Almost Equivariance via Lie Algebra Convolutions

Daniel McNeela MCNEELA@WISC.EDU
Department of Computer Sciences
University of Wisconsin, Madison

Abstract

Recently, the *equivariance* of models with respect to a group action has become an important topic of research in machine learning. Analysis of the built-in equivariance of existing neural network architectures, as well as the study of methods for building model architectures that explicitly "bake in" equivariance, have become significant research areas in their own right. However, imbuing an architecture with a specific group equivariance imposes a strong prior on the types of data transformations that the model expects to see. While strictly-equivariant models enforce symmetries, such as those due to rotations or translations, real-world data does not always follow such strict equivariances, be it due to noise in the data or underlying physical laws that encode only approximate or partial symmetries. In such cases, the prior of strict equivariance can actually prove too strong and cause models to underperform on real-world data. Therefore, in this work we study a closely related topic, that of *almost equivariance.* We give a practical method for encoding almost equivariance in models by appealing to the Lie algebra of a Lie group and defining *Lie algebra convolutions.* We demonstrate that Lie algebra convolutions offer several benefits over Lie group convolutions, including being computationally tractable and well-defined for non-compact groups. Finally, we demonstrate the validity of our approach by benchmarking against datasets in fully equivariant and almost equivariant settings.

Keywords: Equivariance, partial equivariance, approximate equivariance, almost equivariance, soft equivariance

1. Introduction

The past few years have shown a surge in interest in *equivariant* model architectures, those that explicitly impose symmetry with respect to a particular group acting on the model inputs. While it was long-believed that data augmentation strategies could take the place of equivariant model architectures, recent work has demonstrated that this is not the case (Gerken et al., 2022; Lafarge et al., 2020; Wang et al., 2022b). As such, developing methods for building neural network layers that are equivariant to general group actions is of great importance.

More recently, *almost equivariance*, also referred to variously as *approximate, soft,* or *partial equivariance*, has become a rich topic of study. The idea is that the symmetry constraints imposed by full equivariance are not always completely conformed to in real-world systems. For example, the motion of a pendulum in a vacuum is fully symmetric about the midpoint of its arc, but when outside forces such as wind resistance are introduced, only partial equivariance is achieved on each pendulum swing. Accurately modeling real-world physical systems therefore requires building model architectures that have a built-in notion of symmetry but that are not so constrained by it as to be incapable of fully modeling the underlying system dynamics.

2. Related Work

2.1. Strict Equivariance

Much of the work in developing strictly-equivariant model architectures began with the seminal paper of Cohen and Welling (2016), which introduced the group-equivariant convolutional neural network layer. Kondor and Trivedi (2018) generalized this notion of equivariance and convolution to the action of an arbitrary compact group. Further generalizations followed, with the creation of convolutions (Finzi et al., 2020) and efficient MLP layers (Finzi et al., 2021a) equivariant to arbitrary Lie groups. Other neural network types have also been studied through the lens of equivariance, for example, graph neural networks (Satorras et al., 2021), (Batzner et al., 2022), transformers (Hutchinson et al., 2021), and graph transformers (Liao and Smidt, 2023). Cohen et al. (2019) consolidated much of this work into a general framework via which equivariant layers can be understood as maps between spaces of sections of vector bundles. Similar to our work, Dehmamy et al. (2021) devised a convolutional layer on the Lie algebra designed to approximate group convolutional layers. However, their objective was to make the layer as close to equivariant as possible whereas our layer is designed to be flexible so as to be capable of modelling almost equivariances. Finally, rather than devising a new equivariant layer type, Gruver et al. (2023) developed a method based on the Lie derivative which can be used to detect the degree of equivariance learned by an arbitrary model architecture.

2.2. Almost Equivariance

One of the first works on almost equivariance was Finzi et al. (2021b), which introduced the *Residual Pathway Prior* model. Their idea is to construct a neural network layer f that is the sum of two components, A and B, where A is a strictly equivariant layer and B is a more flexible, non-equivariant layer. Furthermore, they place priors on the sizes of A and B such that a model trained using maximum a posteriori estimation is incentivized to favor the strict equivariance of A while relying on B only to explain the difference between f and the fully symmetric architecture determined by A. The priors on A and B can be defined so as to weight the layer towards favoring the use of A.

