Learning Symmetrization for Equivariance with Orbit Distance Minimization

Tien Dat Nguyen* TIENDAT@KAIST.AC.KR
School of Computing, KAIST

Jinwoo Kim* JINWOO-KIM@KAIST.AC.KR
School of Computing, KAIST

Hongseok Yang HONGSEOK.YANG@KAIST.AC.KR
School of Computing, KAIST

Seunghoon Hong SEUNGHOON.HONG@KAIST.AC.KR
School of Computing, KAIST

Editors: Sophia Sanborn, Christian Shewmake, Simone Azeglio, Nina Miolane

Abstract

We present a general framework for symmetrizing an arbitrary neural-network architecture and making it equivariant with respect to a given group. We build upon the proposals of Kim et al. (2023); Kaba et al. (2023) for symmetrization, and improve them by replacing their conversion of neural features into group representations, with an optimization whose loss intuitively measures the distance between group orbits. This change makes our approach applicable to a broader range of matrix groups, such as the Lorentz group $\mathrm{O}(1,3)$, than these two proposals. We experimentally show our method's competitiveness on the $\mathrm{SO}(2)$ image classification task, and also its increased generality on the task with $\mathrm{O}(1,3)$. Our implementation will be made accessible at https://github.com/tiendatnguyen-vision/Orbit-symmetrize.

Keywords: Equivariance, Symmetrization, Rotation Equivariance, Lorentz Equivariance

1. Introduction

Exploiting symmetries is a popular principle for developing an efficient learning system, which is typically realized by defining a hypothesis class of functions equivariant to a given group G of symmetries. While a dominant approach to define such a hypothesis class has been to design a specific G equivariant neural-network architecture (Finzi et al., 2021a; Villar et al., 2021), *architecture-agnostic* alternatives are explored recently (Puny et al., 2022; Basu et al., 2023; Kaba et al., 2023; Kim et al., 2023). These alternatives are based on *symmetrization*, where any unconstrained function $\phi_\theta : \mathcal{X} \to \mathcal{Y}$ is made G equivariant by averaging it over transformations of inputs and outputs induced by certain group elements $g \in G$. In this work, we improve one of the most powerful symmetrization methods from (Kim et al., 2023; Kaba et al., 2023), which symmeterizes ϕ_θ to the following function $\Phi_{\theta,\omega}$:

$$\Phi_{\theta,\omega}(\mathbf{x}) = \mathbb{E}_{\boldsymbol{\epsilon}}[g \cdot \phi_\theta(g^{-1} \cdot \mathbf{x})] \quad \text{with} \quad \rho(g) = r(q_\omega(\mathbf{x}, \boldsymbol{\epsilon})), \tag{1}$$

where $q_\omega : (\mathbf{x}, \boldsymbol{\epsilon}) \mapsto \mathbf{h} \in \mathbb{R}^{n \times n}$ is a G equivariant network, and $r : \mathbf{h} \mapsto \rho(g)$ is a G equivariant *contraction* operator producing the representation $\rho(g)$ of some element $g \in G$.

*These authors contributed equally.

A major issue with Equation (1) is that designing the contraction r is often non-trivial; r should produce a valid group representation $\rho(g)$ from an unstructured feature $\mathbf{h}$ while being G equivariant itself. Prior works employed hand-designed algorithms, such as Gram-Schmidt process for $\mathrm{O}(n)$ (Kaba et al., 2023), but such algorithms are available only for certain groups (Kim et al., 2023). Our goal is to overcome this bottleneck and making the symmetrization work for broader group symmetries, such as the Lorentz group $\mathrm{O}(1,3)$.

Our idea is to design a differentiable objective on q_ω of which gradient-based optimization makes $q_\omega(\mathbf{x}, \boldsymbol{\epsilon})$ directly output valid group representations $\mathbf{h} \approx \rho(g)$, thereby removing the need for contraction r. We design the objective in a principled manner as distance minimization on group orbit space. Compared to symmetrization algorithms of Kim et al. (2023); Kaba et al. (2023), this makes our approach applicable to a much broader range of matrix groups where orbit separating invariants are available. We implement our method for the special orthogonal group $\mathrm{SO}(2)$ and the Lorentz group $\mathrm{O}(1,3)$, and find that our objective can replace the known contraction r for $\mathrm{SO}(2)$ with a negligible performance drop, and successfully achieves symmetrization based equivariance on the Lorentz group $\mathrm{O}(1,3)$.

2. Orbit Distance Minimization

Problem Definition Let $\rho : G \to \mathrm{GL}(n)$ be a group representation that associates each group element $g \in G$ an invertible matrix $\rho(g) \in \mathbb{R}^{n\times n}$. For our G equivariant neural network $q_\omega : (\mathbf{x}, \boldsymbol{\epsilon}) \mapsto \mathbf{h} \in \mathbb{R}^{n\times n}$, the group G acts on the output space through the matrix multiplication of the representation $\mathbf{h} \mapsto g \cdot \mathbf{h} = \rho(g)\mathbf{h}$. Our goal is to train q_ω such that its output is always a valid group representation $\mathbf{h} \in \rho(G)$, where $\rho(G)$ denotes the image of ρ.

Orbit Distance Minimization We will now present a training objective that contracts the output space of q_ω to valid group representations $\rho(G) \subset \mathbb{R}^{n\times n}$. Our key idea is that, instead of working on $\mathbb{R}^{n\times n}$ directly, working on the orbit space (quotient) $\mathbb{R}^{n\times n}/G$ greatly simplifies the problem. Let us write the orbit of an element $\mathbf{h} \in \mathbb{R}^{n\times n}$ under the action of G as $[\mathbf{h}] = \{g \cdot \mathbf{h} : g \in G\}$. The orbit space $\mathbb{R}^{n\times n}/G$ is defined accordingly as $\{[\mathbf{h}] : \mathbf{h} \in \mathbb{R}^{n\times n}\}$.

We now provide an observation that all valid group representations $\rho(G)$ precisely map onto a single point in the orbit space, which is the orbit of the identity matrix $\mathbf{I} \in \mathbb{R}^{n\times n}$ since we have $\rho(G) = \{\rho(g) : g \in G\} = \{g \cdot \mathbf{I} : g \in G\} = [\mathbf{I}]$. This implies, on the orbit space, our objective is understood as contracting all orbits of neural network outputs $[\mathbf{h}]$ towards a fixed point target $[\mathbf{I}]$. Thus, if we can endow the orbit space with a distance metric $d : \mathbb{R}^{n\times n}/G \times \mathbb{R}^{n\times n}/G \to \mathbb{R}^+$, we can frame our training objective as distance minimization:

$$w^* = \arg\min_\omega d([\mathbf{h}], [\mathbf{I}]), \tag{2}$$

where we remark that $q_\omega : (\mathbf{x}, \boldsymbol{\epsilon}) \mapsto \mathbf{h} \in \mathbb{R}^{n\times n}$ is our G equivariant neural network. We now show that the objective indeed contracts the output of q_ω exactly to $\rho(G)$ (proof in A.1):

Theorem 1 *The training objective in Equation (2) achieves the global minimum with the value of* 0 *if and only if q_ω always outputs valid group representations $\mathbf{h} \in \rho(G)$.*

Our problem now reduces to defining a proper distance metric d on the orbit space $\mathbb{R}^{n\times n}/G$. The closest concept we could find in literature is the quotient metric (Burago et al., 2001), but it is intractable as it involves infimum over an infinite set. Instead, we propose a simple

distance metric based on a class of functions called *orbit separating invariants*: G invariant functions f that separate orbits $f(\mathbf{h}) \neq f(\mathbf{h}') \iff [\mathbf{h}] \neq [\mathbf{h}']$ (Dym and Gortler, 2022). In detail, our distance metric can be defined as vector distance on outputs of f (proof in A.1):

Theorem 2 *Let $f : \mathbb{R}^{n\times n} \to \mathbb{R}^k$ be an orbit separating invariant and $\|\cdot\|$ be vector norm. Then, $d([\mathbf{h}], [\mathbf{h}']) = \|f(\mathbf{h}) - f(\mathbf{h}')\|$ is a distance metric on the orbit space $\mathbb{R}^{n\times n}/G$.*

With Theorem 2, we can use the below objective for optimization problem in Equation (2)[1]:

$$w^* = \arg\min_{\omega} \|f(\mathbf{h}) - f(\mathbf{I})\|. \tag{3}$$

In practice, it is desirable to have a differentiable objective such that we can perform a gradient-based optimization. Since $q_\omega : (\mathbf{x}, \boldsymbol{\epsilon}) \mapsto \mathbf{h}$ is already a neural network, the objective in Equation (3) would be differentiable almost everywhere with respect to ω if we choose f and $\|\cdot\|$ to be differentiable almost everywhere.

Discussion on Generality We now discuss the results from invariant theory implying orbit separating invariants $f : \mathbb{R}^{n\times n} \to \mathbb{R}^k$ of bounded dimension $k \leq 2n^2 + 1$ exist for a very general class of matrix groups and they are differentiable everywhere in general.

