when the orbits $t \mapsto \theta_t(p) = \theta_{\exp(t1)}(p)$ are only defined for small $t \in (-\varepsilon, \varepsilon)$. In Darboux coordinates, this constraint is equivalent to the set of equations

$$\frac{\partial V(x^i)}{\partial x^j} + \frac{\partial V(\xi_j)}{\partial \xi_i} = 0, \qquad \frac{\partial V(\xi_i)}{\partial x^j} - \frac{\partial V(\xi_j)}{\partial x^i} = 0, \qquad \frac{\partial V(x^i)}{\partial \xi_j} - \frac{\partial V(x^j)}{\partial \xi_i} = 0$$

for all $1 \leq i, j \leq m$. When the first de Rham cohomology group satisfies $H^1_{\mathrm{dR}}(\mathcal{S}) = 0$, for example when $\mathcal{S}$ is contractible, local Hamiltonicity implies the existence of a global Hamiltonian for V, unique on each component of $\mathcal{S}$ up to addition of a constant by Lee (2013, Proposition 22.17).

Of course, our approach also makes it possible to promote Hamiltonian structure with respect to candidate symplectic forms when learning a vector field V. To do this, we can penalize the nuclear norm of $\mathcal{L}_V$ restricted to a subspace $\tilde{\Omega}$ of candidate closed 2-forms using the regularization function

$$R_{\tilde{\Omega},*}(V) = \big\| \, \mathcal{L}_V|_{\tilde{\Omega}} \, \big\|_*.$$

The strength of this penalty can be increased when solving a regression problem for V until there is a nondegenerate 2-form in the nullspace $\mathrm{Null}(\mathcal{L}_V) \cap \tilde{\Omega}$. This gives a symplectic form with respect to which V is locally Hamiltonian.

Another option is to learn a (globally defined) Hamiltonian function H directly by fitting V_H to data. In this case, we can regularize the learning problem by penalizing the breaking of conservation laws. The time derivative of a quantity, that is, a smooth function $f \in C^\infty(\mathcal{S})$

48

under the flow of V_H is given by the Poisson bracket

$$\{f, H\} := \omega(V_f, V_H) = \mathrm{d}\, f(V_H) = V_H(f) = -V_f(H).$$

Hence, f is a conserved quantity if and only if H is invariant under the flow of V_f—this is Noether's theorem. It is also evident that the Poisson bracket is linear with respect to both of its arguments. In fact, the Poisson bracket turns $C^\infty(\mathcal{S})$ into a Lie algebra with $f \mapsto V_f$ being a Lie algebra homomorphism, i.e., $V_{\{f,g\}} = [V_f, V_g]$.

As a result of these basic properties of the Poisson bracket, the quantities conserved by a given Hamiltonian vector field V_H form a Lie subalgebra given by the nullspace of a linear operator $L_H : C^\infty(\mathcal{S}) \mapsto C^\infty(\mathcal{S})$ defined by

$$L_H : f \mapsto \{f, H\}.$$

To promote conservation of quantities in a given subalgebra $\mathfrak{g} \subset C^\infty(\mathcal{S})$ when learning a Hamiltonian H, we can penalize the nuclear norm of L_H restricted to $\mathfrak{g}$, that is

$$R_{\mathfrak{g},*}(H) = \big\| \, L_H\big|_{\mathfrak{g}} \, \big\|_*.$$

For example, we might expect a mechanical system to conserve angular momentum about some axes, but not others, due to applied torques. In the absence of data to the contrary, it often makes sense to assume that various linear and angular momenta are conserved.

11.4 Multilinear Integral Operators

In this section, we provide machinery to study the symmetries of linear and nonlinear integral operators acting on sections of vector bundles, yielding far-reaching generalizations of our results in Section 6.2. Such operators can form the layers of neural networks acting on various vector and tensor fields supported on manifolds.

Let $\pi_0 : E_0 \to \mathcal{M}_0$ and $\pi_j : E_j \to \mathcal{M}_j$ be vector bundles with $\mathcal{M}_j$ being d_j-dimensional orientable Riemannian manifolds with volume forms $\mathrm{dV}_j \in \Omega^{d_j}(T^*\mathcal{M}_j)$, $j = 1, \ldots, r$. Note that here dV_j does not denote the exterior derivative of a $(d_j - 1)$-form. A continuous section K of the bundle

$$E = E_0 \otimes E_1^* \otimes \cdots \otimes E_r^* := \coprod_{(p,q_1,\ldots,q_r) \in \mathcal{M}_0 \times \cdots \times \mathcal{M}_r} E_{0,p} \otimes E_{1,q_1}^* \otimes \cdots \otimes E_{r,q_r}^*$$

can be viewed as a continuous family of r-multilinear maps

$$K(p, q_1, \ldots, q_r) : \bigoplus_{j=1}^r E_{j,q_j} \to E_{0,p}.$$

The section K can serve as the kernel to define an r-multilinear integral operator $\mathcal{T}_K :$ $D(\mathcal{T}_K) \subset \bigoplus_{j=1}^r \Sigma(E_j) \to \Sigma(E_0)$ with action on $(F_1, \ldots, F_r) \in D(\mathcal{T}_K)$ given by

$$
\mathcal{T}_K[F_1, \ldots, F_r](p) =
$$
$$
\int_{\mathcal{M}_1 \times \cdots \times \mathcal{M}_r} K(p, q_1, \ldots, q_r)[F_1(q_1), \ldots, F_r(q_r)] \, \mathrm{dV}_1(q_1) \wedge \cdots \wedge \mathrm{dV}_r(q_r). \quad (64)
$$

This operator is linear when $r = 1$. When $r > 1$ and $E_1 = \cdots = E_r$, (64) can be used to define a nonlinear integral operator $\Sigma(E_1) \to \Sigma(E_0)$ with action $F \mapsto \mathcal{T}_K[F, \ldots, F]$.

Given fiber-linear right G-actions $\Theta_j : E_j \times G \to E_j$, there is an induced fiber-linear right G-action $\Theta : E \times G \to E$ on E defined by

$$
\Theta_g(K_{p,q_1,\ldots,q_r})[v_1, \ldots, v_r] = \Theta_{0,g}\left(K_{p,q_1,\ldots,q_r}[\Theta_{1,g^{-1}}(v_1), \ldots, \Theta_{r,g^{-1}}(v_r)] \right)
$$

for $K_{p,q_1,\ldots,q_r} \in E$ viewed as an r-multilinear map $E_{1,q_1} \oplus \cdots \oplus E_{r,q_r} \to E_{0,p}$ and $v_j \in E_{j,\theta_{j,g}(q_j)}$. Sections $F_j \in \Sigma(E_j)$ transform according to

$$
\mathcal{K}_g^{E_j} F_j = \Theta_{j,g^{-1}} \circ F_j \circ \theta_{j,g},
$$

with the section defining the integral kernel transforming according to

$$
\mathcal{K}_g^E K(p, q_1, \ldots, q_r)[v_{q_1}, \ldots, v_{q_r}] =
$$
$$
\Theta_{0,g^{-1}}\Big\{ K\big(\theta_{0,g}(p), \theta_{1,g}(q_1), \ldots, \theta_{r,g}(q_r)\big)[\Theta_{1,g}(v_{q_1}), \ldots, \Theta_{r,g}(v_{q_r})] \Big\}. \quad (65)
$$

Using these transformation laws, we define equivariance for the integral operator as follows:

Definition 30 *The integral operator in* (64) *is **equivariant** with respect to $g \in G$ if*

$$
\mathcal{K}_g^{E_0} \mathcal{T}_K\big[\mathcal{K}_{g^{-1}}^{E_1} F_1, \ldots, \mathcal{K}_{g^{-1}}^{E_r} F_r\big] = \mathcal{T}_K[F_1, \ldots, F_r]
$$

for every $(F_1, \ldots, F_r) \in D(\mathcal{T}_K)$.

This definition is equivalent to the following condition on the integral kernel:

Lemma 31 *The integral operator* (64) *is equivariant with respect to $g \in G$ if and only if*

$$
\mathcal{K}_g\big(K \, \mathrm{dV}_1 \wedge \cdots \wedge \mathrm{dV}_r\big) := \mathcal{K}_g^E(K) \, \theta_{1,g}^* \mathrm{dV}_1 \wedge \cdots \wedge \theta_{r,g}^* \mathrm{dV}_r = K \, \mathrm{dV}_1 \wedge \cdots \wedge \mathrm{dV}_r, \quad (66)
$$

where $\mathcal{K}_g^E$ is defined by (65). *A proof is available in Appendix A.*

We note that $\mathcal{K}_g^\Omega(\mathrm{dV}_j) = \theta_{j,g}^* \mathrm{dV}_j$ is the natural transformation of the differential form $\mathrm{dV}_j \in \Omega^{d_j}(T^*\mathcal{M}_j)$ (a covariant tensor field) described in Section 11.2. The Lie derivative of this action on volume forms is given by

$$
\mathcal{L}_\xi^\Omega \, \mathrm{dV} = \mathrm{d}\left(\hat{\theta}(\xi) \lrcorner \, \mathrm{dV}\right) = \mathrm{div}\,\hat{\theta}(\xi)\,\mathrm{dV}
$$

50

thanks to Cartan's magic formula and the definition of divergence (see Lee, 2013). Therefore, differentiating (66) along the curve $g(t) = \exp(t\xi)$ yields the Lie derivative

$$
\mathcal{L}_\xi\big(K\,\mathrm{dV}_1 \wedge \cdots \wedge \mathrm{dV}_r\big) = \left[\mathcal{L}_\xi^E(K) + K\sum_{j=1}^r \operatorname{div}\hat{\theta}_j(\xi)\right]\mathrm{dV}_1 \wedge \cdots \wedge \mathrm{dV}_r\,. \tag{67}
$$

Here, $\mathcal{L}_\xi^E$ is the Lie derivative associated with (65). For the integral operators discussed in Section 6.2, the formulas derived here recover Eqs. 19 and 21.

12. Invariant Submanifolds and Tangency

Studying the symmetries of smooth maps can be cast in a more general framework in which we study the symmetries of submanifolds. Specifically, the symmetries of a smooth map $F : \mathcal{M}_0 \to \mathcal{M}_1$ between manifolds correspond to symmetries of its graph, $\operatorname{gr}(F)$, and the symmetries of a smooth section of a vector bundle $F \in C^\infty(\mathcal{M}, E)$ correspond to symmetries of its image, $\operatorname{im}(F)$ — both of which are properly embedded submanifolds of $\mathcal{M}_0 \times \mathcal{M}_1$ and E, respectively. We show that symmetries of a large class of submanifolds, including the above, are revealed by checking whether the infinitesimal generators of the group action are tangent to the submanifold. In this setting, the Lie derivative of $F \in C^\infty(\mathcal{M}, E)$ has a geometric interpretation as a projection of the infinitesimal generator onto the tangent space of the image $\operatorname{im}(F)$, viewed as a submanifold of the bundle.

12.1 Symmetry of Submanifolds

In this section, we study the infinitesimal conditions for a submanifold to be invariant under the action of a Lie group. Suppose that $\mathcal{N}$ is a manifold and $\theta : \mathcal{N} \times G \to \mathcal{N}$ is a right action of a Lie group G on $\mathcal{N}$. Sometimes we denote this action by $p \cdot g = \theta(p, g)$ when there is no ambiguity. Though our results also hold for left actions, as we discuss later in Remark 37, working with right actions is standard in this context and allows us to leverage results from Lee (2013) more naturally in our proofs. Fixing $p \in \mathcal{N}$, the orbit map of this action is denoted $\theta^{(p)} : G \to \mathcal{N}$. Fixing $g \in G$, the map $\theta_g : \mathcal{N} \to \mathcal{N}$ defined by $\theta_g : p \mapsto \theta(p, g)$ is a diffeomorphism with inverse $\theta_{g^{-1}}$.

Definition 32 *A subset $\mathcal{M} \subset \mathcal{N}$ is G-**invariant** if and only if $\theta(p, g) \in \mathcal{M}$ for every $g \in G$ and $p \in \mathcal{M}$.*

Sometimes we will denote $\mathcal{M} \cdot G = \{\theta(p, g) \ : \ p \in \mathcal{M},\ g \in G\}$, in which case G-invariance of $\mathcal{M}$ can be stated as $\mathcal{M} \cdot G \subset \mathcal{M}$.

We study the group invariance of submanifolds of the following type:

Definition 33 *Let $\mathcal{M}$ be a weakly embedded m-dimensional submanifold of an n-dimensional manifold $\mathcal{N}$. We say that $\mathcal{M}$ is **arcwise-closed** if any smooth curve $\gamma : [a, b] \to \mathcal{N}$ satisfying $\gamma((a, b)) \subset \mathcal{M}$ must also satisfy $\gamma([a, b]) \subset \mathcal{M}$.*

Submanifolds of this type include all properly embedded submanifolds of $\mathcal{N}$ because properly embedded submanifolds are closed subsets (Proposition 5.5; Lee, 2013). More interestingly, we have the following:

Proposition 34 *The leaves of any (nonsingular) foliation of $\mathcal{N}$ are arcwise-closed. We provide a proof in Appendix E.*

This means that the kinds of submanifolds we are considering include all possible Lie subgroups (Theorem 19.25; Lee, 2013) as well as their orbits under free and proper group actions (Proposition 21.7; Lee, 2013). The leaves of singular foliations associated with integrable distributions of nonconstant rank (see Sections 3.18–25; Kolář et al., 1993) can fail to be arcwise-closed. For example, the distribution spanned by the vector field $x\frac{\partial}{\partial x}$ on $\mathbb{R}$ has maximal integral manifolds $(-\infty, 0)$, $\{0\}$, and $(0, \infty)$ forming a singular foliation of $\mathbb{R}$. Obviously, the leaves $(-\infty, 0)$ and $(0, \infty)$ are not arcwise-closed.

Given a submanifold and a candidate group of transformations, the following theorem describes the largest connected Lie subgroup of symmetries of the submanifold. Specifically, these symmetries can be identified by checking tangency conditions between infinitesimal generators and the submanifold.

Theorem 35 *Let $\mathcal{M}$ be an immersed submanifold of $\mathcal{N}$ and let $\theta : \mathcal{N} \times G \to \mathcal{N}$ be a right action of a Lie group G on $\mathcal{N}$ with infinitesimal generator $\hat{\theta} : \mathrm{Lie}(G) \to \mathfrak{X}(\mathcal{N})$. Then*

$$\mathfrak{sym}_G(\mathcal{M}) = \left\{ \xi \in \mathrm{Lie}(G) \ : \ \hat{\theta}(\xi)_p \in T_p\mathcal{M} \quad \forall p \in \mathcal{M} \right\} \tag{68}$$

is the Lie subalgebra of a unique connected Lie subgroup $H \subset G$. If $\mathcal{M}$ is weakly embedded and arcwise-closed in $\mathcal{N}$, then this subgroup has the following properties:

(i) $\mathcal{M} \cdot H \subset \mathcal{M}$

(ii) If $\tilde{H}$ is a connected Lie subgroup of G such that $\mathcal{M} \cdot \tilde{H} \subset \mathcal{M}$, then $\tilde{H} \subset H$.

If $\mathcal{M}$ is properly embedded in $\mathcal{N}$ then H is the identity component of the closed, properly embedded Lie subgroup

$$\mathrm{Sym}_G(\mathcal{M}) = \{g \in G \ : \ \mathcal{M} \cdot g \subset \mathcal{M}\}. \tag{69}$$

A proof is provided in Appendix E.

Since the infinitesimal generator $\hat{\theta}$ is a linear map and $T_p\mathcal{M}$ is a subspace of $T_p\mathcal{N}$, the tangency conditions defining the Lie subalgebra (68) can be viewed as a set of linear equality constraints on the elements $\xi \in \mathrm{Lie}(G)$. Hence, $\mathfrak{sym}_G(\mathcal{M})$ can be computed as the nullspace of a positive semidefinite operator on $\mathrm{Lie}(G)$, defined analogously to the case described earlier in Section 7.1.

The following theorem provides necessary and sufficient conditions for arcwise-closed weakly embedded submanifolds to be G-invariant. These are generally nonlinear constraints on the submanifold, regarded as the zero section of its normal bundle under identification with a tubular neighborhood. However, we will show in Section 12.2 that these become linear constraints recovering the Lie derivative when the submanifold in question is the image of a section of a vector bundle and the group action is fiber-linear.

