Equivariant Representation Learning for Symmetry-Aware Inference with Guarantees

Anonymous Author(s)
Affiliation
Address
email

Abstract: In many real-world applications of regression, conditional probability estimation, and uncertainty quantification, exploiting symmetries rooted in physics or geometry can dramatically improve generalization and sample efficiency. While geometric deep learning has made significant empirical advances by incorporating group-theoretic structure, less attention has been given to statistical learning guarantees. In this paper, we introduce an equivariant representation learning framework that simultaneously addresses regression, conditional probability estimation, and uncertainty quantification while providing first-of-its-kind non-asymptotic statistical learning guarantees. Grounded in operator and group representation theory, our framework approximates the spectral decomposition of the conditional expectation operator, building representations that are both equivariant and disentangled along independent symmetry subgroups. Empirical evaluations on synthetic datasets and real-world robotics applications confirm the potential of our approach, matching or outperforming existing equivariant baselines in regression while additionally providing well-calibrated parametric uncertainty estimates.

Keywords: Representation learning, uncertainty quantification, deep learning, geometric deep learning

1 Introduction

A central problem in machine learning is modeling conditional probabilities—understanding how the distribution of a target variable $\mathbf{y}$ changes with an observed variable $\mathbf{x}$. This underpins robust reasoning under uncertainty in critical applications such as medicine, finance, robotics, and physics [1, 2, 3]. However, estimating conditional distributions remains challenging in high-dimensional settings without strong inductive biases [4, 5, 6].

Symmetry priors, in the form of principled assumptions about invariance or equivariance in the underlying data-generating process, offer a compelling way to reduce sample complexity and improve generalization [7, 8, 9, 10]. These priors naturally arise in inference tasks in chemistry and particle physics [11], set-&-graph structured data [9], computer graphics [12, 13], and dynamical systems with group-invariant/equivariant laws of motion, which are ubiquitous in fields like physics [11], fluid dynamics [14], and robotics [15, 16].

Over the past few years, Geometric Deep Learning (GDL) has produced a rich ecosystem of architectures that encode symmetries, achieving strong empirical performance across various supervised [9, 17, 18, 19] and unsupervised tasks [20, 21, 22]. However, the field remains focused on application specific designs and architectural innovation, with limited understanding of how symmetry priors can be leveraged to *learn representations with provable generalization guarantees*.

In this work, we take a different route: rather than proposing new architectures or solving specific inference tasks, we ask *how to systematically learn symmetry-aware representations that best capture conditional structure in the data*. Specifically, how should equivariant networks be trained so that their

Submitted to the 9th Conference on Robot Learning (CoRL 2025). Do not distribute.

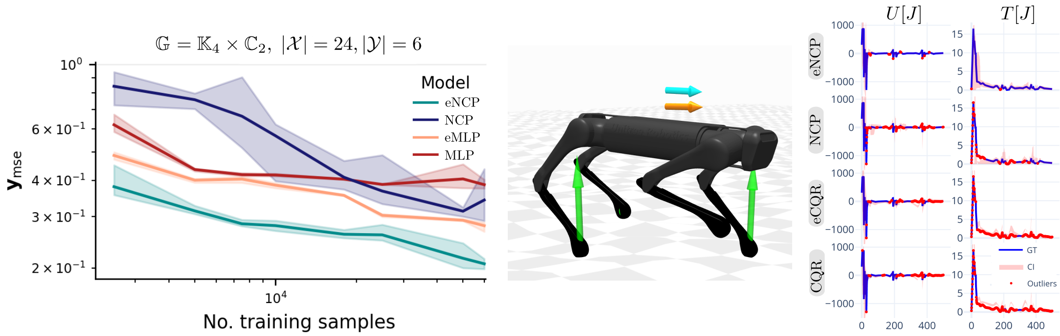

Figure 1: **Left**: Test set sample efficiency for $\mathbb{G}$-equivariant regression (MSE vs. training samples) when predicting the $\mathbb{G}$-equivariant linear and angular momentum of a quadruped robot's center of mass (CoM) from noisy joint positions and velocities. **Right**: Uncertainty quantification via $\mathbb{G}$-equivariant prediction of 90% confidence intervals (CI, light-red area) for the robot's instantaneous work U_t and kinetic energy T_t during locomotion over rough terrain for our method (eNCP) and competitors. The figure shows a trajectory with a strong initial disturbance, where blue markers denote samples within the predicted CI and red markers denote those outside. Note that only eNCP is able to predict well-calibrated CI intervals that cover both the disturbance and recovery phases.

learned features reveal conditional distributions, and how does the quality of these representations affect performance in downstream tasks such as regression and uncertainty quantification?

To answer these questions, we bridge two fields rarely studied together: spectral contrastive learning [23], a self-supervised approach that learns deep representations of data via operator-theoretic modeling of conditional expectations [24, 25], and GDL [9], which enforces symmetry priors as architectural constraints in Neural Networks (NNs). Our approach shows how symmetry constraints shape the representation space and enhance generalization, opening new avenues for cross-fertilization between these fields. Concretely, we demonstrate that our method outperforms GDL techniques on regression tasks (see Fig. 1-left) while providing reliable uncertainty quantification on a challenging robot locomotion task (see Fig. 1-right).

Contributions (1) Methodological framework: We introduce Equivariant Neural Conditional Probability (eNCP), the first framework to combine equivariant neural networks with operator-theoretic estimation of conditional distributions. (2) Task-agnostic representation learning: We show that any $\mathbb{G}$-equivariant architecture can be used to learn *disentangled, symmetry-respecting representations* that generalize across diverse downstream inference tasks. (3) Learning guarantees: By linking the representation quality directly to sample complexity, we provide the *first non-asymptotic statistical learning guarantees* for equivariant conditional models, including regression and uncertainty quantification. (4) Empirical results: On both synthetic and real-world robotics tasks, eNCP consistently outperforms baselines, including contrastive methods Neural Conditional Probability (NCP) [25] and current equivariant models. In particular, eNCP achieves state-of-the-art performance in the challenging task of contact force inference in quadruped locomotion.

Paper structure Sec. 2 reviews modeling conditional probabilities with linear operators and NCP. Sec. 3 formally presents the symmetry priors we consider. Sec. 4 introduces our eNCP learning framework. Sec. 5 outlines our theoretical learning guarantees. Sec. 6 showcases experiments on synthetic and real-world data. Furthermore, because the paper involves complex notation from probability, operator theory, and group theory, the appendices include a glossary of notation (App. A) as well as detailed expositions on representation theory (App. I), symmetric function spaces (App. J), and equivariant linear operators (App. K). Finally, App. B offers an in-depth discussion of related work, contrasting our framework with the literature across these rich fields.

2 Background

We briefly review the operator-theoretic framework for modeling conditional probabilities, which underpins both NCP and our proposed eNCP method. We denote a random variable by $\mathbf{x}$, its

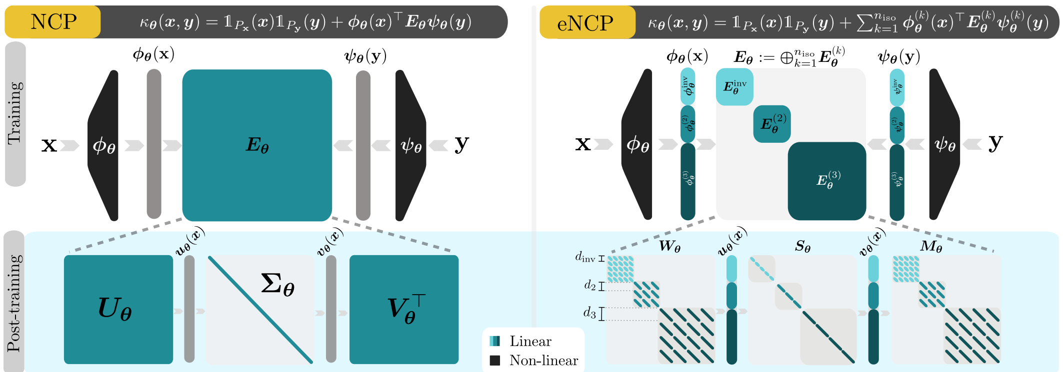

Figure 2: **Left:** NCP's bilinear NN architecture. **Right:** eNCP's $\mathbb{G}$-equivariant bilinear NN architecture, featuring $\boldsymbol{\phi_\theta}$ and $\boldsymbol{\psi_\theta}$ as $\mathbb{G}$-equivariant NNs and $\boldsymbol{E_\theta}$ as a $\mathbb{G}$-equivariant block-diagonal matrix. Each block is equivariant to a subgroup $\mathbb{G}^{(k)} \leq \mathbb{G}$ and is constrained to have singular spaces of dimension at least d_k—the minimal dimension for a faithful representation of the action of $\mathbb{G}^{(k)}$.

realizations by $\boldsymbol{x} \in \mathcal{X}$, its probability distribution by $\mathbb{P}(\mathbf{x})$ and measure by $P_{\mathbf{x}}$. We write expectations as $\mathbb{E}_{\mathbf{x}}[f(\mathbf{x})] = \int_{\mathcal{X}} f(\boldsymbol{x})\, P_{\mathbf{x}}(d\boldsymbol{x})$. The same notations apply to other random variables such as $\mathbf{y}$.

Operator-theoretic modeling of conditional probabilities Kostic et al. [25] proposed to model conditional probabilities by approximating the *conditional expectation operator* [26, 27, 28], $\mathsf{E}_{\mathbf{y}|\mathbf{x}} \colon \mathcal{L}^2_{\mathbf{y}} \to \mathcal{L}^2_{\mathbf{x}}$, a linear integral operator acting on the Hilbert spaces $\mathcal{L}^2_{\mathbf{x}} := \mathcal{L}^2_{P_{\mathbf{x}}}(\mathcal{X}, \mathbb{R})$ and $\mathcal{L}^2_{\mathbf{y}} := \mathcal{L}^2_{P_{\mathbf{y}}}(\mathcal{Y}, \mathbb{R})$ of square-integrable functions of the random variables $\mathbf{x}$ and $\mathbf{y}$, respectively. The action of this operator on any function $h \in \mathcal{L}^2_{\mathbf{y}}$ returns the function's conditional expectation:

$$[\mathsf{E}_{\mathbf{y}|\mathbf{x}} h](\boldsymbol{x}) = \mathbb{E}[h(\mathbf{y})|\mathbf{x}=\boldsymbol{x}] := \int_{\mathcal{Y}} h(\boldsymbol{y}) P_{\mathbf{y}|\mathbf{x}}(d\boldsymbol{y}|\boldsymbol{x}) = \int_{\mathcal{Y}} h(\boldsymbol{y}) \tfrac{P_{\mathbf{yx}}(d\boldsymbol{y},\boldsymbol{x})}{P_{\mathbf{x}}(d\boldsymbol{x})} = \int_{\mathcal{Y}} h(\boldsymbol{y}) \kappa(\boldsymbol{x},\boldsymbol{y}) P_{\mathbf{y}}(d\boldsymbol{y}), \quad (1)$$

where $P_{\mathbf{y}|\mathbf{x}}$ is the conditional probability measure, and $\kappa(\boldsymbol{x},\boldsymbol{y}) := \frac{P_{\mathbf{xy}}(d\boldsymbol{x}, d\boldsymbol{y})}{P_{\mathbf{x}}(d\boldsymbol{x})\, P_{\mathbf{y}}(d\boldsymbol{y})}$ is the kernel of $\mathsf{E}_{\mathbf{y}|\mathbf{x}}$, also known as the Pointwise Mutual Dependency (PMD) [29] (see Fig. 3 and App. H).

The conditional expectation operator is significant because it provides an infinite-dimensional linear model—in a nonlinear representation space—for computing conditional probabilities and expectations. To see this, note that for any $\boldsymbol{x} \in \mathcal{X}$ and any measurable set $\mathbb{B} \subset \mathcal{Y}$ we have that:

$$\mathbb{P}(\mathbf{y} \in \mathbb{B}|\mathbf{x}=\boldsymbol{x}) := \int_{\mathcal{Y}} \mathbb{1}_{\mathbb{B}}(\boldsymbol{y}) P_{\mathbf{y}|\mathbf{x}}(d\boldsymbol{y}|\boldsymbol{x}) = [\mathsf{E}_{\mathbf{y}|\mathbf{x}} \mathbb{1}_{\mathbb{B}}](\boldsymbol{x}), \quad \text{and} \quad \mathbb{E}[\mathbf{y}|\mathbf{x}=\boldsymbol{x}] := [\mathsf{E}_{\mathbf{y}|\mathbf{x}} \mathbf{y}](\boldsymbol{x}). \quad (2)$$

Therefore, to estimate conditional probabilities and expectations, NCP seeks the best finite-dimensional approximation of $\mathsf{E}_{\mathbf{y}|\mathbf{x}}$. As we explain next, this gives rise to a representation learning problem [30], in which the optimal representations of $\mathbf{x}$ and $\mathbf{y}$ are given by the top left and right singular functions of $\mathsf{E}_{\mathbf{y}|\mathbf{x}}$.

Spectral representation learning The problem of approximating the conditional expectation operator $\mathsf{E}_{\mathbf{y}|\mathbf{x}}$ as a rank-r operator $\mathsf{E}_{\boldsymbol{\theta}}$ with matrix representation $\boldsymbol{E_\theta} \in \mathbb{R}^{r \times r}$ is defined as

$$\arg\min_{\boldsymbol{\theta}} \|\mathsf{E}_{\mathbf{y}|\mathbf{x}} - \mathsf{E}_{\boldsymbol{\theta}}\|^2_{\mathrm{HS}} = \mathbb{E}_{\mathbf{x}} \mathbb{E}_{\mathbf{y}} (\kappa(\mathbf{x},\mathbf{y}) - \kappa_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{y}))^2, \quad \text{s.t. } \mathbb{E}_{\mathbf{x}} \mathbb{E}_{\mathbf{y}} \kappa_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{y}) = 1 \text{ and } \operatorname{rank}(\mathsf{E}_{\boldsymbol{\theta}}) \leq r. \quad (3)$$

The optimal solution, denoted $\mathsf{E}_\star$, is the r-truncated Singular Value Decomposition (SVD) of $\mathsf{E}_{\mathbf{y}|\mathbf{x}}$ [31, 25], namely

$$[\mathsf{E}_\star f](\boldsymbol{x}) = \textstyle\sum_{i=0}^{r} \sigma_i \langle f, v_i \rangle_{P_{\mathbf{y}}} u_i(\boldsymbol{x}), \qquad \text{with} \quad \sigma_i u_i(\boldsymbol{x}) = [\mathsf{E}_{\mathbf{y}|\mathbf{x}} v_i](\boldsymbol{x}),\ \forall i \in [r], \quad (4)$$

where (σ_i, u_i, v_i) denotes the i^{th} singular value and left/right singular functions of $\mathsf{E}_{\mathbf{y}|\mathbf{x}}$, with $(\sigma_0{=}1, u_0{=}\mathbb{1}_{P_{\mathbf{x}}}, v_0{=}\mathbb{1}_{P_{\mathbf{y}}})$ being the constant functions supported on $P_{\mathbf{x}}$ and $P_{\mathbf{y}}$, respectively [26, 25].

Consequently, NCP parameterizes $\mathsf{E}_{\boldsymbol{\theta}}$ by a bilinear model $\kappa_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{y}) = 1 + \boldsymbol{\phi_\theta}(\boldsymbol{x})^\top \boldsymbol{E_\theta} \boldsymbol{\psi_\theta}(\boldsymbol{y})$, composed of two encoder NNs $\boldsymbol{\phi_\theta} \colon \mathcal{X} \to \mathbb{R}^r$ and $\boldsymbol{\psi_\theta} \colon \mathcal{Y} \to \mathbb{R}^r$ that aim to approximate the *span* of the top r (non-constant) left and right singular functions of $\mathsf{E}_{\mathbf{y}|\mathbf{x}}$. See Fig. 2-left.

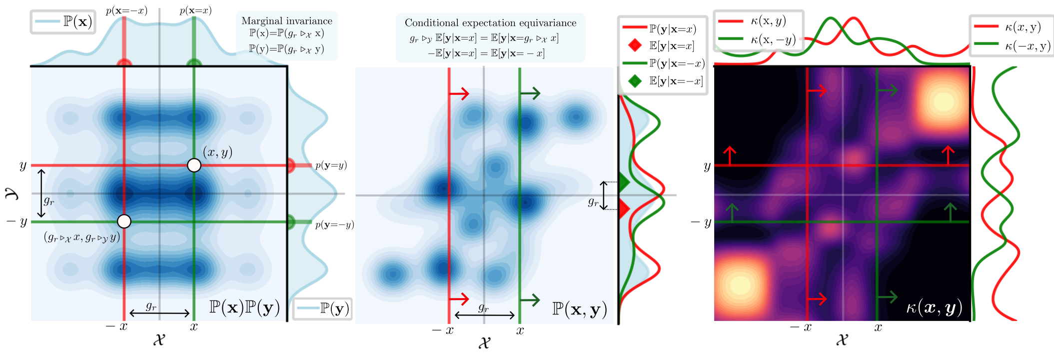

Figure 3: Example of symmetric random variables $(\mathrm{x}, \mathrm{y}) \sim \mathcal{X} \times \mathcal{Y} \subset \mathbb{R} \times \mathbb{R}$, whose marginals $\mathbb{P}(\mathrm{x})$ and $\mathbb{P}(\mathrm{y})$; joint $\mathbb{P}(\mathrm{x}, \mathrm{y})$; and conditional $\mathbb{P}(\mathrm{y}|\mathrm{x})$ distributions are invariant to reflections of the data: $g_r \triangleright_{\mathcal{X}} x = -x$ and $g_r \triangleright_{\mathcal{Y}} y = -y$, where g_r denotes the reflection element of the reflection symmetry group $\mathbb{C}_2 := \{e, g_r | g_r^2 = e\}$. Consequently, the PMD $\kappa(x, y)$ is $\mathbb{C}_2$-invariant.