The existence of orbit separating invariants has been mainly shown for *linearly reductive groups* including $\mathrm{GL}(n)$, semi-simple groups $\mathrm{SL}(n)$, $\mathrm{O}(n)$, $\mathrm{SO}(n)$, finite group S_n, and also $\mathrm{O}(s, n-s)$. Most of the results are derived from the concept of *invariant polynomials*, which are polynomials on matrix entries that are G invariant. To elaborate, consider a group G acting on $\mathbb{R}^{n\times n}$, and let $\mathcal{S}$ be the set of all invariant polynomials $f' : \mathbb{R}^{n\times n} \to \mathbb{R}$. We call a set of invariant polynomials $\{f_1, ..., f_k\}$ the *generating set* if every $f' \in \mathcal{S}$ can be written as $f'(\cdot) = h(f_1(\cdot), ..., f_k(\cdot))$ using some polynomial $h : \mathbb{R}^k \to \mathbb{R}$. For every linearly reductive group, Weyl's theorem (Weyl, 1946) guarantees the existence of a finite generating set. Furthermore, for many subclasses of these groups, it has been shown that this set separates orbits in $\mathbb{R}^{n\times n}$ whose closures do not intersect[2] (Dym and Gortler, 2022; Derksen and Kemper, 2015), allowing us to use their stack $f : \mathbb{R}^{n\times n} \to \mathbb{R}^k$ as our orbit separating invariant. Such f is differentiable everywhere as it is a stack of polynomials. For many groups, the generating set is known from the invariant theory, and so is f (see A.2).

Now consider the dimension k of the separating invariant $f : \mathbb{R}^{n\times n} \to \mathbb{R}^k$ which is the size of the generating set $\{f_1, ..., f_k\}$ in our context. While this can be large for some groups, Dym and Gortler (2022) has shown that random linear projection can almost always reduce it to a set of $2n^2 + 1$ polynomials that still separates orbits (see A.3), which also reduces f.

Final Model Our original goal is to symmetrize a function $\phi_\theta : \mathcal{X} \to \mathcal{Y}$ to be G equivariant. We define our symmetrization as follows by removing contraction r from Equation (1):

$$\Phi_{\theta,\omega}(\mathbf{x}) = \mathbb{E}_{\mathbf{h}}[\mathbf{h} \cdot \phi_\theta(\mathbf{h}^{-1} \cdot \mathbf{x})] \quad \text{where} \quad \mathbf{h} = q_\omega(\mathbf{x}, \boldsymbol{\epsilon}). \tag{4}$$

Given training pairs $(\mathbf{x}, \mathbf{y})$ of input $\mathbf{x} \in \mathcal{X}$ and label $\mathbf{y} \in \mathcal{Y}$, we train for the joint objective of task loss $\mathcal{L}$ and the orbit distance loss (Equation (3)) weighted by a hyperparameter λ:

$$\theta^*, \omega^* = \arg\min_{\theta,\omega} \mathcal{L}(\mathbf{y}, \Phi_{\theta,\omega}(\mathbf{x})) + \lambda\, \mathbb{E}_{\mathbf{h}} \|f(\mathbf{h}) - f(\mathbf{I})\|. \tag{5}$$

[1] For some groups, orbit separation of f is guaranteed only for full-rank inputs (Dym and Gortler, 2022). In this case, the optimality condition in Theorem 1 holds if we assume $\mathbf{h}$ to be full-rank.

[2] For compact groups this guarantees orbit separation; for closed non-compact groups this guarantees orbit separation for full-rank inputs (Dym and Gortler, 2022) which relates to footnote 1.

Table 1: Experimental results on SO(2) and O(1, 3) group symmetries.

(a) Rotated MNIST, SO(2).

Method	Test Error % ↓
GCNN (p4)	2.36 ± 0.15
GCNN (p64)	2.28 ± 0.10
CNN	4.90 ± 0.20
CNN-Aug.	3.30 ± 0.20
CNN-Canonical.	2.32 ± 0.18
CNN-PS	2.21 ± 0.28
CNN-Canonical.-Orbit **(Ours)**	2.44 ± 0.12
CNN-PS-Orbit **(Ours)**	2.37 ± 0.35

(b) Particle Scattering, O(1, 3).

Method	Test MSE ↓
Scalar MLP	0.00171 ± 0.00004
MLP	0.65381 ± 0.23663
MLP-Aug.	0.09101 ± 0.03107
MLP-Canonical.	N/A
MLP-PS	N/A
MLP-Canonical.-Orbit **(Ours)**	0.01027 ± 0.00082
MLP-PS-Orbit **(Ours)**	0.00887 ± 0.00070

Intuitively, if the orbit loss is ≈ 0, we would have $\mathbf{h} \approx \rho(g)$ and the model would closely achieve the symmetrization in Equation (1) while not requiring contraction $r : \mathbf{h} \mapsto \rho(g)$. A formal theoretical analysis on G equivariance and universality of $\Phi_{\theta,\omega}$ is provided in A.4.

3. Experiments

We evaluate our approach on two selected matrix groups: the special orthogonal group in two dimensions SO(2) and the Lorentz group O(1, 3). Experimental details are in A.5.

Image Classification For SO(2), we use the Rotated MNIST (Larochelle et al., 2007), a common benchmark for equivariant models (Cohen and Welling, 2016; Finzi et al., 2020) with randomly rotated 62,000 digits (SO(2) invariance). We follow the setup of Kaba et al. (2023) and use the same 7-layer CNN as our base function ϕ_θ. For the symmetrizer q_ω, we use 2-layer EMLP (Finzi et al., 2021a) of 64 hidden dimensions. For training (Equation (5)), we use cross entropy for task loss $\mathcal{L}$; for orbit distance loss $\lambda \|f(\mathbf{h}) - f(\mathbf{I})\|$ we use the orbit separating invariant $f(\mathbf{h}) = [\text{vec}(\mathbf{h}^\top \mathbf{h}), \det(\mathbf{h})]$ (Dym and Gortler, 2022), L1 norm, $\lambda = 1$.

We use the following baselines: CNN, CNN with SO(2) data augmentation, equivariant GCNN (Cohen and Welling, 2016), and CNN made SO(2) equivariant with symmetrization methods Canonicalization (Kaba et al., 2023) and Probabilistic Symmetrization (PS) (Kim et al., 2023) that follow Equation (1) except Canonicalization drops noise $\boldsymbol{\epsilon}$. We take the performances of CNN, data augmentation, and GCNN from Kaba et al. (2023), and train symmetrization baselines using SO(2) contraction of Kim et al. (2023). The results are in Table 1(a). Symmetrization improves CNN overall, as all symmetrized CNNs outperform data augmentation and perform on par with GCNN. Within symmetrization, replacing contraction $r : \mathbf{h} \mapsto \rho(g)$ with our orbit distance loss leads to a negligible drop in performance. This indicates orbit distance minimization can replace the role of contraction operator, with a slight tradeoff as q_ω takes an extra role of producing valid group representation $\mathbf{h} \approx \rho(g)$.

Particle Scattering For the Lorentz group O(1, 3), we use Particle Scattering synthetic regression dataset from Finzi et al. (2021a) for matrix element in electron muon scattering (O(1, 3) invariance). We use 10,000 train data, and use 1,000 validation and 1,000 test data that are randomly O(1, 3) transformed. We use 3-layer MLP of 128 hidden dimensions as our base function ϕ_θ, and our symmetrizer q_ω is based on 3-layer Scalar MLP (Villar et al., 2021) of 128 hidden dimensions. For training (Equation (5)), we use mean squared error for task loss $\mathcal{L}$; for orbit distance loss $\lambda \|f(\mathbf{h}) - f(\mathbf{I})\|$ we use the orbit separating invariant $f(\mathbf{h}) = [\text{vec}(\mathbf{h}^\top \boldsymbol{\Lambda} \mathbf{h})]$, $\boldsymbol{\Lambda} = \text{diag}([+1, -1, -1, -1])$ (Dym and Gortler, 2022), L1 norm, $\lambda = 1$.

We run the following baselines: MLP, MLP with $\mathrm{O}(1,3)$ data augmentation, and invariant Scalar MLP (Villar et al., 2021), with 3 layers of 128 hidden dimensions. Symmetrization baselines in Equation (1) cannot be built as $\mathrm{O}(1,3)$ contraction is not known. The results are in Table 1(*b*). $\mathrm{O}(1,3)$ symmetrized MLPs based on our method significantly outperform MLP as well as data augmentation. To our knowledge, this is the first successful result in symmetrization based equivariance for $\mathrm{O}(1,3)$, implying symmetrization can be applied to groups where contraction $r : \mathbf{h} \mapsto \rho(g)$ is not available. Yet, our method has performance gap from invariant Scalar MLP. An explanation is that Scalar MLP gets Minkowsky inner product $\mathbf{x}^\top \mathbf{\Lambda} \mathbf{x}$ as input which is heavily correlated to label function of particle scattering, but our base MLP gets transformed input $\approx \rho(g)^{-1} \cdot \mathbf{x}$ which requires additional processing. We leave closing this gap as an important future research direction.

Conclusion We proposed orbit distance minimization, a framework for symmetrization based equivariant learning. Our method is competitive on $\mathrm{SO}(2)$ invariant classification and successfully achieves symmetrization based equivariance on Lorentz group $\mathrm{O}(1,3)$.

Acknowledgements This work was supported in part by the National Research Foundation of Korea (NRF) grant (NRF2021R1C1C1012540) and Institute of Information and communications Technology Planning and evaluation (IITP) grant (2021-0-00537 and 2021-0-02068) funded by the Korea government (MSIT).