Theorem 36 *Let $\mathcal{M}$ be an arcwise-closed weakly embedded submanifold of $\mathcal{N}$ and let $\theta : \mathcal{N} \times G \to \mathcal{N}$ be a right action of a Lie group G on $\mathcal{N}$ with infinitesimal generator*

$\hat{\theta} : \mathrm{Lie}(G) \to \mathfrak{X}(\mathcal{N})$. *Let* $\xi_1, \ldots, \xi_q$ *generate* $\mathrm{Lie}(G)$ *and let* $g_1, \ldots, g_{n_G-1}$ *be elements from each non-identity component of* G. *Then* $\mathcal{M}$ *is* G_0-*invariant if and only if*

$$\hat{\theta}(\xi_i)_p \in T_p \mathcal{M} \qquad \forall p \in \mathcal{M}, \quad i = 1, \ldots, q. \tag{70}$$

If, in addition, we have $\mathcal{M} \cdot g_j \subset \mathcal{M}$ *for every* $j = 1, \ldots, n_G - 1$, *then* $\mathcal{M}$ *is* G-*invariant. A proof is provided in Appendix E.*

Remark 37 (Left actions) *When the group* G *acts on* $\mathcal{N}$ *from the left according to* $\theta^L : G \times \mathcal{N} \to \mathcal{N}$, *we can always construct an equivalent right action* $\theta^R : \mathcal{N} \times G \to \mathcal{N}$ *by setting* $\theta^R(p, g) = \theta^L(g^{-1}, p)$. *The corresponding infinitesimal generators are related by* $\hat{\theta}^R = -\hat{\theta}^L$. *Since* $\hat{\theta}^L(\xi)_p \in T_p \mathcal{M}$ *if and only if* $\hat{\theta}^R(\xi)_p \in T_p \mathcal{M}$, *Theorems 36 and 35 hold without modification for left* G-*actions.*

12.2 The Lie Derivative as a Projection

We provide a geometric interpretation of the Lie derivative in (53) by expressing it in terms of a projection of the infinitesimal generator of the group action onto the tangent space of $\mathrm{im}(F)$ for smooth sections $F \in C^\infty(\mathcal{M}, E)$. This allows us to connect the Lie derivative to the tangency conditions for symmetry of submanifolds presented in Section 12.1.

The Lie derivative $\mathcal{L}_\xi F(p)$ lies in E_p, while $T_{F(p)} \mathrm{im}(F)$ is a subspace of $T_{F(p)} E$. To relate quantities in these different spaces, the following lemma lifts each E_p to a subspace of $T_{F(p)} E$.

Lemma 38 *For every smooth section* $F \in C^\infty(\mathcal{M}, E)$ *there is a well-defined injective vector bundle homomorphism* $\imath_F : E \to TE$ *that is expressed in any local trivialization* $\Phi : \pi^{-1}(\mathcal{U}) \to \mathcal{U} \times \mathbb{R}^k$ *as*

$$\begin{aligned} \mathrm{d}\, \Phi \circ \imath_F \circ \Phi^{-1} : \mathcal{U} \times \mathbb{R}^k &\to T(\mathcal{U} \times \mathbb{R}^k) \\ (p, \boldsymbol{v}) &\mapsto (0, \boldsymbol{v})_{\Phi(F(p))}. \end{aligned} \tag{71}$$

We give a proof in Appendix E.

This is a special case of the *vertical lift* of E into the vertical bundle $VE = \{v \in TE : \mathrm{d}\,\pi v = 0\}$ described by Kolář et al. (1993) in Section 6.11. The *vertical projection* $\mathrm{vpr}_E : VE \to E$ provides a left-inverse satisfying $\mathrm{vpr}_E \circ \imath_F = \mathrm{Id}_E$.

The following result relates the Lie derivative to a projection via the vertical lifting.

Theorem 39 *Given a smooth section* $F \in C^\infty(\mathcal{M}, E)$ *and* $p \in \mathcal{M}$, *the map* $P_{F(p)} := \mathrm{d}(F \circ \pi)(F(p)) : T_{F(p)} E \to T_{F(p)} E$ *is a linear projection onto* $T_{F(p)} \mathrm{im}(F)$ *and for every* $\xi \in \mathrm{Lie}(G)$ *we have*

$$\imath_F \circ (\mathcal{L}_\xi F)(p) = -\big[I - P_{F(p)}\big] \hat{\Theta}(\xi)_{F(p)}. \tag{72}$$

We give a proof in Appendix E.

For the special case of smooth maps $F : \mathcal{V} \to \mathcal{W}$ between finite-dimensional vector spaces viewed as sections $x \mapsto (x, F(x))$ of the trivial bundle $\pi : \mathcal{V} \times \mathcal{W} \to \mathcal{V}$, this theorem reproduces (25). The following corollary provides a link between our main results for sections of vector bundles and our main results for symmetries of submanifolds.

Corollary 40 *For every smooth section $F \in C^\infty(\mathcal{M}, E)$, $\xi \in \mathrm{Lie}(G)$, and $p \in \mathcal{M}$ we have*

$$(\mathcal{L}_\xi F)(p) = 0 \qquad \Leftrightarrow \qquad \hat{\Theta}(\xi)_{F(p)} \in T_{F(p)} \mathrm{im}(F).$$

In particular, this means that for smooth sections, Theorems 26 and 25 are special cases of Theorems 36 and 35.

13. Conclusion

This paper provides a unified theoretical approach to enforce, discover, and promote symmetries in machine learning models. In particular, we provide theoretical foundations for Lie group symmetry in machine learning from a linear-algebraic viewpoint. This perspective unifies and generalizes several leading approaches in the literature, including approaches for incorporating and uncovering symmetries in neural networks and more general machine learning models. The central objects in this work are linear operators describing the finite and infinitesimal transformations of smooth sections of vector bundles with fiber-linear Lie group actions. To make the paper accessible to a wide range of practitioners, Sections 4–10 deal with the special case where the machine learning models are built using smooth functions between vector spaces. Our main results establish that the infinitesimal operators—the Lie derivatives—fully encode the connected subgroup of symmetries for sections of vector bundles (resp. functions between vector spaces). In other words, the Lie derivatives encode symmetries with respect to which the machine learning models are equivariant.

We illustrate that promoting and enforcing continuous symmetries in large classes of machine learning models are dual problems with respect to the bilinear structure of the Lie derivative. Moreover, these ideas extend naturally to identify continuous symmetries of arbitrary submanifolds, recovering the Lie derivative when the submanifold is the image of a section of a vector bundle (resp., the graph of a function between vector spaces). Using the fundamental operators, we also describe how symmetries can be promoted as inductive biases during training of machine learning models using convex penalties. We also provide rigorous data-driven methods for discretizing and approximating the fundamental operators to accomplish the tasks of enforcing, promoting, and discovering symmetry. We note that these tasks admit computational implementations using standard linear algebra and convex optimization.

We evaluate our new approach to promote symmetry using three numerical examples including symmetric polynomial recovery, extrapolatory recovery of a linear multi-spring-mass dynamical system, and training a new Equivariance-Promoting Neural Network (EPNN) to forecast the dynamics of a double-spring pendulum system. Our numerical results for polynomial recovery show that minimizing convex symmetry-promoting penalties can be used to recover highly symmetric polynomial functions using fewer samples than are required to determine the polynomial coefficients directly as the solution of a linear system. This reduction in sample complexity becomes more pronounced in higher dimensions and with increasing symmetry of the underlying function being recovered. Our numerical results on extrapolatory recovery of the linear spring-mass system illustrate how symmetry-promoting techniques can learn to extrapolate from restricted data, while common sparsity-promoting techniques can fail to do so. The results obtained using EPNNs to model a double-spring

pendulum system indicate that promoting symmetry can provide accuracy benefits in low-data regimes. Smaller candidate groups containing correct symmetries were shown to provide the greatest benefits, with enough tolerance for incorrect symmetries in the candidate group to potentially allow the method to be useful in practice. We note that very large candidate groups were neither beneficial nor harmful compared to EPNNs with no symmetry penalty. Interestingly, EPNNs had significantly better performance than the baseline MLP in low-data regimes, even when no symmetry-promoting penalty was used.

The main limitation of our approach to promote symmetry is the need to choose a sufficiently small candidate group G containing correct symmetries of the underlying system. A second limitation is the significant computational cost to work with the discretized linear operators encoding the symmetries of a model. Another challenge and potentially significant limitation is identifying appropriate spaces of functions $\mathcal{F}$ and model architectures capable of expressing nontrivial symmetric functions. For example, it is possible that the only G-symmetric functions in $\mathcal{F}$ are trivial, meaning that enforcing symmetry results in learning only trivial models. One open question is whether our framework can be used in such cases to learn relaxed symmetries, as described by Wang et al. (2022). In other words, it might be possible to find elements in $\mathcal{F}$ that are nearly symmetric, and to bound the degree of asymmetry based on the quantities derived from the fundamental operators, such as their norms. Finally, our reliance on the Lie algebra to study continuous symmetries also limits the ability of our proposed methods to account for partial symmetries, such as the invariance in classifying the characters "Z" and "N" to rotations by small angles, but not large angles.

In follow-up work, we aim to apply the proposed methods to a wider range of examples to provide detailed comparisons with state-of-the-art techniques and to assess performance trade-offs associated with problem dimension, candidate symmetry group, model size, amount of training data, and noise level. Since our proposed methods remain to be tested in the more general setting of vector bundles, it will also be important to assess their practical usefulness by implementing, evaluating, and comparing them to existing techniques. To enable these experiments, a major goal in future work must be to improve the computational efficiency and implementations of techniques to enforce, discover, and promote symmetry within our framework. Work is currently underway to leverage the sparse structure of the discretized operators in certain bases to enable the use of efficient Krylov subspace algorithms and approximations based on matrix sketching (Halko et al., 2011; Woodruff et al., 2014). It will also be useful to identify efficient optimization algorithms for training symmetry-constrained or symmetry-regularized machine learning models. Promising candidates include projected gradient descent, the Alternating Direction Method of Multipliers (ADMM; Boyd et al., 2011; Fazel et al., 2013), variations of Bregman learning and mirror descent (Bungert et al., 2022; Beck and Teboulle, 2003), as well as Iteratively Reweighted Least Squares (IRLS) algorithms (Mohan and Fazel, 2012). Using iterative reweighting could enable symmetry-promoting penalties to be based on nonconvex Schatten p-"norms" with $0 < p < 1$, potentially improving the recovery of underlying symmetry groups compared to the nuclear norm where $p = 1$.

Additional avenues for future work include applying the proposed techniques to continuum problems and partial differential equations through principled implementation in neural operators and through jet-bundle prolongation (Olver, 1986). Combining analogues of our proposed methods with techniques using the weak formulation proposed by Messenger and Bortz (2021); Reinbold et al. (2020) could also provide robust ways to identify symmetric

PDEs in the presence of high noise and limited training data. It will also be important to study the perturbative effects of noisy data in algorithms to discover and promote symmetry, with the goal of understanding the effects of problem dimension, noise level, and amount of data. Another area of future work could be to build on the work of Peitz et al. (2023); Steyert (2022) by studying symmetry in the setting of Koopman operators for dynamical systems. To do this, one might follow the program set forth by Colbrook (2023), where the measure-preserving property of certain dynamical systems is exploited to enhance the Extended Dynamic Mode Decomposition (EDMD) algorithm of Williams et al. (2015).

Acknowledgements

Appendix A. Proofs of Miscellaneous Results

Proposition 6 *Let K be continuous and suppose that $\mathcal{T}_K$ acts on a function space containing all smooth, compactly supported fields. Then*

$$\mathrm{Sym}_G(\mathcal{T}_K) = \{g \in G \ : \ \mathcal{K}_g K = K\}. \tag{20}$$

Proof Obviously, if $\mathcal{K}_g K = K$ then $\mathcal{K}_g^{(\mathcal{W})} \circ \mathcal{T}_K \circ \mathcal{K}_{g^{-1}}^{(\mathcal{V})} = \mathcal{T}_K$. On the other hand, suppose that $\mathcal{K}_g K(x_0, y_0) \neq K(x_0, y_0)$ for some $(x_0, y_0) \in \mathbb{R}^n \times \mathbb{R}^m$. Hence, there are vectors $v \in \mathcal{V}$ and $w \in \mathcal{W}^*$ such that $\langle w, \ \mathcal{K}_g K(x_0, y_0) v - K(x_0, y_0) v \rangle > 0$. This remains true for all y in a neighborhood $\mathcal{U}$ of y_0 by continuity of K and $\mathcal{K}_g K$. Letting $F(x) = v\varphi(x)$ where φ is a smooth, nonnegative function with $\varphi(y_0) > 0$ and support in $\mathcal{U}$, we obtain

$$\langle w, \ \mathcal{K}_g^{(\mathcal{W})} \circ \mathcal{T}_K \circ \mathcal{K}_{g^{-1}}^{(\mathcal{V})} F(x_0) - \mathcal{T}_K F(x_0) \rangle = \int_{\mathbb{R}^m} \langle w, \ \mathcal{K}_g K(x_0, y) v - K(x_0, y) v \rangle \varphi(y) \, \mathrm{d}y > 0,$$

meaning $\mathcal{K}_g^{(\mathcal{W})} \circ \mathcal{T}_K \circ \mathcal{K}_{g^{-1}}^{(\mathcal{V})} \neq \mathcal{T}_K$. Therefore, $\mathcal{K}_g K = K$ if and only if $\mathcal{K}_g^{(\mathcal{W})} \circ \mathcal{T}_K \circ \mathcal{K}_{g^{-1}}^{(\mathcal{V})} = \mathcal{T}_K$. ∎

Proposition 16 *The subgroup in (30) is closed and embedded in G; its Lie subalgebra is*

$$\bigcap_{l=1}^{L} \mathfrak{sym}_G \left(F^{(l)}\right) = \left\{\xi \in \mathrm{Lie}(G) \ : \ \mathcal{L}_\xi F^{(l)} = 0, \ l = 1, \ldots, L\right\}. \tag{32}$$

Proof As an intersection of closed subgroups, $H := \bigcap_{l=1}^{L} \mathrm{Sym}_G \left(F^{(l)}\right)$ is a closed subgroup of G. By the closed subgroup theorem (Theorem 20.12; Lee, 2013), H is an embedded Lie subgroup, whose Lie subalgebra we denote by $\mathfrak{h}$. If $\xi \in \mathfrak{h}$ then $\exp(t\xi) \in \mathrm{Sym}_G \left(F^{(l)}\right)$ for all $t \in \mathbb{R}$ and every $l = 1, \ldots, L$. Differentiating $\mathcal{K}_{\exp(t\xi)} F^{(l)} = F^{(l)}$ at $t = 0$ proves that $\mathcal{L}_\xi F^{(l)} = 0$, i.e., $\xi \in \mathfrak{sym}_G(F^{(l)})$ by Theorem 4. Conversely, if $\mathcal{L}_\xi F^{(l)} = 0$ for every $l = 1, \ldots, L$, then by Theorem 4, $\exp(t\xi) \in H$. Since H is a Lie subgroup, differentiating

$\exp(t\xi)$ at $t = 0$ proves that $\xi \in \mathfrak{h}$. ∎

Lemma 31 *The integral operator* (64) *is equivariant with respect to* $g \in G$ *if and only if*

$$\mathcal{K}_g(K \, dV_1 \wedge \cdots \wedge dV_r) := \mathcal{K}_g^E(K) \, \theta_{1,g}^* \, dV_1 \wedge \cdots \wedge \theta_{r,g}^* \, dV_r = K \, dV_1 \wedge \cdots \wedge dV_r, \quad (66)$$

where $\mathcal{K}_g^E$ *is defined by* (65).

Proof Using the fact that the integral is invariant under pullbacks by diffeomorphisms, we can express the left-hand side of the equivariance condition in Definition 30 as

$$\mathcal{K}_{0,g}\mathcal{T}_K[\mathcal{K}_{1,g^{-1}}F_1, \ldots, \mathcal{K}_{r,g^{-1}}F_r](p) =$$

$$\Theta_{0,g^{-1}}\left\{ \int_{\mathcal{M}_1 \times \cdots \times \mathcal{M}_r} K(\theta_{0,g}(p), q_1, \ldots, q_r)[\Theta_{1,g} \circ F_1 \circ \theta_{1,g^{-1}}(q_1), \ldots, \Theta_{r,g} \circ F_r \circ \theta_{r,g^{-1}}(q_r)] \right.$$

$$\left. dV_1(q_1) \wedge \cdots \wedge dV_r(q_r). \right\} =$$

$$\Theta_{0,g^{-1}}\left\{ \int_{\mathcal{M}_1 \times \cdots \times \mathcal{M}_r} K(\theta_{0,g}(p), \theta_{1,g}(q_1), \ldots, \theta_{r,g}(q_r))[\Theta_{1,g} \circ F_1(q_1), \ldots, \Theta_{r,g} \circ F_r(q_r)] \right.$$

$$\left. \theta_{1,g}^* \, dV_1(q_1) \wedge \cdots \wedge \theta_{r,g}^* \, dV_r(q_r). \right\} =$$

$$\int_{\mathcal{M}_1 \times \cdots \times \mathcal{M}_r} \mathcal{K}_g^E K(p, q_1, \ldots, q_r)[F_1(q_1), \ldots, F_r(q_r)]\theta_{1,g}^* \, dV_1(q_1) \wedge \cdots \wedge \theta_{r,g}^* \, dV_r(q_r).$$

Hence, by comparing the integrand to (64), it is clear that (66) implies that $\mathcal{T}_K$ is equivariant in the sense of Definition 30. Conversely, if $\mathcal{T}_K$ is g-equivariant, then

$$\int_{\mathcal{M}_1 \times \cdots \times \mathcal{M}_r} \mathcal{K}_g^E K[F_1, \ldots, F_r]\theta_{1,g}^* \, dV_1 \wedge \cdots \wedge \theta_{r,g}^* \, dV_r$$

$$= \int_{\mathcal{M}_1 \times \cdots \times \mathcal{M}_r} K[F_1, \ldots, F_r] \, dV_1 \wedge \cdots \wedge dV_r$$

holds for every $(F_1, \ldots, F_r) \in D(\mathcal{T}_K)$. Since the domain contains all smooth, compactly-supported fields $(F_1, \ldots, F_r)$, it follows that (66) holds. ∎

Appendix B. Proofs on Discretizing the Fundamental Operators

We use the following lemma in the proof of Proposition 11.