Since κ is generally unavailable analytically, (3) is solved via the regularized contrastive loss[1]:

$$\begin{aligned} \mathcal{L}_\gamma(\boldsymbol{\theta}) \;=\; & -2\mathbb{E}_{\mathbf{xy}}\kappa_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}) + \mathbb{E}_{\mathbf{x}}\mathbb{E}_{\mathbf{y}}\kappa_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y})^2 + 2\gamma\big(\|\mathbb{E}_{\mathbf{x}}\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{x})\|_F^2 + \|\mathbb{E}_{\mathbf{y}}\boldsymbol{\psi}_{\boldsymbol{\theta}}(\mathbf{y})\|_F^2 \\ & + \|\mathrm{Cov}(\boldsymbol{\phi}_{\boldsymbol{\theta}}) - \boldsymbol{I}_r\|_F^2 + \|\mathrm{Cov}(\boldsymbol{\psi}_{\boldsymbol{\theta}}) - \boldsymbol{I}_r\|_F^2\big), \end{aligned} \tag{5}$$

where the first two regularization terms center the learned representations, ensuring that $\mathbb{E}_{\mathbf{x}}\mathbb{E}_{\mathbf{y}}\kappa_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}) \approx 1$ [25], while the last two enforce approximate orthonormality of the learned bases in $\mathcal{F}_{\mathbf{x}}^{\boldsymbol{\theta}} := \mathrm{span}(\boldsymbol{\phi}_{\boldsymbol{\theta}}) \subset \mathcal{L}_{\mathbf{x}}^2$ and $\mathcal{F}_{\mathbf{y}}^{\boldsymbol{\theta}} := \mathrm{span}(\boldsymbol{\psi}_{\boldsymbol{\theta}}) \subset \mathcal{L}_{\mathbf{y}}^2$ [6]. A key property of NCP is that the learned representations enables reliable regression and conditional probability estimation—and thus uncertainty quantification—via (2) (see Tab. 3 in the appendix and [25]).

3 Problem formulation

This paper tackles the problem of estimating the conditional expectation $\mathbb{E}[\mathbf{y}|\mathbf{x} = \cdot]$, and, more generally, conditional distribution $\mathbb{P}(\mathbf{y}|\mathbf{x})$, for random variables $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{Y}$, under the assumption that[2] $\mathbb{P}(\mathbf{y}|\mathbf{x})$ and $\mathbb{P}(\mathbf{x})$ are $\mathbb{G}$-invariant under symmetry transformations of the data (see Fig. 3), i.e.:

$$\mathbb{P}(\mathbf{y}|\mathbf{x}) = \mathbb{P}(g \triangleright_{\mathcal{Y}} \mathbf{y} | g \triangleright_{\mathcal{X}} \mathbf{x}), \qquad \mathbb{P}(\mathbf{x}) = \mathbb{P}(g \triangleright_{\mathcal{X}} \mathbf{x}), \quad \forall\, g \in \mathbb{G}, \tag{6}$$

where $\mathbb{G}$ denotes a finite symmetry group (Def. I.1) acting on the data spaces $\mathcal{X}$ and $\mathcal{Y}$ via the group actions, $\triangleright_{\mathcal{X}}$: $\mathbb{G} \times \mathcal{X} \to \mathcal{X}$, and $\triangleright_{\mathcal{Y}}$: $\mathbb{G} \times \mathcal{Y} \to \mathcal{Y}$, with $g \triangleright_{\mathcal{X}} \boldsymbol{x} \in \mathcal{X}$ and $g \triangleright_{\mathcal{Y}} \boldsymbol{y} \in \mathcal{Y}$ denoting linear, invertible transformations of $\boldsymbol{x}$ and $\boldsymbol{y}$ defined by $g \in \mathbb{G}$ (see Fig. 3 and Def. I.2).

These priors imply the $\mathbb{G}$-invariance of the joint distribution $\mathbb{P}(\mathbf{x}, \mathbf{y})$ and of $\mathbf{y}$'s marginal distribution $\mathbb{P}(\mathbf{y})$, as well as the $\mathbb{G}$-equivariance of conditional expectations (see Fig. 3-middle and Prop. D.1):

$$g \triangleright_{\mathcal{Y}} \mathbb{E}[\mathbf{y}|\mathbf{x}{=}\boldsymbol{x}] = \mathbb{E}[\mathbf{y}|\mathbf{x}{=}g \triangleright_{\mathcal{X}} \boldsymbol{x}] \quad \forall\, g \in \mathbb{G}, \boldsymbol{x} \in \mathcal{X}. \tag{7}$$

Note that (7) implies the $\mathbb{G}$-equivariance of the regression function $\boldsymbol{x} \mapsto \mathbb{E}[\mathbf{y}|\mathbf{x} = \boldsymbol{x}]$. Therefore, the symmetry priors (6) are satisfied whenever we approximate an equivariant/invariant function—that is, in virtually *all applications of* GDL [9].

The above symmetry priors represent a strong inductive bias for the conditional expectation operator (2), as they lead the PMD kernel defining the operator to be $\mathbb{G}$-invariant (see Fig. 3-right):

$$\kappa(\boldsymbol{x}, \boldsymbol{y}) = \kappa(g \triangleright_{\mathcal{X}} \boldsymbol{x}, g \triangleright_{\mathcal{Y}} \boldsymbol{y}) \quad \forall\, g \in \mathbb{G}, \boldsymbol{x} \in \mathcal{X}, \boldsymbol{y} \in \mathcal{Y}. \tag{8}$$

In Sec. 4, we extend the NCP framework [25] by leveraging (8) to incorporate the above symmetry priors. As we shall see, this enables efficient use of GDL architectures to estimate the $\mathbb{G}$-invariant conditional probabilities in (6) and $\mathbb{G}$-equivariant regression in (7), via (2), with strong learning

[1] Used in density-ratio fitting [29], representation learning [32, 33], and mutual information estimation [34].

[2] Throughout, with some abuse of notation we denote by $\mathbb{P}(\mathbf{x})$ and $\mathbb{P}(\mathbf{y}|\mathbf{x})$ both the probability and conditional probability, respectively, as well as the corresponding densities, when they exist.

guarantees. In Sec. 4, we extend the NCP framework [25] by leveraging (8) to incorporate symmetry priors. This enables efficient estimation of $\mathbb{G}$-invariant conditional probabilities (6) and $\mathbb{G}$-equivariant regression (7) using GDL architectures, via (2), with strong learning guarantees.

4 ENCP method for equivariant representation learning

In this section, we show how to incorporate the symmetry priors (6) into NCP's representation learning framework. First, we analyze the symmetry constraints on the infinite-dimensional conditional expectation operator and prove that, for symmetric random variables $\mathbf{x}$ and $\mathbf{y}$, the optimal solution of (3) yields $\mathbb{G}$-equivariant representations $\boldsymbol{\phi}_{\boldsymbol{\theta}}$ and $\boldsymbol{\psi}_{\boldsymbol{\theta}}$ and approximates the operator with a $\mathbb{G}$-equivariant matrix $\boldsymbol{E}_{\boldsymbol{\theta}}$. Then, we explain how to embed these structural constraints into the bilinear neural network architecture of NCP using any type of equivariant NNs.

Symmetric function spaces The assumption of $\mathbb{G}$-invariance of the marginal probabilities (Sec. 3) implies that the function spaces $\mathcal{L}^2_{\mathbf{x}}$ and $\mathcal{L}^2_{\mathbf{y}}$ are symmetric Hilbert spaces of $\mathbb{G}$-equivariant functions, as these inherit unitary group actions $\triangleright_{\mathcal{L}^2_{\mathbf{x}}} : \mathbb{G} \times \mathcal{L}^2_{\mathbf{x}} \to \mathcal{L}^2_{\mathbf{x}}$ and $\triangleright_{\mathcal{L}^2_{\mathbf{y}}} : \mathbb{G} \times \mathcal{L}^2_{\mathbf{y}} \to \mathcal{L}^2_{\mathbf{y}}$ defined via the push-forward of symmetry transformations of the data spaces (see details in App. J and in Fig. 13):

$$g \triangleright_{\mathcal{L}^2_{\mathbf{x}}} f(\cdot) := f(g^{-1} \triangleright_{\mathcal{X}} \cdot) \in \mathcal{L}^2_{\mathbf{x}}, \qquad g \triangleright_{\mathcal{L}^2_{\mathbf{y}}} h(\cdot) := h(g^{-1} \triangleright_{\mathcal{Y}} \cdot) \in \mathcal{L}^2_{\mathbf{y}}, \quad \forall g \in \mathbb{G}. \tag{9}$$

A fundamental property of $\mathbb{G}$-symmetric Hilbert spaces is their orthogonal decomposition into $n_{\text{iso}} \leq |\mathbb{G}|$ subspaces referred to as *isotypic subspaces*: $\mathcal{L}^2_{\mathbf{x}} = \bigoplus^{\perp}_{k \in [1, n_{\text{iso}}]} \mathcal{L}^{2(k)}_{\mathbf{x}}$, and $\mathcal{L}^2_{\mathbf{y}} = \bigoplus^{\perp}_{k \in [1, n_{\text{iso}}]} \mathcal{L}^{2(k)}_{\mathbf{y}}$ (see Thm. I.8). Where each $\mathcal{L}^{2(k)}_{\mathbf{x}}$ and $\mathcal{L}^{2(k)}_{\mathbf{y}}$ denote the spaces of $\mathbb{G}^{(k)}$-equivariant functions of $\mathbf{x}$ and $\mathbf{y}$, with $\mathbb{G}^{(k)}$ being a **subgroup** of $\mathbb{G}$. This standard result from harmonic analysis [35] enables us to express any $\mathbb{G}$-equivariant function as a sum of its projections onto the isotypic subspaces:

$$f(\cdot) = f^{\text{inv}}(\cdot) + \sum_{k=2}^{n_{\text{iso}}} f^{(k)}(\cdot), \quad h(\cdot) = h^{\text{inv}}(\cdot) + \sum_{k=2}^{n_{\text{iso}}} h^{(k)}(\cdot), \quad \text{s.t } f^{(k)} \in \mathcal{L}^{2(k)}_{\mathbf{x}}, h^{(k)} \in \mathcal{L}^{2(k)}_{\mathbf{y}}, \forall k \in [n_{\text{iso}}], \tag{10}$$

where $f^{(k)}$ and $h^{(k)}$ denote the $\mathbb{G}^{(k)}$-equivariant components of f and h, which are by construction invariant to all $g \notin \mathbb{G}^{(k)}$. Moreover, by convention, we associate the first subspace ($k = 1$) with the space of $\mathbb{G}$-invariant functions, i.e., $\mathbb{G}^{(1)} = \mathbb{G}^{\text{inv}} = \{e\}$ (see Example J.4 in the Appendix).

Equivariant conditional expectation operator The $\mathbb{G}$-invariance of the PMD kernel (8), implies that $\mathsf{E}_{\mathbf{y}|\mathbf{x}}$ is a $\mathbb{G}$-equivariant linear operator (see Def. K.1). This means that $\mathsf{E}_{\mathbf{y}|\mathbf{x}}$ commutes with the group action on the function spaces, and consequently, can be decomposed (disentangled) into a direct sum of operators acting on the corresponding isotypic subspaces (see details in App. K):

$$g \triangleright_{\mathcal{L}^2_{\mathbf{x}}} [\mathsf{E}_{\mathbf{y}|\mathbf{x}} h](\cdot) = \mathsf{E}_{\mathbf{y}|\mathbf{x}}[g \triangleright_{\mathcal{L}^2_{\mathbf{y}}} h](\cdot) \quad \iff \quad [\mathsf{E}_{\mathbf{y}|\mathbf{x}} h](\cdot) = \sum_{k=1}^{n_{\text{iso}}} [\mathsf{E}^{(k)}_{\mathbf{y}|\mathbf{x}} h^{(k)}](\cdot) \quad \forall\, h \in \mathcal{L}^2_{\mathbf{y}}, g \in \mathbb{G}, \tag{11}$$

where each $\mathsf{E}^{(k)}_{\mathbf{y}|\mathbf{x}} : \mathcal{L}^{2(k)}_{\mathbf{y}} \to \mathcal{L}^{2(k)}_{\mathbf{x}}$ models the conditional expectation for $\mathbb{G}^{(k)}$-equivariant functions.

Equivariant disentangled representation learning The $\mathbb{G}$-equivariant structure of $\mathsf{E}_{\mathbf{y}|\mathbf{x}}$ and its disentanglement (11) into isotypic components suggests that computing the conditional expectation of a $\mathbb{G}$-equivariant function is equivalent to summing the conditional expectations of its $\mathbb{G}^{(k)}$-equivariant components for all $k \in [n_{\text{iso}}]$. Therefore, the loss function of problem (3), where $\mathsf{E}_{\mathbf{y}|\mathbf{x}}$ is approximated in finite dimensional spaces $\mathcal{F}^{\boldsymbol{\theta}}_{\mathbf{x}}$ and $\mathcal{F}^{\boldsymbol{\theta}}_{\mathbf{y}}$, decouples into n_{iso} independent (disentangled) components:

$$\begin{aligned} &\arg\min_{\boldsymbol{\theta}} \ \|\mathsf{E}_{\mathbf{y}|\mathbf{x}} - \mathsf{E}_{\boldsymbol{\theta}}\|^2_{\text{HS}} = \textstyle\sum_{k=1}^{n_{\text{iso}}} \|\mathsf{E}^{(k)}_{\mathbf{y}|\mathbf{x}} - \mathsf{E}^{(k)}_{\boldsymbol{\theta}}\|^2_{\text{HS}} = \mathbb{E}_{\mathbf{x}}\mathbb{E}_{\mathbf{y}} \textstyle\sum_{k=1}^{n_{\text{iso}}} (\kappa^{(k)}(\mathbf{x},\mathbf{y}) - \kappa^{(k)}_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{y}))^2, \\ &\quad \text{s.t. } \mathbb{E}_{\mathbf{x}}\mathbb{E}_{\mathbf{y}} \kappa_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{y}) = 1, \qquad \text{and} \ \ \kappa_{\boldsymbol{\theta}}(g \triangleright_{\mathcal{X}} \boldsymbol{x}, g \triangleright_{\mathcal{Y}} \boldsymbol{y}) = \kappa_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{y}), \quad \forall g \in \mathbb{G}, (\boldsymbol{x},\boldsymbol{y}) \in \mathcal{X} \times \mathcal{Y}. \end{aligned} \tag{12}$$

$\boldsymbol{E}_{\boldsymbol{\theta}} = \bigoplus_{k=1}^{n_{\text{iso}}} \boldsymbol{E}^{(k)}_{\boldsymbol{\theta}}$, with each block an $r_k \times r_k$ $\mathbb{G}^{(k)}$-equivariant matrix [3]. The corresponding approximated PMD kernel is given by:

$$\kappa_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{y}) = \mathbb{1}_{P_{\mathbf{x}}}(\boldsymbol{x}) \mathbb{1}_{P_{\mathbf{y}}}(\boldsymbol{y}) + \textstyle\sum_{k=1}^{n_{\text{iso}}} \kappa^{(k)}_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{y}), \qquad \kappa^{(k)}_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{y}) := \boldsymbol{\phi}^{(k)}_{\boldsymbol{\theta}}(\boldsymbol{x})^{\top} \boldsymbol{E}^{(k)}_{\boldsymbol{\theta}} \boldsymbol{\psi}^{(k)}_{\boldsymbol{\theta}}(\boldsymbol{y}), \tag{13}$$

[3]We chose to use square matrices for notational convenience, however the dimensions can vary

Task	$\boldsymbol{f}(\boldsymbol{x}) := \mathbb{E}_{\mathbf{y}}[\mathbf{y}\vert\mathbf{x}{=}\boldsymbol{x}] \approx \hat{\boldsymbol{f}}_{\boldsymbol{\theta}}(\boldsymbol{x})$	$\mathbb{P}[\mathbf{y}{\in}\mathbb{B}\vert\mathbf{x} \in \mathbb{A}] \approx \widehat{\mathbb{P}}_{\boldsymbol{\theta}}[\mathbf{y}{\in}\mathbb{B}\vert\mathbf{x}{\in}\mathbb{A}]$
Estimate	$\widehat{\mathbb{E}}_{\mathbf{y}}[\mathbf{y}]{+}\boldsymbol{\phi}_{\boldsymbol{\theta}}(\boldsymbol{x})^{\top}\boldsymbol{E}_{\theta}\widehat{\mathbb{E}}_{\mathbf{y}}[\boldsymbol{\psi}_{\boldsymbol{\theta}}(\mathbf{y}) \otimes \mathbf{y}]$	$\widehat{\mathbb{E}}_{\mathbf{y}}[\mathbb{1}_{\mathbb{B}}]{+}\frac{\widehat{\mathbb{E}}_{\mathbf{x}}[\mathbb{1}_{\mathbb{A}}(\mathbf{x})\otimes\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{x})]^{\top}\boldsymbol{E}_{\boldsymbol{\theta}}\widehat{\mathbb{E}}_{\mathbf{y}}[\mathbb{1}_{\mathbb{B}}(\mathbf{y})\otimes\boldsymbol{\psi}_{\boldsymbol{\theta}}(\mathbf{y})]}{\widehat{\mathbb{E}}_{\mathbf{x}}[\mathbb{1}_{\mathbb{A}}(\mathbf{x})]}$
Learning Guarantees	$\Vert\boldsymbol{f}{-}\hat{\boldsymbol{f}}_{\boldsymbol{\theta}}\Vert_{\mathcal{L}^2_{\mathbf{x}}}{\lesssim}\sqrt{\mathrm{Var}[\Vert\mathbf{y}\Vert]}\left(\mathcal{E}^r_{\boldsymbol{\theta}}{+}\frac{\ln(n_{\text{iso}}/\delta)}{(d_{\text{iso}}N)^{\frac{\alpha}{1+2\alpha}}}\right)$	$\vert\mathbb{P}{-}\widehat{\mathbb{P}}_{\boldsymbol{\theta}}\vert{\lesssim}\sqrt{\frac{\mathbb{P}[\mathbf{y}{\in}\mathbb{B}]}{\mathbb{P}[\mathbf{x}{\in}\mathbb{G}\triangleright_{\mathcal{X}}\mathbb{A}]}}\left(\mathcal{E}^r_{\boldsymbol{\theta}}{+}\frac{\ln(n_{\text{iso}}/\delta)}{(d_{\text{iso}}N)^{\frac{\alpha}{1+2\alpha}}}\right)$

Table 1: Statistical guarantees for eNCP. The error bounds are shaped by (i) the structure of the symmetry group $\mathbb{G}$—the number of isotypic subspaces n_{iso} and their minimum singular space dimension $d_{\text{iso}} = \sum_{k=1}^{n_{\text{iso}}} d_k$ (see Fig. 2), which enlarge the effective sample size—, (ii) the quality of the learned representations $\mathcal{E}^r_{\boldsymbol{\theta}} = \Vert\mathsf{E}_{\mathsf{y}\vert\mathsf{x}} - \mathsf{E}_{\boldsymbol{\theta}}\Vert_{\text{op}} \leq \sqrt{\mathcal{L}_\gamma(\boldsymbol{\theta}) - \mathcal{L}_\gamma(\star)}$, and (iii) the operator's singular-value decay rate $\alpha > 0$. Note that $\mathbb{G} \triangleright_{\mathcal{X}} \mathbb{A} := \cup_{g\in\mathbb{G}}\ g \triangleright_{\mathcal{X}} \mathbb{A}$ denotes the group orbit of $\mathbb{A}$.

where $\mathbb{1}_{P_{\mathbf{x}}}(\boldsymbol{x})\mathbb{1}_{P_{\mathbf{y}}}(\boldsymbol{y})$ arises since the first singular functions of $\mathsf{E}_{\mathbf{y}\vert\mathbf{x}}$ are constant, see (4).