The approach taken in Wang et al. (2022a) is somewhat different. They give an explicit definition of approximate equivariance then model it via a *relaxed group convolutional layer*, wherein the single kernel Ψ of a strictly equivariant group convolutional layer is replaced with a set of kernels $\{\Psi_l\}_{l=1}^{L}$. This introduces a specific, symmetry-breaking dependence on a pair of group elements (g, h).

Romero and Lohit (2022) take an altogether different approach. They introduce a model, which they call the *Partial G-CNN*, and show how to train it to learn layer-wise levels of equivariance from data. A key differentiator in their approach is the learning of a probability distribution over group elements at each group convolutional layer, allowing them to sample group elements during group convolutions.

van der Ouderaa et al. (2022) relax equivariance constraints by defining a non-stationary group convolution. They parameterize the kernel for the convolution by choosing a basis for the Lie algebra, $\mathfrak{g}$, of G and defining elements $g \in G$ as exponential maps of Lie algebra elements.

Finally, Petrache and Trivedi (2023) provide a take on approximate equivariance rooted in statistical learning theory and provide generalization and error bounds on approximately equivariant architectures.

3. Method

3.1. Equivariance & Almost Equivariance

We first give the definitions of equivariance and almost equivariance upon which this paper is based. In defining almost equivariance of a model with respect to the action of some Lie group, G, we seek a definition that offers both theoretical insight as well as practical significance. We start by addressing the abstract case, in which we define almost equivariance for general functions on a Riemannian manifold. We then drop to the level of practice and give a method for encoding almost equivariance into a machine learning model taking inputs on some data manifold.

Definition 1 (equivariant function) *Let G be a Lie group acting smoothly on smooth Riemannian manifolds (M, g) and (N, h) via the left actions $G \times M \to M$ and $G \times N \to N$ given by $(g, x) \mapsto g \cdot x$. Furthermore, let f be a mapping of smooth manifolds, $f : M \to N$. Then we say f is equivariant with respect to the action of G if it commutes with the actions of G on M and N, i.e.*

$$g \cdot f(x) = f(g \cdot x)$$

Definition 2 (ε-almost equivariant function) *Now, consider the same setup as in the previous definition. We say a function $f : M \to N$ is ε-almost equivariant if the following is satisfied*

$$d(f(g \cdot x), g \cdot f(x)) < \varepsilon$$

for all $g \in G$ and $x \in M$, where d is the distance metric on N. We can think of such a function as commuting with the actions of G on M and N to within some ε.

3.2. Lie Algebra Convolutions

Similar to the approach taken in van der Ouderaa et al. (2022), we build an almost equivariant neural network layer based on the Lie algebra, $\mathfrak{g}$, of a matrix Lie group, $G \leqslant GL_n(\mathbb{R})$. However, our model makes use of a few, key differences. First, rather than parametrizing our kernel in a finite-dimensional random Fourier features basis, we instead encode the Lie algebra basis explicitly. For most matrix Lie groups, the corresponding Lie algebra basis has an easily calculated set of generators, i.e. a set of basis elements, $\{x_i\}$. Second, instead of mapping elements of $\mathfrak{g}$ directly to G via the exponential map, we train a neural network, $\mathcal{N}_\theta : \mathfrak{g} \to \mathbb{R}^{n \times n}$, to learn an approximation to this mapping directly from data. This confers some key benefits over previous approaches. For one, the kernels used in past work are still constrained to take as input only group elements, $v \in G$, which to some extent limits the flexibility with which they can model partial equivariances. In contrast, our kernel can take any $x \in \mathbb{R}^{n \times n}$ as an input, allowing us to model a more flexible class of functions while still maintaining the interpretability achieved by parameterizing this function class via elements of the Lie algebra.

Furthermore, whereas van der Ouderaa et al. (2022) relax equivariance constraints by letting their kernel depend on an absolute group element, v, we define a simpler convolution that still allows us to relax equivariance constraints.