Lemma 41 *Suppose that* $S_m \to S$ *is a convergent sequence of matrices and* $\mathrm{Null}(S) \subset \mathrm{Null}(S_n) \subset \mathrm{Null}(S_m)$ *whenever* $n \geq m$. *Then there is an integer* M_0 *such that for every* $m \geq M_0$ *we have* $\mathrm{Null}(S_m) = \mathrm{Null}(S)$.

Proof Since the sequence of matrices acts on a finite-dimensional space, the sequence $\dim \text{Null}(S_m)$ is a bounded, monotone nonincreasing sequence of integers. Therefore, there exists an integer M_0 such that $\dim \text{Null}(S_m) = \dim \text{Null}(S_{M_0})$ for every $m \geq M_0$. Since $\text{Null}(S_m) \subset \text{Null}(S_{M_0})$, we must have $\text{Null}(S_m) = \text{Null}(S_{M_0})$ for every $m \geq M_0$. Since $\text{Null}(S) \subset \text{Null}(S_{M_0})$ by assumption, it remains to show the reverse containment. Suppose $\xi \in \text{Null}(S_{M_0})$, then

$$S\xi = \lim_{m \to \infty} S_m \xi = 0$$

meaning that $\text{Null}(S_{M_0}) \subset \text{Null}(S)$. ∎

Proposition 11 *Let μ be a strictly positive probability measure on a smooth manifold $\mathcal{M}$ such that $\langle \xi, S_{\mathcal{M}} \xi \rangle < \infty$ for every $\xi \in \text{Lie}(G)$. Let z_i be drawn independently from the distribution μ and let $S_m : \text{Lie}(G) \to \text{Lie}(G)$ be defined by*

$$\langle \eta, \ S_m \xi \rangle_{\text{Lie}(G)} = \frac{1}{m} \sum_{i=1}^{m} \hat{\theta}(\eta)_{z_i}^T (I - P_{z_i})^T (I - P_{z_i}) \hat{\theta}(\xi)_{z_i}. \tag{23}$$

Then there is almost surely an integer M_0 such that for every $m \geq M_0$ we have $\text{Null}(S_m) = \mathfrak{sym}_G(\mathcal{M})$.

Proof By the Cauchy-Schwarz inequality, our assumption means that $\langle \eta, S_{\mathcal{M}} \xi \rangle < \infty$ for every $\eta, \xi \in \text{Lie}(G)$. Let $\xi_1, \ldots, \xi_{\dim G}$ be a basis for $\text{Lie}(G)$. By the strong law of large numbers, specifically Theorem 7.7 in Koralov and Sinai (2012), we have

$$\langle \xi_j, \ S_m \xi_k \rangle_{\text{Lie}(G)} \to \langle \xi_j, S_{\mathcal{M}} \xi_k \rangle$$

for every j, k almost surely. Consequently, $S_m \to S_{\mathcal{M}}$ almost surely. By nonnegativity of each term in the sum defining $\langle \xi, \ S_m \xi \rangle_{\text{Lie}(G)}$, it follows that $\text{Null}(S_n) \subset \text{Null}(S_m)$ when $n \geq m$. Moreover, if $\xi \in \text{Null}(S_{\mathcal{M}})$, then it follows from the continuity of $z \mapsto (I - P_z)\hat{\theta}(\xi)_z$ that $(I - P_z)\hat{\theta}(\xi)_z = 0$ for every $z \in \mathcal{M}$. Hence, $\xi \in \text{Null}(S_m)$ for every m. Therefore, S_m and $S_{\mathcal{M}}$ obey the hypotheses of Lemma 41 almost surely and the conclusion follows. ∎

Proposition 17 *Suppose that $\mathcal{F}'$ is finite-dimensional and $\langle f, \ g \rangle_{L^2(\mu_M)} \to \langle f, \ g \rangle_{L^2(\mu)}$ as $M \to \infty$ for every $f, g \in \mathcal{F}'$. Then there is an M_0 such that (39) is an inner product on $\mathcal{F}'$ for every $M \geq M_0$.*

Proof Let $f_1, \ldots, f_N$ be a basis for $\mathcal{F}'$. Consider the sequence of Gram matrices $\boldsymbol{G}_M$ with entries

$$[\boldsymbol{G}_M]_{i,j} = \langle f_i, \ f_j \rangle_{L^2(\mu_M)}.$$

It suffices to show that $\boldsymbol{G}_M$ is positive-definite for sufficiently large M. Since the $L^2(\mu)$ inner product is positive-definite on $\mathcal{F}'$, it follows that the Gram matrix $\boldsymbol{G}$ with entries

$$[\boldsymbol{G}]_{i,j} = \langle f_i, \ f_j \rangle_{L^2(\mu)}$$

is positive-definite. Hence, its smallest eigenvalue $\lambda_{\min}(\boldsymbol{G})$ is positive. Since the ordered eigenvalues of symmetric matrices are continuous with respect to their entries (Corollary 4.3.15; Horn and Johnson, 2013) and $[\boldsymbol{G}_M]_{i,j} \to [\boldsymbol{G}]_{i,j}$ for all $1 \le i, j \le N$ by assumption, we have $\lambda_{\min}(\boldsymbol{G}_M) \to \lambda_{\min}(\boldsymbol{G})$ as $M \to \infty$. Therefore, there is an M_0 so that for every $M \ge M_0$ we have $\lambda_{\min}(\boldsymbol{G}_M) > 0$, i.e., $\boldsymbol{G}_M$ is positive-definite. ∎

Our proof of Proposition 18 relies on the following lemma:

Lemma 42 *Let $\mathcal{P}$ denote a finite-dimensional vector space of polynomials $\mathbb{R}^m \to \mathbb{R}$. If $M \ge \dim(\mathcal{P})$ then the evaluation map $T : \mathcal{P} \to \mathbb{R}^M$ defined by*

$$T_{(x_1,\ldots,x_M)} : P \mapsto (P(x_1), \ldots, P(x_M))$$

is injective for almost every $(x_1, \ldots, x_M) \in (\mathbb{R}^m)^M$ with respect to Lebesgue measure.

Proof Letting $M_0 = \dim(\mathcal{P})$ and choosing a basis $P_1, \ldots, P_{M_0}$ for $\mathcal{P}$, injectivity of $T_{(x_1,\ldots,x_M)}$ is equivalent to injectivity of the $M \times M_0$ matrix

$$\boldsymbol{T}_{(x_1,\ldots,x_M)} = \begin{bmatrix} P_1(x_1) & \cdots & P_{M_0}(x_1) \\ \vdots & \ddots & \vdots \\ P_1(x_M) & \cdots & P_{M_0}(x_M) \end{bmatrix}.$$

Finally, this is equivalent to

$$\varphi(x_1, \ldots, x_M) = \det\left((\boldsymbol{T}_{(x_1,\ldots,x_M)})^T \boldsymbol{T}_{(x_1,\ldots,x_M)} \right)$$

taking a nonzero value. We observe that φ is a polynomial on the Euclidean space $(\mathbb{R}^m)^M$.

Suppose there exists a set of points $(\bar{x}_1, \ldots, \bar{x}_M) \in (\mathbb{R}^m)^M$ such that $T_{(\bar{x}_1,\ldots,\bar{x}_M)}$ is injective. Then for this set $\varphi(\bar{x}_1, \ldots, \bar{x}_M) \neq 0$. Obviously, $\varphi(0, \ldots, 0) = 0$, meaning that φ cannot be constant. Thanks to the main result in Caron and Traynor (2005), this means that each level set of φ has zero Lebesgue measure in $(\mathbb{R}^m)^M$. In particular, the level set $\varphi^{-1}(0)$, consisting of those $x_1, \ldots, x_M$ for which $T_{(x_1,\ldots,x_M)}$ fails to be injective, has zero Lebesgue measure. Therefore, it suffices to prove that there exists $(\bar{x}_1, \ldots, \bar{x}_M) \in (\mathbb{R}^m)^M$ such that $T_{(\bar{x}_1,\ldots,\bar{x}_M)}$ is injective. We do this by induction.

It is clear that there exists $\bar{x}_1$ so that the 1×1 matrix

$$\boldsymbol{T}_1 = \begin{bmatrix} P_1(\bar{x}_1) \end{bmatrix}$$

has full rank since P_1 cannot be the zero polynomial. Proceeding by induction, we assume that there exists $\bar{x}_1, \ldots, \bar{x}_s$ so that

$$\boldsymbol{T}_s = \begin{bmatrix} P_1(\bar{x}_1) & \cdots & P_s(\bar{x}_1) \\ \vdots & \ddots & \vdots \\ P_1(\bar{x}_s) & \cdots & P_s(\bar{x}_s) \end{bmatrix}$$

has full rank. Suppose that the matrix

$$\tilde{\boldsymbol{T}}_{s+1}(x) = \begin{bmatrix} P_1(\bar{x}_1) & \cdots & P_s(\bar{x}_1) & P_{s+1}(\bar{x}_1) \\ \vdots & \ddots & \vdots & \vdots \\ P_1(\bar{x}_s) & \cdots & P_s(\bar{x}_s) & P_{s+1}(\bar{x}_s) \\ P_1(x) & \cdots & P_s(x) & P_{s+1}(x) \end{bmatrix}$$

has rank $< s+1$ for every $x \in \mathbb{R}^m$. Since the upper left $s \times s$ block of $\tilde{\boldsymbol{T}}_{s+1}(x)$ is $\boldsymbol{T}_s$, we must always have $\mathrm{rank}(\tilde{\boldsymbol{T}}_{s+1}(x)) = s$. The nullspace of $\tilde{\boldsymbol{T}}_{s+1}(x)$ is contained in the nullspace of the upper $s \times (s+1)$ block of $\tilde{\boldsymbol{T}}_{s+1}(x)$. Since both nullspaces are one-dimensional, they are equal. The upper $s \times (s+1)$ block of $\tilde{\boldsymbol{T}}_{s+1}(x)$ does not depend on x, so there is a fixed nonzero vector $\boldsymbol{v} \in \mathbb{R}^{s+1}$ so that $\tilde{\boldsymbol{T}}_{s+1}(x)\boldsymbol{v} = \boldsymbol{0}$ for every $x \in \mathbb{R}^m$. The last row of this expression reads

$$v_1 P_1(x) + \cdots + v_{s+1} P_{s+1}(x) = 0 \qquad \forall x \in \mathbb{R}^m,$$

contradicting the linear independence of $P_1, \ldots, P_{s+1}$. Therefore there exists $\bar{x}_{s+1}$ so that $\boldsymbol{T}_{s+1} = \tilde{\boldsymbol{T}}_{s+1}(\bar{x}_{s+1})$ has full rank. It follows by induction on s that there exists $\bar{x}_1, \ldots \bar{x}_{M_0} \in \mathbb{R}^m$ so that $\boldsymbol{T}_{(\bar{x}_1, \ldots \bar{x}_{M_0})} = \boldsymbol{T}_{M_0}$ has full rank. Choosing any $M - M_0$ additional vectors yields an injective $\boldsymbol{T}_{(\bar{x}_1, \ldots \bar{x}_M)}$, which completes the proof. ■

Proposition 18 *Let $\mathcal{F}'$ be a space of real polynomial functions $\mathbb{R}^m \to \mathbb{R}^n$ and let $\pi_i : \mathbb{R}^n \to \mathbb{R}$ be the ith coordinate projection $\pi(c_1, \ldots, c_n) = c_i$. Let*

$$M \geq M_0 := \max_{1 \leq i \leq n} \dim(\pi_i(\mathcal{F}')) \tag{40}$$

and let $w_1, \ldots, w_M > 0$ be positive weights. Then for almost every set of points $(x_1, \ldots, x_M) \in (\mathbb{R}^m)^M$ with respect to Lebesgue measure, (39) is an inner product on $\mathcal{F}'$.

Proof The sum in (39) clearly defines a symmetric, positive-semidefinite bilinear form on $\mathcal{F}'$. It remains to show that this bilinear form is positive-definite. Suppose that there is a function $f \in \mathcal{F}'$ such that $\langle f, f \rangle_{L^2(\mu_M)} = 0$. Thanks to Lemma 42, our assumption that $M \geq \dim(\pi_i(\mathcal{F}'))$ means that the evaluation operator $T_{(x_1, \ldots, x_M)}$ is injective on $\pi_i(\mathcal{F}')$ for almost every $(x_1, \ldots, x_M) \in (\mathbb{R}^m)^M$ with respect to Lebesgue measure. Since a countable (in this case finite) union of sets of measure zero has measure zero, it follows that for almost every $(x_1, \ldots, x_M) \in (\mathbb{R}^m)^M$ with respect to Lebesgue measure, $T_{(x_1, \ldots, x_M)}$ is injective on every $\pi_i(\mathcal{F}')$, $i = 1, \ldots, n$. Defining the positive diagonal matrix

$$\boldsymbol{D} = \frac{1}{\sqrt{M}} \begin{bmatrix} \sqrt{w_1} & & \\ & \ddots & \\ & & \sqrt{w_M} \end{bmatrix},$$

and using (39) yields

$$0 = \langle f, f \rangle_{L^2(\mu_M)} = \sum_{j=1}^{n} \left(\boldsymbol{D} T_{(x_1, \ldots, x_M)} \pi_j f \right)^T \boldsymbol{D} T_{(x_1, \ldots, x_M)} \pi_j f.$$

This implies that $T_{(x_1,\ldots,x_M)}\pi_j f = \mathbf{0}$ for $j = 1,\ldots,n$. Since $T_{(x_1,\ldots,x_M)}$ is injective on each $\pi_j(\mathcal{F}')$ it follows that each $\pi_j f = 0$, meaning that $f = 0$. This completes the proof. ∎

Appendix C. Proofs on Symmetric Polynomials

Proposition 19 *Let $r \leq n$ and suppose that $\{c_k - c_1\}_{k=2}^{r}$ and $\{u_k\}_{k=1}^{r}$ are sets of linearly-independent vectors in $\mathbb{R}^n$. Then, $\mathfrak{sym}_{\mathrm{SE}(n)}(F_{rad})$ contains the $\frac{1}{2}(n-r)(n-r+1)$-dimensional subalgebra*

$$\mathfrak{g}_{rad} = \left\{ \begin{bmatrix} S & v \\ 0 & 0 \end{bmatrix} : S^T = -S \text{ and } Sc_1 = \cdots = Sc_r = -v \right\} \tag{44}$$

and $\mathfrak{sym}_{\mathrm{SE}(n)}(F_{lin})$ contains the $\frac{1}{2}(n-r)(n-r+1)$-dimensional subalgebra

$$\mathfrak{g}_{lin} = \left\{ \begin{bmatrix} S & v \\ 0 & 0 \end{bmatrix} : S^T = -S, \ Su_1 = \cdots = Su_r = 0, \text{ and } u_1^T v = \cdots = u_r^T v = 0 \right\}. \tag{45}$$

Either every polynomial φ_{rad} with degree $\leq d$ gives $\mathfrak{sym}_{\mathrm{SE}(n)}(F_{rad}) \neq \mathfrak{g}_{rad}$, or else the set of polynomials φ_{rad} with degree $\leq d$ satisfying $\mathfrak{sym}_{\mathrm{SE}(n)}(F_{rad}) \neq \mathfrak{g}_{rad}$ is a set of measure zero. Likewise, for φ_{lin}, F_{lin}, and $\mathfrak{g}_{lin}$.