References

Sourya Basu, Prasanna Sattigeri, Karthikeyan Natesan Ramamurthy, Vijil Chenthamarakshan, Kush R. Varshney, Lav R. Varshney, and Payel Das. Equi-tuning: Group equivariant fine-tuning of pretrained models. In *AAAI*, 2023.

D. Burago, I.U.D. Burago, and S. Ivanov. *A Course in Metric Geometry*. American Mathematical Society, 2001.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *ICML*, 2016.

Harm Derksen and Gregor Kemper. *Computational Invariant Theory*. Springer, 2015.

Nadav Dym and Steven J. Gortler. Low dimensional invariant embeddings for universal geometric learning. *arXiv*, 2022.

Nadav Dym and Haggai Maron. On the universality of rotation equivariant point cloud networks. *arXiv preprint arXiv:2010.02449*, 2020.

Marc Finzi, Samuel Stanton, Pavel Izmailov, and Andrew Gordon Wilson. Generalizing convolutional neural networks for equivariance to lie groups on arbitrary continuous data. In *ICML*, 2020.

Marc Finzi, Max Welling, and Andrew Gordon Wilson. A practical method for constructing equivariant multilayer perceptrons for arbitrary matrix groups. In *ICML*, 2021a.

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, 2021b.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv*, 2016.

Sékou-Oumar Kaba, Arnab Kumar Mondal, Yan Zhang, Yoshua Bengio, and Siamak Ravanbakhsh. Equivariance with learned canonicalization functions. In *ICML*, 2023.

Jinwoo Kim, Tien Dat Nguyen, Ayhan Suleymanzade, Hyeokjun An, and Seunghoon Hong. Learning probabilistic symmetrization for architecture agnostic equivariance. *arXiv*, 2023.

Hugo Larochelle, Dumitru Erhan, Aaron C. Courville, James Bergstra, and Yoshua Bengio. An empirical evaluation of deep architectures on problems with many factors of variation. In *ICML*, 2007.

Omri Puny, Matan Atzmon, Edward J. Smith, Ishan Misra, Aditya Grover, Heli Ben-Hamu, and Yaron Lipman. Frame averaging for invariant and equivariant network design. In *ICLR*, 2022.

Edgar Riba, Dmytro Mishkin, Daniel Ponsa, Ethan Rublee, and Gary Bradski. Kornia: an open source differentiable computer vision library for pytorch. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3674–3683, 2020.

Nimrod Segol and Yaron Lipman. On universal equivariant set networks. *arXiv preprint arXiv:1910.02421*, 2019.

Soledad Villar, David W. Hogg, Kate Storey-Fisher, Weichi Yao, and Ben Blum-Smith. Scalars are universal: Equivariant machine learning, structured like classical physics. In *NeurIPS*, 2021.

Hermann Weyl. *The Classical Groups: Their Invariants and Representations*. Princeton University Press, 1946.

Appendix A. Proofs and Supplementary Discussions

A.1. Proof of Theorem 1 and Theorem 2

Before the proofs, we provide a formal definition of distance metric on the orbit space.

Definition 3 *A function $d : \mathbb{R}^{n\times n}/G \times \mathbb{R}^{n\times n}/G \to \mathbb{R}^+$ is a distance metric on the orbit space $\mathbb{R}^{n\times n}/G$ if it satisfies the following conditions for all orbits $[\mathbf{h}], [\mathbf{h}'], [\mathbf{h}''] \in \mathbb{R}^{n\times n}/G$:*

1. $d([\mathbf{h}], [\mathbf{h}']) \geq 0$ *(non-negativity),*
2. $d([\mathbf{h}], [\mathbf{h}']) = 0 \iff [\mathbf{h}] = [\mathbf{h}']$ *(identity of indiscernibles),*
3. $d([\mathbf{h}], [\mathbf{h}']) = d([\mathbf{h}'], [\mathbf{h}])$ *(symmetry),*
4. $d([\mathbf{h}], [\mathbf{h}']) \leq d([\mathbf{h}], [\mathbf{h}'']) + d([\mathbf{h}''], [\mathbf{h}'])$ *(triangle inequality).*

We now provide the proofs.

Theorem 1 *The training objective in Equation (2) achieves the global minimum with the value of 0 if and only if q_ω always outputs valid group representations $\mathbf{h} \in \rho(G)$.*

Proof Recall the definition of an orbit $[\mathbf{h}] = \{g\cdot\mathbf{h} : g \in G\}$. ($\Longrightarrow$) If $d([\mathbf{h}], [\mathbf{I}]) = 0$, since d is a distance metric, we have $[\mathbf{h}] = [\mathbf{I}]$ from identity of indiscernibles. This implies $[\mathbf{h}] = \rho(G)$ since $[\mathbf{I}] = \{g \cdot \mathbf{I} : g \in G\} = \{\rho(g) : g \in G\} = \rho(G)$. On the orbit $[\mathbf{h}] = \{g \cdot \mathbf{h} : g \in G\}$, by selecting the identity element $\mathrm{id} \in G$ we get $\mathbf{h} \in [\mathbf{h}]$, thus $\mathbf{h} \in \rho(G)$. ($\Longleftarrow$) If q_ω always outputs valid group representations $\mathbf{h} \in \rho(G)$, we can write $\mathbf{h} = \rho(h)$ for some $h \in G$. Then we have $[\mathbf{h}] = \{g \cdot \mathbf{h} : g \in G\} = \{g \cdot \rho(h) : g \in G\} = \{(gh) \cdot \mathbf{I} : g \in G\} = \{g' \cdot \mathbf{I} : g'h^{-1} \in G\} = \{g' \cdot \mathbf{I} : g' \in G\} = [\mathbf{I}]$. Note that we used the associativity and invertibility of group elements as well as the fact that right operation by $h \in G$ maps a group G to itself. Since $[\mathbf{h}] = [\mathbf{I}]$ and d is a distance metric, we have that $d([\mathbf{h}], [\mathbf{I}]) = 0$ due to identity of indiscernibles, which is a global minimum due to non-negativity. ■

Theorem 2 *Let $f : \mathbb{R}^{n\times n} \to \mathbb{R}^k$ be an orbit separating invariant and $\|\cdot\|$ be vector norm. Then, $d([\mathbf{h}], [\mathbf{h}']) = \|f(\mathbf{h}) - f(\mathbf{h}')\|$ is a distance metric on the orbit space $\mathbb{R}^{n\times n}/G$.*

Proof We first note that a vector norm $\|\cdot\|$ induces a distance metric $d'(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} - \mathbf{v}\|$, which is called the induced metric. We now explicitly show that $d([\mathbf{h}], [\mathbf{h}']) = \|f(\mathbf{h}) - f(\mathbf{h}')\|$ satisfies the four conditions in Definition 3. Since $\|\cdot\|$ is a vector norm, non-negativity clearly holds. Since f is an orbit separating invariant $[\mathbf{h}] \neq [\mathbf{h}'] \iff f(\mathbf{h}) \neq f(\mathbf{h}')$, by invoking the identity of indiscernibles of the induced metric we have $[\mathbf{h}] \neq [\mathbf{h}'] \iff \|f(\mathbf{h}) - f(\mathbf{h}')\| \neq 0$, which proves the identity of indiscernibles for $d([\mathbf{h}], [\mathbf{h}'])$. Symmetry and triangle inequality of $d([\mathbf{h}], [\mathbf{h}'])$ are inherited from the symmetry and triangle inequality of the induced metric, as we have $d([\mathbf{h}], [\mathbf{h}']) = \|f(\mathbf{h}) - f(\mathbf{h}')\| = \|f(\mathbf{h}') - f(\mathbf{h})\| = d([\mathbf{h}'], [\mathbf{h}])$ for symmetry and $d([\mathbf{h}], [\mathbf{h}']) = \|f(\mathbf{h}) - f(\mathbf{h}')\| \leq \|f(\mathbf{h}) - f(\mathbf{h}'')\| + \|f(\mathbf{h}'') - f(\mathbf{h}')\| = d([\mathbf{h}], [\mathbf{h}'']) + d([\mathbf{h}''], [\mathbf{h}'])$ for triangle inequality. Therefore, $d([\mathbf{h}], [\mathbf{h}'])$ is a distance metric on the orbit space. ■

Table 2: Orbit separating invariants for some group actions from (Dym and Gortler (2022)), along with the domain on which orbit separation is guaranteed.