Definition 3 (Almost Equivariant Lie Algebra Convolution) *We construct an almost equivariant Lie algebra convolution by letting $u, x = \sum_{i=1}^{\dim \mathfrak{g}} c_i x_i \in \mathfrak{g}$ and defining*

$$h(u) = (k_\omega \star f)(u) = \int_{x \in \mathfrak{g}} k_\omega \left(\mathcal{N}_\theta(x)^{-1} \exp(u) \right) f(x) d\mu(x)$$

Here, instead of integrating with respect to the Haar measure, as would be required if we were integrating over the Lie group, G, we are able to instead integrate with respect to the Lebesgue measure, μ, defined on $\mathbb{R}^{n \times n}$. This is because we are integrating over the Lie algebra, $\mathfrak{g}$, which is a vector subspace of $\mathbb{R}^{n \times n}$. Furthermore, this allows us to generalize beyond compact groups, because while the Haar measure is defined only for compact groups, the Lesbegue measure is defined for the Lie algebra of *any* Lie group, compact or not. While we still ultimately convolve with group elements (in the case of compact groups, for which $\exp : \mathfrak{g} \to G$ is surjective), our inputs, u, are taken from the Lie algebra, $\mathfrak{g}$, and then pushed onto the Lie group, G, via the exp map.

Additionally, because the exp map is surjective only for compact Lie groups (Hall, 2015), the approach of parameterizing Lie group elements by applying the exp map to elements of the Lie algebra only works in the compact case. Because we model the mapping function, $\mathcal{N}_\theta : \mathfrak{g} \to G$, using a neural network (in our case, a single-layer MLP), our approach extends to non-compact Lie groups.

Finally, our approach easily interpolates between full equivariance, partial equivariance, and non-equivariance. When presented with fully equivariant training data, our neural network over Lie algebra elements can learn the exponential map. When presented with almost equivariant training data, this same neural network will learn an approximation to the exponential map that is justified by said data. And finally, when presented with a task for which equivariance is not beneficial, the neural network is free to learn an arbitrary function over the Lie algebra that best models the training data.

4. Results

For each task, we benchmark against the Residual Pathway Prior model of Finzi et al. (2021b), the Appoximately Equivariant GCNN of (Wang et al., 2022a), the $E(2)$-equivariant E2CNN of (Weiler and Cesa, 2019), and a Standard CNN exhibiting only translational equivariance.

4.1. Image Classification

We first test our model on an image classification task. We focus on the Rot-MNIST dataset. The images in Rot-MNIST are taken from the MNIST dataset and subjected to random rotations. We would expect rotational equivariance to be beneficial for classifying these images. However, the dataset is not fully rotationally equivariant in the sense that applying a 180 degree rotation to the digit 6 causes it to look like the digit 9 and vice versa. We find that our model outperforms all baselines for this task.

Group	Num Samples	Model	Rot-MNIST	Pendulum RMSE	Pendulum Average RMSE
SE(2)	10	Almost Equivariant CNN	**93.55**	$\mathbf{0.0350 \pm 0.0001}$	$\mathbf{0.5963 \pm 2.1581}$
		Residual Pathway Prior	85.20	0.0350 ± 0.0001	14.4018 ± 26.8171
	N/A	Approximate GCNN	85.51	0.0350 ± 0.0001	0.8241 ± 1.4568
T(2)	N/A	Standard CNN	92.05	0.0354 ± 0.0009	0.6573 ± 1.0565
E(2)	10	E2CNN	92.81	0.0350 ± 0.0000	3.5987 ± 2.8203

Table 1: Classification accuracies (%) for Rot-MNIST as well as RMSE values and average RMSE across hyperparameter configurations for pendulum prediction. The first column gives the Lie group with respect to which (almost) equivariance is imposed. Our model is the Almost Equivariant CNN. The `Num Samples` column gives the number of elements drawn from the Lie algebra when computing the convolution.