Proof Consider the function $\tilde{F}_{\mathrm{rad}} : \mathbb{R}^n \to \mathbb{R}^r$ defined by

$$\tilde{F}_{\mathrm{rad}}(x) = (\|x - c_1\|_2^2, \ \ldots, \ \|x - c_r\|_2^2)$$

with standard action of $\mathrm{SE}(n)$ on its domain and the trivial action on its codomain. The symmetries of $\tilde{F}_{\mathrm{rad}}$ are shared by F_{rad}. By Theorem 4, the Lie algebra of $\tilde{F}_{\mathrm{rad}}$'s symmetry group is characterized by

$$\xi = \begin{bmatrix} S & v \\ 0 & 0 \end{bmatrix} \in \mathfrak{sym}_{\mathrm{SE}(n)} \tilde{F}_{\mathrm{rad}} \quad \Leftrightarrow \quad 0 = \mathcal{L}_\xi \tilde{F}_{\mathrm{rad}}(x) = \frac{\partial \tilde{F}_{\mathrm{rad}}(x)}{\partial x}(Sx + v) \quad \forall x \in \mathbb{R}^n.$$

This means the generators ξ are characterized by the equations

$$0 = (x - c_i)^T (Sx + v) = x^T Sx - c_i^T Sx + x^T v - c_i^T v, \qquad \forall x \in \mathbb{R}^n \quad i = 1,\ldots,r.$$

Since $\xi \in \mathfrak{se}(n)$, we have $S^T = -S$, called *skew symmetry*, giving $x^T Sx = 0$. The above is satisfied if and only if

$$Sc_i = -v,$$

which automatically yields $c_i^T v = -c_i^T Sc_i = 0$. Therefore, $\mathfrak{sym}_{\mathrm{SE}(n)} \tilde{F}_{\mathrm{rad}} = \mathfrak{g}_{rad} \subset \mathfrak{sym}_{\mathrm{SE}(n)} F_{\mathrm{rad}}$. To determine the dimension of the symmetry group, we observe that S must satisfy

$$S(c_2 - c_1) = \cdots = S(c_r - c_1) = 0, \tag{73}$$

and any such S uniquely determines $v = -Sc_1$. Therefore, the dimension of $\mathfrak{g}_{rad}$ equals the dimension of the space of skew-symmetric matrices satisfying (73). Let the columns of

$W = \begin{bmatrix} W_1 & W_2 \end{bmatrix}$ form an orthonormal basis for $\mathbb{R}^n$ with the $r - 1$ columns of W_1 being a basis for $\mathrm{span}\{(c_2 - c_1), \ldots, (c_r - c_1)\}$. The above constraints, together with skew-symmetry, mean that S takes the form

$$S = \begin{bmatrix} W_1 & W_2 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & \tilde{S} \end{bmatrix} \begin{bmatrix} W_1^T \\ W_2^T \end{bmatrix},$$

where $\tilde{S}$ is an $(n - r + 1) \times (n - r + 1)$ skew-symmetric matrix. Therefore, the dimension of $\mathfrak{g}_{\mathrm{rad}}$ equals the dimension of the space of $(n - r + 1) \times (n - r + 1)$ skew-symmetric matrices, which is $\frac{1}{2}(n - r)(n - r + 1)$.

The argument for F_{lin} is similar, with the symmetries of

$$\tilde{F}_{\mathrm{lin}}(x) = (u_1^T x, \ \ldots, \ u_r^T x) = U^T x$$

also being symmetries of F_{lin}. The condition $0 = \mathcal{L}_\xi \tilde{F}_{\mathrm{lin}}$ is equivalent to

$$U^T S x + U^T v = 0 \qquad \forall x \in \mathbb{R}^n,$$

which occurs if and only if $U^T S = 0$ and $U^T v = 0$. This immediately yields $\mathfrak{sym}_{\mathrm{SE}(n)} \tilde{F}_{\mathrm{lin}} = \mathfrak{g}_{\mathrm{lin}} \subset \mathfrak{sym}_{\mathrm{SE}(n)} F_{\mathrm{lin}}$. Per our earlier argument, the skew-symmetric matrices S satisfying

$$S u_1 = \cdots = S u_r = 0$$

form a vector space with dimension $\frac{1}{2}(n - r)(n - r - 1)$. The subspace of vectors $v \in \mathbb{R}^n$ satisfying $U^T v = 0$ is $(n - r)$-dimensional. Adding these gives the dimension of $\mathfrak{g}_{\mathrm{lin}}$, which is $\frac{1}{2}(n - r)(n - r + 1)$.

Suppose there exists a polynomial φ_0 with degree $\leq d$ such that $\mathfrak{sym}_{\mathrm{SE}(n)}(F_{\mathrm{rad}}) = \mathfrak{g}_{\mathrm{rad}}$ when $\varphi_{\mathrm{rad}} = \varphi_0$. Let $\mathfrak{g}_\perp$ be a complementary subspace to $\mathfrak{g}_{\mathrm{rad}}$ in $\mathfrak{se}(n)$, that is, $\mathfrak{g}_\perp \oplus \mathfrak{g}_{\mathrm{rad}} = \mathfrak{se}(n)$. We observe that $\mathfrak{sym}_{\mathrm{SE}(n)}(F_{\mathrm{rad}}) \neq \mathfrak{g}_{\mathrm{rad}}$ if and only if there is a nonzero $\xi_\perp \in \mathfrak{g}_\perp$ satisfying $\mathcal{L}_{\xi_\perp} F_{\mathrm{rad}} = 0$. The "if" part of this claim is obvious. The "only if" part follows from the fact that $\mathfrak{sym}_{\mathrm{SE}(n)}(F_{\mathrm{rad}}) \neq \mathfrak{g}_{\mathrm{rad}}$ means that $\mathcal{L}_\xi F_{\mathrm{rad}} = 0$ for some nonzero $\xi \notin \mathfrak{g}_{\mathrm{rad}}$. Using the direct-sum decomposition of $\mathfrak{se}(n)$, there are unique $\xi_\perp \in \mathfrak{g}_\perp$ and $\xi_{\mathrm{rad}} \in \mathfrak{g}_{\mathrm{rad}}$ such that $\xi = \xi_\perp + \xi_{\mathrm{rad}}$, yielding

$$0 = \mathcal{L}_\xi F_{\mathrm{rad}} = \mathcal{L}_{\xi_\perp} F_{\mathrm{rad}} + \mathcal{L}_{\xi_{\mathrm{rad}}} F_{\mathrm{rad}} = \mathcal{L}_{\xi_\perp} F_{\mathrm{rad}}.$$

Moreover, $\xi_\perp \neq 0$ because $\xi \notin \mathfrak{g}_{\mathrm{rad}}$. Letting $\xi_1, \ldots, \xi_D$ form a basis for $\mathfrak{g}_\perp$, we consider the $D \times D$ Gram matrix $\boldsymbol{G}(F_{\mathrm{rad}})$ with entries

$$\left[\boldsymbol{G}(F_{\mathrm{rad}}) \right]_{i,j} = \int_{[0,1]^n} \mathcal{L}_{\xi_i} F_{\mathrm{rad}}(x) \mathcal{L}_{\xi_j} F_{\mathrm{rad}}(x) \ \mathrm{d}x.$$

This matrix is singular if and only if there is a nonzero $\xi_\perp \in \mathfrak{g}_\perp$ satisfying $\mathcal{L}_{\xi_\perp} F_{\mathrm{rad}}(x) = 0$ for every x in the cube $[0,1]^n$. Since

$$\mathcal{L}_{\xi_\perp} F_{\mathrm{rad}}(x) = \frac{\partial F_{\mathrm{rad}}(x)}{\partial x}(Sx + v), \qquad \xi_\perp = \begin{bmatrix} S & v \\ 0 & 0 \end{bmatrix},$$

is a polynomial function of x, it vanishes in the cube if and only if it vanishes everywhere. Hence $\boldsymbol{G}(F_{\mathrm{rad}})$ is singular if and only if $\mathfrak{sym}_{\mathrm{SE}(n)}(F_{\mathrm{rad}}) \neq \mathfrak{g}_{\mathrm{rad}}$. Letting $\boldsymbol{c}$ denote the vector of coefficients defining φ_{rad} in a basis for the polynomials of degree $\leq d$ on $\mathbb{R}^r$, we observe that

$$f : \boldsymbol{c} \mapsto \det\left(\boldsymbol{G}(F_{\mathrm{rad}})\right)$$

is a polynomial function of $\boldsymbol{c}$. The set of polynomials φ_{rad} with degree $\leq d$ for which $\mathfrak{sym}_{\mathrm{SE}(n)}(F_{\mathrm{rad}}) \neq \mathfrak{g}_{\mathrm{rad}}$ corresponds to the zero level set of f, i.e., those $\boldsymbol{c}$ such that $f(\boldsymbol{c}) = 0$. Obviously, $f(0) = 0$, and taking the coefficients $\boldsymbol{c}_0$ corresponding to φ_0 gives $f(\boldsymbol{c}_0) \neq 0$, meaning f is a nonconstant polynomial. Since each level set of a nonconstant polynomial is a set of measure zero (Caron and Traynor, 2005), it follows that the zero level set of f has measure zero. Precisely the same argument works for φ_{lin}, F_{lin}, and $\mathfrak{g}_{\mathrm{lin}}$. ∎

Corollary 20 *With the same hypotheses as Proposition 19, let $d \geq r$. The set of polynomials φ_{rad} with degree $\leq d$ satisfying $\mathfrak{sym}_{\mathrm{SE}(n)}(F_{rad}) \neq \mathfrak{g}_{rad}$ is a set of measure zero.*

Proof Thanks to Proposition 19, it suffices to find a degree-r polynomial φ_{rad} such that $\mathfrak{sym}_{\mathrm{SE}(n)}(F_{\mathrm{rad}}) \subset \mathfrak{g}_{\mathrm{rad}}$. We choose $\varphi_{\mathrm{rad}}(z_1, \ldots, z_r) = z_1 + z_2^2 + \cdots + z_r^r$, giving

$$F_{\mathrm{rad}}(x) = \sum_{k=1}^{r} \|x - c_k\|_2^{2k}.$$

If $\xi = \begin{bmatrix} S & v \\ 0 & 0 \end{bmatrix} \in \mathfrak{se}(n)$ generates a symmetry of F_{rad} then

$$0 = \mathcal{L}_\xi F_{\mathrm{rad}}(x) = \sum_{k=1}^{r} k\|x - c_k\|_2^{2(k-1)} \underbrace{(x - c_k)^T (Sx + v)}_{x^T(Sc_k + v) - c_k^T v} \qquad \forall x \in \mathbb{R}^n. \tag{74}$$

The terms in this expression with highest degree in x must vanish, yielding

$$0 = r\|x\|_2^{2(r-1)} x^T(Sc_r + v) \qquad \forall x \in \mathbb{R}^n.$$

This implies that $Sc_r + v = 0$. Proceeding inductively, suppose that $Sc_l + v = 0$ for every $l > k$. Then, vanishing the highest-degree term in (74) gives

$$0 = k\|x\|_2^{2(k-1)} x^T(Sc_k + v) \qquad \forall x \in \mathbb{R}^n,$$

implying that $Sc_k + v = 0$. It follows that

$$Sc_k + v = 0 \qquad \forall k = 1, \ldots, r,$$

by induction, meaning that $\xi \in \mathfrak{g}_{\mathrm{rad}}$. Hence, $\mathfrak{sym}_{\mathrm{SE}(n)}(F_{\mathrm{rad}}) \subset \mathfrak{g}_{\mathrm{rad}}$, which completes the proof. ∎

Corollary 21 *With the same hypotheses as Proposition 19, let $d \geq 2$. The set of polynomials φ_{lin} with degree $\leq d$ satisfying $\mathfrak{sym}_{\mathrm{SE}(n)}(F_{lin}) \neq \mathfrak{g}_{lin}$ is a set of measure zero.*

Proof Thanks to Proposition 19, it suffices to find a quadratic polynomial φ_{lin} such that $\mathfrak{sym}_{\mathrm{SE}(n)}(F_{\mathrm{lin}}) \subset \mathfrak{g}_{\mathrm{lin}}$. Letting $D = \mathrm{diag}[1, 2, \ldots, r]$ and $U = \begin{bmatrix} u_1 & \cdots & u_r \end{bmatrix}$, consider the quadratic function $\varphi_{\mathrm{lin}}(z) = \frac{1}{2} z^T D z$, giving

$$F_{\mathrm{lin}}(x) = \frac{1}{2} x^T U D U^T x.$$

If $\xi = \begin{bmatrix} S & v \\ 0 & 0 \end{bmatrix} \in \mathfrak{se}(n)$ generates a symmetry of F_{lin} then

$$0 = \mathcal{L}_\xi F_{\mathrm{lin}}(x) = x^T U D U^T S x + x^T U D U^T v \qquad \forall x \in \mathbb{R}^n.$$

Differentiating the above with respect to x at $x = 0$ yields $U D U^T v = 0$, which, because UD is injective, means that $U^T v = 0$. The fact that $x^T U D U^T S x = 0$ for every x means that $U D U^T S + S^T U D U^T = 0$, i.e.,

$$U D U^T S = S U D U^T.$$

Letting the columns of $U_\perp$ span the orthogonal complement to columns of U and expressing

$$S = U \tilde{S}_{1,1} U^T + U \tilde{S}_{1,2} U_\perp^T + U_\perp \tilde{S}_{2,1} U^T + U_\perp \tilde{S}_{2,2} U_\perp^T,$$

the above commutation relation with UDU gives

$$U D \tilde{S}_{1,1} U^T + U D \tilde{S}_{1,2} U_\perp^T = U \tilde{S}_{1,1} D U^T + U_\perp \tilde{S}_{2,1} D U^T.$$

Multiplying on left and right by combinations of U^T or $U_\perp^T$ and U or $U_\perp$ extracts the relations

$$D \tilde{S}_{1,1} = \tilde{S}_{1,1} D, \qquad D \tilde{S}_{1,2} = 0, \qquad \tilde{S}_{2,1} D = 0.$$

Since D is invertible, we must have $\tilde{S}_{1,2} = 0$ and $\tilde{S}_{2,1} = 0$. Since $S^T = -S$, we must also have $\tilde{S}_{1,1}^T = -\tilde{S}_{1,1}$, meaning that its diagonal entries are identically zero. Considering the (j, k) element of $\tilde{S}_{1,1}$ with $j \neq k$, we have

$$j[\tilde{S}_{1,1}]_{j,k} = [D\tilde{S}_{1,1}]_{j,k} = [\tilde{S}_{1,1}D]_{j,k} = k[\tilde{S}_{1,1}]_{j,k},$$

meaning that $[\tilde{S}_{1,1}]_{j,k} = 0$. Therefore, only $\tilde{S}_{2,2}$ can be nonzero, which gives $SU = 0$. Combined with the fact that $U^T v = 0$, we conclude that $\mathfrak{sym}_{\mathrm{SE}(n)}(F_{\mathrm{lin}}) \subset \mathfrak{g}_{\mathrm{lin}}$, completing the proof. $\blacksquare$

Appendix D. Proofs on Sections of Vector Bundles

Proposition 24 *For every $\xi, \eta \in \mathrm{Lie}(G)$, and $\alpha, \beta, t \in \mathbb{R}$, we have*

$$\frac{\mathrm{d}}{\mathrm{d}\,t} \mathcal{K}_{\exp(t\xi)} F = \mathcal{L}_\xi \mathcal{K}_{\exp(t\xi)} F = \mathcal{K}_{\exp(t\xi)} \mathcal{L}_\xi F \qquad \forall F \in D(\mathcal{L}_\xi), \tag{54}$$

$$\mathcal{L}_{\alpha\xi+\beta\eta} F = (\alpha\mathcal{L}_\xi + \beta\mathcal{L}_\eta) F \qquad \forall F \in C^1(\mathcal{M}, E), \tag{55}$$

and

$$\mathcal{L}_{[\xi,\eta]} F = (\mathcal{L}_\xi \mathcal{L}_\eta - \mathcal{L}_\eta \mathcal{L}_\xi) F \qquad \forall F \in C^2(\mathcal{M}, E). \tag{56}$$

Proof We begin by proving

$$\frac{\mathrm{d}}{\mathrm{d}\,t} \mathcal{K}_{\exp(t\xi)} F = \mathcal{K}_{\exp(t\xi)} \mathcal{L}_\xi F = \mathcal{L}_\xi \mathcal{K}_{\exp(t\xi)} F, \tag{75}$$

for every $F \in D(\mathcal{L}_\xi)$. To prove the first equality, we choose $p \in \mathcal{M}$, let $p' = \theta_{\exp(t_0\xi)}(p)$, and compute

$$\frac{1}{t}\left[(\mathcal{K}_{\exp(t_0\xi)}\mathcal{K}_{\exp(t\xi)}F)(p) - (\mathcal{K}_{\exp(t_0\xi)}F)(p) \right] = \frac{1}{t}\Theta_{\exp(-t_0\xi)} \circ \left(\mathcal{K}_{\exp(t\xi)}F - F \right) \circ \theta_{\exp(t_0\xi)}(p)$$

$$= \Theta_{\exp(-t_0\xi)}\left(\frac{1}{t}\left[(\mathcal{K}_{\exp(t\xi)}F)(p') - F(p') \right] \right).$$

Here, we have used the composition law for the operators $\mathcal{K}_{gh} = \mathcal{K}_g\mathcal{K}_h$ and the fact that $\Theta_{\exp(-t_0\xi)}$ is fiber-linear. Taking the limit at $t \to 0$ yields

$$\left.\frac{\mathrm{d}}{\mathrm{d}\,t}\right|_{t=t_0} (\mathcal{K}_{\exp(t\xi)}F)(p) = \Theta_{\exp(-t_0\xi)}\left(\mathcal{L}_\xi F(p') \right) = (\mathcal{K}_{\exp(t_0\xi)}\mathcal{L}_\xi F)(p),$$

which is the first equality in (75).