Group	Domain	Orbit separating invariant	Dimension
S_n	$\mathbb{R}^{n\times n}$	$[\phi_\alpha(\mathbf{h}) = \sum_{j=1}^n \mathbf{h}_j^\alpha], \quad \alpha \in \mathbb{Z}^d_{\geq 0}, \lvert\alpha\rvert \leq n$	$\binom{2n}{n}$
$\mathrm{O}(n)$	$\mathbb{R}^{n\times n}$	$[\mathrm{vec}(\mathbf{h}^\top \mathbf{h})]$	n^2
$\mathrm{SO}(n)$	$\mathbb{R}^{n\times n}$	$[\mathrm{vec}(\mathbf{h}^\top \mathbf{h}), \det \mathbf{h}]$	n^2+1
$\mathrm{O}(1, n-1)$	$\mathbb{R}^{n\times n}_{\text{full}}$	$[\mathrm{vec}(\mathbf{h}^\top \mathbf{\Lambda} \mathbf{h})], \quad \mathbf{\Lambda} = \mathrm{diag}([+1, -1, ..., -1])$	n^2
$\mathrm{SL}(n)$	$\mathbb{R}^{n\times n}_{\text{full}}$	$[\det \mathbf{h}]$	1
$\mathrm{GL}(n)$	$\mathbb{R}^{n\times n}_{\text{full}}$	$[\det^2(\mathbf{h}\mathbf{W}_i)/\det^{-1}(\mathbf{h}\mathbf{h}^\top)], \quad \mathbf{W}_i \in \mathbb{R}^{n\times n}, i = 1, ...2n^2+1$	$2n^2+1$

A.2. Supplementary Discussion on Orbit Separating Invariants

In this section, we provide examples of known orbit separating invariants for certain linearly reductive groups in Table 2 along with supplementary discussion for S_n and $\mathrm{GL}(n)$ groups. For the symmetric group S_n, the orbit separating invariant is implemented upon a set of invariant functions known as multi-dimensional power sum polynomials (Dym and Maron, 2020; Segol and Lipman, 2019). For a given input $\mathbf{h} \in \mathbb{R}^{n\times n}$, this invariant is written as:

$$\phi_\alpha(\mathbf{h}) = \sum_{j=1}^{n} \mathbf{h}_j^\alpha, \quad \alpha \in \mathbb{Z}^d_{\geq 0}, |\alpha| \leq n, \tag{6}$$

where $\mathbf{h}_j$ is jth row of $\mathbf{h}$, $\alpha = (\alpha_1, ...\alpha_n) \in \mathbb{N}^n$ is a multi-index, and $\mathbf{h}_j^\alpha = \mathbf{h}_{j,\alpha_1} \times ... \times \mathbf{h}_{j,\alpha_n}$. For the general linear group $\mathrm{GL}(n)$, on the contrary to other groups in Table 2 whose orbit separating invariants are based on the generating set of invariant polynomials, the generating set with respect to action of $\mathrm{GL}(n)$ consists solely of constant polynomial, which cannot be used to implement the orbit separating invariant f. Instead, Dym and Gortler (2022) has shown that f can be built by adopting a family of rational invariants as follows:

$$q(\mathbf{h}, \mathbf{W}) = \frac{\det(\mathbf{h}\mathbf{W})^2}{\det(\mathbf{h}\mathbf{h}^\top)}, \quad \mathbf{h}, \mathbf{W} \in \mathbb{R}^{n\times n}, \tag{7}$$

where, Dym and Gortler (2022) shows that, for almost every $\mathbf{W}_1, ...\mathbf{W}_{2n^2+1}$ randomly sampled from $\mathbb{R}^{n\times n}$ the set of functions $f = \{q(\mathbf{h}, \mathbf{W}_i) : i = 1, ..2n^2+1\}$ separates orbits of invertible matrices in $\mathbb{R}^{n\times n}$. This example demonstrates that our method has the possibility to be applied to linearly reductive groups even when non-trivial generating set of invariant polynomial is not available, as long as an alternative orbit separating set is found.

A.3. Supplementary Discussion on Random Projection for Scalability

In this section, we discuss the method for controlling the dimension of separating invariants $f : \mathbb{R}^{n\times n} \to \mathbb{R}^k$ to be $k \leq 2n^2+1$ as we have discussed in Section 2. Specifically, we summarize the projection technique suggested in Dym and Gortler (2022), which produce a smaller set of orbit separating invariants from the generating set. The idea is summarized below, which is an immediate consequence of corollary 1.9 in Dym and Gortler (2022):

Lemma 4 *Consider a group G that acts on $\mathbb{R}^{n\times n}$ and let $f : \mathbb{R}^{n\times n} \to \mathbb{R}^k$ be an orbit separating invariant for this action. For almost every vectors $\mathbf{w}^{(1)}, ...\mathbf{w}^{(2n^2+1)} \in \mathbb{R}^k$ sampled randomly, the function $\hat{f} : \mathbb{R}^{n\times n} \to \mathbb{R}^{2n^2+1}$ with component functions defined as:*

$$\hat{f}_j(\mathbf{x}) = \sum_{i=1}^{k} \mathbf{w}_i^{(j)} f_i(\mathbf{x}), \quad j = 1, ...2n^2 + 1, \tag{8}$$

is also an orbit separating invariant.

With this technique, given an orbit separating invariant f in arbitrary high dimension k, we can get a new orbit separating invariant $\hat{f}$ in dimension $2n^2+1$. Furthermore, this technique preserves differentiability as $\hat{f}_j$ is merely a linear combination of f_i. If f is composed of polynomials, as in the many cases of linearly reductive groups (Dym and Gortler, 2022), we can avoid k intermediate variables $f_i(\mathbf{x})$ in Equation (8) by fixing linear projections $\mathbf{w}$ and contracting k polynomials f_i into $2n^2 + 1$ polynomials $\hat{f}_j$ before computing $\hat{f}(\mathbf{x})$.

A.4. Proof of Equivariance and Universality

In this section, we formally prove the G equivariance and universality of our symmetrized model $\Phi_{\theta,\omega}$ in Equation (4). We define some notations and assumptions beforehand.

Definitions In the proofs, we set our group to be a matrix group $G \subset \mathrm{GL}(n)$, and define representations $\rho : G \to \mathrm{GL}(n)$ and $\rho^{gl} : \mathrm{GL}(n) \to \mathrm{GL}(n)$ under the restriction of being *aligned* $\rho(g) = \rho^{gl}(g)$ for all $g \in G$ without the loss of generality. For example, ρ and ρ^{gl} can be chosen as identity maps. Recall our G equivariant neural network $q_\omega : (\mathbf{x}, \boldsymbol{\epsilon}) \mapsto \mathbf{h} \in \mathbb{R}^{n\times n}$ in Equation (4). Given that $\boldsymbol{\epsilon}$ is a random variable, we can consider the implicit probabilistic distribution characterized by q_ω, which we denote as $p_\omega(\mathbf{h}|\mathbf{x})$. Further assuming that $\mathbf{h}$ is full-rank, we have $\mathbf{h} = g \in \mathrm{GL}(n)$, and we write our distribution as $p_\omega(g|\mathbf{x})$. Based on the notations, we can rewrite Equation (4) more formally as follows:

$$\Phi_{\theta,\omega}(\mathbf{x}) = \mathbb{E}_{p_\omega(g|\mathbf{x})}[\rho_{\mathcal{Y}}^{gl}(g)\phi_\theta(\rho_{\mathcal{X}}^{gl}(g)^{-1}\mathbf{x})], \tag{9}$$

where $\rho_{\mathcal{X}}^{gl} : \mathrm{GL}(n) \to \mathrm{GL}(\mathcal{X})$ and $\rho_{\mathcal{Y}}^{gl} : \mathrm{GL}(n) \to \mathrm{GL}(\mathcal{Y})$ are the representations of $\mathrm{GL}(n)$ on the input space and output space of the base function $\phi_\theta : \mathcal{X} \to \mathcal{Y}$, respectively.

A.4.1. Proof of Equivariance

In this section, we prove that even if our G equivariant parameterized distribution $p_\omega(g|\mathbf{x})$ in Equation (9) over which the base function ϕ_θ is symmetrized is not strictly supported on G, the entire symmetrized function $\Phi_{\theta,\omega}$ would still retain G equivariance.

Definition 5 *Consider a group $G \subset \mathrm{GL}(n)$ acting on a vector space $\mathcal{X}$. The conditional probabilistic distribution $p_\omega(g|\mathbf{x})$ for $g \in \mathrm{GL}(n)$ and $\mathbf{x} \in \mathcal{X}$ is G equivariant if it satisfies:*

$$p_\omega(g|\mathbf{x}) = p_\omega(g'g|\rho_{\mathcal{X}}(g')\mathbf{x}), \ \forall g' \in G, g \in \mathrm{GL}(n), \mathbf{x} \in \mathcal{X}, \tag{10}$$

where $\rho_{\mathcal{X}} : G \to \mathrm{GL}(n)$ is the representation of G on $\mathcal{X}$.

This generalizes the notion of probabilistic G equivariance in Kim et al. (2023) to our framework, since the support of $p_\omega(\cdot|\mathbf{x})$ is not limited to G but extended to $\mathrm{GL}(n)$. We first prove that G equivariance of p_ω is achieved with an appropriate choice of q_ω and $\boldsymbol{\epsilon}$:

Lemma 6 *If q_ω is G equivariant and $p(\boldsymbol{\epsilon})$ is G invariant under a representation $\rho_\mathcal{E}$ that satisfies $\det(\rho_\mathcal{E}(\boldsymbol{\epsilon})) = 1 \forall \boldsymbol{\epsilon} \in \mathcal{E}$, then the probabilistic distribution $p_\omega(g|\mathbf{x})$ characterized by q_ω is G equivariant (Definition 5).*

Proof Our proof is inspired by the proof of Theorem 3 in Kim et al. (2023). Firstly, we interpret the probability $p_\omega(g|\mathbf{x}, \boldsymbol{\epsilon})$ as a delta distribution:

$$p_\omega(g|\mathbf{x}, \boldsymbol{\epsilon}) = \delta(\rho^{gl}(g) = q_\omega(\mathbf{x}, \boldsymbol{\epsilon})). \tag{11}$$