4.2. Damped Pendulum

The second task is to predict the angle, $\theta \in [0, \pi]$, made with the vertical at time $t \in \mathbb{R}^+$ of a pendulum undergoing simple harmonic motion and subjected to wind resistance. The pendulum is modeled as a mass m connected to a massless rod of length L subjected to an acceleration due to gravity of $g = -9.8\text{m}/\text{sec}^2$ and position function $\theta(t)$. The differential equation governing such motion is $\frac{\partial^2 \theta}{\partial t^2} + \frac{\lambda}{m}\frac{\partial \theta}{\partial t} + \frac{g}{L}\theta = 0$ where λ is the coefficient of friction governing the wind resistance which is modeled as a force $F_w = -\lambda L \frac{\partial \theta}{\partial t}$. We simulate the trajectory of the pendulum using the Runge-Kutta method to obtain an iterative, approximate solution to the above, second-order differential equation. We sample $\theta(t)$ for 6000 values of $t \in (0, 60)$ using a $dt = 0.01$ and setting $m = L = 1$, $\theta(0) = \pi/3$, $\frac{\partial \theta}{\partial t}(0) = 0$, and $\lambda = 0.2$. We partition this data into a 90%/10% train-test split and train using k-fold cross validation a series of models to predict angular position from the time, $t \in (0, 60)$. We find that while all the models exhibit comparable performance on this task, ours exhibits the lowest variance across different hyperparameter settings.

5. Discussion

In this work, we introduced a convolution on the elements of a Lie algebra, for which Lie algebra elements are sampled using the Lebesgue measure on the algebra, that approximates a fully equivariant group convolution. We then showed that such a convolution can model almost equivariance relative to *any* group action, even those of non-compact groups. Finally, we validated our assumptions by testing our model on a 2D image classification task having $SO(2)$ almost equivariance and a 1D sequence regression task exhibiting full $SO(2)$ equivariance.

References

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. doi: 10.1038/s41467-022-29939-5. URL https://doi.org/10.1038/s41467-022-29939-5.

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, and Maurice Weiler. A general theory of equivariant cnns on homogeneous spaces. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/b9cfe8b6042cf759dc4c0cccb27a6737-Paper.pdf.

Nima Dehmamy, Robin Walters, Yanchen Liu, Dashun Wang, and Rose Yu. Automatic symmetry discovery with lie algebra convolutional network. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 2503–2515. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/148148d62be67e0916a833931bd32b26-Paper.pdf.

Pavel Etingof, Oleg Golberg, Sebastian Hensel, Tiankai Liu, Alex Schwendner, Dmitry Vaintrob, and Elena Yudovina. Introduction to representation theory, 2011.

Marc Finzi, Samuel Stanton, Pavel Izmailov, and Andrew Gordon Wilson. Generalizing convolutional neural networks for equivariance to lie groups on arbitrary continuous data. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3165–3176. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/finzi20a.html.

Marc Finzi, Max Welling, and Andrew Gordon Gordon Wilson. A practical method for constructing equivariant multilayer perceptrons for arbitrary matrix groups. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 3318–3328. PMLR, 18–24 Jul 2021a. URL https://proceedings.mlr.press/v139/finzi21a.html.

Marc Anton Finzi, Gregory Benton, and Andrew Gordon Wilson. Residual pathway priors for soft equivariance constraints. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021b. URL https://openreview.net/forum?id=k505ekjMzww.

William Fulton and Joe Harris. *Representation theory: A first course*. Springer, 2004.

Jan Gerken, Oscar Carlsson, Hampus Linander, Fredrik Ohlsson, Christoffer Petersson, and Daniel Persson. Equivariance versus augmentation for spherical images. In Kamalika

Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 7404–7421. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/gerken22a.html.

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*, 2023. URL https://openreview.net/forum?id=JL7Va5Vy15J.

Brian Hall. *Lie Groups, Lie Algebras, and Representations: An Elementary Introduction.* Springer International Publishing, Cham, 2015. ISBN 978-3-319-13467-3. doi: 10.1007/978-3-319-13467-3_1. URL https://doi.org/10.1007/978-3-319-13467-3.

Michael J Hutchinson, Charline Le Lan, Sheheryar Zaidi, Emilien Dupont, Yee Whye Teh, and Hyunjik Kim. Lietransformer: Equivariant self-attention for lie groups. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4533–4543. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/hutchinson21a.html.

Risi Kondor and Shubhendu Trivedi. On the generalization of equivariance and convolution in neural networks to the action of compact groups. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2747–2755. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/kondor18a.html.

Maxime W. Lafarge, Erik J. Bekkers, Josien P. W. Pluim, Remco Duits, and Mitko Veta. Roto-translation equivariant convolutional networks: Application to histopathology image analysis. *CoRR*, abs/2002.08725, 2020. URL https://arxiv.org/abs/2002.08725.