The second equality in (75) follows from

$$\lim_{t\to 0} \frac{1}{t}\left[\mathcal{K}_{\exp(t\xi)}\mathcal{K}_{\exp(t_0\xi)}F - \mathcal{K}_{\exp(t_0\xi)}F \right] = \lim_{t\to 0} \frac{1}{t}\left[\mathcal{K}_{\exp(t_0\xi)}\mathcal{K}_{\exp(t\xi)}F - \mathcal{K}_{\exp(t_0\xi)}F \right]$$

$$= \mathcal{K}_{\exp(t_0\xi)}\mathcal{L}_\xi F.$$

This shows that $\mathcal{K}_{\exp(t_0\xi)}F \in D(\mathcal{L}_\xi)$ and $\mathcal{L}_\xi \mathcal{K}_{\exp(t_0\xi)}F = \mathcal{K}_{\exp(t_0\xi)}\mathcal{L}_\xi F$.

Next, we prove

$$\mathcal{L}_{\alpha\xi+\beta\eta} F = \alpha\mathcal{L}_\xi F + \beta\mathcal{L}_\eta F, \tag{76}$$

when $F \in C^1(\mathcal{M}, E)$. To do this, we choose $p \in \mathcal{M}$, and define the map $h : G \to E_p$ by

$$h : g \mapsto \mathcal{K}_g F(p) = \Theta\big(F(\theta(p, g)), g^{-1} \big). \tag{77}$$

As a composition of C^1 maps, h is C^1, and its derivative at the identity is

$$\mathrm{d}\,h(e)\xi_e = \left.\frac{\mathrm{d}}{\mathrm{d}\,t}\right|_{t=0} h(\exp(t\xi)) = \mathcal{L}_\xi F(p) \tag{78}$$

65

for every $\xi_e \in T_e G \cong \mathrm{Lie}(G)$. Since the derivative is linear, it follows that $\xi \mapsto \mathcal{L}_\xi F(p)$ is linear.

Finally, we prove that

$$\mathcal{L}_{[\xi,\eta]} F = \frac{1}{2} \left. \frac{\mathrm{d}^2}{\mathrm{d}\,t^2} \right|_{t=0} \mathcal{K}_{\exp(t\xi)} \mathcal{K}_{\exp(t\eta)} \mathcal{K}_{\exp(-t\xi)} \mathcal{K}_{\exp(-t\eta)} F = \mathcal{L}_\xi \mathcal{L}_\eta F - \mathcal{L}_\eta \mathcal{L}_\xi F, \qquad (79)$$

when $F \in C^2(\mathcal{M}, E)$. Recall that $\mathrm{Fl}_\xi^t : g \mapsto g \cdot \exp(t\xi)$ gives the flow of the left-invariant vector field $\xi \in \mathrm{Lie}(G)$ (see Theorem 4.18(3) in Kolář et al., 1993). By Theorem 3.16 in Kolář et al. (1993) the curve $\gamma : \mathbb{R} \to G$ given by

$$\gamma(t) = \mathrm{Fl}_{-\eta}^t \circ \mathrm{Fl}_{-\xi}^t \circ \mathrm{Fl}_\eta^t \circ \mathrm{Fl}_\xi^t(e) = \exp(t\xi) \exp(t\eta) \exp(-t\xi) \exp(-t\eta).$$

satisfies $\gamma(0) = e$, $\gamma'(0) = 0$, and

$$\frac{1}{2} \gamma''(0) = [\xi, \eta]_e \in T_e G \qquad (80)$$

in the sense that $\gamma''(0) : f \mapsto (f \circ \gamma)''(0)$ is a derivation on $C^\infty(G)$, hence an element of $T_e G$. Composing with the map in (77) yields

$$0 = \mathrm{d}\,h(e)\gamma'(0) = (h \circ \gamma)'(0) = \left. \frac{\mathrm{d}}{\mathrm{d}\,t} \right|_{t=0} \mathcal{K}_{\gamma(t)} F(p). \qquad (81)$$

Combining (80) and (78) (noting the definition of the tangent map $\mathrm{d}\,h(e)$ acting on derivations, as in Kolář et al. (1993), Lee (2013)) gives

$$\mathcal{L}_{[\xi,\eta]} F(p) = \frac{1}{2} \mathrm{d}\,h(e)\gamma''(0) = \frac{1}{2}(h \circ \gamma)''(0) = \frac{1}{2} \left. \frac{\mathrm{d}^2}{\mathrm{d}\,t^2} \right|_{t=0} \mathcal{K}_{\gamma(t)} F(p).$$

This proves the first equality in (79) thanks to the composition law

$$\mathcal{K}_{\gamma(t)} = \mathcal{K}_{\exp(t\xi)} \mathcal{K}_{\exp(t\eta)} \mathcal{K}_{\exp(-t\xi)} \mathcal{K}_{\exp(-t\eta)}.$$

To differentiate the above expression, we use the following observations. If $F_t \in \Sigma(E)$ is such that $(t, p) \mapsto F_t(p)$ is $C^2(\mathbb{R} \times \mathcal{M}, E)$, then obviously $\frac{\mathrm{d}}{\mathrm{d}\,t} F_t \in C^1(\mathcal{M}, E)$ with the usual identification $TE_p \cong E_p$. Moreover, we have

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\,t} \mathcal{K}_g F_t(p) &= \frac{\mathrm{d}}{\mathrm{d}\,t} \Theta_{g^{-1}} \big(F_t(\theta_g(p)) \big) \\
&= \Theta_{g^{-1}} \Big(\frac{\mathrm{d}}{\mathrm{d}\,t} F_t(\theta_g(p)) \Big) = \mathcal{K}_g \Big(\frac{\mathrm{d}}{\mathrm{d}\,t} F_t \Big)(p)
\end{aligned} \qquad (82)$$

because $F_t(\theta_g(p)) \in E_{\theta_g(p)}$ for all $t \in \mathbb{R}$ and $\Theta_{g^{-1}}$ is linear on $E_{\theta_g(p)}$. Using this, we obtain

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{L}_\xi F_t(p) &= \frac{\mathrm{d}}{\mathrm{d}t}\frac{\mathrm{d}}{\mathrm{d}\tau}\Big|_{\tau=0}\mathcal{K}_{\exp(\tau\xi)}F_t(p) \\
&= \frac{\mathrm{d}}{\mathrm{d}\tau}\Big|_{\tau=0}\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{K}_{\exp(\tau\xi)}F_t(p) \\
&= \frac{\mathrm{d}}{\mathrm{d}\tau}\Big|_{\tau=0}\mathcal{K}_{\exp(\tau\xi)}\Big(\frac{\mathrm{d}}{\mathrm{d}t}F_t\Big)(p) = \mathcal{L}_\xi\Big(\frac{\mathrm{d}}{\mathrm{d}t}F_t\Big)(p)
\end{aligned}
\tag{83}
$$

because $(t,\tau) \mapsto \mathcal{K}_{\exp(\tau\xi)}F_t(p)$ lies in the vector space E_p, allowing us to exchanged the order of differentiation. Since $(t_1,t_2,t_3,t_4) \mapsto \mathcal{K}_{\exp(t\xi)}\mathcal{K}_{\exp(t\eta)}\mathcal{K}_{\exp(-t\xi)}\mathcal{K}_{\exp(-t\eta)}F(p)$ lies in the vector space E_p for all $(t_1,t_2,t_3,t_4) \in \mathbb{R}^4$, we can apply the chain rule and (82) to obtain

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{K}_{\gamma(t)}F = {}& \frac{\partial}{\partial t_1}\Big|_{t_1=t}\mathcal{K}_{\exp(t_1\xi)}\mathcal{K}_{\exp(t\eta)}\mathcal{K}_{\exp(-t\xi)}\mathcal{K}_{\exp(-t\eta)}F \\
&+ \mathcal{K}_{\exp(t\xi)}\frac{\partial}{\partial t_2}\Big|_{t_2=t}\mathcal{K}_{\exp(t_2\eta)}\mathcal{K}_{\exp(-t\xi)}\mathcal{K}_{\exp(-t\eta)}F \\
&+ \mathcal{K}_{\exp(t\xi)}\mathcal{K}_{\exp(t\eta)}\frac{\partial}{\partial t_3}\Big|_{t_3=t}\mathcal{K}_{\exp(-t_3\xi)}\mathcal{K}_{\exp(-t\eta)}F \\
&\qquad\qquad + \mathcal{K}_{\exp(t\xi)}\mathcal{K}_{\exp(t\eta)}\mathcal{K}_{\exp(-t\xi)}\frac{\partial}{\partial t_4}\Big|_{t_4=t}\mathcal{K}_{\exp(-t_4\eta)}F.
\end{aligned}
$$

Using (75) gives

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{K}_{\gamma(t)}F = {}& \mathcal{L}_\xi\overbrace{\mathcal{K}_{\exp(t\xi)}\mathcal{K}_{\exp(t\eta)}\mathcal{K}_{\exp(-t\xi)}\mathcal{K}_{\exp(-t\eta)}}^{\mathcal{K}_{\gamma(t)}}F \\
&+ \mathcal{K}_{\exp(t\xi)}\mathcal{L}_\eta\mathcal{K}_{\exp(t\eta)}\mathcal{K}_{\exp(-t\xi)}\mathcal{K}_{\exp(-t\eta)}F \\
&+ \mathcal{K}_{\exp(t\xi)}\mathcal{K}_{\exp(t\eta)}\mathcal{L}_{-\xi}\mathcal{K}_{\exp(-t\xi)}\mathcal{K}_{\exp(-t\eta)}F \\
&\qquad + \underbrace{\mathcal{K}_{\exp(t\xi)}\mathcal{K}_{\exp(t\eta)}\mathcal{K}_{\exp(-t\xi)}\mathcal{K}_{\exp(-t\eta)}}_{\mathcal{K}_{\gamma(t)}}\mathcal{L}_{-\eta}F.
\end{aligned}
$$

Applying the same technique to differentiate a second time and using the linearity in (76) to cancel terms yields

$$
\frac{\mathrm{d}^2}{\mathrm{d}t^2}\Big|_{t=0}\mathcal{K}_{\gamma(t)}F = \mathcal{L}_\xi\underbrace{\frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0}\mathcal{K}_{\gamma(t)}F}_{0} + \underbrace{\mathcal{L}_\xi\mathcal{L}_\eta F + \mathcal{L}_\eta\mathcal{L}_{-\xi}F + \mathcal{L}_\eta\mathcal{L}_{-\xi}F + \mathcal{L}_{-\xi}\mathcal{L}_{-\eta}F}_{2\big(\mathcal{L}_\xi\mathcal{L}_\eta F - \mathcal{L}_\eta\mathcal{L}_\xi F\big)}
$$
$$
+ \underbrace{\frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0}\mathcal{K}_{\gamma(t)}\mathcal{L}_{-\eta}F}_{0},
$$

which completes the proof. ∎

Theorem 25 *If $F \in C(\mathcal{M}, E)$ is a continuous section, then $\mathrm{Sym}_G(F)$ is a closed, embedded Lie subgroup of G whose Lie subalgebra is*

$$\mathfrak{sym}_G(F) = \{\xi \in \mathrm{Lie}(G) \; : \; F \in D(\mathcal{L}_\xi) \text{ and } \mathcal{L}_\xi F = 0\}. \tag{57}$$

Proof We begin by showing that $\mathrm{Sym}_G(F)$ is a closed subgroup of G. It is obviously a subgroup, for if $g_1, g_2 \in \mathrm{Sym}_G(F)$ then

$$\mathcal{K}_{g_1 g_2} F = \mathcal{K}_{g_1} \mathcal{K}_{g_2} F = \mathcal{K}_{g_1} F = F,$$

meaning that $g_1 g_2 \in \mathrm{Sym}_G(F)$. To show that $\mathrm{Sym}_G(F)$ is closed, we observe that for each $p \in \mathcal{M}$, the map $h_p : G \to E$ defined by

$$h_p : g \mapsto \mathcal{K}_g F(p) = \Theta\big(F(\theta(p, g)), g^{-1}\big)$$

is continuous, as it is a composition of continuous maps. Since $F(p)$ is a single point in E, the preimage set $h_p^{-1}(F(p))$ is closed in G. Since $\mathrm{Sym}_G(F)$ is an intersection,

$$\mathrm{Sym}_G(F) = \bigcap_{p \in \mathcal{M}} h_p^{-1}(F(p)),$$

of closed sets, it follows that $\mathrm{Sym}_G(F)$ is closed in G. By the closed subgroup theorem (Theorem 20.12; Lee, 2013) it follows that $\mathrm{Sym}_G(F)$ is an embedded Lie subgroup of G.

Let $\mathfrak{h} = \mathrm{Lie}(\mathrm{Sym}_G(F))$ be the Lie algebra of $\mathrm{Sym}_G(F)$. Choosing any $\xi \in \mathfrak{h}$ we have $\exp(t\xi) \in \mathrm{Sym}_G(F)$ for every $t \in \mathbb{R}$, yielding

$$\lim_{t \to 0} \frac{1}{t} \left[\mathcal{K}_{\exp(t\xi)} F - F \right] = 0.$$

Hence, $F \in D(\mathcal{L}_\xi)$ and $\mathcal{L}_\xi F = 0$, meaning that $\mathfrak{h} \subset \mathfrak{sym}_G(F)$, as defined by (57).

To show the reverse containment, choose $\xi \in \mathfrak{sym}_G(F)$, meaning that $F \in D(\mathcal{L}_\xi)$ and $\mathcal{L}_\xi F = 0$. We observe that (54) in Proposition 24 yields

$$\frac{\mathrm{d}}{\mathrm{d}t} \mathcal{K}_{\exp(t\xi)} F = \mathcal{K}_{\exp(t\xi)} \mathcal{L}_\xi F = 0 \qquad \forall t \in \mathbb{R}.$$

It follows that $\mathcal{K}_{\exp(t\xi)} F = F$, that is, $\exp(t\xi) \in \mathrm{Sym}_G(F)$ for all $t \in \mathbb{R}$. Differentiating at $t = 0$ proves that $\xi \in \mathfrak{h}$. Therefore, $\mathfrak{h} = \mathfrak{sym}_G(F)$, which completes the proof. ∎

Theorem 26 *Suppose that G has n_G connected components with G_0 being the one that contains the identity element. Let $\xi_1, \ldots, \xi_q$ generate $\mathrm{Lie}(G)$ and let $g_1, \ldots, g_{n_G - 1}$ be elements from each non-identity component of G. A continuous section $F \in C(\mathcal{M}, E)$ is G_0-equivariant if and only if*

$$F \in D(\mathcal{L}_{\xi_i}) \quad and \quad \mathcal{L}_{\xi_i} F = 0, \qquad i = 1, \ldots, q. \tag{58}$$

If, in addition, we have

$$\mathcal{K}_{g_j} F = F, \qquad j = 1, \ldots, n_G - 1, \tag{59}$$

then F is G-equivariant.

Proof If $F \in C(\mathcal{M}, E)$ is G_0-equivariant, then $\mathcal{K}_{\exp(t\xi)} F = F$ for all $\xi \in \mathrm{Lie}(G)$ and $t \in \mathbb{R}$. Differentiating with respect to t at $t = 0$ gives $\mathcal{L}_\xi F = 0$.

Conversely, suppose that $\mathcal{L}_{\xi_i} F = 0$ for a collection of generators $\xi_1, \ldots, \xi_q$ of $\mathrm{Lie}(G)$. By Theorem 25, $\mathrm{Sym}_G(F)$ is a closed Lie subgroup of G whose Lie subalgebra $\mathfrak{sym}_G(F)$ contains $\xi_1, \ldots, \xi_q$. Since $\xi_1, \ldots, \xi_q$ generate $\mathrm{Lie}(G)$, it follows that $\mathfrak{sym}_G(F) = \mathrm{Lie}(G)$. This means that $G_0 \subset \mathrm{Sym}_G(F)$ due to the correspondence between connected Lie subgroups and their Lie subalgebras established by Theorem 19.26 in Lee (2013). Specifically, the identity component of $\mathrm{Sym}_G(F)$ must correspond to G_0 since both are connected Lie subgroups of G with identical Lie subalgebras.