This parameterization inherently preserves the symmetry constraints of each operator's singular functions, which we leverage in both theory and practice (see Apps. E and K.2.1 for details).

Disentangled training loss Having introduced the equivariance constraints on the truncated operator matrix, to solve (12) we follow the NCP approach and rewrite it using the contrastive loss (5), which, reflecting the operator's isotypic decomposition in (11) becomes separable:

$$\mathcal{L}_\gamma(\boldsymbol{\theta}) := \sum_{k=1}^{n_{\text{iso}}}\left(-2\mathbb{E}_{\mathbf{xy}}\kappa^{(k)}_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{y}) + \mathbb{E}_{\mathbf{x}}\mathbb{E}_{\mathbf{y}}\kappa^{(k)}_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{y})^2 + \gamma\Omega^{(k)}(\boldsymbol{\theta})\right) + 2\gamma\left(\Vert\mathbb{E}_{\mathbf{x}}\boldsymbol{\phi}^{\text{inv}}_{\boldsymbol{\theta}}(\mathbf{x})\Vert^2_F + \Vert\mathbb{E}_{\mathbf{y}}\boldsymbol{\psi}^{\text{inv}}_{\boldsymbol{\theta}}(\mathbf{y})\Vert^2_F\right). \tag{14}$$

This decomposes the problem of learning $\mathbb{G}$-equivariant representations of $\mathbf{x}$ and $\mathbf{y}$ into learning n_{iso} less constrained $\mathbb{G}^{(k)}$-equivariant representations transforming according to distinct subgroups of $\mathbb{G}$. Such representations are known in the literature as *disentangled* representations [18] (see Def. I.9).

Moreover, we improve the estimates of the regularization terms in (5) by leveraging our symmetry priors to: (i) tighten the centering regularization (14) given that functions in $\mathcal{F}^{(k)}_{\mathbf{x}}$ and $\mathcal{F}^{(k)}_{\mathbf{y}}$ are centered by construction for $k \neq \text{inv}$ (see Cor. L.4)—and (ii) exploit the orthogonality between isotypic subspaces (10) to independently regularize orthonormality for each isotypic subspace (see example in Fig. 10 in the appendix), leading to better covariance estimates [36]:

$$\Omega^{(k)}(\boldsymbol{\theta}) := \sum_{k=1}^{n_{\text{iso}}}\Vert\mathrm{Cov}(\boldsymbol{\phi}^{(k)}) - \boldsymbol{I}_{r_k}\Vert^2_F + \Vert\mathrm{Cov}(\boldsymbol{\psi}^{(k)}) - \boldsymbol{I}_{r_k}\Vert^2_F. \tag{15}$$

Given a batch $\{(\boldsymbol{x}_n, \boldsymbol{y}_n)\}_{n=1}^N$ and their corresponding embeddings $\{(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\boldsymbol{x}_n), \boldsymbol{\psi}_{\boldsymbol{\theta}}(\boldsymbol{y}_n))\}_{n=1}^N$, the empirical unregularized loss is estimated via U-statistics, yielding an unbiased estimate with an effective sample size of N^2 [32, 34].

$$\widehat{\mathcal{L}}_0(\boldsymbol{\theta}) = \sum_{k\in[n_{\text{iso}}]}\left[\tfrac{1}{N}\sum_{n\in[N]}\kappa^{(k)}_{\boldsymbol{\theta}}(\boldsymbol{x}_n,\boldsymbol{y}_n) + \tfrac{1}{N(N-1)}\sum_{a\in[N]}\sum_{b\in[N]\setminus\{a\}}\kappa^{(k)}_{\boldsymbol{\theta}}(\boldsymbol{x}_a,\boldsymbol{y}_b)^2\right]. \tag{16}$$

Similarly, we use U-statistics to obtain unbiased estimates for orthonormal regularization in (15), achieving an effective sample size of $d_k N^2$ per isotypic subspace (see App. F.2). Consequently, standard NN optimization methods can be employed to learn equivariant representations via the approximate model of $\mathsf{E}_{\mathbf{y}\vert\mathbf{x}}$, enabling downstream inference tasks described in the next section.

5 Inference and learning guarantees

Once training is complete, the learned $\mathbb{G}$-invariant PMD from (13) can be used, via (2), for $\mathbb{G}$-equivariant regression and $\mathbb{G}$-invariant conditional probability estimation. In summary, these estimates are obtained using a NN architecture composed of $\boldsymbol{\phi}_{\boldsymbol{\theta}}$, $\boldsymbol{E}_{\boldsymbol{\theta}}$, and a final linear layer that delivers the basis expansion coefficients of the target variable in the $\mathbf{y}$ representation space $\mathcal{F}^{\boldsymbol{\theta}}_{\mathbf{y}} = \mathrm{span}(\boldsymbol{\psi}_{\boldsymbol{\theta}})$.

Crucially, these parametric estimators come with tight statistical guarantees—summarized in Tab. 1 and Thm. C.1 and derived in Apps. C and M. These guarantees show that the contrastive objectives (5) and (14) serve as faithful surrogates for the standard Mean Squared Error (MSE) regression objective. (i) The bounds are governed by a regularity parameter α satisfying $\sum_{i\in\mathbb{N}}\sigma_i^{1/\alpha} < \infty$ (with $\alpha = \infty$ for finite-rank operators and $\alpha = 0$ for merely compact ones). In particular, the operator is trace class when $\alpha = 1$ and Hilbert-Schmidt when $\alpha = \frac{1}{2}$—the latter equivalent to $\kappa \in \mathcal{L}^2_{P_{\mathbf{x}} \times P_{\mathbf{y}}}(\mathcal{X} \times \mathcal{Y})$ (see App. M). Accordingly, the learning rates range from arbitrarily slow as $\alpha \to 0$ to the fast rate $(d_{\text{iso}} N)^{-1/2}$ as $\alpha \to \infty$; (ii) equivariant disentangled representations boost the effective sample size to $d_{\text{iso}} N \geq n_{\text{iso}} N \gg N$, providing not only the expected n_{iso} gain from disentanglement but also an additional $d_{\text{iso}} = \sum_{k=1}^{n_{\text{iso}}} d_k$ boost (see Fig. 2-right) by fully exploiting the equivariant structure within each isotypic–singular space; (iii) for applications requiring pointwise control, (20) provides a set-wise learning bound that quantifies how symmetries mitigate bottlenecks in estimating observables tied to rare events—here the effective rarity of $\mathbf{x} \in \mathbb{A}$ is captured by $\gamma_{\mathbb{G}'}(\mathbb{A})$, yielding gains up to $|\mathbb{G}|, \mathbb{P}[\mathbf{x} \in \mathbb{A}] \gg \mathbb{P}[\mathbf{x} \in \mathbb{A}]$ when $\mathbb{A}$ is asymmetric; and (iv) in the absence of symmetry priors—i.e., when $\mathbb{G} = e$ and $|\mathbb{G}| = n_{\text{iso}} = d_{\text{iso}} = 1$—our framework recovers the baseline results of [37], whereas leveraging symmetries amplifies the effective sample size and fundamentally alleviates the intrinsic bottlenecks of rare-event estimation.

6 Experiments

We present three experiments evaluating our method in (i) approximating the conditional expectation operator and the use of the learned operator for (ii) $\mathbb{G}$-equivariant regression and (iii) symmetry-aware uncertainty quantification. For additional empirical evidence, and specific details refer to App. G.

Conditional expectation operator learning This experiment *directly* quantifies the MSE of approximating $\mathsf{E}_{\mathbf{y}|\mathbf{x}}$, i.e., $\kappa_{\text{mse}} := \mathbb{E}_{\mathbf{x}} \mathbb{E}_{\mathbf{y}} (\kappa(\mathbf{x}, \mathbf{y}) - \kappa_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}))^2$. To achieve this, we extend the Conditional Gaussian Mixture Model (cGMM) of Gilardi et al. [38] to parametrically construct symmetric vector-valued random variables $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{Y}$ that satisfy the symmetry priors in (6) for arbitrary finite symmetry groups (see a 2D example in Fig. 3). This provides an analytical form of the PMD ratio κ, enabling direct estimation of κ_{mse}, usually impossible for real-world datasets.

The results in Fig. 5 compare our model eNCP against its symmetry-agnostic counterpart NCP and two baselines—a standard Multi-Layer Perceptron (MLP) and an Equivariant MLP (eMLP)—all with equivalent architectural footprint. Where NCP and eNCP are trained using (5) and (14), respectively, while MLP and eMLP are trained using standard MSE.

The results in Fig. 5 demonstrate that our eNCP model outperforms all other baselines in both performance and sample complexity. Consistent with [37], the NCP model shows poorer sample complexity than MLP and eMLP due to its indirect approach to regression, via approximation of $\mathsf{E}_{\mathbf{y}|\mathbf{x}}$. However, by incorporating symmetry priors our eNCP model appears to mitigate this limitation.

$\mathbb{G}$-Equivariant regression To test our model's potential for performing $\mathbb{G}$-equivariant regression, we address the robot's Center of Mass (CoM) momenta regression task of [15]. The goal is to predict a quadruped robot's CoM linear $\boldsymbol{l} \in \mathbb{R}^3$ and angular momenta $\boldsymbol{k} \in \mathbb{R}^3$ given the noisy observations of the robot's generalized positions $\boldsymbol{q} \in \mathbb{R}^{12}$ and velocity coordinates $\dot{\boldsymbol{q}} \in \mathbb{R}^{12}$, i.e., $[\boldsymbol{l}^\top, \boldsymbol{k}^\top]^\top = h_{\text{CoM}}(\boldsymbol{q} + \epsilon_{\boldsymbol{q}}, \dot{\boldsymbol{q}} + \epsilon_{\dot{\boldsymbol{q}}})$ (see details in App. G.2 and Fig. 7 in the appendix). We compare eNCP against NCP and two baselines—a standard MLP and an eMLP—all with equivalent architectural footprint. Where NCP and eNCP are trained using (5) and (14), respectively, while MLP and eMLP are trained using standard MSE.

The results in Fig. 1 demonstrate that our eNCP model outperforms all other baselines in both performance and sample complexity. Consistent with [37], the NCP model shows poorer sample complexity than MLP and eMLP due to its indirect approach to regression, via approximation of $\mathsf{E}_{\mathbf{y}|\mathbf{x}}$. However, by incorporating symmetry priors our eNCP model appears to mitigate this limitation.

Symmetry aware uncertainty quantification Finally, we demonstrate the practical impact of our approach on a core robotics problem: providing robust uncertainty quantification for unavailable yet crucial state observables for robot control and state estimation [39, 40]. Specifically, we use proprioceptive sensor readings to provide 90% confidence intervals for the robot's Ground Reaction Forces (GRF) $\boldsymbol{\tau}_{\text{grf}} \in \mathbb{R}^{12}$, the instantaneous work exerted or subtracted to the robot $U(\boldsymbol{q}, \dot{\boldsymbol{q}}, \boldsymbol{\tau}) \in \mathbb{R}$, and the kinetic energy $T(\boldsymbol{q}, \dot{\boldsymbol{q}}) \in \mathbb{R}$, while the robot traverses rough terrain (see App. G.3.2). Reliable

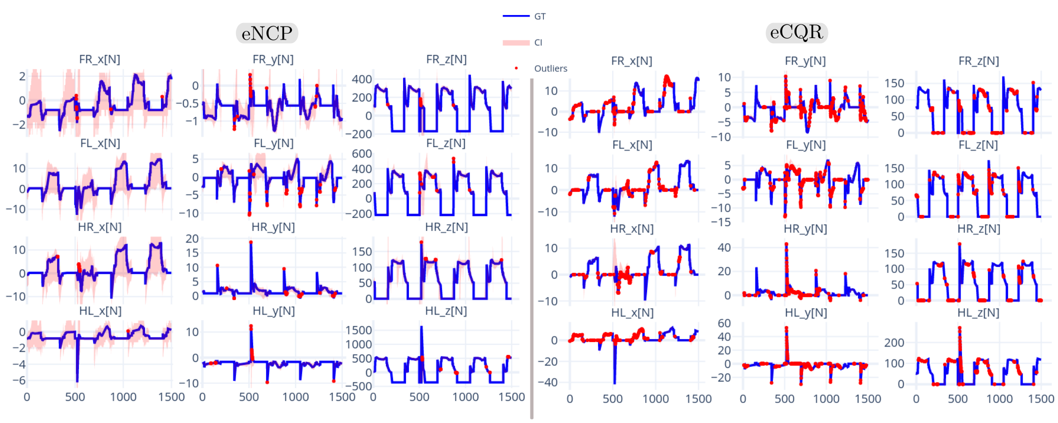

Figure 4: Prediction of 90% confidence intervals (CI) for the ground-reaction forces $\boldsymbol{\tau}_{\text{grf}} \in \mathbb{R}^{12}$ of a quadruped robot on rough terrain with varying friction. We compare the eNCP vs. eCQR (see NCP and CQR in Fig. 4) models based on relaxed coverage and set size (see Tab. 4). CIs are computed for each leg—front-right (FR), front-left (FL), hind-right (HR), and hind-left—along the x, y, and z axes. Forces outside the CI are highlighted in red, while those within appear in blue.

probabilistic estimates of these quantities are of crucial relevance for optimal control [39], contact detection [40], state estimation [41], and system identification [42].

	r-Coverage ↑	Coverage ↑
eNCP	99.5±0.1%	95.0±0.4%
NCP	99.5±0.0%	56.9±0.3%
eCQR	84.2±0.7%	6.7±1.2%
CQR	80.5±3.7%	8.5±0.9%

Table 2: Relaxed coverage, see (31), and Coverage, see (30), for the test-set confidence intervals in quadruped locomotion uncertainty estimation of $\mathbf{y} = [\boldsymbol{\tau}_{\text{grf}}^{\top}, U, T]^{\top}$. Target coverage is: 90%.

This task tests our model's ability to learn conditional distributions from high-dimensional data, considering that for the eNCP and NCP models, quantile estimation is done by regressing the Conditional Cumulative Distribution Function (CCDF) for each dimension of $\mathbf{y} = [\mathrm{y}_1, \dots]$ and then applying a linear search to extract quantiles (see Fig. 9 in the appendix). This is achieved by discretizing the range of each y_i into N_b bins and estimating $\mathbb{P}(\mathrm{y}_i \in \mathbb{A}_{i,n} | \mathbf{x} = \cdot) := [\mathsf{E}_{\mathrm{y}|\mathrm{x}} \mathbb{1}_{\mathbb{A}_{i,n}}](\cdot)$ for all $n \in [N_b]$ (see Sec. 5), where $\mathbb{A}_{i,n}$ consists of the first n bins. In practice, this means regressing $|\mathcal{Y}| \times N_b$ conditional probabilities corresponding to sets of varying sizes in a single forward pass. By contrast, the baseline CQR [43] and its equivariant adaptation eCQR directly regress quantiles for a fixed coverage level (i.e., the probability that an event lies within the predicted confidence interval) and need retraining for different coverage values.

The results in Tab. 2, Fig. 1 (for U and T) and in Fig. 12 in the appendix (for $\boldsymbol{\tau}_{\text{grf}}$) show eNCP as the only model capable of providing robust uncertainty quantification, as it is the only model with an empirical coverage on the test set close to the desired value, rendering other models unreliable for practical applications. This underscores eNCP's potential for conditional probability estimation.

7 Conclusions

We introduce a novel framework for equivariant representation learning that enables estimation of equivariant regression and conditional probabilities with statistical learning guarantees. Our approach builds on a recent contrastive representation learning method that approximates the spectral decomposition of the conditional expectation operator. By incorporating symmetry priors, we impose additional structural constraints that further decompose the conditional expectation operator and enhance the effective sample size. We demonstrate the benefits of our approach through both theoretical learning bounds and empirical experiments. Notably, we provide the first theoretical learning guarantees for equivariant regression using neural network features, thereby bridging spectral representation learning and geometric deep learning.

References

[1] R. Izbicki. *Machine Learning Beyond Point Predictions: Uncertainty Quantification*. imprint, 1st edition, 2025. ISBN 978-65-01-20272-3.

[2] R. C. Smith. *Uncertainty Quantification: Theory, Implementation, and Applications*. Society for Industrial and Applied Mathematics, Jan. 2013. URL http://dx.doi.org/10.1137/1.9781611973228.

[3] L. Wasserman. *All of Nonparametric Statistics*. Springer Texts in Statistics. Springer, New York, NY, 1 edition, May 2007.

[4] D. W. Scott. Feasibility of multivariate density estimates. *Biometrika*, 78(1):197–205, 1991. URL http://dx.doi.org/10.1093/biomet/78.1.197.

[5] T. Nagler and C. Czado. Evading the curse of dimensionality in nonparametric density estimation with simplified vine copulas. *Journal of Multivariate Analysis*, 151:69–89, Oct. 2016. URL http://dx.doi.org/10.1016/j.jmva.2016.07.003.

[6] R. Izbicki and A. B. Lee. Converting high-dimensional regression to high-dimensional conditional density estimation. 2017.

[7] K. Kashinath, M. Mustafa, A. Albert, J. Wu, C. Jiang, S. Esmaeilzadeh, K. Azizzadenesheli, R. Wang, A. Chattopadhyay, A. Singh, et al. Physics-informed machine learning: case studies for weather and climate modelling. *Philosophical Transactions of the Royal Society A*, 379 (2194):20200093, 2021.

[8] R. Wang. Incorporating symmetry into deep dynamics models for improved generalization. In *International Conference on Learning Representations (ICLR)*, 2021.

[9] M. M. Bronstein, J. Bruna, T. Cohen, and P. Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

[10] B. Elesedy. *Symmetry and generalisation in machine learning*. PhD thesis, University of Oxford, 2023.

[11] M. S. Dresselhaus, G. Dresselhaus, and A. Jorio. *Group theory: application to the physics of condensed matter*. Springer Science & Business Media, 2007.

[12] N. J. Mitra, M. Pauly, M. Wand, and D. Ceylan. Symmetry in 3d geometry: Extraction and applications. *Computer Graphics Forum*, 32(6):1–23, Feb. 2013. URL http://dx.doi.org/10.1111/cgf.12010.

[13] L. Kovar, M. Gleicher, and F. Pighin. Motion graphs. *ACM Trans. Graph.*, 21(3):473–482, July 2002. URL https://doi.org/10.1145/566654.566605.

[14] P. J. Olver. *Applications of Lie groups to differential equations*, volume 107. Springer Science & Business Media, 1993.

[15] D. Ordoñez-Apraez, G. Turrisi, V. Kostic, M. Martin, A. Agudo, F. Moreno-Noguer, M. Pontil, C. Semini, and C. Mastalli. Morphological symmetries in robotics. *The International Journal of Robotics Research*, 0(0):02783649241282422, 0. doi:10.1177/02783649241282422. URL https://doi.org/10.1177/02783649241282422.