We marginalize over $p(\boldsymbol{\epsilon})$ to get $p_\omega(g|\mathbf{x})$:

$$\begin{aligned} p_\omega(g|\mathbf{x}) &= \int_{\boldsymbol{\epsilon}} p_\omega(g|\mathbf{x}, \boldsymbol{\epsilon}) p(\boldsymbol{\epsilon}) d\boldsymbol{\epsilon} \\ &= \int_{\boldsymbol{\epsilon}} \delta(\rho^{gl}(g) = q_\omega(\mathbf{x}, \boldsymbol{\epsilon})) p(\boldsymbol{\epsilon}) d\boldsymbol{\epsilon}. \end{aligned} \tag{12}$$

Moreover, we have:

$$p_\omega(g'g|\rho_\mathcal{X}(g')\mathbf{x}) = \int_{\boldsymbol{\epsilon}} \delta(\rho^{gl}(g'g) = q_\omega(\rho_\mathcal{X}(g')\mathbf{x}, \boldsymbol{\epsilon})) p(\boldsymbol{\epsilon}) d\boldsymbol{\epsilon}. \tag{13}$$

Since ρ is a restriction of ρ^{gl} into G, we automatically have $\rho^{gl}(g) = \rho(g)\ \forall g \in G$. Together with the equivariance of q_ω, we have:

$$\begin{aligned} q_\omega(\rho_\mathcal{X}(g')\mathbf{x}, \boldsymbol{\epsilon}) &= \rho(g') q_\omega(\mathbf{x}, \rho_\mathcal{E}(g')^{-1}\boldsymbol{\epsilon}) \\ &= \rho^{gl}(g') q_\omega(\mathbf{x}, \rho_\mathcal{E}(g')^{-1}\boldsymbol{\epsilon}). \end{aligned} \tag{14}$$

This leads to:

$$\begin{aligned} p_\omega(g'g|\rho_\mathcal{X}(g')\mathbf{x}) &= \int_{\boldsymbol{\epsilon}} \delta(\rho^{gl}(g'g) = \rho^{gl}(g') q_\omega(\mathbf{x}, \rho_\mathcal{E}(g')^{-1}\boldsymbol{\epsilon})) p(\boldsymbol{\epsilon}) d\boldsymbol{\epsilon} \\ &= \int_{\boldsymbol{\epsilon}} \delta(\rho^{gl}(g) = q_\omega(\mathbf{x}, \rho_\mathcal{E}(g')^{-1}\boldsymbol{\epsilon})) p(\boldsymbol{\epsilon}) d\boldsymbol{\epsilon}. \end{aligned} \tag{15}$$

Next, to compute Equation (15), we introduce a change of variable $\boldsymbol{\epsilon}' = \rho_\mathcal{E}(g')^{-1}\boldsymbol{\epsilon}$:

$$p_\omega(g'g|\rho_\mathcal{X}(g')\mathbf{x}) = \int_{\boldsymbol{\epsilon}'} \delta(\rho^{gl}(g) = q_\omega(\mathbf{x}, \boldsymbol{\epsilon}')) p(\rho_\mathcal{E}(g')\boldsymbol{\epsilon}') \frac{1}{|\det \rho_\mathcal{E}(g')^{-1}|} d\boldsymbol{\epsilon}'. \tag{16}$$

Since $\rho_\mathcal{E}(\cdot)$ always has determinant 1, we have $|\det \rho_\mathcal{E}(g')^{-1}| = 1$. Furthermore, the invariance of $p(\boldsymbol{\epsilon})$ with respect to G gives $p(\rho_\mathcal{E}(g')\boldsymbol{\epsilon}') = p(\boldsymbol{\epsilon}')$. Eventually, we get:

$$\begin{aligned} p_\omega(g'g|\rho_\mathcal{X}(g')\mathbf{x}) &= \int_{\boldsymbol{\epsilon}'} \delta(\rho^{gl}(g) = q_\omega(\mathbf{x}, \boldsymbol{\epsilon}')) p(\boldsymbol{\epsilon}') d\boldsymbol{\epsilon}' \\ &= p_\omega(g|\mathbf{x}). \end{aligned} \tag{17}$$

This finishes the proof. ■

Next, we show the G equivariance of the symmetrized model $\Phi_{\theta,\omega}$ (Equation (9)):

Theorem 7 *If p_ω is G equivariant, then $\Phi_{\theta,\omega}$ is G equivariant for arbitrary ϕ_θ.*

Proof We prove $\Phi_{\theta,\omega}(\rho_\mathcal{X}(g')\mathbf{x}) = \rho_\mathcal{Y}(g')\Phi_{\theta,\omega}(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$, $g' \in G$. From Equation (9):

$$\Phi_{\theta,\omega}(\rho_\mathcal{X}(g')\mathbf{x}) = \mathbb{E}_{p_\omega(g|\rho_\mathcal{X}(g')\mathbf{x})}[\rho^{gl}_\mathcal{Y}(g)\phi_\theta(\rho^{gl}_\mathcal{X}(g)^{-1}\rho_\mathcal{X}(g')\mathbf{x})]. \tag{18}$$

Let us define $h = g'^{-1}g \in \mathrm{GL}(n)$, then we have $g = g'h$. Because p_ω is G equivariant, we have $p_\omega(g|\rho_\mathcal{X}(g')\mathbf{x}) = p_\omega(g'h|\rho_\mathcal{X}(g')\mathbf{x}) = p_\omega(h|\mathbf{x})$. Thus, Equation (18) becomes:

$$\begin{aligned}
\Phi_{\theta,\omega}(\rho_\mathcal{X}(g')\mathbf{x}) &= \mathbb{E}_{p_\omega(h|\mathbf{x})}[\rho^{gl}_\mathcal{Y}(g'h)\phi_\theta(\rho^{gl}_\mathcal{X}(g'h)^{-1}\rho^{gl}_\mathcal{X}(g')\mathbf{x})] \\
&= \rho^{gl}_\mathcal{Y}(g')\mathbb{E}_{p_\omega(h|\mathbf{x})}[\rho^{gl}_\mathcal{Y}(h)\phi_\theta(\rho^{gl}_\mathcal{X}(h)^{-1}\mathbf{x})] \\
&= \rho^{gl}_\mathcal{Y}(g')\Phi_{\theta,\omega}(\mathbf{x}) \\
&= \rho_\mathcal{Y}(g')\Phi_{\theta,\omega}(\mathbf{x}).
\end{aligned} \tag{19}$$

The last equality is from the fact that ρ is a restriction of ρ^{gl} to G. This finishes the proof. ■

A.4.2. Proof of Universality

In this section, we prove that even if our G equivariant parameterized distribution $p_\omega(g|\mathbf{x})$ in Equation (9) is not strictly supported on G, the entire symmetrized function $\Phi_{\theta,\omega}$ is still a universal approximator of arbitrary G equivariant functions as long as $p_\omega(g|\mathbf{x})$ can approximate a G equivariant distribution $h(g|\mathbf{x})$ which is compactly supported on G.

Definition 8 *We say that a G equivariant distribution $h(g|\mathbf{x})$ on $g \in \mathrm{GL}(n)$ and $\mathbf{x} \in \mathcal{X}$ is* compactly supported on G *if (1) $h(g|\mathbf{x})$ is supported on G for all $\mathbf{x}$, and (2) for any compact set $\mathcal{K} \subset \mathcal{X}$, the union of the support $\cup_{\mathbf{x}\in\mathcal{K}} \operatorname{supp} h(g|\mathbf{x})$ is compact.*

Definition 9 *We say that a parameterized G equivariant distribution $p_\omega(g|\mathbf{x})$ on $g \in \mathrm{GL}(n)$ and $\mathbf{x} \in \mathcal{X}$ is* approximately compactly supported on G *if there exists a distribution $h(g|\mathbf{x})$ compactly supported on G such that, for any compact set $\mathcal{K} \subset \mathcal{X}$ and any $0 < \alpha < 1$, there exists a choice of parameters ω that (1) the following holds for all $g \in G$ and $\mathbf{x} \in \mathcal{K}$:*

$$p_\omega(g|\mathbf{x}) \geq (1-\alpha)h(g|\mathbf{x}), \tag{20}$$

and (2) the union of the support $\cup_{\mathbf{x}\in\mathcal{K}} \operatorname{supp} p_\omega(g|\mathbf{x})$ is a subset of a compact set $\mathcal{H}$ that only depend on the given h and $\mathcal{K}$.

Lemma 10 *Let $h(g|\mathbf{x})$ be a G equivariant distribution compactly supported on G. For any function $\phi_\theta : \mathcal{X} \to \mathcal{Y}$, we define its symmetrization $\kappa_\theta : \mathcal{X} \to \mathcal{Y}$ over $h(g|\mathbf{x})$ as follows:*

$$\kappa_\theta(\mathbf{x}) = \mathbb{E}_{h(g|\mathbf{x})}[\rho_\mathcal{Y}(g)\phi_\theta(\rho_\mathcal{X}(g)^{-1}\mathbf{x})]. \tag{21}$$

Then, κ_θ is G equivariant. Furthermore, κ_θ is a universal approximator of G equivariant functions if ϕ_θ is a universal approximator.