Now, let us suppose in addition that $\mathcal{K}_{g_j} F = F$ for an element g_j from each non-identity component G_j, $j = 1, \ldots, n_G - 1$ of G. By Proposition 7.15 in Lee (2013), G_0 is a normal subgroup of G and every connected component G_j of G is diffeomorphic to G_0. In fact in the proof of this result it is shown that every connected component of G_j is a coset of G_0, meaning that $G_j = G_0 \cdot g_j$. Choosing any $g \in G_j$ there is an element $g_0 \in G_0$ such that $g = g_0 \cdot g_j$ and we obtain

$$\mathcal{K}_g F = \mathcal{K}_{g_0} \mathcal{K}_{g_j} F = \mathcal{K}_{g_0} F = F.$$

This completes the proof because $G = \bigcup_{j=0}^{n_G - 1} G_j$. ∎

Appendix E. Proofs on Invariant Submanifolds and Tangency

Proposition 34 *The leaves of any (nonsingular) foliation of $\mathcal{N}$ are arcwise-closed.*

Proof Consider a leaf $\mathcal{M}$ of an m-dimensional foliation on the n-dimensional manifold $\mathcal{N}$ and let $\gamma : [a, b] \to \mathcal{N}$ be a smooth curve satisfying $\gamma((a, b)) \subset \mathcal{M}$. First, it is clear that $\mathcal{M}$ is a weakly embedded submanifold of $\mathcal{N}$ since $\mathcal{M}$ is an integral manifold of an involutive distribution (Proposition 19.19; Lee, 2013) and the local structure theorem for integral manifolds (Proposition 19.16; Lee, 2013) shows that they are weakly embedded.

By continuity of γ, any neighborhood of $\gamma(b)$ in $\mathcal{N}$ must have nonempty intersection with $\mathcal{M}$. By definition of a foliation (see Lee, 2013), there is a coordinate chart $(\mathcal{U}, \boldsymbol{x})$ for $\mathcal{N}$ with $\gamma(b) \in \mathcal{U}$ such that $\boldsymbol{x}(\mathcal{U})$ is a coordinate-aligned cube in $\mathbb{R}^n$ and $\mathcal{M} \cap \mathcal{U}$ consists of countably many slices of the form $x^{m+1} = c^{m+1}, \ldots, x^n = c^n$ for constants $c^{m+1}, \ldots, c^n$. Since γ is continuous, there is a $\delta > 0$ so that $\gamma((b - \delta, b]) \subset \mathcal{U}$, and in particular, $\gamma((b - \delta, b)) \subset \mathcal{M} \cap \mathcal{U}$. By continuity of γ, there are constants $c^{m+1}, \ldots, c^n$ such that $x^i(\gamma(t)) = c^i$ for every $i = m + 1, \ldots, n$ and $t \in (b - \delta, b)$. Hence, we have

$$x^i(\gamma(b)) = \lim_{t \to b} x^i(\gamma(t)) = c^i, \qquad i = m + 1, \ldots, n,$$

meaning that $\gamma(b) \in \mathcal{M}$. An analogous argument shows that $\gamma(a) \in \mathcal{M}$, completing the proof that $\mathcal{M}$ is arcwise-closed. ∎

Our proof of Theorem 35 relies on the following technical lemma concerning the integral curves of vector fields tangent to weakly embedded, arcwise-closed submanifolds.

Lemma 43 *Let $\mathcal{M}$ be an arcwise-closed weakly embedded submanifold of a manifold $\mathcal{N}$. Let $V \in \mathfrak{X}(\mathcal{N})$ be a vector field tangent to $\mathcal{M}$, that is*

$$V_p \in T_p\mathcal{M} \qquad \forall p \in \mathcal{M}.$$

If $\gamma : I \to \mathcal{N}$ is a maximal integral curve of V that intersects $\mathcal{M}$, then γ lies in $\mathcal{M}$.

Proof By the translation lemma (Lemma 9.4; Lee, 2013), we can assume without loss of generality that $0 \in I$ and $p_0 = \gamma(0) \in \mathcal{M}$. Let $\imath_\mathcal{M} : \mathcal{M} \hookrightarrow \mathcal{N}$ denote the inclusion map. Since $\mathcal{M}$ is an immersed submanifold of $\mathcal{N}$ and V is tangent to $\mathcal{M}$, there is a unique smooth vector field $V|_\mathcal{M} \in \mathfrak{X}(\mathcal{M})$ that is $\imath_\mathcal{M}$-related to V thanks to Proposition 8.23 in Lee (2013). Let $\tilde{\gamma} : \tilde{I} \to \mathcal{M}$ be the maximal integral curve of $V|_\mathcal{M}$ with $\tilde{\gamma}(0) = p_0$. By the naturality of integral curves (Proposition 9.6; Lee, 2013) $\imath_\mathcal{M} \circ \tilde{\gamma}$ is an integral curve of V with $\imath_\mathcal{M} \circ \tilde{\gamma}(0) = p_0$. Since integral curves of smooth vector fields starting at the same point are unique (Theorem 9.12(a); Lee, 2013) we have $\tilde{I} \subset I$ and

$$\imath_\mathcal{M} \circ \tilde{\gamma}(t) = \gamma(t) \qquad \forall t \in \tilde{I}. \tag{84}$$

Therefore, it remains to show that $\tilde{I} = I$.

By the local existence of integral curves (Proposition 9.2; Lee, 2013), the domains I and $\tilde{I}$ of the maximal integral curves γ and $\tilde{\gamma}$ are open intervals in $\mathbb{R}$. Suppose, for the sake of producing a contradiction, that there exists $t \in I$ with $t > \tilde{I}$. Then it follows that the least upper bound $b = \sup \tilde{I}$ is an element of I. By (84) and continuity of γ we have

$$q_0 = \gamma(b) = \lim_{t \to b} \imath_\mathcal{M} \circ \tilde{\gamma}(t). \tag{85}$$

Since $\mathcal{M}$ is arcwise-closed, it follows that $q_0 \in \mathcal{M}$.

To complete the proof, we use the local existence of an integral curve for $V|_\mathcal{M}$ starting at q_0 to contradict the maximality of $\tilde{\gamma}$. By the local existence of integral curves (Proposition 9.2; Lee, 2013) and the translation lemma (Lemma 9.4; Lee, 2013), there is an $\varepsilon > 0$ and an integral curve $\hat{\gamma} : (b - \varepsilon, b + \varepsilon) \to \mathcal{M}$ of $V|_\mathcal{M}$ such that $\hat{\gamma}(b) = q_0 = \gamma(b)$. Shrinking the interval, we take $0 < \varepsilon < b - a$. Again, by naturality and uniqueness of integral curves we must have $\imath_\mathcal{M} \circ \hat{\gamma}(t) = \gamma(t)$ for all $t \in (b - \varepsilon, b + \varepsilon)$. Hence, by (84) and injectivity of $\imath_\mathcal{M}$ it follows that $\hat{\gamma}(t) = \tilde{\gamma}(t)$ for all $t \in (b - \varepsilon, b)$. Applying the gluing lemma (Corollary 2.8; Lee, 2013) to $\tilde{\gamma}$ and $\hat{\gamma}$ yields an extension of $\tilde{\gamma}$ to the larger open interval $\tilde{I} \cup (b - \varepsilon, b + \varepsilon)$. Since this contradicts the maximality of $\tilde{\gamma}$, there is no $t \in I$ with $t > \tilde{I}$. The same argument shows that there is no $t \in I$ with $t < \tilde{I}$, and so we must have $\tilde{I} = I$. ∎

We also use the following lemma describing the elements of a Lie group that can be constructed from products of exponentials.

Lemma 44 *Let G_0 be the identity component of a Lie group G. Then every element $g \in G_0$ can be expressed as a finite product $g = h_m \cdots h_1$ of elements $h_i = \exp(\xi_i)$ for $\xi_i \in \mathrm{Lie}(G)$.*

Let G_i be a connected component of G and let $g_i \in G_i$. Then every element $g \in G_i$ can be expressed as $g = g_0 g_i$ for some $g_0 \in G_0$.

Proof By the inverse function theorem (more specifically by Proposition 20.8(f) in Lee (2013)), the range of the exponential map contains an open, connected neighborhood $\mathcal{U}$ of the identity element $e \in G$. The inverses of the elements in $\mathcal{U}$ also belong to the range of the exponential map thanks to Proposition 20.8(c) in Lee (2013). By Proposition 7.14(b) and Proposition 7.15 in Lee (2013), it follows that $\mathcal{U}$ generates the identity component G_0 of G. That is, any element $g \in G_0$ can be written as a finite product of elements in $\mathcal{U}$ and their inverses, which proves the first claim.

By Proposition 7.15 in Lee (2013), G_0 is a normal subgroup of G and every connected component of G is diffeomorphic to G_0. In fact in the proof of this result it is shown that every connected component of G is a coset of G_0. Therefore, if G_i is a nonidentity connected component of G and $g_i \in G_i$ then $G_i = G_0 \cdot g_i$, which proves the second claim. ∎

Theorem 35 *Let $\mathcal{M}$ be an immersed submanifold of $\mathcal{N}$ and let $\theta : \mathcal{N} \times G \to \mathcal{N}$ be a right action of a Lie group G on $\mathcal{N}$ with infinitesimal generator $\hat{\theta} : \mathrm{Lie}(G) \to \mathfrak{X}(\mathcal{N})$. Then*

$$\mathfrak{sym}_G(\mathcal{M}) = \{\xi \in \mathrm{Lie}(G) \ : \ \hat{\theta}(\xi)_p \in T_p\mathcal{M} \quad \forall p \in \mathcal{M}\} \tag{68}$$

is the Lie subalgebra of a unique connected Lie subgroup $H \subset G$. If $\mathcal{M}$ is weakly embedded and arcwise-closed in $\mathcal{N}$, then this subgroup has the following properties:

(i) $\mathcal{M} \cdot H \subset \mathcal{M}$

(ii) If $\tilde{H}$ is a connected Lie subgroup of G such that $\mathcal{M} \cdot \tilde{H} \subset \mathcal{M}$, then $\tilde{H} \subset H$.

If $\mathcal{M}$ is properly embedded in $\mathcal{N}$ then H is the identity component of the closed, properly embedded Lie subgroup

$$\mathrm{Sym}_G(\mathcal{M}) = \{g \in G \ : \ \mathcal{M} \cdot g \subset \mathcal{M}\}. \tag{69}$$

Proof The set $\mathfrak{sym}_G(\mathcal{M})$ is a subspace of $\mathrm{Lie}(G)$, for if $\xi_1, \xi_2 \in \mathfrak{sym}_G(\mathcal{M})$ and $a_1, a_2 \in \mathbb{R}$ then

$$\hat{\theta}(a_1\xi_1 + a_2\xi_2)_p = a_1 \underbrace{\hat{\theta}(\xi_1)}_{\in T_p\mathcal{M}} + a_2 \underbrace{\hat{\theta}(\xi_2)}_{\in T_p\mathcal{M}} \in T_p\mathcal{M}$$

thanks to linearity of the infinitesimal generator $\hat{\theta}$. To show that $\mathfrak{sym}_G(\mathcal{M})$ is a Lie subalgebra, we must show that it is also closed under the Lie bracket. Recall that $\hat{\theta}$ is a Lie algebra homomorphism (Theorem 20.15; Lee, 2013), and so $\hat{\theta}([\xi_1, \xi_2]) = [\hat{\theta}(\xi_1), \hat{\theta}(\xi_2)]$. Since the Lie bracket of two vector fields tangent to an immersed submanifold is also tangent to the submanifold (Corollary 8.32; Lee, 2013), it follows that $[\hat{\theta}(\xi_1), \hat{\theta}(\xi_1)]$ is tangent to $\mathcal{M}$. Hence, $\mathfrak{sym}_G(\mathcal{M})$ is closed under the Lie bracket and is therefore a Lie subalgebra of $\mathrm{Lie}(G)$. By Theorem 19.26 in Lee (2013), there is a unique connected Lie subgroup of $H \subset G$ whose Lie subalgebra is $\mathfrak{sym}_G(\mathcal{M})$.

Now suppose that $\mathcal{M}$ is weakly embedded and arcwise-closed in $\mathcal{N}$. We aim to show that $\mathcal{M} \cdot H \subset \mathcal{M}$. Choosing any $\xi \in \mathfrak{sym}_G(\mathcal{M})$, Lemma 20.14 in Lee (2013) shows that ξ,

regarded as a left-invariant vector field on G, and $\hat{\xi} = \hat{\theta}(\xi)$ are $\theta^{(p)}$-related for every $p \in \mathcal{N}$. By the naturality of integral curves (Proposition 9.6; Lee, 2013) it follows that $\gamma_{\xi}^{(p)} : \mathbb{R} \to \mathcal{N}$ defined by

$$\gamma_{\xi}^{(p)}(t) = p \cdot \exp(t\xi)$$

is the unique maximal integral curve of $\hat{\xi}$ passing through p at $t = 0$. When $p \in \mathcal{M}$, this integral curve lies in $\mathcal{M}$ thanks to Lemma 43. This means that $\mathcal{M}$ is invariant under the action of any group element in the range of the exponential map restricted to $\mathfrak{sym}_G(\mathcal{M})$. Proceeding by induction, suppose that $\mathcal{M}$ is invariant under the action of any product of m such elements. If $g = h_1 \cdots h_m \cdot h_{m+1}$ is a product of $m + 1$ elements $h_i \in \exp\left(\mathfrak{sym}_G(\mathcal{M})\right) \subset H$, then it follows from associativity and the induction hypothesis that

$$\mathcal{M} \cdot (h_1 \cdots h_m \cdot h_{m+1}) = (\mathcal{M} \cdot h_1 \cdots h_m) \cdot h_{m+1} \subset \mathcal{M} \cdot h_{m+1} \subset \mathcal{M}.$$

Therefore, $\mathcal{M}$ is invariant under the action of any finite product of group elements in $\exp\left(\mathfrak{sym}_G(\mathcal{M})\right)$ by induction on m. By Lemma 44, it follows that $\mathcal{M}$ is H-invariant, proving claim (i).

To prove claim (ii), suppose that $\tilde{H}$ is another connected Lie subgroup of G such that $\mathcal{M} \cdot \tilde{H} \subset \mathcal{M}$. Choosing any $p \in \mathcal{M}$ and $\xi \in \mathrm{Lie}(\tilde{H})$, we have

$$p \cdot \exp(t\xi) \in \mathcal{M} \qquad \forall t \in \mathbb{R}.$$

Since $\mathcal{M}$ is weakly embedded in $\mathcal{N}$, this defines a smooth curve $\gamma : \mathbb{R} \to \mathcal{M}$ such that $\imath_{\mathcal{M}} \circ \gamma(t) = p \cdot \exp(t\xi)$, where $\imath_{\mathcal{M}} : \mathcal{M} \hookrightarrow \mathcal{N}$ is the inclusion map. Differentiating and using the definition of the infinitesimal generator gives

$$\hat{\theta}(\xi)_p = \left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0} p \cdot \exp(t\xi) = \mathrm{d}\,\imath_{\mathcal{M}}(p) \left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0} \gamma(t) \in T_p\mathcal{M}.$$

Therefore, $\mathrm{Lie}(\tilde{H}) \subset \mathfrak{sym}_G(\mathcal{M})$ which implies that $\tilde{H} \subset H$ by Theorem 19.26 in Lee (2013), establishing claim (ii).

Now suppose that $\mathcal{M}$ is properly embedded in $\mathcal{N}$ and denote

$$\mathrm{Sym}_G(\mathcal{M}) = \{g \in G \ : \mathcal{M} \cdot g \subset \mathcal{M}\} = \bigcap_{p \in \mathcal{M}} (\theta^{(p)})^{-1}(\mathcal{M}).$$

The equality of these expressions is a simple matter of unwinding their definitions. It is clear that $\mathrm{Sym}_G(\mathcal{M})$ is a subgroup of G, for if $g_1, g_2 \in \mathrm{Sym}_G(\mathcal{M})$ then the composition law for the group action gives $\mathcal{M} \cdot (g_1 \cdot g_2) = (\mathcal{M} \cdot g_1) \cdot g_2 \subset \mathcal{M} \cdot g_1 \subset \mathcal{M}$. Since $\mathcal{M}$ is properly embedded, it is closed in $\mathcal{N}$ (see Proposition 5.5 in Lee, 2013), meaning that each preimage set $(\theta^{(p)})^{-1}(\mathcal{M})$ is closed in G by continuity of $\theta^{(p)}$. As an intersection of closed subsets, it follows that $\mathrm{Sym}_G(\mathcal{M})$ is closed in G. By the closed subgroup theorem (Theorem 20.12; Lee, 2013), $\mathrm{Sym}_G(\mathcal{M})$ is a properly embedded Lie subgroup of G. The same holds for the identity component $\mathrm{Sym}_G(\mathcal{M})_0$ of $\mathrm{Sym}_G(\mathcal{M})$ since $\mathrm{Sym}_G(\mathcal{M})_0$ is closed in $\mathrm{Sym}_G(\mathcal{M})$, which implies that $\mathrm{Sym}_G(\mathcal{M})_0$ is closed in G.