[16] X. Zhu, D. Wang, O. Biza, G. Su, R. Walters, and R. Platt. Sample Efficient Grasp Learning Using Equivariant Models. In *Proceedings of Robotics: Science and Systems*, New York City, NY, USA, June 2022.

[17] M. Weiler, P. Forré, E. Verlinde, and M. Welling. *Equivariant and Coordinate Independent Convolutional Networks*. World Scientific, 2023. URL https://maurice-weiler.gitlab.io/cnn_book/EquivariantAndCoordinateIndependentCNNs.pdf.

[18] I. Higgins, D. Amos, D. Pfau, S. Racaniere, L. Matthey, D. Rezende, and A. Lerchner. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018.

[19] E. van der Pol, D. Worrall, H. van Hoof, F. Oliehoek, and M. Welling. Mdp homomorphic networks: Group symmetries in reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 4199–4210. Curran Associates, Inc., 2020.

[20] R. Dangovski, L. Jing, C. Loh, S. Han, A. Srivastava, B. Cheung, P. Agrawal, and M. Soljacic. Equivariant self-supervised learning: Encouraging equivariance in representations. In *International Conference on Learning Representations*.

[21] H. Keurti, H.-R. Pan, M. Besserve, B. F. Grewe, and B. Schölkopf. Homomorphism autoencoder–learning group structured representations from observed transitions. In *International Conference on Machine Learning*, pages 16190–16215. PMLR, 2023.

[22] X. Wang, H. Chen, S. Tang, Z. Wu, and W. Zhu. Disentangled representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[23] J. Y. Zou, D. J. Hsu, D. C. Parkes, and R. P. Adams. Contrastive learning using spectral methods. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[24] Y.-H. H. Tsai, M. Q. Ma, M. Yang, H. Zhao, L.-P. Morency, and R. Salakhutdinov. Self-supervised representation learning with relative predictive coding. In *International Conference on Learning Representations*, 2021.

[25] V. R. Kostic, P. Novelli, R. Grazzi, K. Lounici, and M. Pontil. Learning invariant representations of time-homogeneous stochastic dynamical systems. In *The Twelfth International Conference on Learning Representations*, 2024.

[26] C. R. Baker. Joint measures and cross-covariance operators. *Trans. Am. Math. Soc.*, 186: 273–273, 1973.

[27] L. Song, J. Huang, A. Smola, and K. Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 961–968, 2009.

[28] K. Fukumizu, F. R. Bach, and M. I. Jordan. Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. *Journal of Machine Learning Research*, 5(Jan):73–99, 2004.

[29] M. Sugiyama, T. Suzuki, and T. Kanamori. *Density ratio estimation in machine learning*. Cambridge University Press, 2012.

[30] A. v. d. Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[31] C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

[32] Z. Wang, Y. Luo, Y. Li, J. Zhu, and B. Schölkopf. Spectral representation learning for conditional moment models. *arXiv preprint arXiv:2210.16525*, 2022.

[33] J. Z. HaoChen, C. Wei, A. Gaidon, and T. Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. In *Advances in Neural Information Processing Systems*, volume 34, pages 5000–5011. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/27debb435021eb68b3965290b5e24c49-Paper.pdf.

[34] Y.-H. H. Tsai, H. Zhao, M. Yamada, L.-P. Morency, and R. R. Salakhutdinov. Neural methods for point-wise dependency estimation. In *Advances in Neural Information Processing Systems*, volume 33, pages 62–72. Curran Associates, Inc., 2020.

[35] G. W. Mackey. Harmonic analysis as the exploitation of symmetry–a historical survey. *Bulletin (New Series) of the American Mathematical Society*, 3(1.P1):543 – 698, 1980.

[36] P. Shah and V. Chandrasekaran. Group symmetry and covariance regularization. In *2012 46th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6. IEEE, 2012.

[37] V. R. Kostic, G. Pacreau, G. Turri, P. Novelli, K. Lounici, and M. Pontil. Neural conditional probability for uncertainty quantification. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[38] N. Gilardi, S. Bengio, and M. Kanevski. Conditional gaussian mixture models for environmental risk mapping. In *Proceedings of the 12th IEEE workshop on neural networks for signal processing*, pages 777–786. IEEE, 2002.

[39] G. Bledt, P. M. Wensing, S. Ingersoll, and S. Kim. Contact model fusion for event-based locomotion in unstructured terrains. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4399–4406. IEEE, 2018.

[40] M. Maravgakis, D.-E. Argiropoulos, S. Piperakis, and P. Trahanias. Probabilistic contact state estimation for legged robots using inertial information. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 12163–12169. IEEE, 2023.

[41] Y. Nisticò, J. C. V. Soares, L. Amatucci, G. Fink, and C. Semini. Muse: A real-time multi-sensor state estimator for quadruped robots. *IEEE Robotics and Automation Letters*, 2025.

[42] M. Gautier. Dynamic identification of robots with power model. In *Proceedings of international conference on robotics and automation*, volume 3, pages 1922–1927. IEEE, 1997.

[43] S. Feldman, S. Bates, and Y. Romano. Calibrated multiple-output quantile regression with representation learning. *Journal of Machine Learning Research*, 24(24):1–48, 2023.

[44] P. H. Le-Khac, G. Healy, and A. F. Smeaton. Contrastive representation learning: A framework and review. *Ieee Access*, 8:193907–193934, 2020.

[45] H. Waida, Y. Wada, L. Andéol, T. Nakagawa, Y. Zhang, and T. Kanamori. Towards understanding the mechanism of contrastive learning via similarity structure: A theoretical analysis. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 709–727. Springer, 2023.

[46] H. Bao, Y. Nagano, and K. Nozawa. On the surrogate gap between contrastive and supervised losses. In *International conference on machine learning*, pages 1585–1606. PMLR, 2022.

[47] D. D. Johnson, A. E. Hanchi, and C. J. Maddison. Contrastive learning can find an optimal basis for approximately view-invariant functions. *arXiv preprint arXiv:2210.01883*, 2022.

[48] E. Cole, X. Yang, K. Wilber, O. Mac Aodha, and S. Belongie. When does contrastive visual representation learning work? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14755–14764, June 2022.

[49] C. Tosh, A. Krishnamurthy, and D. Hsu. Contrastive learning, multi-view redundancy, and linear models. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pages 1179–1206. PMLR, 16–19 Mar 2021.

[50] L. Lin, J. Zhang, and J. Liu. Mutual information driven equivariant contrastive learning for 3d action representation learning. *IEEE Transactions on Image Processing*, 2024.

[51] O. Henaff. Data-efficient image recognition with contrastive predictive coding. In *International conference on machine learning*, pages 4182–4192. PMLR, 2020.

[52] S. Ozair, C. Lynch, Y. Bengio, A. van den Oord, S. Levine, and P. Sermanet. Wasserstein dependency measure for representation learning. 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/f9209b7866c9f69823201c1732cc8645-Paper.pdf.

[53] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[54] J. Z. HaoChen, C. Wei, A. Kumar, and T. Ma. Beyond separability: Analyzing the linear transferability of contrastive representations to related subpopulations. *Advances in neural information processing systems*, 35:26889–26902, 2022.

[55] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020.

[56] T. Yerxa, J. Feather, E. Simoncelli, and S. Chung. Contrastive-equivariant self-supervised learning improves alignment with primate visual area it. *Advances in neural information processing systems*, 37:96045–96070, 2024.

[57] R. Dangovski, L. Jing, C. Loh, S. Han, A. Srivastava, B. Cheung, P. Agrawal, and M. Soljačić. Equivariant contrastive learning. In *International Conference on Learning Representations*, 2022.

[58] Y. Wang, K. Hu, S. Gupta, Z. Ye, Y. Wang, and S. Jegelka. Understanding the role of equivariance in self-supervised learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[59] G. L. Marchetti, G. Tegnér, A. Varava, and D. Kragic. Equivariant representation learning via class-pose decomposition. In *International Conference on Artificial Intelligence and Statistics*, pages 4745–4756. PMLR, 2023.

[60] S. Gupta, J. Robinson, D. Lim, S. Villar, and S. Jegelka. Structuring representation geometry with rotationally equivariant contrastive learning. In *The Twelfth International Conference on Learning Representations*, 2023.

[61] M. Golubitsky, I. Stewart, and D. G. Schaeffer. *Singularities and Groups in Bifurcation Theory: Volume II*, volume 69. Springer Science & Business Media, 2012.

[62] É. Cartan. *La théorie des groupes finis et continus et l'analysis situs*. Number 42 in Mémorial des sciences mathématiques. Gauthier-Villars, 1952.

[63] J. Yang, N. Dehmamy, R. Walters, and R. Yu. Latent space symmetry discovery. In *International Conference on Machine Learning*, 2023.

[64] Y. Mroueh, S. Voinea, and T. A. Poggio. Learning with group invariant features: A kernel perspective. *Advances in neural information processing systems*, 28, 2015.

[65] B. Tahmasebi and S. Jegelka. The exact sample complexity gain from invariances for kernel regression. *Advances in Neural Information Processing Systems*, 36, 2023.

[66] B. Elesedy. Provably strict generalisation benefit for invariance in kernel methods. In *Advances in Neural Information Processing Systems*, volume 34, pages 17273–17283. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/8fe04df45a22b63156ebabbb064fcd5e-Paper.pdf.

[67] B. Elesedy and S. Zaidi. Provably strict generalisation benefit for equivariant models. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2959–2969. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/elesedy21a.html.

[68] D. Pal, A. Kannan, G. Arakalgud, and M. Savvides. Max-margin invariant features from transformed unlabelled data. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[69] S. Mei, T. Misiakiewicz, and A. Montanari. Learning with invariances in random features and kernel models. In *Conference on Learning Theory*, pages 3351–3418. PMLR, 2021.

[70] A. Bietti, L. Venturi, and J. Bruna. On the sample complexity of learning under geometric stability. *Advances in neural information processing systems*, 34:18673–18684, 2021.

[71] K. Donhauser, M. Wu, and F. Yang. How rotational invariance of common kernels prevents generalization in high dimensions. In *International Conference on Machine Learning*, pages 2804–2814. PMLR, 2021.

[72] R. Wang, R. Walters, and R. Yu. Incorporating symmetry into deep dynamics models for improved generalization. *arXiv preprint arXiv:2002.03061*, 2020.

[73] R. Wang, R. Walters, and R. Yu. Approximately equivariant networks for imperfectly symmetric dynamics. In *International Conference on Machine Learning*, pages 23078–23091. PMLR, 2022.

[74] J. Brandstetter, R. v. d. Berg, M. Welling, and J. K. Gupta. Clifford neural layers for pde modeling. *arXiv preprint arXiv:2209.04934*, 2022.

[75] G. Turrisi, V. Modugno, L. Amatucci, D. Kanoulas, and C. Semini. On the benefits of gpu sample-based stochastic predictive controllers for legged locomotion. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 13757–13764, 2024. doi:10.1109/IROS58592.2024.10801698.

[76] Z. Liu, S. J. Gortler, and M. F. Cohen. Hierarchical spacetime control. In *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*, pages 35–42, 1994.

[77] A. W. Knapp. *Representation Theory of Semisimple Groups, An Overview Based on Examples (PMS-36)*. Princeton University Press, Princeton, 1986.

[78] N. Kovachki, Z. Li, B. Liu, K. Azizzadenesheli, K. Bhattacharya, A. Stuart, and A. Anandkumar. Neural operator: Learning maps between function spaces with applications to pdes. *Journal of Machine Learning Research*, 24(89):1–97, 2023.

[79] B. Bercu, B. Delyon, and E. Rio. *Concentration Inequalities for Sums and Martingales*. SpringerBriefs in Mathematics. Springer, 2015.

[80] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv:1011.3027*, 2011.

Part I

Appendix

A Symbols and notation

Numbers and Arrays

x	A scalar, or scalar function $x(\cdot)$
$\boldsymbol{x}$	A vector, or vector-valued function $\boldsymbol{x}(\cdot)$
$\boldsymbol{x}_1 \oplus \boldsymbol{x}_2$	Direct sum (stacking) of vectors, such that $\boldsymbol{x}_1 \oplus \boldsymbol{x}_2 := \left[\begin{smallmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{smallmatrix}\right]$
$\boldsymbol{K}$	A matrix
$\boldsymbol{A} \oplus \boldsymbol{B}$	Direct sum of matrices, such that $\boldsymbol{A} \oplus \boldsymbol{B} := \left[\begin{smallmatrix} \boldsymbol{A} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{B} \end{smallmatrix}\right]$
K	A linear operator
$\boldsymbol{I}$	Identity matrix
$\delta_{i,j}$	The Kronecker function, equal to 1 when $i = j$, and 0 when $i \neq j$

Sets, Vector Spaces, and Function Spaces

$\mathcal{X}, \mathcal{Z}, \mathcal{H}, \mathcal{F}$	A vector or Hilbert space
$\mathbb{R}, \mathbb{C}$	The set of real and complex numbers
$\mathcal{X} \oplus \mathcal{Y}$	Direct sum of vector spaces $\mathcal{X}$ and $\mathcal{Y}$, such that if $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{y} \in \mathcal{Y}$, then $\boldsymbol{x} \oplus \boldsymbol{y} \in \mathcal{X} \oplus \mathcal{Y}$
$\mathcal{L}^2_{\mathbf{x}} := \mathcal{L}^2_{P_{\mathbf{x}}}(\mathcal{X}, \mathbb{R})$	The Hilbert space of square-integrable functions on $\mathcal{X}$ with respect to the measure $P_{\mathbf{x}}$, defined as $\mathcal{L}^2_{P_{\mathbf{x}}}(\mathcal{X}) := \{f \vert \int_{\mathcal{X}} \lvert f(\boldsymbol{x}) \rvert^2 P_{\mathbf{x}}(d\boldsymbol{x}) < \infty\}$
$\langle f, f' \rangle_{P_{\mathbf{x}}}$	Inner product between f an f' in $\mathcal{L}^2_{P_{\mathbf{x}}}\mathcal{X}$, defined as $\langle f, f' \rangle_{P_{\mathbf{x}}} := \int_{\mathcal{X}} f(\boldsymbol{x}) f'(\boldsymbol{x}) P_{\mathbf{x}}(d\boldsymbol{x})$

Group and Representation theory

$\mathbb{G}$	A symmetry group
g, g_1, g_a	A symmetry group element
$g \triangleright \boldsymbol{x}$	The (left) group action of g on $\boldsymbol{x}$ defined by $g \triangleright \boldsymbol{x} := \boldsymbol{\rho}_{\mathcal{X}}(g)\mathcal{X}$
$\boldsymbol{\rho}_{\mathcal{X}}$	A representation of the group $\mathbb{G}$ on the vector space $\mathcal{X}$, defined for a chosen basis of $\mathcal{X}$
$\bar{\boldsymbol{\rho}}_k$	An irreducible representation Def. I.7 of the group $\mathbb{G}$
$\boldsymbol{\rho}_{\mathcal{X}}(g)$	Representation of the group element g on the vector space $\mathcal{X}$, defined for a chosen basis $\mathcal{X}$
$\boldsymbol{\rho}_{\mathcal{X}} \oplus \boldsymbol{\rho}_{\mathcal{Y}}$	Direct sum of group representations, such that $\boldsymbol{\rho}_{\mathcal{X}}(g) \oplus \boldsymbol{\rho}_{\mathcal{Y}}(g) := \begin{bmatrix} \boldsymbol{\rho}_{\mathcal{X}}(g) & \\ & \boldsymbol{\rho}_{\mathcal{Y}}(g) \end{bmatrix}$
$\mathbb{G}\boldsymbol{x}$	The group orbit of $\boldsymbol{x}$, defined as $\mathbb{G}\boldsymbol{x} := \{g \triangleright \boldsymbol{x} \mid g \in \mathbb{G}\}$
$\gamma_{\mathbb{G}'}(A)$	The symmetry index of a set $A \subseteq \mathcal{X}$ w.r.t. probability distribution on $\mathcal{X}$ and group elements $\mathbb{G}' \subseteq \mathbb{G}$
$\mathbb{G}_a \times \mathbb{G}_b$	Direct product of groups $\mathbb{G}_a$ and $\mathbb{G}_b$
$\mathbb{U}(\mathcal{X})$	Unitary group on the vector space $\mathcal{X}$
$\mathbb{GL}(\mathcal{X})$	General Linear group on the vector space $\mathcal{X}$, a.k.a the space of invertible matrices in $\mathbb{R}^{\vert\mathcal{X}\vert \times \vert\mathcal{X}\vert}$
$\mathbb{C}_n$	Cyclic group of order n
$\mathbb{K}_4$	Klein four-group

Probability Theory

$\mathbf{x} \sim \mathbb{P}(\mathbf{x})$	Random vector $\mathbf{x} \in \mathcal{X}$ has distribution $\mathbb{P}(\mathbf{x})$
$P_{\mathbf{x}}$	A probability measure on the space $\mathcal{X}$
$\mathbb{E}_{\mathbf{x}}[f(\mathbf{x})]$	Expectation of $f(\mathbf{x})$ with respect to $P_{\mathbf{x}}$
$\mathrm{Cov}(f(\mathbf{x}))$	Variance of $f(\mathbf{x})$ with respect to $P_{\mathbf{x}}$, define as $\mathbb{E}_{\mathbf{x}}(f(\mathbf{x}) - \mathbb{E}_{\mathbf{x}} f(\mathbf{x}))^2$
$\mathrm{Cov}(f(\mathbf{x}), h(\mathbf{y}))$	Covariance of $f(\mathbf{x})$ and $h(\mathbf{y})$ with respect to the joint distribution $P_{\mathbf{xy}}$, defined as $\mathbb{E}_{\mathbf{xy}}(f(\mathbf{x}) - \mathbb{E}_{\mathbf{x}} f(\mathbf{x}))(h(\mathbf{y}) - \mathbb{E}_{\mathbf{y}} h(\mathbf{y}))$
$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$	Gaussian distribution over $\boldsymbol{x}$ with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$

B Related work

B.1 Contrastive representation learning

Contrastive representation learning obtains high-dimensional representations from unlabeled data by contrasting positive and negative sample pairs via a noise contrastive loss (similar to Eq. (5)) [44, 45, 46]. Most works in this field aim to learn representations in a self-supervised fashion that transfer well to downstream classification tasks [47, 48, 49, 30, 24, 33]. In contrast, our approach targets representations that effectively transfer to (equivariant) regression and uncertainty quantification, as in [25]. Given a dataset $\mathbb{D} = \{(\boldsymbol{x}_n, \boldsymbol{y}_n)\}_{n=1}^{N}$ from a target (stochastic) function $\mathbf{y} = \boldsymbol{f}(\mathbf{x})$, we treat positive pairs as drawn from the joint distribution $(\boldsymbol{x}, \boldsymbol{y}) \sim \mathbb{P}(\mathbf{x}, \mathbf{y})$ and negative pairs as drawn from the product of the marginals $(\boldsymbol{x}, \boldsymbol{y}) \sim \mathbb{P}(\mathbf{x})\mathbb{P}(\mathbf{y})$. In this setting, our contrastive loss aims to learn representations that approximate the PMD ratio $\kappa(\boldsymbol{x}, \boldsymbol{y}) = \frac{\mathbb{P}(\boldsymbol{x},\boldsymbol{y})}{\mathbb{P}(\boldsymbol{x})\mathbb{P}(\boldsymbol{y})}$, [25] or equivalently, the pointwise mutual information $\ln(\kappa(\boldsymbol{x}, \boldsymbol{y}))$ [30, 50, 51, 52]. Crucially, our work is the first study this problem when there is prior knowledge of the invariance of κ under the action of a compact symmetry group, which occurs in most applications of GDL.