John M. Lee. *Introduction to Smooth Manifolds.* Springer New York, New York, NY, 2003. ISBN 978-0-387-21752-9. doi: 10.1007/978-0-387-21752-9. URL https://doi.org/10.1007/978-0-387-21752-9.

John M. Lee. *Introduction to Riemannian Manifolds.* Springer International Publishing, Cham, 2018. ISBN 978-3-319-91755-9. doi: 10.1007/978-3-319-91755-9. URL https://doi.org/10.1007/978-3-319-91755-9.

Yi-Lun Liao and Tess Smidt. Equiformer: Equivariant graph attention transformer for 3d atomistic graphs. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=KwmPfARgOTD.

Mircea Petrache and Shubhendu Trivedi. Approximation-generalization trade-offs under (approximate) group equivariance, 2023.

David W. Romero and Suhas Lohit. Learning partial equivariances from data. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 36466–36478. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/ec51d1fe4bbb754577da5e18eb54e6d1-Paper-Conference.pdf.

Víctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 9323–9332. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/satorras21a.html.

Tycho F.A. van der Ouderaa, David W. Romero, and Mark van der Wilk. Relaxing equivariance constraints with non-stationary continuous filters. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=5oEk8fvJxny.

Rui Wang, Robin Walters, and Rose Yu. Approximately equivariant networks for imperfectly symmetric dynamics. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 23078–23091. PMLR, 17–23 Jul 2022a. URL https://proceedings.mlr.press/v162/wang22aa.html.

Rui Wang, Robin Walters, and Rose Yu. Data augmentation vs. equivariant networks: A theory of generalization on dynamics forecasting, 2022b.

Maurice Weiler and Gabriele Cesa. General E(2)-Equivariant Steerable CNNs. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

Appendix A. Appendix

A.1. Mathematical Background

We give brief introductions to the subjects of representation theory, differential topology and geometry, and Lie theory, stating only those definitions, propositions, and theorems needed to understand the paper. For more comprehensive background, we encourage readers to consult any of Fulton and Harris (2004); Etingof et al. (2011); Hall (2015) for representation theory, any of Lee (2003, 2018) for differential topology and geometry, and Hall (2015) for Lie theory.

A.1.1. Representation Theory

Definition 4 (Representation of an associative algebra) *We define a representation (ρ, V) of an associative algebra A to be a vector space V with an associated homomorphism $\rho : A \to End(V)$ where $End(V)$ denotes the set of endomorphisms of V, i.e. linear operators from V to itself.*

Definition 5 (Lie group representation) *A representation (ρ, V) of a Lie group G is a homomorphism $\rho : G \to GL(V)$ where V is a vector space.*

Definition 6 (Lie algebra representation) *A representation (ρ, V) of a Lie algebra $\mathfrak{g}$ is a homomorphism $\rho : \mathfrak{g} \to \mathfrak{gl}(V)$ where V is a vector space.*

Definition 7 (Morphism of representations) *A morphism of representations $(\rho_1, V), (\rho_2, W)$ is a map $\phi : V \to W$ satisfying*

$$\phi(\rho_1(a)(v)) = \rho_2(a)\phi(v)$$

for all $a \in A, v \in V$.

We can view morphisms as the set of transformations on V that preserve *equivariance* with respect to some pair of representations. ϕ is also sometimes called an *intertwining map.* In other words, in equivariant deep learning we seek to learn neural networks $\mathcal{N}$ that are morphisms of representations. In almost equivariant deep learning, we seek models $\mathcal{N}$ that are almost morphisms in the sense described in the paper intro.

Definition 8 (Subrepresentation) *A subrepresentation of (ρ, V) is a subspace $U \subseteq V$ such that $\rho(a)(u) \in U$ for all $a \in A, u \in U$.*

A.1.2. Differential Topology & Geometry, Lie Groups, and Lie Algebras

Definition 9 (Smooth manifold) *A smooth manifold is a Hausdorff, second countable, locally Euclidean topological space, M, equipped with a smooth structure.*

Definition 10 (Riemannian manifold) *A Riemannian manifold is a pair (M, g) where M is a smooth manifold and g is a choice of Riemannian metric on M.*