Finally, we show that $H = \mathrm{Sym}_G(\mathcal{M})_0$ is the identity component of $\mathrm{Sym}_G(\mathcal{M})$. First, we observe that $H \subset \mathrm{Sym}_G(\mathcal{M})_0$ because H is connected and contained in $\mathrm{Sym}_G(\mathcal{M})$. The

reverse containment follows from the fact that $\mathrm{Sym}_G(\mathcal{M})_0$ is a connected Lie subgroup satisfying $\mathcal{M} \cdot \mathrm{Sym}_G(\mathcal{M})_0 \subset \mathcal{M}$, which by our earlier result implies that $\mathrm{Sym}_G(\mathcal{M})_0 \subset H$. ■

Theorem 36 *Let $\mathcal{M}$ be an arcwise-closed weakly embedded submanifold of $\mathcal{N}$ and let $\theta : \mathcal{N} \times G \to \mathcal{N}$ be a right action of a Lie group G on $\mathcal{N}$ with infinitesimal generator $\hat{\theta} : \mathrm{Lie}(G) \to \mathfrak{X}(\mathcal{N})$. Let $\xi_1, \ldots, \xi_q$ generate $\mathrm{Lie}(G)$ and let $g_1, \ldots, g_{n_G-1}$ be elements from each non-identity component of G. Then $\mathcal{M}$ is G_0-invariant if and only if*

$$\hat{\theta}(\xi_i)_p \in T_p\mathcal{M} \qquad \forall p \in \mathcal{M}, \quad i = 1, \ldots, q. \tag{70}$$

If, in addition, we have $\mathcal{M} \cdot g_j \subset \mathcal{M}$ for every $j = 1, \ldots, n_G - 1$, then $\mathcal{M}$ is G-invariant.

Proof First, suppose that $\mathcal{M}$ is G_0-invariant. In particular, this means that for every $p \in \mathcal{M}$ and $\xi \in \mathrm{Lie}(G)$, the smooth curve $\gamma_\xi^{(p)} : \mathbb{R} \to \mathcal{N}$ defined by

$$\gamma_\xi^{(p)}(t) = p \cdot \exp(t\xi)$$

lies in $\mathcal{M}$. Since $\mathcal{M}$ is weakly embedded in $\mathcal{N}$, $\gamma_\xi^{(p)}$ is also smooth as a map into $\mathcal{M}$. Specifically, there is a smooth curve $\tilde{\gamma}_\xi^{(p)} : \mathbb{R} \to \mathcal{M}$ so that $\gamma_\xi^{(p)} = \imath_\mathcal{M} \circ \tilde{\gamma}_\xi^{(p)}$ where $\imath_\mathcal{M} : \mathcal{M} \hookrightarrow \mathcal{N}$ is the inclusion map. Differentiating at $t = 0$ yields

$$\hat{\theta}(\xi)_p = \frac{\mathrm{d}}{\mathrm{d}\,t} \gamma_\xi^{(p)}(t) \Big|_{t=0} = \mathrm{d}\,\imath_\mathcal{M}(p) \frac{\mathrm{d}}{\mathrm{d}\,t} \tilde{\gamma}_\xi^{(p)}(t) \Big|_{t=0},$$

which lies in $T_p\mathcal{M} = \mathrm{Range}\,(\mathrm{d}\,\imath_\mathcal{M}(p))$. In particular, $\hat{\theta}(\xi_i)_p \in T_p\mathcal{M}$ for every $p \in \mathcal{M}$ and $i = 1, \ldots, q$.

Conversely, suppose that the tangency condition expressed in (70) holds. By Theorem 35, the elements $\xi_1, \ldots, \xi_q$ belong to the Lie subalgebra $\mathfrak{sym}_G(\mathcal{M})$ of the largest connected Lie subgroup $H \subset G$ of symmetries of $\mathcal{M}$. Since $\xi_1, \ldots, \xi_q$ generate $\mathrm{Lie}(G)$, it follows that $\mathfrak{sym}_G(\mathcal{M}) = \mathrm{Lie}(G)$. Therefore, by Theorem 19.26 in Lee (2013), we obtain $H = G_0$ because both are connected Lie subgroups of G with identical Lie subalgebras.

Finally, suppose, in addition, that $\mathcal{M} \cdot g_j \subset \mathcal{M}$ for an element g_j from each non-identity component G_j of G. By Lemma 44, if $g \in G_j$ then there is an element $g_0 \in G_0$ such that $g = g_0 \cdot g_j$. Therefore, we obtain

$$\mathcal{M} \cdot g = \mathcal{M} \cdot g_0 \cdot g_j \subset \mathcal{M} \cdot g_j \subset \mathcal{M},$$

which completes the proof because $G = \bigcup_{j=0}^{n_G-1} G_j$. ■

Lemma 38 *For every smooth section $F \in C^\infty(\mathcal{M}, E)$ there is a well-defined injective vector bundle homomorphism $\imath_F : E \to TE$ that is expressed in any local trivialization*

$\Phi : \pi^{-1}(\mathcal{U}) \to \mathcal{U} \times \mathbb{R}^k$ *as*

$$
\begin{aligned}
\mathrm{d}\,\Phi \circ \imath_F \circ \Phi^{-1} : \mathcal{U} \times \mathbb{R}^k &\to T(\mathcal{U} \times \mathbb{R}^k) \\
(p, \boldsymbol{v}) &\mapsto (0, \boldsymbol{v})_{\Phi(F(p))}.
\end{aligned}
\tag{71}
$$

Proof The map $\imath_F$ defined in a local trivialization Φ by (71) is injective. It is a vector bundle homomorphism because $\mathrm{d}\,\Phi \circ \imath_F \circ \Phi^{-1}$, Φ, and $\mathrm{d}\,\Phi$ are vector bundle homomorphisms and Φ and $\mathrm{d}\,\Phi$ are invertible. It remains to show that the definition of $\imath_F$ does not depend on the choice of local trivialization. Given two local trivializations Φ and $\tilde{\Phi}$ defined on $\pi^{-1}(\mathcal{U}) \subset E$ where $\mathcal{U}$ is an open subset of $\mathcal{M}$, it suffices to show that the following diagram commutes:

$$
\begin{array}{ccccc}
\mathcal{U} \times \mathbb{R}^k & \xleftarrow{\quad \Phi \quad} & \pi^{-1}(\mathcal{U}) \subset E & \xrightarrow{\quad \tilde{\Phi} \quad} & \mathcal{U} \times \mathbb{R}^k \\
\downarrow{\scriptstyle \imath_{\Phi \circ F}} & & \downarrow{\scriptstyle \imath_F} & & \downarrow{\scriptstyle \imath_{\tilde{\Phi} \circ F}} \\
T(\mathcal{U} \times \mathbb{R}^k) & \xleftarrow{\quad \mathrm{d}\,\Phi \quad} & T(\pi^{-1}(\mathcal{U})) \subset TE & \xrightarrow{\quad \mathrm{d}\,\tilde{\Phi} \quad} & T(\mathcal{U} \times \mathbb{R}^k)
\end{array}
$$

Since $\tilde{\Phi} \circ \Phi^{-1}$ is a bundle homomorphism descending to the identity, it can be written as

$$
\tilde{\Phi} \circ \Phi^{-1} : (p, \boldsymbol{v}) \mapsto (p, \boldsymbol{T}(p)\boldsymbol{v})
$$

for a matrix-valued function $\boldsymbol{T} : \mathcal{U} \to \mathbb{R}^{k \times k}$. Moreover, the matrices are invertible because the local trivializations are bundle isomorphisms. Differentiating, we obtain

$$
\mathrm{d}\,\tilde{\Phi} \circ \mathrm{d}\,\Phi^{-1} : (w_p, \boldsymbol{w})_{(p, \boldsymbol{v})} \mapsto (w_p, \mathrm{d}\,\boldsymbol{T}(p)w_p + \boldsymbol{T}(p)\boldsymbol{w})_{\tilde{\Phi} \circ \Phi^{-1}(p, \boldsymbol{v})},
$$

where $w_p \in T_p\mathcal{U}$. Composing this with $\imath_{\Phi \circ F} : (p, \boldsymbol{v}) \mapsto (0, \boldsymbol{v})_{\Phi(F(p))}$, we obtain

$$
\mathrm{d}\,\tilde{\Phi} \circ \mathrm{d}\,\Phi^{-1} \circ \imath_{\Phi \circ F}(p, \boldsymbol{v}) = (0, \boldsymbol{T}(p)\boldsymbol{v})_{\tilde{\Phi}(F(p))} = \imath_{\tilde{\Phi} \circ F}(p, \boldsymbol{T}(p)\boldsymbol{v}) = \imath_{\tilde{\Phi} \circ F} \circ \tilde{\Phi} \circ \Phi^{-1}(p, \boldsymbol{v}),
$$

proving that the diagram commutes. ∎

Theorem 39 *Given a smooth section $F \in C^\infty(\mathcal{M}, E)$ and $p \in \mathcal{M}$, the map $P_{F(p)} := \mathrm{d}(F \circ \pi)(F(p)) : T_{F(p)}E \to T_{F(p)}E$ is a linear projection onto $T_{F(p)} \mathrm{im}(F)$ and for every $\xi \in \mathrm{Lie}(G)$ we have*

$$
\imath_F \circ (\mathcal{L}_\xi F)(p) = -\big[I - P_{F(p)}\big]\hat{\Theta}(\xi)_{F(p)}.
\tag{72}
$$

Proof We observe that $F \circ \pi : E \to E$ is a smooth idempotent map whose image is $\mathrm{im}(F) \subset E$. By differentiating the expression $(F \circ \pi) \circ (F \circ \pi) = F \circ \pi$ at a point $F(p) \in \mathrm{im}(F)$, we obtain

$$
\mathrm{d}(F \circ \pi)(F(p))\,\mathrm{d}(F \circ \pi)(F(p)) = \mathrm{d}(F \circ \pi)(F(p)),
$$

meaning that $\mathrm{d}(F \circ \pi)(F(p)) : T_{F(p)}E \to T_{F(p)}E$ is a linear projection. Since

$$
\mathrm{d}(F \circ \pi)(F(p)) = \mathrm{d}\,F(p)\,\mathrm{d}\,\pi(F(p)),
$$

we have $\text{Range}\left(\mathrm{d}(F \circ \pi)(F(p))\right) \subset \text{Range}(\mathrm{d}\,F(p)) = T_{F(p)}\,\text{im}(F)$. Differentiating $F = (F \circ \pi) \circ F$ yields

$$\mathrm{d}\,F(p) = \mathrm{d}(F \circ \pi)(F(p))\,\mathrm{d}\,F(p),$$

meaning that $T_{F(p)}\,\text{im}(F) \subset \text{Range}\left(\mathrm{d}(F \circ \pi)(F(p))\right)$. Since $\text{Range}\left(\mathrm{d}(F \circ \pi)(F(p))\right) = T_{F(p)}\,\text{im}(F)$ it follows that $\mathrm{d}(F \circ \pi)(F(p))$ is a linear projection onto $T_{F(p)}\,\text{im}(F)$.

The generalized Lie derivative in (53) can be expressed as

$$
\begin{aligned}
(\mathcal{L}_\xi F)(p) &= \lim_{t \to 0} \frac{1}{t}\left[\Theta_{\exp(-t\xi)}(F(\theta_{\exp(t\xi)}(p))) - F(p)\right] \\
&= \lim_{t \to 0} \Theta\left(\exp(-t\xi),\ \frac{1}{t}\left[F(\theta_{\exp(t\xi)}(p)) - \Theta_{\exp(t\xi)}(F(p))\right]\right) \\
&= \lim_{t \to 0} \frac{1}{t}\left[F(\theta_{\exp(t\xi)}(p)) - \Theta_{\exp(t\xi)}(F(p))\right].
\end{aligned}
$$

The first equality follows because $\Theta_{g^{-1}}$ is a vector bundle homomorphism, meaning that the restricted map $\Theta_{g^{-1}}|_{E_{p\cdot g}} : E_{p\cdot g} \to E_p$ is linear; here $g = \exp(t\xi)$. The second equality follows because $\Theta : E \times G \to E$ is continuous. Note that in the first expression the limit is taken in the vector space E_p, whereas in the last expression the limit must be taken in E.

We proceed by expressing everything in a local trivialization $\Phi : \pi^{-1}(\mathcal{U}) \to \mathcal{U} \times \mathbb{R}^k$ of an open neighborhood $\mathcal{U} \subset \mathcal{M}$ of $p \in \mathcal{M}$. Since the maps Θ_g, Φ, and Φ^{-1} are vector bundle homomorphisms, there is a matrix-valued function $\boldsymbol{T}_g : \mathcal{U} \to \mathbb{R}^{k \times k}$ such that

$$\tilde{\Theta}_g = \Phi \circ \Theta_g \circ \Phi^{-1} : (p, \boldsymbol{v}) \mapsto (\theta_g(p),\ \boldsymbol{T}_g(p)\boldsymbol{v}).$$

Differentiating $\tilde{\Theta}_{\exp(t\xi)}(p, \boldsymbol{v})$ with respect to t yields the generator

$$\hat{\tilde{\Theta}}(\xi)_{(p,\boldsymbol{v})} = \left.\frac{\mathrm{d}}{\mathrm{d}\,t}\right|_{t=0} \tilde{\Theta}_{\exp(t\xi)}(p, \boldsymbol{v}) = \left(\hat{\theta}(\xi)_p,\ \hat{\boldsymbol{T}}(\xi)_p\boldsymbol{v}\right)_{(p,\boldsymbol{v})},$$

where $\hat{\boldsymbol{T}}(\xi)_p = \left.\frac{\mathrm{d}}{\mathrm{d}\,t}\right|_{t=0} \boldsymbol{T}_{\exp(t\xi)}(p)$. We define the function $\tilde{\boldsymbol{F}} : \mathcal{U} \to \mathbb{R}^k$ by

$$(p, \tilde{\boldsymbol{F}}(p)) = \Phi \circ F(p) \qquad \forall p \in \mathcal{U}.$$

Using the above definitions, we can express the generalized Lie derivative in the local trivialization:

$$
\begin{aligned}
\Phi \circ (\mathcal{L}_\xi F)(p) &= \lim_{t \to 0}\left(\theta_{\exp(t\xi)}(p),\ \frac{1}{t}\left[\tilde{\boldsymbol{F}}(\theta_{\exp(t\xi)}(p)) - \boldsymbol{T}_{\exp(t\xi)}\tilde{\boldsymbol{F}}(p)\right]\right) \\
&= \left(p,\ \mathrm{d}\,\tilde{\boldsymbol{F}}(p)\hat{\theta}(\xi)_p - \hat{\boldsymbol{T}}(\xi)_p\tilde{\boldsymbol{F}}(p)\right).
\end{aligned}
\tag{86}
$$

Applying Lemma 38 allows us to express the left-hand side of (72) as

$$\mathrm{d}\,\Phi\left[\imath_F \circ (\mathcal{L}_\xi F)(p)\right] = \left(0,\ \mathrm{d}\,\tilde{\boldsymbol{F}}(p)\hat{\theta}(\xi)_p - \hat{\boldsymbol{T}}(\xi)_p\tilde{\boldsymbol{F}}(p)\right)_{\Phi(F(p))}. \tag{87}$$

We can also express the quantities on the right-hand side of (72) in the local trivialization. To do this, we compute

$$
\begin{aligned}
\mathrm{d}\,\Phi(F(p))\,\mathrm{d}(F\circ\pi)(F(p))\hat{\Theta}(\xi)_{F(p)} &= \frac{\mathrm{d}}{\mathrm{d}\,t}\bigg|_{t=0}\Phi\circ F\circ\pi\circ\Theta_{\exp(t\xi)}(F(p)) \\
&= \frac{\mathrm{d}}{\mathrm{d}\,t}\bigg|_{t=0}\Phi\circ F(\theta_{\exp(t\xi)}(p)) \\
&= \left(\hat{\theta}(\xi)_p,\ \mathrm{d}\,\tilde{\boldsymbol{F}}(p)\hat{\theta}(\xi)_p\right)_{\Phi(F(p))}
\end{aligned}
$$

and

$$
\begin{aligned}
\mathrm{d}\,\Phi(F(p))\hat{\Theta}(\xi)_{F(p)} &= \frac{\mathrm{d}}{\mathrm{d}\,t}\bigg|_{t=0}\Phi\circ\Theta_{\exp(t\xi)}\circ\Phi^{-1}\circ\Phi\circ F(p) \\
&= \hat{\tilde{\Theta}}(\xi)_{\Phi(F(p))} = \left(\hat{\theta}(\xi)_p,\ \hat{\boldsymbol{T}}(\xi)_p\tilde{\boldsymbol{F}}(p)\right)_{\Phi(F(p))}.
\end{aligned}
$$

Subtracting these yields

$$
\mathrm{d}\,\Phi\left[\left(\mathrm{d}(F\circ\pi)(F(p))-\mathrm{Id}_{T_{F(p)}E}\right)\hat{\Theta}(\xi)_{F(p)}\right] = \left(0,\ \mathrm{d}\,\tilde{\boldsymbol{F}}(p)\hat{\theta}(\xi)_p - \hat{\boldsymbol{T}}(\xi)_p\tilde{\boldsymbol{F}}(p)\right)_{\Phi(F(p))},
$$

which, upon comparison with (87) completes the proof. ■

References

R. Abraham, J. E. Marsden, and T. Ratiu. *Manifolds, Tensor Analysis, and Applications*, volume 75 of *Applied Mathematical Sciences*. Springer-Verlag, 1988.