Linear transferability The goal of contrastive representation learning is to acquire representations that transfer to diverse downstream inference tasks [53, 45]. While empirical studies demonstrate that contrastive learning can outperform supervised methods [48, 30, 51], theoretical works aim to establish *linear separability/transferability* guarantees [54] [4]. That is, showing that linear functionals of the (frozen) learned representations suffice for regression/classification inference.

In the context of **classification**, [45, 46, 47] show that contrastive learning losses serve as surrogates for standard supervised classification losses (e.g., the cross-entropy). Where the gap between the surrogate and supervised loss diminishes with the number of negative samples [46] (N^2 for the loss in Eq. (16)). To provide these transferability guarantees, these work assume $\mathcal{X} = \mathcal{Y}$, so that the PMD ratio κ becomes a positive definite kernel. Consequently, kernel method guarantees can be transferred to the classification task, even when the representations are parameterized by NNs [47, 46, 54].

Considerably fewer works have studied contrastive representation learning in the context of downstream **regression** tasks [56, 25]. Crucially, Kostic et al. [25] show that a contrastive learning loss serves as surrogate to the MSE regression loss (A summary of this method appears in Sec. 2 and in Tab. 3). While, to the best of our knowledge, [56] is the only work empirically studying contrastive learning for regression in the presence of symmetries.

Task	$\boldsymbol{f}(\boldsymbol{x}) := \mathbb{E}_{\mathbf{y}}[\mathbf{y} \vert \mathbf{x}{=}\boldsymbol{x}] \approx \hat{\boldsymbol{f}}_{\boldsymbol{\theta}}(\boldsymbol{x})$	$\mathbb{P}[\mathbf{y} {\in} \mathbb{B} \vert \mathbf{x} \in \mathbb{A}] \approx \widehat{\mathbb{P}}_{\boldsymbol{\theta}}[\mathbf{y} {\in} \mathbb{B} \vert \mathbf{x} {\in} \mathbb{A}]$
Estimate	$\widehat{\mathbb{E}}_{\mathbf{y}}[\mathbf{y}] + \boldsymbol{\phi}_{\boldsymbol{\theta}}(\boldsymbol{x})^{\top} \boldsymbol{E}_{\boldsymbol{\theta}} \widehat{\mathbb{E}}_{\mathbf{y}}[\boldsymbol{\psi}_{\boldsymbol{\theta}}(\mathbf{y}) \otimes \mathbf{y}]$	$\widehat{\mathbb{E}}_{\mathbf{y}}[\mathbb{1}_{\mathbb{B}}] + \frac{\widehat{\mathbb{E}}_{\mathbf{x}}[\mathbb{1}_{\mathbb{A}}(\mathbf{x}) \otimes \boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{x})]^{\top} \boldsymbol{E}_{\boldsymbol{\theta}} \widehat{\mathbb{E}}_{\mathbf{y}}[\mathbb{1}_{\mathbb{B}}(\mathbf{y}) \otimes \boldsymbol{\psi}_{\boldsymbol{\theta}}(\mathbf{y})]}{\widehat{\mathbb{E}}_{\mathbf{x}}[\mathbb{1}_{\mathbb{A}}(\mathbf{x})]}$
Guarantees	$\Vert \boldsymbol{f} - \hat{\boldsymbol{f}}_{\boldsymbol{\theta}} \Vert_{\mathcal{L}^2_{\mathbf{x}}} \lesssim \sqrt{\mathrm{Var}[\Vert \mathbf{y} \Vert]} \left(\mathcal{E}^r_{\boldsymbol{\theta}} + \frac{\ln(1/\delta)}{N^{\frac{\alpha}{1+2\alpha}}} \right)$	$\vert \mathbb{P} - \widehat{\mathbb{P}}_{\boldsymbol{\theta}} \vert \lesssim \sqrt{\frac{\mathbb{P}[\mathbf{y} \in \mathbb{B}]}{\mathbb{P}[\mathbf{x} \in \mathbb{A}]}} \left(\mathcal{E}^r_{\boldsymbol{\theta}} + \frac{\ln(1/\delta)}{N^{\frac{\alpha}{1+2\alpha}}} \right)$

Table 3: Statistical learning guarantees of NCP [25] for regression and conditional probability estimation. The bounds are shaped by the quality of the learned representations $\mathcal{E}^r_{\boldsymbol{\theta}} = \Vert \mathsf{E}_{\mathsf{y}|\mathsf{x}} - \mathsf{E}_{\boldsymbol{\theta}} \Vert_{\mathrm{op}} \leq \sqrt{\mathcal{L}_\gamma(\boldsymbol{\theta}) - \mathcal{L}_\gamma(\star)}$ (see (5)), the sample size N, and the decay rate of $\mathsf{E}_{\mathsf{y}|\mathsf{x}}$ singular-values $\alpha > 0$, which quantifies the difficulty of the problem.

B.2 Equivariant representation learning

Equivariant contrastive representation learning [57, 58] aims to learn representations that are equivariant—instead of invariant—to data transformations. For example, Marchetti et al. [59], Gupta et al. [60], Lin et al. [50] provide empirical evidence that representations of 3D scenes, images, and human body poses that are equivariant to translations, rotations, or reflections yield improved performance

[4] Also refeered to as linear evaluation protocol [55]

in *classification* tasks. Additionally, Yerxa et al. [56] show that rotation- and reflection-aware image representations enhance the *regression* of neural responses in the macaque inferior temporal cortex, while also providing theoretical justification that such equivariant representations mirror the known structure of animal visual perception. By introducing these transformations via data-augmentation of the training set, these methods inherently enforce symmetries in the data distributions, which are the fundamental priors assumed in Sec. 3.

Disentangled representations In equivariant representation learning, disentangled representations have been extensively studied [22]. Initially, [53] defined disentanglement as decomposing representations into components that capture distinct, independently varying factors. Later, using group theory, Higgins et al. [18] formalized that a representation is disentangled if its space decomposes into orthogonal subspaces reflecting a symmetry group decomposition, with each subspace influenced exclusively by one subgroup (see Def. I.9). As discussed in App. I, this aligns with the isotypic decomposition of a Hilbert space [35]: $\mathcal{H} = \bigoplus_{k=1}^{\perp} \mathcal{H}^{(k)}$—known in dynamical systems [61]—when the symmetry group decomposes as $\mathbb{G} = \prod_{k=1}^{n_{\text{iso}}} \mathbb{G}^{(k)}$. Orthogonality between subspaces follows from Schur's orthogonality relations via Cartan's and Peter-Weyl's theorems [62]. This symmetric structure is the cause of the achitectural constraints imposed in the eNCP architecture Fig. 2.

Several empirical works have explored disentanglement in representation learning. For instance, Keurti et al. [21] proposed an autoencoder-based method to learn disentangled equivariant representations by using loss regularization to enforce latent space equivariance and sparsity for separating latent group actions. Unlike our approach, their method does not assume prior knowledge of the symmetry group and relies entirely on loss regularization rather than architectural constraints. Similarly, works such as [see e.g. 63, 20] have investigated various symmetry priors in latent space by examining the emergence of disentangled structures and enforcing algebraic constraints. Notably, in fields like molecular dynamics, physics, computer graphics, and robotics, symmetry priors are intrinsic to the task or system [15, 50, 59], making them natural assumptions. In a similar spirit to our work, Marchetti et al. [59] leverage the known $\mathbb{SO}_3$ symmetries of the 3D world to learn $\mathbb{SO}_3$-disentangled equivariant representations using contrastive learning, thereby demonstrating the empirical advantages of symmetry-aware, disentangled representations for object classification.

B.3 Symmetry-aware statistical learning theory

Existing literature on symmetry-aware learning focuses on group-invariant regression via kernel methods [64, 65, 66, 67, 10, 68, 69, 70, 71]. Most of these methods cannot be directly transferred to modern GDL architectures. In contrast, in deep learning and GDL, while many works offer a group-theoretical analysis and empirical evidence of the benefits of incorporating symmetry priors [7, 72, 73, 74], none, to our knowledge, provide statistical learning guarantees that analytically quantify these benefits in terms of the structure of the compact/finite symmetry group.

C Inference and learning guarantees

Consider vector-valued regression with an observable $\boldsymbol{h} \in \mathcal{L}^2_{P_{\mathbf{y}}}(\mathcal{Y}, \mathcal{Z})$, where $\mathcal{Z}$ is a symmetry-endowed vector space. The target function $\boldsymbol{z}\colon \mathcal{X} \to \mathcal{Z}$ is the conditional expectation of of $\boldsymbol{h}$, that is: $\boldsymbol{z}(\boldsymbol{x}) := \mathbb{E}_{\mathbf{y}}[\boldsymbol{h}(\mathbf{y})|\mathbf{x} = \boldsymbol{x}] = [\mathsf{E}_{\mathsf{y}|\mathsf{x}}\boldsymbol{h}](\boldsymbol{x})$. Then, using the learned model, we estimate $\boldsymbol{z}$ by

$$[\mathsf{E}_{\mathsf{y}|\mathsf{x}}\boldsymbol{h}](\boldsymbol{x}) \approx \widehat{z}_{\boldsymbol{\theta}}(\boldsymbol{x}) := \widehat{\mathbb{E}}_{\mathbf{y}}[\boldsymbol{h}(\mathbf{y})] + \boldsymbol{\phi}_{\boldsymbol{\theta}}^{\top}(\boldsymbol{x})^{\top} \boldsymbol{E}_{\theta}\, \widehat{\mathbb{E}}_{\mathbf{y}}[\boldsymbol{\psi}_{\boldsymbol{\theta}}(\mathbf{y}) \otimes \boldsymbol{h}(\mathbf{y})],$$

where $\widehat{\mathbb{E}}_{\mathbf{y}}[\boldsymbol{\psi}_{\boldsymbol{\theta}}(\mathbf{y}) \otimes \boldsymbol{h}(\mathbf{y})]$ denote the basis expansion coefficients of $\boldsymbol{h}$ in the learned basis of $\mathcal{F}^{\boldsymbol{\theta}}_{\mathbf{y}} \subset \mathcal{L}^2_{\mathbf{y}}$. With $\widehat{\mathbb{E}}_{\mathbf{x}}\colon \mathcal{L}^2_{\mathbf{x}} \to \mathbb{R}$ and $\widehat{\mathbb{E}}_{\mathbf{y}}\colon \mathcal{L}^2_{\mathbf{y}} \to \mathbb{R}$ being the $\mathbb{G}$-invariant empirical expectations defined by:

$$\widehat{\mathbb{E}}_{\mathbf{x}}[f(\mathbf{x})] = \frac{1}{|\mathbb{G}|N} \sum_{g \in \mathbb{G}} \sum_{n \in [N]} f(g \triangleright_{\mathcal{X}} \boldsymbol{x}_n) \quad \text{and} \quad \widehat{\mathbb{E}}_{\mathbf{y}}[h(\mathbf{y})] = \frac{1}{|\mathbb{G}|N} \sum_{g \in \mathbb{G}} \sum_{n \in [N]} h(g \triangleright_{\mathcal{Y}} \boldsymbol{y}_n). \tag{17}$$

Hence, our method learns representations of $\mathbf{x}$ and $\mathbf{y}$ that transform *nonlinear regression of observables* into a simple *linear regression* in the learned space. For example, assuming $\mathbf{y}$ has bounded variance and setting $\boldsymbol{h}(\boldsymbol{y}) = \boldsymbol{y}$, we recover the standard ($\mathbb{G}$-equivariant) regression solution (see

Tab. 1-left). Equally important, by letting $\boldsymbol{h} = \mathbb{1}_{\mathbb{B}}$—the indicator of a measurable set $\mathbb{B} \subseteq \mathcal{Y}$—the model estimates conditional probabilities (see Tab. 1-right), thereby supporting both regression and uncertainty quantification (e.g., conditional quantiles, covariances; see Sec. 6 and [25]).

To further illustrate the impact of symmetries in conditional probability estimation, we consider conditioning on measurable sets $\mathbb{A} \subseteq \mathcal{X}$, leading to the estimate $\boldsymbol{z}(\mathbb{A}) := \mathbb{E}_{\mathbf{y}}[\boldsymbol{h}(\mathbf{y})|\mathbf{x} \in \mathbb{A}] \approx \widehat{\mathbb{E}}_{\mathbf{x}}[\widehat{\boldsymbol{z}}_{\boldsymbol{\theta}}(\boldsymbol{x})]/\widehat{\mathbb{E}}_{\mathbf{x}}[\mathbb{1}_{\mathbb{A}}(\mathbf{x})]$. In this context, symmetry is crucial in alleviating the bottlenecks of rare event estimation, as described next in Thm. C.1. To capture this effect, we introduce the *symmetry index* of $\mathbb{A}$ with respect to the probability distribution of $\mathbf{x}$, which quantifies the degree of symmetry in $\mathbb{A}$:

$$\gamma_{\mathbb{G}}(\mathbb{A}) = \frac{1}{|\mathbb{G}|-1} \sum_{g \in \mathbb{G} \setminus \{e\}} \frac{\mathbb{P}(\mathbf{x} \in \mathbb{A} \cap \; g \triangleright_{\mathcal{X}} \mathbb{A})}{\mathbb{P}(\mathbf{x} \in \mathbb{A})}. \tag{18}$$

Observe that $\gamma_{\mathbb{G}'}(\mathbb{A}) \in [0,1]$. In particular, $\gamma_{\mathbb{G}}(\mathbb{A}) = 1$ if $\mathbb{A}$ is $\mathbb{G}$-invariant (e.g., the vertical and horizontal reflection planes in Fig. 3), while $\gamma_{\mathbb{G}}(\mathbb{A}) = 0$ if $g \triangleright_{\mathcal{X}} \mathbb{A} \cap \mathbb{A} = \emptyset$ for all $g \in \mathbb{G}$ (e.g., any set disjoint from the reflection planes in Fig. 3). We refer to the latter as a $\mathbb{G}$-asymmetric set.

The following learning bounds cover the general setting presented above.

Theorem C.1. *Let $\mathbb{P}_{\mathbf{xy}}$ and $\mathbb{P}_{\mathbf{x}}$ be $\mathbb{G}$-invariant, and let $\mathsf{E}_{\mathbf{y}|\mathbf{x}}$ be a $(1/\alpha)$-Schatten-class operator. Given $\boldsymbol{\theta} \in \Theta$, let $\kappa_{\boldsymbol{\theta}}$ be the kernel given in (13) that defines a rank $r = d_{\text{iso}} m$ $\mathbb{G}$-equivariant operator $\mathsf{E}_{\boldsymbol{\theta}}$, where m is the number of distinct singular spaces, and $d_{\text{iso}} = \sum_{k \in [n_{\text{iso}}]} d_k \geq n_{\text{iso}}$ denotes the "total dimensionality" associated to the group $\mathbb{G}$. Given $\delta \in [0,1)$, let $\mathcal{E}^r_{\boldsymbol{\theta}} = \|\mathsf{E}_{\mathbf{y}|\mathbf{x}} - \mathsf{E}_{\boldsymbol{\theta}}\|_{op}$ be the representation learning error. Let $\varepsilon^\star_N(\delta) = [d_{\text{iso}} N]^{-\frac{\alpha}{1+2\alpha}} \ln(n_{\text{iso}}/\delta)$ and $m \asymp [N d_{\text{iso}}^{-2\alpha}]^{\frac{1}{1+2\alpha}}$.*

If $\boldsymbol{h} \in \mathcal{L}^2_{P_{\mathbf{y}}}(\mathcal{Y}, \mathcal{Z})$ is either $\mathbb{G}$-invariant or $\mathbb{G}$-equivariant, $\mathbb{A} \subset \mathcal{X}$ is a measurable set and $\mathbb{G}' \leq \mathbb{G}$, then with probability at least $1-\delta$ w.r.t. an iid draw of $\mathbb{D}_N = \{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^N$ from $P_{\mathbf{xy}}$ it holds

$$\|\boldsymbol{z} - \hat{\boldsymbol{z}}_{\boldsymbol{\theta}}\|_{\mathcal{L}^2_{P_{\mathbf{x}}}(\mathcal{X},\mathcal{Z})} \lesssim \sqrt{\operatorname{Var}[\|\boldsymbol{h}(\mathbf{y})\|_{\mathcal{Z}}]} \left[\mathcal{E}^r_{\boldsymbol{\theta}} + \varepsilon^\star_N(\delta)\right] \tag{19}$$

and

$$\|\boldsymbol{z}(\mathbb{A}) - \hat{\boldsymbol{z}}_{\boldsymbol{\theta}}(\mathbb{A})\|_{\mathcal{Z}} \lesssim \frac{\sqrt{1 + (|\mathbb{G}'|-1)\gamma_{\mathbb{G}'}(\mathbb{A})}\,\sqrt{\operatorname{Var}[\|\boldsymbol{h}(\mathbf{y})\|_{\mathcal{Z}}]}}{\sqrt{|\mathbb{G}'|\mathbb{P}(\mathbf{x} \in \mathbb{A})}} \left[\mathcal{E}^r_{\boldsymbol{\theta}} + \varepsilon^\star_N(\delta)\right]. \tag{20}$$

Proof. $\mathbb{G}$-invariance of $P_{\mathbf{x}}$ and $P_{\mathbf{y}}$ allows us to control both bias (Thm. M.2) and variance (Prop. M.3) of $\hat{\boldsymbol{z}}_{\boldsymbol{\theta}}$. A simple balancing of m yields the final bound on the error. $\square$

We conclude by highlighting key implications of the theorem. The parameter α quantifies the problem regularity via $\sum_{i \in \mathbb{N}} \sigma_i^{1/\alpha} < \infty$, with $\alpha = \infty$ for finite-rank operators and $\alpha = 0$ for merely compact operators. The operator is trace class for $\alpha = 1$ and Hilbert–Schmidt for $\alpha = 1/2$, which is equivalent to $\kappa \in \mathcal{L}^2_{P_{\mathbf{x}} \times P_{\mathbf{y}}}(\mathcal{X} \times \mathcal{Y})$ (see App. M). Hence, our results cover learning rates ranging from arbitrarily slow (as $\alpha \to 0$) to fast rates $[d_{\text{iso}} N]^{-1/2}$ as $\alpha \to \infty$; (ii) Equivariant disentangled representations boost the effective sample size to $d_{\text{iso}} N \geq n_{\text{iso}} N \gg N$, providing not only the expected n_{iso} gain from disentanglement but also a remarkable $d_{\text{iso}} = \sum_{k=1}^{n_{\text{iso}}} d_k$ boost (see Fig. 2-right)—achieved by fully exploiting the equivariant structure within each isotypic-singular space. (iii) Because point-wise guarantees are essential in some applications, (20) offers a set-wise learning bound that quantifies how symmetries help overcome bottlenecks in estimating observables associated with rare events. In particular, the effective rarity of $\mathbf{x} \in \mathbb{A}$ is captured by $\gamma_{\mathbb{G}'}(\mathbb{A})$, yielding a maximal gain of $|\mathbb{G}|\mathbb{P}[\mathbf{x} \in \mathbb{A}] \gg \mathbb{P}[\mathbf{x} \in \mathbb{A}]$ when $\mathbb{A}$ is asymmetric. (iv) When no symmetry prior exist, that is, when $\mathbb{G} = \{e\}$ and $|\mathbb{G}| = n_{\text{iso}} = d_{\text{iso}} = 1$, our framework recovers the baseline results of [37]. In contrast, exploiting symmetries yields substantial statistical gains: it amplifies the effective sample size and fundamentally mitigates the inherent bottlenecks of rare event estimation.