Proof Our proof is inspired by the proof of theorem 2 of Kim et al. (2023). We first prove the G equivariance of κ_θ by showing $\kappa_\theta(\rho_\mathcal{X}(g')\mathbf{x}) = \rho_\mathcal{Y}(g')\kappa_\theta(\mathbf{x})\forall g \in G, \mathbf{x} \in \mathcal{X}$. We write:

$$\kappa_\theta(\rho_\mathcal{X}(g')\mathbf{x}) = \mathbb{E}_{h(g|\rho_\mathcal{X}(g')\mathbf{x})}[\rho_\mathcal{Y}(g)\phi_\theta(\rho_\mathcal{X}(g)^{-1}\rho_\mathcal{X}(g')\mathbf{x})]. \tag{22}$$

Let $m = g'^{-1}g$, then we have $g = g'm$. Since h is G equivariant, we have $h(g|\rho_\mathcal{X}(g')\mathbf{x}) = h(g'm|\rho_\mathcal{X}(g')\mathbf{x}) = h(m|\mathbf{x})$. Therefore, Equation (22) becomes the following:

$$\begin{aligned}\kappa_\theta(\rho_\mathcal{X}(g')\mathbf{x}) &= \mathbb{E}_{h(m|\mathbf{x})}[\rho_\mathcal{Y}(g'm)\phi_\theta(\rho_\mathcal{X}(g'm)^{-1}\rho_\mathcal{X}(g')\mathbf{x}] \\ &= \rho_\mathcal{Y}(g')\mathbb{E}_{h(m|\mathbf{x})}[\rho_\mathcal{Y}(m)\phi_\theta(\rho_\mathcal{X}(m)^{-1}\mathbf{x})] \\ &= \rho_\mathcal{Y}(g')\kappa_\theta(\mathbf{x}),\end{aligned} \tag{23}$$

showing the G equivariance of κ_θ.

We now prove the universality of κ_θ. Assume a compact set $\mathcal{K} \subset \mathcal{X}$ is given. Let us denote $\mathcal{M}_\mathcal{K} = \cup_{\mathbf{x}\in\mathcal{K}} \operatorname{supp} h(g|\mathbf{x})$ and $\mathcal{N}_\mathcal{K} = \{g^{-1}|g \in \mathcal{M}_\mathcal{K}\}$. Since h is compactly supported on G, by definition $\mathcal{M}_\mathcal{K}$ is compact. Furthermore, $\mathcal{N}_\mathcal{K}$ is also compact as it is image of a compact set $\mathcal{M}_\mathcal{K}$ under matrix inversion operator $g \mapsto g^{-1}$ which is continuous on $\mathrm{GL}(n)$. Let $\psi : \mathcal{X} \to \mathcal{Y}$ be an arbitrary G equivariant function. By equivariance of ψ, we have:

$$\begin{aligned}\|\psi(\mathbf{x}) - \kappa_\theta(\mathbf{x})\| &= \|\psi(\mathbf{x}) - \mathbb{E}_{h(g|\mathbf{x})}[\rho_\mathcal{Y}(g)\phi_\theta(\rho_\mathcal{X}(g)^{-1}\mathbf{x})]\| \\ &= \|\mathbb{E}_{h(g|\mathbf{x})}[\psi(\mathbf{x})] - \mathbb{E}_{h(g|\mathbf{x})}[\rho_\mathcal{Y}(g)\phi_\theta(\rho_\mathcal{X}(g)^{-1}\mathbf{x})]\| \\ &= \|\mathbb{E}_{h(g|\mathbf{x})}[\rho_\mathcal{Y}(g)\psi(\rho_\mathcal{X}(g)^{-1}\mathbf{x})] - \mathbb{E}_{h(g|\mathbf{x})}[\rho_\mathcal{Y}(g)\phi_\theta(\rho_\mathcal{X}(g)^{-1}\mathbf{x})]\|.\end{aligned} \tag{24}$$

Since the union of the support $\mathcal{M}_\mathcal{K} = \cup_{\mathbf{x}\in\mathcal{K}} \operatorname{supp} h(g|\mathbf{x})$ is compact and $\mathcal{Y}$ is finite-dimension, there exist $c > 0$ such that $\|\rho_\mathcal{Y}(g)\| \leq c \forall g \in \mathcal{M}_\mathcal{K}$. Therefore, Equation (24) becomes:

$$\begin{aligned}\|\psi(\mathbf{x}) - \kappa_\theta(\mathbf{x})\| &\leq \max_{g\in\mathcal{M}_\mathcal{K}} \|\rho_\mathcal{Y}(g)\| \; \mathbb{E}_{h(g|\mathbf{x})}\|\psi(\rho_\mathcal{X}(g)^{-1}\mathbf{x}) - \phi_\theta(\rho_\mathcal{X}(g)^{-1}\mathbf{x})\| \\ &\leq c \; \mathbb{E}_{h(g|\mathbf{x})}\|\psi(\rho_\mathcal{X}(g^{-1})\mathbf{x}) - \phi_\theta(\rho_\mathcal{X}(g^{-1})\mathbf{x})\|.\end{aligned} \tag{25}$$

Let us define the set $\mathcal{K}_\mathrm{sym} = \cup_{g\in\mathcal{N}_\mathcal{K}}\rho_\mathcal{X}(g)\mathcal{K}$. Since $\mathcal{N}_\mathcal{K}$ is compact and $\mathcal{X}$ is finite-dimension, the set $\{\rho_\mathcal{X}(g)|g \in \mathcal{N}_\mathcal{K}\}$ is compact, which implies the compactness of $\mathcal{K}_\mathrm{sym}$. Since ϕ_θ is a universal approximator, for any $\epsilon > 0$, there exists a choice of parameters θ such that:

$$\max_{g\in\mathcal{N}_\mathcal{K}} \|\psi(\rho_\mathcal{X}(g)\mathbf{x}) - \phi_\theta(\rho_\mathcal{X}(g)\mathbf{x})\| \leq \epsilon/c, \tag{26}$$

for all $\mathbf{x} \in \mathcal{K}$. Therefore, Equation (25) becomes:

$$\|\psi(\mathbf{x}) - \kappa_\theta(\mathbf{x})\| \leq c \max_{g\in\mathcal{N}_\mathcal{K}} \|\psi(\rho_\mathcal{X}(g)\mathbf{x}) - \phi_\theta(\rho_\mathcal{X}(g)\mathbf{x})\| \leq \epsilon, \tag{27}$$

for all $\mathbf{x} \in \mathcal{K}$. This finishes the proof. ■

So far, we have proven universality of symmetrized function κ_θ over $h(g|\mathbf{x})$ compactly supported on G (Definition 8). We now prove for symmetrized function $\Phi_{\theta,\omega}$ over parameterized distribution $p_\omega(g|\mathbf{x})$ which is approximately compactly supported on G (Definition 9).

Theorem 11 *The symmetrized function $\Phi_{\theta,\omega}$ in Equation (9) is a universal approximator of G equivariant functions if ϕ_θ is a continuous universal approximator and the parameterized probabilistic distribution p_ω is approximately compactly supported on G.*

Proof We prove that for arbitrary G equivariant function $\psi : \mathcal{X} \to \mathcal{Y}$, for any given compact set $\mathcal{K} \subset \mathcal{X}$ and $\epsilon > 0$, there exists a choice of parameters θ, ω such that $\|\psi(\mathbf{x}) - \Phi_{\theta,\omega}(\mathbf{x}))\| \leq \epsilon$ holds for all $\mathbf{x} \in \mathcal{K}$. First, given that p_ω is approximately compactly supported on G, from Definition 9 we obtain the distribution $h(g|\mathbf{x})$ which is compactly supported on G, and a compact set $\mathcal{H}$ that includes the union of the support $\cup_{\mathbf{x}\in\mathcal{K}} \operatorname{supp} p_\omega(g|\mathbf{x})$. Based on that, we define $\kappa_\theta : \mathcal{X} \to \mathcal{Y}$ following Lemma 10 as $\kappa_\theta(\mathbf{x}) = \mathbb{E}_{h(g|\mathbf{x})}[\rho_\mathcal{Y}(g)\phi_\theta(\rho_\mathcal{X}(g)^{-1}\mathbf{x})]$. With the triangle inequality of the metric induced by $\|\cdot\|$, we have:

$$\|\psi(\mathbf{x}) - \Phi_{\theta,\omega}(\mathbf{x}))\| \leq \|\psi(\mathbf{x}) - \kappa_\theta(\mathbf{x})\| + \|\kappa_\theta(\mathbf{x}) - \Phi_{\theta,\omega}(\mathbf{x}))\|. \tag{28}$$

According to Lemma 10, κ_θ is G equivariant and there exists a choice of parameter θ_* such that $\|\psi(\mathbf{x}) - \kappa_{\theta_*}(\mathbf{x})\| \leq \epsilon/2,\ \forall \mathbf{x} \in \mathcal{K}$. So we only need to prove there exists a choice of parameter ω such that $\|\kappa_{\theta_*}(\mathbf{x}) - \phi_{\theta_*,\omega}(\mathbf{x}))\| \leq \epsilon/2,\ \forall \mathbf{x} \in \mathcal{K}$. We first write:

$$\|\kappa_{\theta_*}(\mathbf{x}) - \Phi_{\theta_*,\omega}(\mathbf{x}))\| = \|\mathbb{E}_{h(g|\mathbf{x})}[\rho_\mathcal{Y}(g) f_{\theta_*}(\rho_\mathcal{X}(g)^{-1}\mathbf{x})] - \mathbb{E}_{p_\omega(g|\mathbf{x})}[\rho_\mathcal{Y}^{gl}(g) f_{\theta_*}(\rho_\mathcal{X}^{gl}(g)^{-1}\mathbf{x})]\|. \tag{29}$$