Ralph Abraham and Jerrold E Marsden. *Foundations of mechanics*. AMS Chelsea Publishing, 2 edition, 2008.

Akshay Agrawal, Robin Verschueren, Steven Diamond, and Stephen Boyd. A rewriting system for convex optimization problems. *Journal of Control and Decision*, 5(1):42–60, 2018.

Amir Ali Ahmadi and Bachir El Khadir. Learning dynamical systems with side information. In *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, volume 120 of *Proceedings of Machine Learning Research*, pages 718–727. PMLR, 10–11 Jun 2020. URL https://proceedings.mlr.press/v120/ahmadi20a.html.

Amir Ali Ahmadi and Bachir El Khadir. Learning dynamical systems with side information. *SIAM Review*, 65(1):183–223, 2023. doi: 10.1137/20M1388644.

Tara Akhound-Sadegh, Laurence Perreault-Levasseur, Johannes Brandstetter, Max Welling, and Siamak Ravanbakhsh. Lie point symmetry and physics-informed networks. *Advances in Neural Information Processing Systems*, 36, 2024.

Peter J Baddoo, Benjamin Herrmann, Beverley J McKeon, J Nathan Kutz, and Steven L Brunton. Physics-informed dynamic mode decomposition. *Proceedings of the Royal Society A*, 479(2271):20220576, 2023.

Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E Smidt, and Boris Kozinsky. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature communications*, 13(1):2453, 2022.

Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

Gregory Benton, Marc Finzi, Pavel Izmailov, and Andrew G Wilson. Learning invariances in neural networks from training data. *Advances in neural information processing systems*, 33:17605–17616, 2020.

Tyrus Berry and Dimitrios Giannakis. Spectral exterior calculus. *Communications on Pure and Applied Mathematics*, 73(4):689–770, 2020.

Nicolas Boullé and Alex Townsend. A mathematical guide to operator learning. *arXiv preprint arXiv:2312.14688*, 2023.

Thierry Bouwmans, Sajid Javed, Hongyang Zhang, Zhouchen Lin, and Ricardo Otazo. On the applications of robust PCA in image and video processing. *Proceedings of the IEEE*, 106(8):1427–1457, 2018.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL `http://github.com/jax-ml/jax`.

Johannes Brandstetter, Max Welling, and Daniel E Worrall. Lie point symmetry data augmentation for neural pde solvers. In *International Conference on Machine Learning*, pages 2241–2256. PMLR, 2022.

S. L. Brunton and J. N. Kutz. *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*. Cambridge University Press, 2nd edition, 2022.

Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the national academy of sciences*, 113(15):3932–3937, 2016.

Steven L. Brunton, Bernd R. Noack, and Petros Koumoutsakos. Machine learning for fluid mechanics. *Annual Review of Fluid Mechanics*, 52:477–508, 2020.

Steven L Brunton, Marko Budišić, Eurika Kaiser, and J Nathan Kutz. Modern Koopman theory for dynamical systems. *SIAM Review*, 64(2):229–340, 2022.

Leon Bungert, Tim Roith, Daniel Tenbrinck, and Martin Burger. A Bregman learning framework for sparse neural networks. *Journal of Machine Learning Research*, 23(192): 1–43, 2022.

Jameson Cahill, Dustin G Mixon, and Hans Parshall. Lie PCA: Density estimation for symmetric manifolds. *Applied and Computational Harmonic Analysis*, 2023.

Jared L Callaham, Georgios Rigas, Jean-Christophe Loiseau, and Steven L Brunton. An empirical mean-field model of symmetry-breaking in a turbulent wake. *Science Advances*, 8(eabm4786), 2022.

Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9:717–772, 2009.

Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):1–37, 2011.

Richard Caron and Tim Traynor. The zero set of a polynomial. *WSMR Report 05-02*, 2005. URL https://www.researchgate.net/profile/Richard-Caron-3/publication/281285245_The_Zero_Set_of_a_Polynomial/links/55df56b608aecb1a7cc1a043/The-Zero-Set-of-a-Polynomial.pdf.

Kathleen Champion, Peng Zheng, Aleksandr Y Aravkin, Steven L Brunton, and J Nathan Kutz. A unified sparse optimization framework to learn parsimonious physics-informed models from data. *IEEE Access*, 8:169259–169271, 2020.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

Shuxiao Chen, Edgar Dobriban, and Jane H. Lee. A group-theoretic framework for data augmentation. *J. Mach. Learn. Res.*, 21(1), jan 2020. ISSN 1532-4435.

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

Taco S Cohen, Mario Geiger, Jonas Köhler, and Max Welling. Spherical CNNs. *arXiv preprint arXiv:1801.10130*, 2018.

Taco S Cohen, Mario Geiger, and Maurice Weiler. A general theory of equivariant CNNs on homogeneous spaces. *Advances in neural information processing systems*, 32, 2019.

Matthew J Colbrook. The mpEDMD algorithm for data-driven computations of measure-preserving dynamical systems. *SIAM Journal on Numerical Analysis*, 61(3):1585–1608, 2023.

Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 113–123, 2019.

DeepMind, Igor Babuschkin, Kate Baumli, Alison Bell, Surya Bhupatiraju, Jake Bruce, Peter Buchlovsky, David Budden, Trevor Cai, Aidan Clark, Ivo Danihelka, Antoine Dedieu, Claudio Fantacci, Jonathan Godwin, Chris Jones, Ross Hemsley, Tom Hennigan, Matteo

Hessel, Shaobo Hou, Steven Kapturowski, Thomas Keck, Iurii Kemaev, Michael King, Markus Kunesch, Lena Martens, Hamza Merzic, Vladimir Mikulik, Tamara Norman, George Papamakarios, John Quan, Roman Ring, Francisco Ruiz, Alvaro Sanchez, Laurent Sartran, Rosalia Schneider, Eren Sezener, Stephen Spencer, Srivatsan Srinivasan, Miloš Stanojević, Wojciech Stokowiec, Luyu Wang, Guangyao Zhou, and Fabio Viola. The DeepMind JAX Ecosystem, 2020. URL `http://github.com/google-deepmind`.

Krish Desai, Benjamin Nachman, and Jesse Thaler. Symmetry discovery with deep learning. *Physical Review D*, 105(9):096031, 2022.

Steven Diamond and Stephen Boyd. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

Carlos Esteves, Christine Allen-Blanchette, Ameesh Makadia, and Kostas Daniilidis. Learning SO(3) equivariant representations with spherical CNNs. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 52–68, 2018.

Maryam Fazel, Ting Kei Pong, Defeng Sun, and Paul Tseng. Hankel matrix rank minimization with applications to system identification and realization. *SIAM Journal on Matrix Analysis and Applications*, 34(3):946–977, 2013.

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

Marc Finzi, Max Welling, and Andrew Gordon Wilson. A practical method for constructing equivariant multilayer perceptrons for arbitrary matrix groups. In *International Conference on Machine Learning*, pages 3318–3328. PMLR, 2021.

Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4):193–202, 1980.

Somdatta Goswami, Aniruddha Bora, Yue Yu, and George Em Karniadakis. Physics-informed deep neural operator networks. In *Machine Learning in Modeling and Simulation: Methods and Applications*, pages 219–254. Springer, 2023.

Morikuni Gotô. Faithful representations of Lie groups II. *Nagoya mathematical journal*, 1: 91–107, 1950.

David Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, 2011.

David J Gross. The role of symmetry in fundamental physics. *Proceedings of the National Academy of Sciences*, 93(25):14256–14259, 1996.

Nate Gruver, Marc Anton Finzi, Micah Goldblum, and Andrew Gordon Wilson. The Lie derivative for measuring learned equivariance. In *The Eleventh International Conference on Learning Representations*, 2022.

Yifei Guan, Steven L Brunton, and Igor Novosselov. Sparse nonlinear models of chaotic electroconvection. *Royal Society Open Science*, 8(8):202367, 2021.

Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.

Brian C. Hall. *Lie Groups, Lie Algebras, and Representations: An Elementary Introduction.* Springer, 2015.

Ryuichiro Hataya, Jan Zdenek, Kazuki Yoshizoe, and Hideki Nakayama. Faster autoaugment: Learning augmentation strategies using backpropagation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pages 1–16. Springer, 2020.

P. J. Holmes, J. L. Lumley, G. Berkooz, and C. W. Rowley. *Turbulence, coherent structures, dynamical systems and symmetry.* Cambridge Monographs in Mechanics. Cambridge University Press, Cambridge, England, 2nd edition, 2012.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis.* Cambridge University Press, 2 edition, 2013.

Eurika Kaiser, J Nathan Kutz, and Steven L Brunton. Discovering conservation laws from data for control. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 6415–6421. IEEE, 2018.

Eurika Kaiser, J Nathan Kutz, and Steven L Brunton. Data-driven discovery of koopman eigenfunctions for control. *Machine Learning: Science and Technology*, 2(3):035023, 2021.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Ivan Kolář, Peter W. Michor, and Jan Slovák. *Natural Operations in Differential Geometry.* Springer-Verlag, 1993.

Risi Kondor and Shubhendu Trivedi. On the generalization of equivariance and convolution in neural networks to the action of compact groups. In *International Conference on Machine Learning*, pages 2747–2755. PMLR, 2018.

B. O. Koopman. Hamiltonian systems and transformations in Hilbert space. *Proceedings of the National Academy of Sciences*, 17:315–318, 1931.

Leonid B. Koralov and Yakov G. Sinai. *Theory of Probability and Random Processes.* Springer, 2 edition, 2012.

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

Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

John M. Lee. *Introduction to Smooth Manifolds: Second Edition.* Springer, 2013.

Mario Lezcano-Casado and David Martınez-Rubio. Cheap orthogonal constraints in neural networks: A simple parametrization of the orthogonal and unitary group. In *International Conference on Machine Learning*, pages 3794–3803. PMLR, 2019.

Ziming Liu and Max Tegmark. Machine learning hidden symmetries. *Phys. Rev. Lett.*, 128: 180201, May 2022. doi: 10.1103/PhysRevLett.128.180201. URL https://link.aps.org/doi/10.1103/PhysRevLett.128.180201.

J.-C. Loiseau and S. L. Brunton. Constrained sparse Galerkin regression. *Journal of Fluid Mechanics*, 838:42–67, 2018.

Haggai Maron, Heli Ben-Hamu, Nadav Shamir, and Yaron Lipman. Invariant and equivariant graph networks. *arXiv preprint arXiv:1812.09902*, 2018.

J. E. Marsden and T. S. Ratiu. *Introduction to mechanics and symmetry.* Springer-Verlag, 2nd edition, 1999.

Alexandre Mauroy, Y Susuki, and I Mezić. *Koopman operator in systems and control.* Springer, 2020.

Daniel A Messenger and David M Bortz. Weak SINDy: Galerkin-based data-driven model selection. *Multiscale Modeling & Simulation*, 19(3):1474–1497, 2021.

Carl D Meyer. *Matrix analysis and applied linear algebra*, volume 71. Siam, 2000.

Igor Mezić. Spectral properties of dynamical systems, model reduction and decompositions. *Nonlinear Dynamics*, 41:309–325, 2005.

Xu Miao and Rajesh PN Rao. Learning the Lie groups of visual invariance. *Neural computation*, 19(10):2665–2693, 2007.

Karthik Mohan and Maryam Fazel. Iterative reweighted algorithms for matrix rank minimization. *Journal of Machine Learning Research*, 13(1):3441–3473, 2012.

Artem Moskalev, Anna Sepliarskaia, Ivan Sosnovik, and Arnold Smeulders. LieGG: Studying learned Lie group generators. *Advances in Neural Information Processing Systems*, 35: 25212–25223, 2022.

E Noether. Invariante variationsprobleme nachr. d. könig. gesellsch. d. wiss. zu göttingen, math-phys. klasse 1918: 235-257. *English Reprint: physics/0503066, http://dx. doi. org/10.1080/00411457108231446*, page 57, 1918.

Peter J. Olver. *Applications of Lie Groups to Differential Equations.* Springer, 1986.

Samuel E Otto and Clarence W Rowley. Koopman operators for estimation and control of dynamical systems. *Annual Review of Control, Robotics, and Autonomous Systems*, 4: 59–87, 2021.

Sebastian Peitz, Hans Harder, Feliks Nüske, Friedrich Philipp, Manuel Schaller, and Karl Worthmann. Partial observations, coarse graining and equivariance in Koopman operator theory for large-scale dynamical systems. *arXiv preprint arXiv:2307.15325*, 2023.

Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving

nonlinear partial differential equations. *Journal of Computational physics*, 378:686–707, 2019.

Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017.

Rajesh Rao and Daniel Ruderman. Learning Lie groups for invariant visual perception. *Advances in neural information processing systems*, 11, 1999.

Benjamin Recht, Maryam Fazel, and Pablo A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.

Patrick AK Reinbold, Daniel R Gurevich, and Roman O Grigoriev. Using noisy or incomplete data to discover models of spatiotemporal dynamics. *Physical Review E*, 101(010203), 2020.

David W Romero and Suhas Lohit. Learning partial equivariances from data. *Advances in Neural Information Processing Systems*, 35:36466–36478, 2022.

Clarence W. Rowley, Ioannis G. Kevrekidis, Jerrold E. Marsden, and Kurt Lust. Reduction and reconstruction for self-similar dynamical systems. *Nonlinearity*, 16(4):1257, 2003.

Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

Vivian T Steyert. *Uncovering Structure with Data-driven Reduced-Order Modeling*. PhD thesis, Princeton University, 2022.

Terence Tao. Two small facts about Lie groups. `https://terrytao.wordpress.com/2011/06/25/two-small-facts-about-lie-groups/`, 6 2011.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

David A Van Dyk and Xiao-Li Meng. The art of data augmentation. *Journal of Computational and Graphical Statistics*, 10(1):1–50, 2001.

V. S. Varadarajan. *Lie groups, Lie algebras, and their representations*. Springer, 1984.

Rui Wang, Robin Walters, and Rose Yu. Approximately equivariant networks for imperfectly symmetric dynamics. In *International Conference on Machine Learning*, pages 23078–23091. PMLR, 2022.

Maurice Weiler and Gabriele Cesa. General E(2)-equivariant steerable CNNs. *Advances in neural information processing systems*, 32, 2019.

Maurice Weiler, Mario Geiger, Max Welling, Wouter Boomsma, and Taco S Cohen. 3D steerable CNNs: Learning rotationally equivariant features in volumetric data. *Advances in Neural Information Processing Systems*, 31, 2018.

Matthew O Williams, Ioannis G Kevrekidis, and Clarence W Rowley. A data-driven approximation of the Koopman operator: extending dynamic mode decomposition. *Journal of Nonlinear Science*, 6:1307–1346, 2015.

David P Woodruff et al. Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science*, 10(1–2):1–157, 2014.

Jianke Yang, Nima Dehmamy, Robin Walters, and Rose Yu. Latent space symmetry discovery. *arXiv preprint arXiv:2310.00105*, 2023a.

Jianke Yang, Robin Walters, Nima Dehmamy, and Rose Yu. Generative adversarial symmetry discovery. *arXiv preprint arXiv:2302.00236*, 2023b.

Jianke Yang, Wang Rao, Nima Dehmamy, Robin Walters, and Rose Yu. Symmetry-informed governing equation discovery. *arXiv preprint arXiv:2405.16756*, 2024.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.