D Symmetry constraints on conditional expectations

Under the assumed symmetry priors in (6) the conditional expectation of $\mathbf{y}$ is a $\mathbb{G}$-equivariant function/map. This property is depicted in Fig. 3-center and proved in the following proposition.

Proposition D.1 ($\mathbb{G}$-equivariant conditional expectations)**.** *Let $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{Y}$ be two vector valued random variables satisfying the symmetry priors of Eq. (6). Then, the conditional expectation of $\mathbf{y}$ given $\mathbf{x}$ is $\mathbb{G}$-equivariant, since, for every $g \in \mathbb{G}, \boldsymbol{x} \in \mathcal{X}$,*

$$\begin{aligned}
\mathbb{E}[\mathbf{y}|\mathbf{x} = g \triangleright_{\mathcal{X}} \boldsymbol{x}] = g \triangleright_{\mathcal{Y}} \mathbb{E}[\mathbf{y}|\mathbf{x} = \boldsymbol{x}] \\
&= \int_{\mathcal{Y}} g \triangleright_{\mathcal{Y}} \boldsymbol{y}\ P_{\mathbf{y}|\mathbf{x}}(d\boldsymbol{y}|\boldsymbol{x}) \\
&= \int_{\mathcal{Y}} \boldsymbol{y}\ P_{\mathbf{y}|\mathbf{x}}\Big(g^{-1} \triangleright_{\mathcal{Y}} d\boldsymbol{y}|\boldsymbol{x}\Big) \\
&= \int_{\mathcal{Y}} \boldsymbol{y}\ P_{\mathbf{y}|\mathbf{x}}(d\boldsymbol{y}|g \triangleright_{\mathcal{X}} \boldsymbol{x}) \quad \textit{(by Eq. (6))} \\
&= \mathbb{E}[\mathbf{y}|\mathbf{x} = g \triangleright_{\mathcal{X}} \boldsymbol{x}].
\end{aligned}$$

E $\mathbb{G}$-Equivariant bilinear NN architecture

This section outlines how to construct a $\mathbb{G}$-equivariant disentangled representation for the random variables $\mathbf{x}$ and $\mathbf{y}$ using **any** type of $\mathbb{G}$-equivariant NN architecture backbone, such as MLP, CNNs, Transformers, and others.

Let $\boldsymbol{f_\theta} : \mathcal{X} \mapsto \mathbb{R}^r$ and $\boldsymbol{h_\theta} : \mathcal{Y} \mapsto \mathbb{R}^r$ be two $\mathbb{G}$-equivariant NNs, whose outputs will be interpreted as the basis functions of the truncated symmetric function spaces $\mathcal{F}_{\mathbf{x}} \subset \mathcal{L}^2_{\mathbf{x}}$ and $\mathcal{F}_{\mathbf{y}} \subset \mathcal{L}^2_{\mathbf{y}}$. Assume, the group representations on $\mathcal{F}_{\mathbf{x}}$ and $\mathcal{F}_{\mathbf{y}}$ are constructed from multiplicities of the group's regular representation, $\boldsymbol{\rho}_{\mathcal{F}_{\mathbf{x}}} = \bigoplus_{n=1}^{r/|\mathbb{G}|} \boldsymbol{\rho}_{\text{reg}}$ and $\boldsymbol{\rho}_{\mathcal{F}_{\mathbf{y}}} = \bigoplus_{n=1}^{r/|\mathbb{G}|} \boldsymbol{\rho}_{\text{reg}}$—as done usually in practice [17]. Since for (most) finite groups, the decomposition of $\boldsymbol{\rho}_{\text{reg}}$ into *irreps* is known or can be computed, we have access to the analytical change of basis $\boldsymbol{Q}_{\mathbf{x}} : \mathcal{F}_{\mathbf{x}} \mapsto \mathcal{F}_{\mathbf{x}}$ and $\boldsymbol{Q}_{\mathbf{y}} : \mathcal{F}_{\mathbf{y}} \mapsto \mathcal{F}_{\mathbf{y}}$ to transition to the isotypic basis. Consequently, we can directly parameterize the representations of the random variables in disentangled form as:

$$\boldsymbol{\phi_\theta}(\cdot) = \boldsymbol{Q}_{\mathbf{x}}^\top(\boldsymbol{f_\theta}(\cdot) - \mathbb{E}_{\mathbf{x}}[\boldsymbol{f_\theta}(\mathbf{x})]), \quad \boldsymbol{\psi_\theta}(\cdot) = \boldsymbol{Q}_{\mathbf{y}}^\top(\boldsymbol{h_\theta}(\cdot) - \mathbb{E}_{\mathbf{y}}[\boldsymbol{h_\theta}(\mathbf{y})]). \tag{21}$$

Given that during training these representations are not orthogonal, the truncated operator is parameterized as the trainable $\mathbb{G}$-equivariant matrix $\boldsymbol{E_\theta} = \oplus_k^{n_{\text{iso}}} \boldsymbol{E}_{\boldsymbol{\theta}}^{(k)} = \oplus_k^{n_{\text{iso}}} \boldsymbol{O}^{(k)} \otimes \boldsymbol{I}_{d_k}$ with parameters $\{\boldsymbol{O}^{(k)} \in \mathbb{R}^{m_k \times m_k}\}_{k=1}^{n_{\text{iso}}}$. Hence, the kernel of each the truncated operator is given in terms of the model free parameters by:

$$\kappa_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y}) = \mathbb{1}_{P_{\mathbf{x}}}(\boldsymbol{x})\mathbb{1}_{P_{\mathbf{y}}}(\boldsymbol{y}) + \sum_{k=1}^{n_{\text{iso}}} \sum_{s,t}^{m_k} \boldsymbol{O}_{s,t}^{(k)} \sum_{i,j}^{d_k} \phi_{s,i}^{\boldsymbol{\theta}(k)}(\boldsymbol{x}) \psi_{t,j}^{\boldsymbol{\theta}(k)}(\boldsymbol{y}). \tag{22}$$

Note that after training, the SVD of the learned operator can be computed by exploiting the constraints imposed by the operator's $\mathbb{G}$-equivariance (see Thm. K.5 and Fig. 2). Importantly, once changed to the spectral basis, the group action on the approximated spectral basis matches that on the isotypic basis (see Cor. K.4).

F Symmetry aware orthonormalization of disentangled representations

This section covers how to compute unbiased empirical estimates of the orthonormalization and centering regularization terms in Eq. (14) in the presence of symmetries.

Let $\mathsf{E}_{\mathbf{y}|\mathbf{x}} : \mathcal{L}^2_{\mathbf{y}} \mapsto \mathcal{L}^2_{\mathbf{x}}$ be the conditional expectation operator and $\mathsf{E}_{\boldsymbol{\theta}} : \mathcal{F}_{\mathbf{y}} \mapsto \mathcal{F}_{\mathbf{x}}$ be its r-rank approximation on the spaces $\mathcal{F}_{\mathbf{x}} = \text{span}(\{\phi_i\}_{i=1}^r)$ and $\mathcal{F}_{\mathbf{y}} = \text{span}(\{\psi_i\}_{i=1}^r)$. Denote by $\kappa(\boldsymbol{x}, \boldsymbol{y}) := \frac{P_{\mathbf{xy}}(\boldsymbol{x},\boldsymbol{y})}{P_{\mathbf{x}}(\boldsymbol{x})P_{\mathbf{y}}(\boldsymbol{y})}$ and $\kappa_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y}) := \sum_{i,j=1}^r [\boldsymbol{E}_\theta]_{i,j} \phi_i(\boldsymbol{x}) \psi_j(\boldsymbol{y}) = \boldsymbol{\phi}(\boldsymbol{x})^\top \boldsymbol{E_\theta} \boldsymbol{\psi}(\boldsymbol{y})$ the kernel functions of the full and restricted operator, respectively. Then we have that:

$$
\begin{aligned}
\|\mathsf{E}_{\mathbf{y}|\mathbf{x}} - \mathsf{E}_{\theta}\|_{\mathrm{HS}}^2 &\leq -2\langle \mathsf{E}_{\mathbf{y}|\mathbf{x}}, \mathsf{E}_{\theta}\rangle_{\mathrm{HS}} + \|\mathsf{E}_{\theta}\|_{\mathrm{HS}}^2, && (23a)\\
&\leq -2\int_{\mathcal{X}\times\mathcal{Y}} \kappa(\boldsymbol{x},\boldsymbol{y})\kappa_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{y}) P_{\mathbf{x}}(d\boldsymbol{x})P_{\mathbf{y}}(d\boldsymbol{y}) + \int_{\mathcal{X}\times\mathcal{Y}} \kappa_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{y})^2 P_{\mathbf{x}}(d\boldsymbol{x})P_{\mathbf{y}}(d\boldsymbol{y})\\
&\leq -2\int_{\mathcal{X}\times\mathcal{Y}} \kappa_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{y}) P_{\mathbf{xy}}(d\boldsymbol{x},d\boldsymbol{y}) + \int_{\mathcal{X}\times\mathcal{Y}} \kappa_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{y})^2 P_{\mathbf{x}}(d\boldsymbol{x})P_{\mathbf{y}}(d\boldsymbol{y})\\
&\leq -2\mathbb{E}_{\mathbf{xy}}\kappa_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{y}) + \mathbb{E}_{\mathbf{x}}\mathbb{E}_{\mathbf{y}}\kappa_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{y})^2. && (23b)
\end{aligned}
$$

For the purpose of our representation learning problem, we consider the scenario in which the chosen basis sets include the constant function, and all other basis functions are centered by construction. That is, $\mathbb{I}_{\mathcal{F}_{\mathbf{x}}} = \{\mathbb{1}_{P_{\mathbf{x}}}\} \cup \{\phi_i \mid \langle \phi_i, \mathbb{1}_{P_{\mathbf{x}}}\rangle_{\mathbf{x}} = 0\}_{i=1}^r$ and $\mathbb{I}_{\mathcal{F}_{\mathbf{y}}} = \{\mathbb{1}_{P_{\mathbf{y}}}\} \cup \{\psi_i \mid \langle \psi_i, \mathbb{1}_{P_{\mathbf{y}}}\rangle_{\mathbf{y}} = 0\}_{i=1}^r$. This results in the $(r+1)$-dimensional matrices:

$$
\boldsymbol{V}_{\mathbf{x}} := \left[\begin{smallmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{C}_{\mathbf{x}} \end{smallmatrix}\right], \quad \boldsymbol{V}_{\mathbf{y}} := \left[\begin{smallmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{C}_{\mathbf{y}} \end{smallmatrix}\right], \tag{24}
$$

where $\boldsymbol{C}_{\mathbf{x}} = \mathrm{Cov}(\boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{x})) \in \mathbb{R}^{r\times r}$, $\boldsymbol{C}_{\mathbf{y}} = \mathrm{Cov}(\boldsymbol{\psi}(\mathbf{y}), \boldsymbol{\psi}(\mathbf{y})) \in \mathbb{R}^{r\times r}$ denote the matrix forms of the truncated covariance operators $\mathsf{C}_{\mathbf{x}} : \mathcal{F}_{\mathbf{x}} \mapsto \mathcal{F}_{\mathbf{x}}$ and $\mathsf{C}_{\mathbf{y}} : \mathcal{F}_{\mathbf{y}} \mapsto \mathcal{F}_{\mathbf{y}}$ (see Def. L.5), respectively. Then the orthonormality regularization of Eq. (5) becomes:

$$
\|\boldsymbol{V}_{\mathbf{x}} - \boldsymbol{I}\|_{\mathrm{F}}^2 = \|\boldsymbol{C}_{\mathbf{x}} - \boldsymbol{I}_r\|_{\mathrm{F}}^2 + 2\|\mathbb{E}_{P_{\mathbf{x}}}\boldsymbol{\phi}(\mathbf{x})\|^2 \quad \|\boldsymbol{V}_{\mathbf{y}} - \boldsymbol{I}\|_{\mathrm{F}}^2 = \|\boldsymbol{C}_{\mathbf{y}} - \boldsymbol{I}_r\|_{\mathrm{F}}^2 + 2\|\mathbb{E}_{P_{\mathbf{y}}}\boldsymbol{\psi}(\mathbf{y})\|^2. \tag{25}
$$

Since $\|\boldsymbol{C}_{\mathbf{x}}\|_{\mathrm{F}}^2 = \mathrm{tr}(\boldsymbol{C}_{\mathbf{xy}}^2)$ involves products of covariance matrices, we compute its empirical value using unbiased estimators. For generality, we present the unbiased estimator for the cross-covariance.

Unbiased estimation of Frobenious norm of cross-covariance operators Since $\|\boldsymbol{C}_{\mathbf{xy}}\|_F^2 = \mathrm{tr}(\boldsymbol{C}_{\mathbf{xy}}^2)$ involves products of covariance matrices, we obtain unbiased estimates from finite samples by computing the metric using two independent sampling sets from $P_{\mathbf{xy}}$. This is observed by:

$$
\begin{aligned}
\|\boldsymbol{C}_{\mathbf{xy}}\|_{\mathrm{F}}^2 &= \mathrm{tr}(\boldsymbol{C}_{\mathbf{xy}}^2) = \sum_{i=1}^r [\boldsymbol{C}_{\mathbf{xy}}^2]_{i,i} = \sum_{i=1}^r\sum_{j=1}^r [\boldsymbol{C}_{\mathbf{xy}}]_{i,j}[\boldsymbol{C}_{\mathbf{xy}}]_{j,i}\\
&= \sum_{i=1}^r\sum_{j=1}^r \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim P_{\mathbf{xy}}}[\phi_{c,i}(\mathbf{x})\psi_{c,j}(\mathbf{y})]\mathbb{E}_{(\mathbf{x}',\mathbf{y}')\sim P_{\mathbf{xy}}}[\phi_{c,j}(\mathbf{x}')\psi_{c,i}(\mathbf{y}')]\\
&= \mathbb{E}_{(\mathbf{x},\mathbf{y},\mathbf{x}',\mathbf{y}')\sim P_{\mathbf{xy}}}[\sum_{i=1}^r \phi_{c,i}(\mathbf{x})\psi_{c,i}(\mathbf{y}') \sum_{j=1}^r \phi_{c,j}(\mathbf{x}')\psi_{c,j}(\mathbf{y})]\\
&= \mathbb{E}_{(\mathbf{x},\mathbf{y},\mathbf{x}',\mathbf{y}')\sim P_{\mathbf{xy}}}[(\boldsymbol{\phi}_c(\mathbf{x})^\top\boldsymbol{\psi}_c(\mathbf{y}'))(\boldsymbol{\phi}_c(\mathbf{x}')^\top\boldsymbol{\psi}_c(\mathbf{y}))]\\
&\approx \frac{1}{N^2}\sum_{n=1}^N\sum_{m=1}^N (\boldsymbol{\phi}_c(\boldsymbol{x}_n)^\top\boldsymbol{\psi}_c(\boldsymbol{y}'_m))(\boldsymbol{\phi}_c(\boldsymbol{x}'_m)^\top\boldsymbol{\psi}_c(\boldsymbol{y}_n)),
\end{aligned} \tag{26}
$$

where $\boldsymbol{\phi}_c(\mathbf{x}) = \boldsymbol{\phi}(\mathbf{x}) - \mathbb{E}_{P_{\mathbf{x}}}\boldsymbol{\phi}(\mathbf{x})$ denotes the centered basis functions, and $((\boldsymbol{x},\boldsymbol{y}),(\boldsymbol{x}',\boldsymbol{y}')) \sim P_{\mathbf{xy}}$ indicates two independent sampling sets from $P_{\mathbf{xy}}$ used for the unbiased estimation of $\|\boldsymbol{C}_{\mathbf{x}}\|_F^2$. The final equation then provides the unbiased empirical estimator computed on a dataset $\mathbb{D} = \{(\boldsymbol{x}_n,\boldsymbol{y}_n) \sim P_{\mathbf{xy}}\}_{n=1}^N$ and any random permutation of it, denoted as $\mathbb{D}' = \{(\boldsymbol{x}'_n,\boldsymbol{y}'_n) \sim P_{\mathbf{xy}}\}_{n=1}^N$.