Definition 11 (Riemannian metric) *A Riemannian metric for a manifold M is a smoothly-varying choice of inner product on the tangent space T_pM. Equivalently, a Riemannian metric on M is a smooth covariant 2-tensor field $g \in \mathcal{T}^2(M)$ whose value g_p at each $p \in M$ is an inner product on T_pM.*

Proposition 12 *Every smooth manifold admits a Riemannian metric.*

Definition 13 (Isometry) *An isometry of Riemannian manifolds (M, g) and $(\tilde{M}, \tilde{g})$ is a diffeomorphism $\varphi : M \to \tilde{M}$ such that $\varphi^* \tilde{g} = g$. Equivalently, φ is a metric-preserving diffeomorphism.*

Definition 14 (Lie group) *A Lie group is a smooth manifold with an algebraic group structure such that the multiplication map $m : G \times G \to G$ and the inversion map $i : G \to G$ are both smooth.*

Definition 15 (Lie algebra) *A Lie algebra is a vector space $\mathfrak{g}$ over a field F, equipped with a map $[\cdot,\cdot] : \mathfrak{g} \times \mathfrak{g} \to \mathfrak{g}$, called the bracket, which satisfies the following three properties:*

1. *Bilinearity*
2. *Antisymmetry*

$$[X, Y] = -[Y, X]$$

3. The Jacobi Identity

$$[X, [Y, Z]] + [Y, [Z, X]] + [Z, [X, Y]] = 0$$

Theorem 16 (Ado's Theorem) *Every finite-dimensional real Lie algebra admits a faithful finite-dimensional representation.*

Definition 17 (Matrix exponential) *Given $A \in \mathbb{R}^{n\times n}$, the matrix exponential is the function* $\exp : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ *given by*

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

Definition 18 (Haar measure) *Let G be a locally compact group. Then the (unique up to scalars, nonzero, left-invariant) Haar measure on G is the Borel measure μ satisfying the following*

1. *$\mu(xE) = \mu(E)$ for all $x \in G$ and all measurable $E \subseteq G$.*
2. *$\mu(U) > 0$ for every non-empty open set $U \subseteq G$.*
3. *$\mu(K) < \infty$ for every compact set $K \subseteq G$.*

Proposition 19 *Every Lie group is locally compact and thus comes equipped with a Haar measure.*

A.2. Model Training & Hyperparameter Tuning

A.2.1. Pendulum Trajectory Prediction

For the pendulum trajectory prediction task, we performed a grid search over the following parameters across all models excluding, to some extent, the standard CNN. For the standard CNN, we used a fixed architecture with three convolutional layers having a kernel size of 2 and having 32, 64, and 128 channels, respectively. This was followed by two linear layers having weight matrices of sizes 128 × 256 and 256 × 2, respectively.

Each model was given a batch size of 16 and trained for 100 epochs. An 80%/10%/10% train-validation-test split was used, with RMSE calculated on the test set after the final epoch. The data was not shuffled due to this being a time series prediction task. Four random seeds were used at each step of the grid search, with average test set RMSE and standard deviations calculated with respect to the four random seeds.

Learning Rate	Optimizer	Kernel Sizes	Hidden Channels	# Hidden Layers
1e-4, 1e-3, 1e-2, 1e-1	Adam, SGD	2, 3, 4, 5	16, 32	1, 2, 3, 4

Table 2: Model hyperparameters used in grid search for the pendulum trajectory prediction task.

A.2.2. Rotated MNIST Classification

For the Rotated MNIST classification task, we performed a grid search over the following parameters across all models excluding the standard CNN. For the standard CNN, we used a fixed architecture with two convolutional layers having hidden channel counts of 32 and 64, respectively, and a kernel size of 3. The convolutional layers are followed by dropout and two linear layers having weight matrices of sizes 9126×128 and 128×10, respectively.

Each model was trained for 200 epochs with a linear learning rate decay schedule. The standard 10k/2k/50k train-validation-test split was used, with classification accuracy calculated on the test set after the final epoch.

Learning Rate	Optimizer	Kernel Sizes	Hidden Channels	# Hidden Layers	Batch Sizes
1e-4, 1e-3, 1e-2, 1e-1	Adam	3, 4, 5	16, 32	1, 2, 3, 4	16, 32, 64

Table 3: Model hyperparameters used in grid search for the Rot-MNIST classification task.