Since p_ω is approximately compactly supported on G, for any $0 < \alpha < 1$, there exists a choice of parameter ω_* such that (1) $p_{\omega_*}(g|\mathbf{x}) \geq (1-\alpha)h(g|\mathbf{x})$ holds for all $g \in G, \mathbf{x} \in \mathcal{K}$, and (2) the union of the support $\cup_{\mathbf{x}\in\mathcal{K}} \operatorname{supp} p_{\omega_*}(g|\mathbf{x})$ is a subset of the compact set $\mathcal{H}$. Given that, we define an unnormalized distribution $p'_{\omega_*}(g|\mathbf{x})$ as follows:

$$\begin{cases} p'_{\omega_*}(g|\mathbf{x}) = p_{\omega_*}(g|\mathbf{x}) - (1-\alpha)h(g|\mathbf{x}) & \forall g \in G, \\ p'_{\omega_*}(g|\mathbf{x}) = p_{\omega_*}(g|\mathbf{x}) & \forall g \notin G. \end{cases} \tag{30}$$

By setting $\omega = \omega_*$, the right side of Equation (29) becomes:

$$\begin{aligned} &\|\kappa_{\theta_*}(\mathbf{x}) - \Phi_{\theta_*,\omega}(\mathbf{x}))\| \\ &= \left\| \int_G [\rho_\mathcal{Y}(g) f_{\theta_*}(\rho_\mathcal{X}(g)^{-1}\mathbf{x})] h(g|\mathbf{x}) d_g - \int_{\mathrm{GL}(n)} [\rho_\mathcal{Y}^{gl}(g) f_{\theta_*}(\rho_\mathcal{X}^{gl}(g)^{-1}\mathbf{x})] p_{\omega_*}(g|\mathbf{x}) d_g \right\| \\ &= \left\| \alpha \int_G [\rho_\mathcal{Y}(g) f_{\theta_*}(\rho_\mathcal{X}(g)^{-1}\mathbf{x})] h(g|\mathbf{x}) d_g - \int_{\mathrm{GL}(n)} [\rho_\mathcal{Y}^{gl}(g) f_{\theta_*}(\rho_\mathcal{X}^{gl}(g)^{-1}\mathbf{x})] p'_{\omega_*}(g|\mathbf{x}) d_g \right\| \end{aligned} \tag{31}$$

By denoting $A(\mathbf{x}, g) = \rho_\mathcal{Y}(g) f_{\theta_*}(\rho_\mathcal{X}(g)^{-1}\mathbf{x})$ and $B(\mathbf{x}, g) = \rho_\mathcal{Y}^{gl}(g) f_{\theta_*}(\rho_\mathcal{X}^{gl}(g)^{-1}\mathbf{x})$, we have:

$$\begin{aligned} \|\kappa_{\theta_*}(\mathbf{x}) - \Phi_{\theta_*,\omega}(\mathbf{x}))\| &= \left\| \alpha \int_G A(\mathbf{x}, g) h(g|\mathbf{x}) d_g - \int_{\mathrm{GL}(n)} B(\mathbf{x}, g) p'_{\omega_*}(g|\mathbf{x}) d_g \right\| \\ &\leq \alpha \int_G \|A(\mathbf{x}, g)\| h(g|\mathbf{x}) d_g + \int_{\mathrm{GL}(n)} \|B(\mathbf{x}, g)\| p'_{\omega_*}(g|\mathbf{x}) d_g. \end{aligned} \tag{32}$$

Since f_{θ_*} is continuous and the set $\mathcal{M}_\mathcal{K} = \cup_{\mathbf{x}\in\mathcal{K}} \operatorname{supp} h(g|\mathbf{x})$ is compact, we have that the set $\cup_{\mathbf{x}\in\mathcal{K}} \{A(\mathbf{x}, g) : g \in \operatorname{supp} h(g|\mathbf{x})\}$ is compact. Thus, there exists $C_1 > 0$ such that

$\|A(\mathbf{x}, g)\| \leq C_1$ for all $\mathbf{x} \in \mathcal{K}, g \in \mathcal{M}_{\mathcal{K}}$. Let us define the set $\mathcal{N}_{\mathcal{K}} = \cup_{\mathbf{x}\in\mathcal{K}} \operatorname{supp} p'_{\omega_*}(g|\mathbf{x})$ and assume that $\mathcal{N}_{\mathcal{K}} \subseteq \mathrm{GL}(n)$. From the property of p_{ω_*} that the union of the support is bounded by the compact set $\mathcal{H}$, and the definition of p'_{ω_*} in Equation (30), we can see that $\mathcal{N}_{\mathcal{K}}$ is bounded by the compact set $\mathcal{H}$. This also leads to the compactness of $\cup_{\mathbf{x}\in\mathcal{K}}\{B(\mathbf{x}, g) : g \in \operatorname{supp} p'_{\omega_*}(g|\mathbf{x})\}$. Therefore, there exists $C_2 > 0$ such that $\|\rho_{\mathcal{Y}}^{gl}(g) f_{\theta_*}(\rho_{\mathcal{X}}^{gl}(g)^{-1}\mathbf{x})\| \leq C_2$ for all $\mathbf{x} \in \mathcal{K}$ and $g \in \mathcal{N}_{\mathcal{K}}$. As a result, Equation (32) becomes:

$$\begin{aligned}\|\kappa_{\theta_*}(\mathbf{x}) - \Phi_{\theta_*,\omega_*}(\mathbf{x}))\| &\leq \alpha \int_G C_1 h(g|\mathbf{x}) d_g + \int_{\mathrm{GL}(n)} C_2 p'_{\omega_*}(g|\mathbf{x}) d_g \\ &= \alpha C_1 + \alpha C_2. \end{aligned} \tag{33}$$

Notice that C_1, C_2 are dependent on distribution $h(g|\mathbf{x})$ and the set $\mathcal{H}$, but not ω. Therefore, by choosing $\alpha = \frac{\epsilon}{2(C_1+C_2)}$, and setting $\omega = \omega_*$ accordingly, we have:

$$\|\kappa_{\theta_*}(\mathbf{x}) - \phi_{\theta_*,\omega_*}(\mathbf{x}))\| \leq \epsilon/2, \ \forall \mathbf{x} \in \mathcal{K} \tag{34}$$

This finishes the proof. ■

A.5. Experimental Details

A.5.1. Rotated MNIST

In this section, we supplement the implementation and training details for the Rotated MNIST experiment. We employ exactly the same CNN architecture as Kaba et al. (2023) to implement our base function ϕ_θ, which has 7 layers with hidden dimensions of 32, 64, 128 for layers 1 – 3, layers 4 – 6, and layer 7 respectively. At layers 4 and 7, a 5×5 convolution with stride 2 are used instead of pooling. Other convolutions use 3×3 filters with stride 1. Each convolution is followed by batch normalization and ReLU activation. Dropout of $p = 0.4$ is used at layers 4 and 7.

For our SO(2) equivariant symmetrizer $q_\omega(\mathbf{x}, \boldsymbol{\epsilon})$, given an input image $\mathbf{x} \in \mathbb{R}^{28\times28\times1}$, we preprocess it into a tensor format and apply an EMLP (Finzi et al., 2021b) on it. In more detail, firstly, we construct a coordinate map $\mathbf{C} \in \mathbb{R}^{28\times28\times2}$ where the central pixel has coordinate $(0, 0)$ and each corner has coordinate $(\pm14, \pm14)$. This coordinate map will be shared for all images. We construct a tensor $\mathbf{v} \in \mathbb{R}^{28\times28\times3}$ by concatenating $\mathbf{x}$ and $\mathbf{C}$ across the channel dimension ($\mathbf{v}[:, :, 0] = \mathbf{x}, \mathbf{v}[:, :, 1:] = \mathbf{C}$). Then, we flat the first two dimension to have tensor $\mathbf{v} \in \mathbb{R}^{784\times3}$. Now, each row of $\mathbf{v}$ corresponds to a pixel in the original image, where the first channel is the pixel value, while the last two channels are pixel's coordinates. Next, we sort rows in $\mathbf{v}$ in an ascending order of the first column ($\mathbf{v}[:, 0]$) and mask out all rows with a pixel value less than a predefined threshold $t = 0.2$. Then, we keep the top $m = 200$ rows and achieve a new tensor $\mathbf{v} \in \mathbb{R}^{m\times3}$. Lastly, we extract two tensors $\mathbf{v}_1, \mathbf{v}_2$ from $\mathbf{v}$ as following: $\mathbf{v}_1 = \mathbf{v}[:, 0]^\top \in \mathbb{R}^{1\times m}, \mathbf{v}_2 = \mathbf{v}[:, 1:]^\top \in \mathbb{R}^{2\times m}$. Notice that these processing steps do not break rotation equivariance and $\mathbf{v}_1$ is SO(2) invariant while $\mathbf{v}_2$ is SO(2) equivariant with respect to input image $\mathbf{x}$.