F.1 Unbiased estimation of orthonormal regularization

The regularization term for optimizing the loss (5) involves encouraging the basis sets to be orthonormal. The metric quantifying the orthogonality of the basis sets is defined by:

$$
\begin{aligned}
\|\boldsymbol{V}_{\mathbf{x}} - \boldsymbol{I}\|_{\mathrm{F}}^2 &= \|\boldsymbol{C}_{\mathbf{x}} - \boldsymbol{I}_r\|_{\mathrm{F}}^2 + 2\|\mathbb{E}_{P_{\mathbf{x}}}\boldsymbol{\phi}(\mathbf{x})\|^2 = \mathrm{tr}(\boldsymbol{C}_{\mathbf{x}}^2) - 2\mathrm{tr}(\boldsymbol{C}_{\mathbf{x}}) + r + 2\|\mathbb{E}_{P_{\mathbf{x}}}\boldsymbol{\phi}(\mathbf{x})\|^2,\\
\|\boldsymbol{V}_{\mathbf{y}} - \boldsymbol{I}\|_{\mathrm{F}}^2 &= \|\boldsymbol{C}_{\mathbf{y}} - \boldsymbol{I}_r\|_{\mathrm{F}}^2 + 2\|\mathbb{E}_{P_{\mathbf{y}}}\boldsymbol{\psi}(\mathbf{y})\|^2 = \mathrm{tr}(\boldsymbol{C}_{\mathbf{y}}^2) - 2\mathrm{tr}(\boldsymbol{C}_{\mathbf{y}}) + r + 2\|\mathbb{E}_{P_{\mathbf{y}}}\boldsymbol{\psi}(\mathbf{y})\|^2.
\end{aligned} \tag{27}
$$

Hence given a dataset of samples $\mathbb{D} = \{(\boldsymbol{x}_n, \boldsymbol{y}_n) \sim P_{\mathbf{xy}}\}_{n=1}^N$, and any random permutation of the dataset order $\mathbb{D}' = \{(\boldsymbol{x}'_n, \boldsymbol{y}'_n) \sim P_{\mathbf{xy}}\}_{n=1}^N$ we can derive unbiased empirical estimates of (27) as:

$$
\begin{aligned}
\|\boldsymbol{V}_{\mathbf{x}} - \boldsymbol{I}\|_{\mathrm{F}}^2 &\approx \widehat{\mathbb{E}}_{(\mathbf{x},\mathbf{x}')\sim P_{\mathbf{x}}}[(\boldsymbol{\phi}_c(\mathbf{x})^\top \boldsymbol{\phi}_c(\mathbf{x}'))^2] - 2\widehat{\mathbb{E}}_{P_{\mathbf{x}}}[\boldsymbol{\phi}_c(\mathbf{x})^\top \boldsymbol{\phi}_c(\mathbf{x})] + r + 2\|\widehat{E}_{P_{\mathbf{x}}}\boldsymbol{\phi}(\mathbf{x})\|^2 \\
&\approx \frac{1}{N^2}\sum_{n=1}^N\sum_{m=1}^N (\boldsymbol{\phi}_c(\boldsymbol{x}_n)^\top \boldsymbol{\phi}_c(\boldsymbol{x}'_m))^2 - 2\frac{1}{N}\sum_{n=1}^N \boldsymbol{\phi}_c(\boldsymbol{x}_n)^\top \boldsymbol{\phi}_c(\boldsymbol{x}_n) + r + 2\|\frac{1}{N}\sum_{n=1}^N \boldsymbol{\phi}(\boldsymbol{x}_n)\|^2, \\
\|\boldsymbol{V}_{\mathbf{y}} - \boldsymbol{I}\|_{\mathrm{F}}^2 &\approx \widehat{\mathbb{E}}_{(\mathbf{y},\mathbf{y}')\sim P_{\mathbf{y}}}[(\boldsymbol{\psi}_c(\mathbf{y})^\top \boldsymbol{\psi}_c(\mathbf{y}'))^2] - 2\widehat{\mathbb{E}}_{P_{\mathbf{y}}}[\boldsymbol{\psi}_c(\mathbf{y})^\top \boldsymbol{\psi}_c(\mathbf{y})] + r + 2\|\widehat{E}_{P_{\mathbf{y}}}\boldsymbol{\phi}(\mathbf{y})\|^2 \\
&\approx \frac{1}{N^2}\sum_{n=1}^N\sum_{m=1}^N (\boldsymbol{\psi}_c(\boldsymbol{y}_n)^\top \boldsymbol{\psi}_c(\boldsymbol{y}'_m))^2 - 2\frac{1}{N}\sum_{n=1}^N \boldsymbol{\psi}_c(\boldsymbol{y}_n)^\top \boldsymbol{\psi}_c(\boldsymbol{y}_n) + r + 2\|\frac{1}{N}\sum_{n=1}^N \boldsymbol{\psi}(\boldsymbol{y}_n)\|^2.
\end{aligned} \tag{28}
$$

F.2 Orthonormal regularization of symmetric Hilbert spaces

Since the covariance operators $\mathsf{C}_{\mathbf{x}} : \mathcal{L}^2_{\mathbf{x}} \mapsto \mathcal{L}^2_{\mathbf{x}}$ and $\mathsf{C}_{\mathbf{y}} : \mathcal{L}^2_{\mathbf{y}} \mapsto \mathcal{L}^2_{\mathbf{y}}$ are $\mathbb{G}$-equivariant (see Prop. L.6), their matrix representations in the isotypic basis are block-diagonal. Hence (27) becomes:

$$
\begin{aligned}
\|\boldsymbol{V}_{\mathbf{x}} - \boldsymbol{I}\|_{\mathrm{F}}^2 &= \|\boldsymbol{C}_{\mathbf{x}} - \boldsymbol{I}_r\|_{\mathrm{F}}^2 + 2\|\mathbb{E}_{P_{\mathbf{x}}}\boldsymbol{\phi}(\mathbf{x})\|^2 \\
&= \| \oplus_{k=1}^{n_{\text{iso}}} \boldsymbol{C}_{\mathbf{x}}^{(k)} - \boldsymbol{I}_r\|_{\mathrm{F}}^2 + 2\|\mathbb{E}_{P_{\mathbf{x}}}\boldsymbol{\phi}^{\text{inv}}(\mathbf{x})\|^2, \\
&= \sum_{k=1}^{n_{\text{iso}}} \|\boldsymbol{C}_{\mathbf{x}}^{(k)} - \boldsymbol{I}_r^{(k)}\|_{\mathrm{F}}^2 + 2\|\mathbb{E}_{P_{\mathbf{x}}}\boldsymbol{\phi}^{\text{inv}}(\mathbf{x})\|^2 \\
&= \sum_{k=1}^{n_{\text{iso}}} \left(\|\boldsymbol{C}_{\mathbf{x}}^{(k)}\|_F^2 - 2\text{tr}(\boldsymbol{C}_{\mathbf{x}}^{(k)}) + r_k\right) + 2\|\mathbb{E}_{P_{\mathbf{x}}}\boldsymbol{\phi}(\mathbf{x})\|^2 \\
&= 2\|\mathbb{E}_{P_{\mathbf{x}}}\boldsymbol{\phi}(\mathbf{x})\|^2 + r + \sum_{k=1}^{n_{\text{iso}}} \|\boldsymbol{Z}_{\mathbf{x}}^{(k)} \otimes \boldsymbol{I}_{|\bar{\boldsymbol{\rho}}_k|}\|_F^2 - 2\text{tr}(\boldsymbol{Z}_{\mathbf{x}}^{(k)} \otimes \boldsymbol{I}_{|\bar{\boldsymbol{\rho}}_k|}) \\
&= 2\|\mathbb{E}_{P_{\mathbf{x}}}\boldsymbol{\phi}(\mathbf{x})\|^2 + r + \sum_{k=1}^{n_{\text{iso}}} |\bar{\boldsymbol{\rho}}_k| \left(\|\boldsymbol{Z}_{\mathbf{x}}^{(k)}\|_F^2 - 2\text{tr}(\boldsymbol{Z}_{\mathbf{x}}^{(k)})\right),
\end{aligned} \tag{29}
$$

where the Frobenius norm of the matrices $\boldsymbol{Z}_{\mathbf{x}}^{(k)}$ and $\boldsymbol{Z}_{\mathbf{y}}^{(k)}$, for all $k \in [1, n_{\text{iso}}]$, admit unbiased estimators as given in equation (26). Similar development follows for the $\mathbf{y}$ case.

G Experimental setup

In this section we provide details on the experimental setup. We first describe general design choices and hyperparameters and then provide details for each experiment.

Sample efficiency experiments For both the conditional expectation operator approximation and the $\mathbb{G}$-equivariant regression experiments, we evaluate model performance by measuring sample efficiency/complexity. To do so, we partition the dataset $\mathbb{D} = \{(\boldsymbol{x}_n, \boldsymbol{y}_n)\}_{n=1}^N$ into training, validation, and testing splits in proportions of 70%, 15%, and 15%, respectively. With fixed validation and testing sets, we iteratively train the models on increasing portions of the training set and report the test performance for each size.

For each training set size, we select the model checkpoint with the best validation loss to compute the test performance. Thus, these experiments quantify the generalization error (or true risk) and its evolution as a function of the training set size.

NNs architectures and hyperparameters To compare our equivariant representation learning framework with other contrastive and supervised methods, all (inference) models share a similar fixed architectural footprint. For the baseline models, the only hyperparameter tuned is the learning rate, whereas for the NCP and eNCP models we additionally tune the regularization weight γ in Eqs. (5) and (14). Further details for each experiment are provided in the corresponding sections below.

Code reproducibility All experiments, plots and examples are provided in the open-access repository and python package *symm_rep_learn*.

G.1 Conditional expectation operator approximation

In this experiment, we extend the conditional Gaussian Mixture Model (GMM) proposed by Gilardi et al. [38] to parametrically construct symmetric random variables taking values in arbitrary data spaces $\mathcal{X}$ and $\mathcal{Y}$ and with arbitrary finite symmetry groups $\mathbb{G}$. The GMM is defined by

$$\mathbf{z} := \mathbf{x} \oplus \mathbf{y} \sim \sum_{g \in \mathbb{G}} \sum_{c=1}^{n_g} \mathcal{N}(\boldsymbol{\rho}_{\mathcal{Z}}(g)\mu_{\mathbf{z},c}\ ,\ \boldsymbol{\rho}_{\mathcal{Z}}(g)\Sigma_{\mathbf{z},c}\boldsymbol{\rho}_{\mathcal{Z}}(g)^\top),$$

where $\boldsymbol{\rho}_{\mathcal{Z}}(g) := \boldsymbol{\rho}_{\mathcal{X}}(g) \oplus \boldsymbol{\rho}_{\mathcal{Y}}(g)$ are arbitrary group representations of $\mathbb{G}$ and n_g is the number of unique Gaussians with randomly sampled means $\mu_{\mathbf{z}} := \mu_{\mathbf{x}} \oplus \mu_{\mathbf{y}}$ and block-diagonal covariances $\Sigma_{\mathbf{z}} := \Sigma_{\mathbf{x}} \oplus \Sigma_{\mathbf{y}}$. Since every Gaussian appears in group orbits, this symmetric GMM has $\mathbb{G}$-invariant marginal distributions and an analytical expression for the conditional expectation operator kernel $\kappa(\boldsymbol{x}, \boldsymbol{y}) = {}^{p_{\mathbf{xy}}(\boldsymbol{x},\boldsymbol{y})}/_{P_{\mathbf{x}}(\boldsymbol{x})P_{\mathbf{y}}(\boldsymbol{y})}$ (see 2D example in Fig. 3). Consequently, we can directly estimate the approximation of the conditional expectation operator (Eq. (5)) as the mean squared error between the true and learned density ratios, i.e., $\kappa_{\text{mse}} := \mathbb{E}_{\mathbf{x}}\mathbb{E}_{\mathbf{y}} \left\|\kappa(\mathbf{x}, \mathbf{y}) - \kappa_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y})\right\|^2$.

To the best of our knowledge, this is the first synthetic experiment that directly estimates the truncation error of the conditional expectation operator in an inference task-agnostic setting, serving as a benchmark for future work.

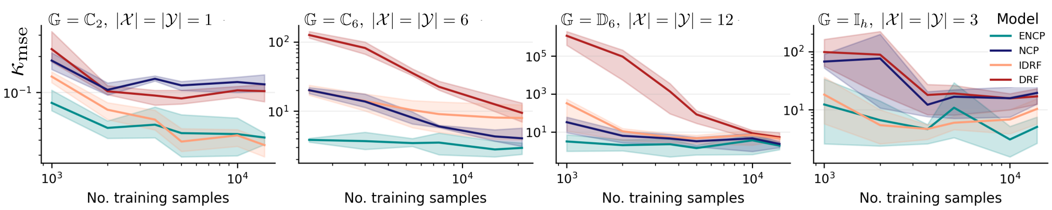

Figure 5: Sample efficiency plots comparing the test set PMD MSE $\kappa_{\text{mse}} := \mathbb{E}_{\mathbf{x}}\mathbb{E}_{\mathbf{y}}(\kappa(\mathbf{x}, \mathbf{y}) - \kappa_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}))^2$ versus the number of training samples, in log scales. Each plot corresponds to a symmetric cGMM with distinct symmetry groups and $(\mathbf{x}, \mathbf{y})$ dimensionality. The tested groups are the cyclic groups $\mathbb{C}_2$ and $\mathbb{C}_6$, the Dihedral group $\mathbb{D}_6$ (order 12), and the Icosahedral group $\mathbb{I}_h$ (order 60).

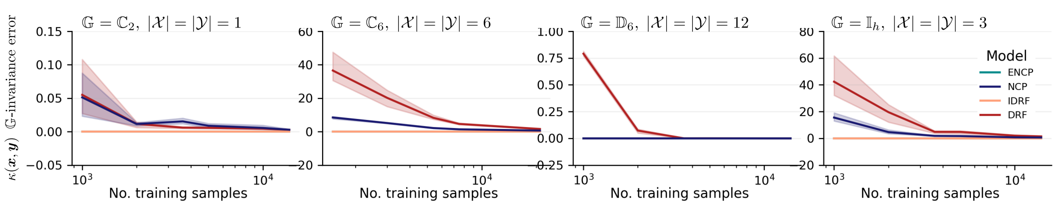

Figure 6: Sample efficiency plots comparing test set regression mean-square-error of the density ratio $\kappa(\boldsymbol{x}, \boldsymbol{y}) = {}^{P_{\mathbf{xy}}(\boldsymbol{x},\boldsymbol{y})}/_{P_{\mathbf{x}}(\boldsymbol{x})P_{\mathbf{y}}(\boldsymbol{y})}$ (log-scale) vs. the number of samples in the training set (log-scale). Each plot represents a different symmetric GMM with varying symmetry groups $\mathbb{G}$ and dimensionalities of the random variables $|\mathcal{X}|$ and $|\mathcal{Y}|$. The groups tested are the cyclicg grups $\mathbb{C}_2$ and $\mathbb{C}_6$, the Dihedral group $\mathbb{D}_6$ of order 12 and the Icosahedral group $\mathbb{I}_h$ of order 60.

Fig. 5 compares sample efficiency using κ_{mse}, while Fig. 6 shows the error in the $\mathbb{G}$-invariant of the learned κ ratio versus sample size, highlighting that symmetry-aware methods encode this property as an architectural constraint, ensuring a strictly $\mathbb{G}$-invariant learned ratio.

G.2 $\mathbb{G}$-equivariant regression of robot's CoM momenta

In this experiment, we evaluate the quality of the learned representations using the contrastive loss Eqs. (5) and (14) alongside supervised learning baselines trained with the standard MSE loss. The task is a $\mathbb{G}$-equivariant benchmark in robotics presented in [15], with the goal of predicting a quadruped robot's CoM linear $\boldsymbol{l} \in \mathbb{R}^3$ and angular momenta $\boldsymbol{k} \in \mathbb{R}^3$ from noisy observations of the robot's generalized positions $\boldsymbol{q} \in \mathbb{R}^{12}$ and velocity coordinates $\dot{\boldsymbol{q}} \in \mathbb{R}^{12}$. Consequently, the random variables are defined as $\boldsymbol{x} = \boldsymbol{q} + \epsilon_{\boldsymbol{q}} \oplus \dot{\boldsymbol{q}} + \epsilon_{\dot{\boldsymbol{q}}}$ and $\boldsymbol{y} = \boldsymbol{l} \oplus \boldsymbol{k}$, where $\epsilon_{\boldsymbol{q}} \in \mathbb{R}^{12}$ and $\epsilon_{\dot{\boldsymbol{q}}} \in \mathbb{R}^{12}$ are independent Gaussian noise terms that model sensor noise. The function computing the CoM momenta from these proprioceptive observations is highly non-linear and $\mathbb{G}$-equivariant whenever $\mathbb{G}$ is a morphological symmetry group of the robot (see Fig. 7 and [15] for details).

The robot considered is the quadruped robot Solo (Fig. 7-right), which possesses a symmetry group of order 8: $\mathbb{G} = \mathbb{K}_4 \times \mathbb{C}_2$, as depicted in this animation showing 8 symmetric robot configurations along with their corresponding linear and angular momenta vectors.

NN architectures We configure all models under consideration (eNCP, NCP, eMLP, and MLP) to have an inference-time NN architecture with a similar footprint. In particular, the encoder network for $\mathbf{x}$ in NCP and eNCP is designed similarly to the NN used in MLP/eMLP. The idea is to test how a model with the same capacity performs on the downstream task of classification when trained using either the representation learning loss or a supervised learning loss. The backbone of all architectures is a standard multilayer perceptron consisting of three hidden layers, each with 512 units, followed by a final hidden layer containing 128 units. This final layer encodes the feature vector r for the NCP and eNCP models. Crucially, since $\mathbb{G}$-equivariance enforces weight sharing in the NN architecture, the encoder NN for eNCP and eMLP comprises $\times 8$ fewer parameters than their symmetry-agnostic counterparts.

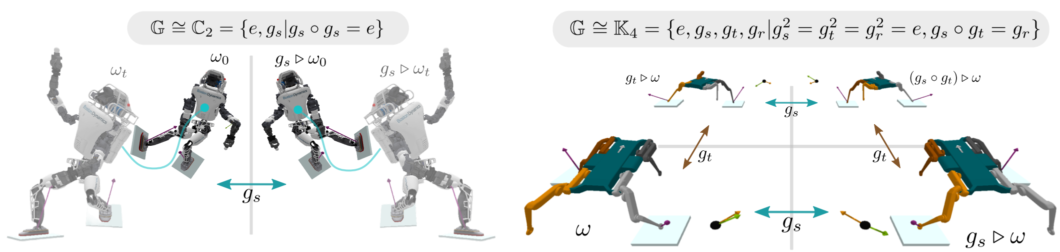

Figure 7: Example of morphological finite symmetry in robotics. **Left**: A humanoid robot with the reflectional symmetry group $\mathbb{G} \equiv \mathbb{C}_2$. **Right**: The quadruped robot Solo with the symmetry group $\mathbb{G} = \mathbb{K}_4 \times \mathbb{C}_2$ (only $\mathbb{K}_4$ is shown for clarity). The robot's center of mass linear $\boldsymbol{l} \in \mathbb{R}^3$ and angular $\boldsymbol{k} \in \mathbb{R}^3$ momentum are depicted as orange and green vectors, respectively, for each symmetric configuration. Images adapted from Ordoñez-Apraez et al. [15] with author approval.

G.3 Uncertainty quantification via conditional quantile regression

The goal of these experiments benchmark is to learn the family of conditional distributions $\mathbb{P}(\mathbf{y} \mid \mathbf{x} = \cdot)$ for a bivariate random variable $\mathbf{y} = [\mathrm{y}_0, \mathrm{y}_1] \in \mathbb{R}^2$ given a scalar covariate $\mathbf{x} \in \mathbb{R}$. Once $\mathbb{P}(\mathbf{y} \mid \mathbf{x})$ is recovered, the practitioner can estimate *conditional* $(1-\alpha)$–confidence regions by regressing the lower and upper conditional quantiles $q_{\alpha/2}(\mathbf{x})$, $q_{1-\alpha/2}(\mathbf{x})$ for any desired miscoverage level $\alpha \in (0, 1)$. In particular, a 95% confidence region corresponds to $\alpha = 0.05$, so the two quantiles of interest are $q_{0.025}(\mathbf{x})$ and $q_{0.975}(\mathbf{x})$. See Fig. 8 for a visual representation of the problem.