To train symmetrization models, we employ faithful noise $\boldsymbol{\epsilon} \in \mathbb{R}^{2\times d_\mathcal{E}}$ with $d_\mathcal{E}$ set to 10. To implement the symmetrizer, we leverage a 2-layers EMLP with hidden dimension 64

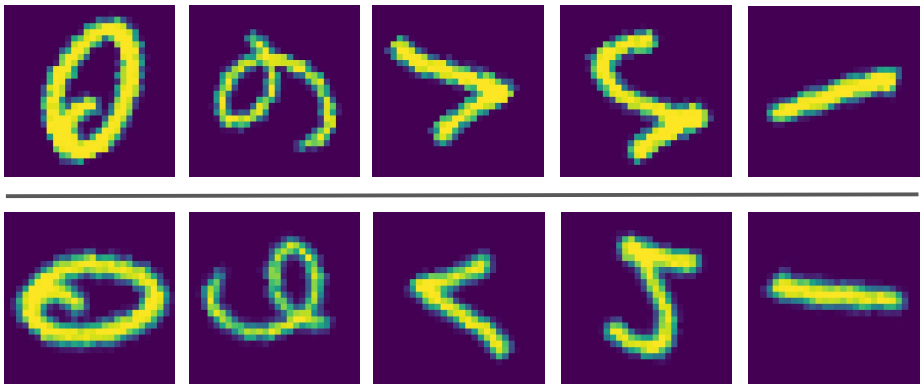

Figure 1: Example transformed images for Rotated MNIST dataset. The first row include images from the original dataset, while the second row are those images transformed by output of our learned symmetrizer $q_\omega(\mathbf{x}, \boldsymbol{\epsilon})$. It is clear from these figures that transformations associated with $q_\omega(\mathbf{x}, \boldsymbol{\epsilon})$ is purely rotation.

that has the following input and output representations in notation of Finzi et al. (2021a):

$$\begin{cases} \text{repin} = (d_{\mathcal{E}} + m)T_{(0)} + mT_{(1)} \\ \text{repout} = 2T_{(1)} \end{cases} \tag{35}$$

With the symmetrizer $q_\omega(\mathbf{x}, \boldsymbol{\epsilon})$ designed in this manner, q_ω is guaranteed to always output 2×2 matrices in $\mathrm{SO}(2)$ equivariant manner with respect to the input image $\mathbf{x}$.

When we compute the input transformation $q_\omega(\mathbf{x}, \boldsymbol{\epsilon})^{-1} \cdot \mathbf{x}$ as in Equation (4), we find that since $q_\omega(\mathbf{x}, \boldsymbol{\epsilon})$ is a neural feature, directly applying matrix inverse to it harms the stability of training. To enhance stability, we assume as if $q_\omega(\mathbf{x}, \boldsymbol{\epsilon})$ is already close to a $\mathrm{SO}(2)$ matrix, and employ the property of $\mathrm{SO}(n)$ matrices $\rho(g)^{-1} = \rho(g)^\top$ to compute an approximation of inverse. This approximation becomes more exact as orbit distance training progresses $q_\omega(\mathbf{x}, \boldsymbol{\epsilon}) \to \rho(g) \in \mathrm{SO}(2)$. Indeed, we observe this significantly improves training stability while not harming the quality of learned transformations $q_\omega(\mathbf{x}, \boldsymbol{\epsilon})$. To apply the transformation $q_\omega(\mathbf{x}, \boldsymbol{\epsilon})^{-1}$ to input image $\mathbf{x}$, we follow implementation of Kaba et al. (2023) and employ computer vision library Kornia (Riba et al., 2020).

We train the models by minimizing the cross entropy loss on classification task jointly with L1 norm for orbit distance minimization using $\lambda = 1.0$. The models are trained for 1000 epochs using Adam optimizer and a learning rate of 0.0003. All symmetrization methods in Table 1(*a*) are trained with this same setting for fair comparison. Some example transformations $q_\omega(\mathbf{x}, \boldsymbol{\epsilon})^{-1} \cdot \mathbf{x}$ learned by our models trained with orbit distance minimization is shown in Figure 1, which verifies that valid group representations are indeed learned.

A.5.2. Particle Scattering

In this section, we supplement the implementation and training details for the Particle Scattering experiment. We use 3-layer MLP with 128 hidden dimensions and SiLU (Hendrycks and Gimpel, 2016) activation function as our base function ϕ_θ.

Given an input $\mathbf{x} \in \mathbb{R}^{4\times 4}$, for the $\mathrm{O}(1,3)$ equivariant symmetrizer $q_\omega(\mathbf{x}, \boldsymbol{\epsilon})$, we use a 3-layer Scalar MLP (Villar et al., 2021) with 28 hidden dimensions and SiLU activation, preceded by an $\mathrm{O}(1,3)$ equivariant featurization procedure. In more detail, we first sample the noise variable $\boldsymbol{\epsilon} \in \mathbb{R}^{4\times d_\mathcal{E}}$ from a compactly supported distribution $p(\boldsymbol{\epsilon})$ which is invariant to $\mathrm{O}(1,3)$ under trivial representation $\rho_\mathcal{E}(g) = \mathbf{I}$. We use elementwise uniform distribution $\boldsymbol{\epsilon}_{ij} \sim \mathrm{Unif}[\mathbf{a}_{ij}, \mathbf{a}_{ij} + \mathbf{b}_{ij}]$ with trainable offset $\mathbf{a} \in \mathbb{R}^{4\times d_\mathcal{E}}$ and scale $\mathbf{b} \in \mathbb{R}^{4\times d_\mathcal{E}}$ initialized as $\mathbf{1}$ and $\mathbf{0}$ respectively. Then, before Scalar MLP, we transform $\boldsymbol{\epsilon}$ into a feature matrix $\mathbf{z} \in \mathbb{R}^{4\times d_\mathcal{E}}$ with a simple procedure equivariant to $\mathrm{O}(1,3)$ transformations of the input $\mathbf{x}$. We interpret the sampled noise $\boldsymbol{\epsilon}$ as the the Minkowsky inner product between input $\mathbf{x}$ and feature $\mathbf{z}$ (which is unknown at this point) $\boldsymbol{\epsilon} = \mathbf{x}^\top \boldsymbol{\Lambda} \mathbf{z}$ where $\boldsymbol{\Lambda} = \mathrm{diag}([+1, -1, -1, -1])$, from which we obtain $\mathbf{z} = (\mathbf{x}^\top \boldsymbol{\Lambda})^{-1} \boldsymbol{\epsilon}$. As Minkowsky inner product $\boldsymbol{\epsilon}$, or space-time interval, is $\mathrm{O}(1,3)$ invariant as can be seen in $\mathbf{x}^\top \boldsymbol{\Lambda} \mathbf{z} = (g \cdot \mathbf{x})^\top \boldsymbol{\Lambda} (g \cdot \mathbf{z}) = \mathbf{x}^\top \rho(g)^\top \boldsymbol{\Lambda} \rho(g) \mathbf{z} = \mathbf{x}^\top \boldsymbol{\Lambda} \mathbf{z}$, having fixed the noise $\boldsymbol{\epsilon}$, transforming $\mathbf{x} \mapsto g \cdot \mathbf{x}$ transforms $\mathbf{z} \mapsto g \cdot \mathbf{z}$ accordingly. For Canonicalization (Kaba et al., 2023) that has to drop stochasticity, we simply use deterministic $\boldsymbol{\epsilon} = \mathbf{a}$ to obtain the feature $\mathbf{z}$. Then, we then use this feature $\mathbf{z}$ to supplement the input $\mathbf{x}$ by addition or channel concatenation, and provide the combined feature as an input to Scalar MLP and obtain the output $\mathbf{h} \in \mathbb{R}^{4\times 4}$. In our experiments, we find that this featurization significantly and consistently improves orbit distance training. Note that our framework and theoretical results are not altered, as the featurization $(\mathbf{x}, \boldsymbol{\epsilon}) \mapsto \mathbf{z}$ is $\mathrm{O}(1,3)$ equivariant and can be interpreted as a part the equivariant symmetrizer $q_\omega : (\mathbf{x}, \boldsymbol{\epsilon}) \mapsto \mathbf{h}$.

Similar as in the $\mathrm{SO}(2)$ experiment, to avoid computing inverse of neural feature during the input transformation $q_\omega(\mathbf{x}, \boldsymbol{\epsilon})^{-1} \cdot \mathbf{x}$ (Equation (4)), we assume as if $q_\omega(\mathbf{x}, \boldsymbol{\epsilon})$ is already close to a $\mathrm{O}(1,3)$ matrix, and employ the property of $\mathrm{O}(1,3)$ matrices $\rho(g)^{-1} = \boldsymbol{\Lambda} \rho(g)^\top \boldsymbol{\Lambda}$ to compute an approximation of inverse. This approximation becomes more exact as orbit distance training progresses $q_\omega(\mathbf{x}, \boldsymbol{\epsilon}) \to \rho(g) \in \mathrm{O}(1,3)$, and we observe this significantly improves training stability while not harming the quality of learned transformations $q_\omega(\mathbf{x}, \boldsymbol{\epsilon})$.

We train the models by minimizing the mean squared error loss for regression task jointly with L1 norm on the orbit distance minimization using $\lambda = 1$. All models are trained for 1,000 epochs with batch size 1,000 using Adam optimizer with a learning rate of 0.003. All methods in Table 1(*b*) are trained with this same setting. Both of our models $q_\omega(\mathbf{x}, \boldsymbol{\epsilon}) \mapsto \mathbf{h}$ consistently achieve orbit distance $\|f(\mathbf{h}) - f(\mathbf{I})\|$ of around 0.02–0.03 on unseen inputs $\mathbf{x}$, while random Gaussian matrices have loss of around ≈ 50. This supports that valid group representations are learned by our approach.