Conditional quantile regression models We compare the NCP and proposed eNCP models to a standard baseline for parametric NN conditional quantile regression, namely CQR [43], which uses two separate NNs to predict the lower and upper quantiles of the conditional distribution, trained with a pinball loss function (see [43] for details). Both models use MLP backbones with similar parameter counts, ensuring that improvements are solely due to the loss functions.

Furthermore, CQR can only be trained for specific confidence intervals, requiring retraining for different quantiles. In contrast, the NCP and eNCP models, trained using the deep representation

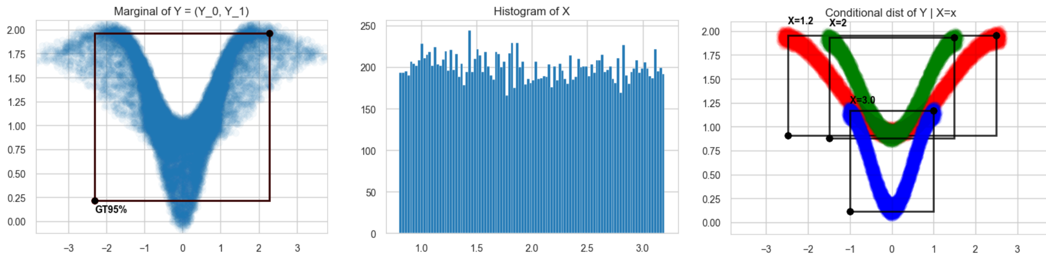

Figure 8: Synthetic experiment in uncertainty quantification, originally proposed by Feldman et al. [43]. The task is to predict the 95% confidence intervals (black bounding boxes) of a random variable $\mathbf{y} \in \mathbb{R}^2$ conditioned on a scalar random variable $\mathbf{x} \in \mathbb{R}$. **Left:** The marginal distribution $\mathbb{P}(\mathbf{y})$. **Middle:** The marginal distribution $\mathbb{P}(\mathbf{x})$. **Right:** Example conditional distributions $\mathbb{P}(\mathbf{y}|\mathbf{x} = \cdot)$ for different conditioning values.

learning approach of Secs. 2 and 4, regress the CCDF of each dimension of $\mathbf{y}$ given $\mathbf{x}$. Thus, they can estimate conditional quantiles for any confidence interval via the quantile estimation algorithm from the CCDF described in Kostic et al. [25] without retraining. See details in Fig. 9.

Evaluation metrics: coverage and set size Let $\mathbb{C}_{1-\alpha}(\mathbf{x}) \subseteq \mathbb{R}^d$ denote a *prediction set* of nominal level $(1-\alpha)$ produced by a conditional quantile regression model for the response $\mathbf{y} \in \mathbb{R}^d$ given the covariate $\mathbf{x} \in \mathbb{R}^p$. In all experiments we assess two complementary metrics.

- **Coverage.** The conditional *coverage* of $\mathbb{C}_{1-\alpha}$ is the probability that the true response is captured by the predicted region,

$$\mathrm{c}_{1-\alpha}(\mathbf{x}) := \mathbb{P}\big(\mathbf{y} \in \mathbb{C}_{1-\alpha}(\mathbf{x}) \mid \mathbf{x}\big), \qquad \text{with the target } \ \mathrm{c}_{1-\alpha}(\mathbf{x}) \approx 1-\alpha \ \ \forall \mathbf{x}. \tag{30}$$

  In practice we report the *marginal* coverage $\widehat{\mathbb{E}}_{\mathbf{x}}[\mathrm{c}_{1-\alpha}(\mathbf{x})]$, estimated on a large held-out sample; values above (resp. below) $1-\alpha$ indicate over- (resp. under-) coverage.

- **Relaxed Coverage (r-Coverage).** The conditional *relaxed coverage* of $\mathbb{C}_{1-\alpha}$ is defined as the probability that each scalar component of the response lies within its corresponding predicted confidence interval. Formally, if $\mathbf{y} = [\mathrm{y}_1, \dots, \mathrm{y}_d]$ and $\mathbb{C}_{1-\alpha}(\mathbf{x})$ has corresponding marginal intervals $\mathbb{C}^{(i)}_{1-\alpha}(\mathbf{x})$ for $i \in \{1, \dots, d\}$, then

$$\mathrm{rc}_{1-\alpha}(\mathbf{x}) := \prod_{i=1}^{d} \mathbb{P}\Big(\mathrm{y}_i \in \mathbb{C}^{(i)}_{1-\alpha}(\mathbf{x}) \,\Big|\, \mathbf{x}\Big), \tag{31}$$

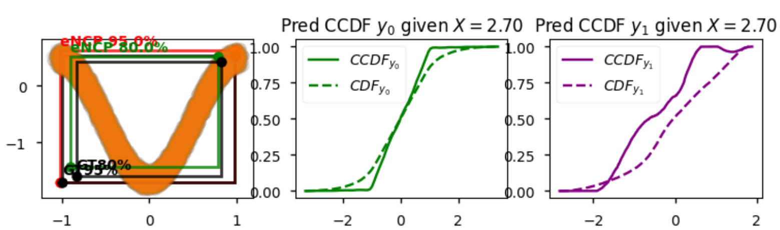

Figure 9: Prediction of the 80% and 95% confidence intervals for the random variable $\mathbf{y}$ in experiment App. G.3 using the proposed eNCP model. The model estimates the CCDF by discretizing each dimension of $\mathbf{y} = [\mathrm{y}_1, \mathrm{y}_2]$ into 100 bins and computing the conditional probabilities $\mathbb{P}(\mathrm{y}_i \in \mathbb{A}_n | \mathbf{x} = \cdot) := [\mathsf{E}_{\mathrm{y}|\mathrm{x}} \mathbb{1}_{\mathbb{A}_n}](\cdot)$ for all $n \in [100]$ based on the learned conditional expectation operator $\kappa_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y})$ (see Sec. 5). Here, $\mathbb{A}_n$ comprises the bins from the 0-th to the n-th. This yields the estimated CCDF for y_1 (center) and y_2 (right) at $\mathbf{x} = 2.7$. The CCDFs can then be used to estimate upper and lower quantiles for any confidence interval [25]. In practice, the eNCP model regresses 2×100 variables in a single forward pass. Thus, the final layer of the conditional quantile regression model is a linear layer of size $r \times (2 \times 100)$, where r is the number of features in the $\mathbf{y}$ representation (see Sec. 2).

with the target $\mathrm{rc}_{1-\alpha}(\mathbf{x}) \approx 1-\alpha$ for all $\mathbf{x}$. As with coverage, we report the *marginal* relaxed coverage $\widehat{\mathbb{E}}_{\mathbf{x}}[\mathrm{rc}_{1-\alpha}(\mathbf{x})]$.

- **Set size.** To quantify how informative the region is, we measure its *size* (volume) under the Lebesgue measure λ^d:

$$\mathrm{Size}_{1-\alpha}(\mathbf{x}) := \mathrm{vol}\big(\mathbb{C}_{1-\alpha}(\mathbf{x})\big). \tag{32}$$

  Smaller sets correspond to sharper uncertainty estimates, provided the required coverage is met. For multidimensional responses the volume is expressed in the natural units of $\mathbb{R}^d$; for $d=1$ it reduces to the interval length. As with coverage, we report the marginal expectation $\widehat{\mathbb{E}}_{\mathbf{x}}[\mathrm{Size}_{1-\alpha}(\mathbf{x})]$ so that models can be compared fairly across the entire input distribution.

G.3.1 Synthetic benchmark

The goal of these experiments is to learn the conditional distributions $\mathbb{P}(\mathbf{y} \mid \mathbf{x} = \cdot)$ for a bivariate random variable $\mathbf{y} = [\mathrm{y}_0, \mathrm{y}_1] \in \mathbb{R}^2$ given a scalar covariate $\mathbf{x} \in \mathbb{R}$. Following Feldman et al. [43], the covariate is sampled uniformly: $\mathbf{x} \sim \mathrm{Unif}\big(0.8,\ 3.2\big)$, and the response variable $\mathbf{y}$ is produced by a non-linear transformation of auxiliary latent variables (see Fig. 8):

$$\begin{aligned} \mathrm{y}_0 &= \frac{\mathrm{z}}{\beta\,\mathbf{x}} \ + \ \mathrm{r}\cos\phi, & \mathrm{z} &\sim \mathrm{Unif}(-\pi,\pi), \\ & & \phi &\sim \mathrm{Unif}(0, 2\pi), \\ \mathrm{y}_1 &= \tfrac{1}{2}\big(-\cos\mathrm{z} + 1\big) \ + \ \mathrm{r}\sin\phi \ + \ \sin\mathbf{x}, & \mathrm{r} &\sim \mathrm{Unif}(-0.1, 0.1). \end{aligned}$$

Here, $\beta > 0$ is a scaling constant.

The additive perturbation $r(\cos\phi, \sin\phi)$ yields heteroskedastic, anisotropic noise, whereas the $\frac{1}{2}(-\cos z + 1)$ and $\sin\mathbf{x}$ terms introduce strong non-monotonicity and interaction effects between $\mathbf{x}$ and $\mathbf{y}$. As a result, the conditional quantile functions $\mathbf{x} \mapsto q_\tau(\mathbf{x})$ are highly non-linear, making this dataset an ideal low-dimensional experiment for conditional quantile regression methods.

Results The experiment results are depicted in Fig. 11. Where the NCP and eNCP models outperform the baseline CQR model in terms of both coverage and set size. Furthermore, Fig. 10 illustrates the basis functions learned by the NCP and eNCP models for the random variable $\mathbf{y} = [\mathrm{y}_0, \mathrm{y}_1]$. In contrast to the standard NCP model, the eNCP model incorporates symmetry priors, enabling a clean separation of its latent representation into two orthogonal subspaces: one corresponding to $\mathbb{C}_2$-invariant functions and the other to functions that change sign under reflection.

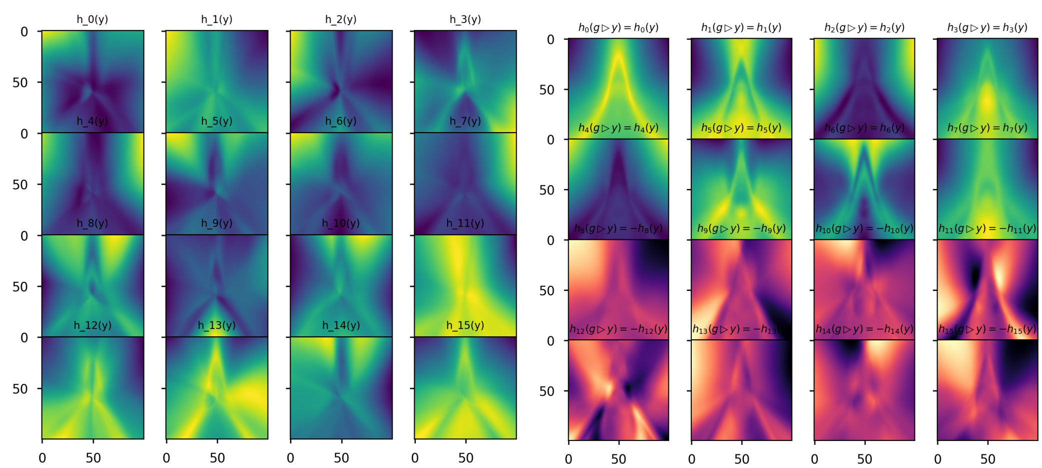

Figure 10: **Left:** Learned basis functions from the NCP model for $\mathbf{y} = [\mathrm{y}_0, \mathrm{y}_1]$. **Right:** Learned basis functions from the eNCP model for $\mathbf{y}$. The marginal distribution of $\mathbf{y}$ exhibits reflection symmetry $g_r \triangleright_{\mathcal{Y}} \mathbf{y} = [-\mathrm{y}_0, \mathrm{y}_1]$ under $\mathbb{G} = \mathbb{C}_2$. Incorporating this prior, the eNCP model decomposes its latent space as $\mathcal{F}_{\mathbf{y}} = \mathcal{F}_{\mathbf{y}}^{\mathrm{inv}} \oplus \mathcal{F}_{\mathbf{y}}^{(2)}$, with the first subspace capturing $\mathbb{C}_2$-invariant functions and the second capturing those that change sign under reflection. The orthogonality of these subspaces allows independent optimization of the basis functions.

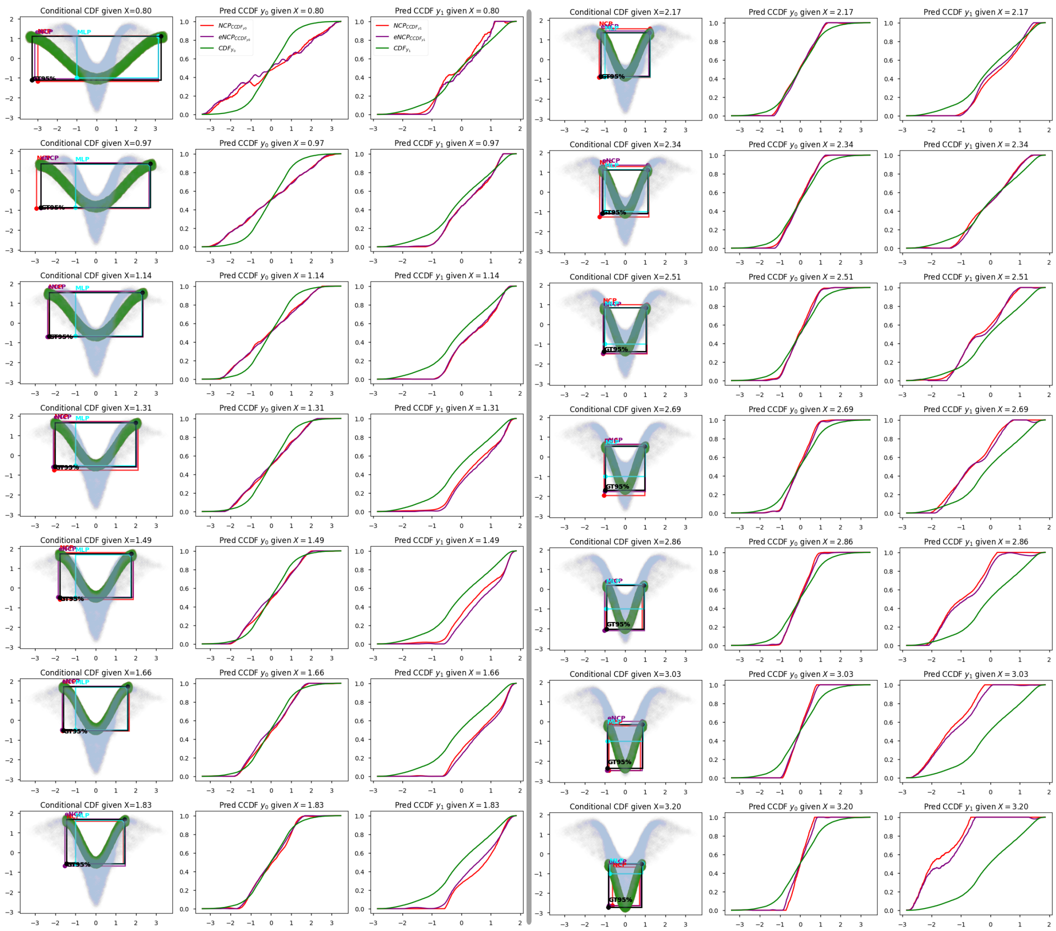

Figure 11: Results of a synthetic experiment in uncertainty quantification comparing CQR, NCP, and eNCP models. The task, originally proposed by Feldman et al. [43], is to predict the 95% confidence intervals of a random variable $\mathbf{y} \in \mathbb{R}^2$ conditioned on a scalar random variable $\mathbf{x} \in \mathbb{R}$. The conditional distributions $\mathbb{P}(\mathbf{y}|\mathbf{x} = \cdot)$ are shown in the left and fourth columns for different conditioning values, while the second-third and fifth-sixth columns display the CCDF predicted by the eNCP and NCP models, respectively. The CQR model directly regresses the upper and lower quantiles for each dimension of $\mathbf{y}$ and must be retrained if the confidence interval probability changes. In contrast, since the NCP and eNCP models estimate the CCDF for each dimension, these predictions can be easily adapted to any confidence interval probability by simply changing the threshold value.

	Validation			Test		
	r-Coverage ↑	Coverage ↑	Set Size ↓	r-Coverage ↑	Coverage ↑	Set Size ↓
eNCP	99.3±0.0%	94.1±0.4%	$2.4{\pm}0.4{\times}10^{10}$	99.5±0.1%	95.0±0.4%	$4.3{\pm}3.6{\times}10^{9}$
NCP	96.4±0.0%	56.9±0.1%	$3.9{\pm}4.5{\times}10^{10}$	99.5±0.0%	56.9±0.3%	$2.6{\pm}1.4{\times}10^{10}$
eCQR	70.7±0.6%	7.3±1.7%	$3.7{\pm}2.6{\times}10^{8}$	84.2±0.7%	6.7±1.2%	$1.7{\pm}1.7{\times}10^{7}$
CQR	67.6±1.8%	7.6±0.4%	$2.5{\pm}2.4{\times}10^{9}$	80.5±3.7%	8.5±0.9%	$1.4{\pm}0.1{\times}10^{8}$

Table 4: Validation and test set metrics for the prediction of 95% confidence intervals on observables of a quadruped robot traversing rough terrains (see App. G.3.2). Model performance is evaluated using three metrics: **(i)** relaxed coverage (r-Coverage) (Eq. (31)), **(ii)** coverage (Eq. (30)), and **(iii)** set size (Eq. (32)). The best results are highlighted in blue. Note that although the confidence interval volumes (set size) of the eCQR and CQR models are significantly smaller than those of the NCP and eNCP models, the former fail to achieve the expected 95% coverage on both the validation and test sets. In contrast, the eNCP model attains the best overall coverage, proving its effectiveness for uncertainty quantification. Importantly, the eNCP and NCP models can be adjusted, **without retraining**, to provide confidence intervals for any desired coverage level, whereas the CQR and eCQR models must be retrained for each new level.

G.3.2 Uncertainty quantification in quadruped legged locomotion

We test how well conditional-quantile models can recover the conditional 95% confidence regions of three physically meaningful observables produced by a simulated AlienGo quadruped walking over rough terrain (see Fig. 1) under varying friction coefficients. The dataset was collected using the Quadruped-PyMPC simulation framework and model predictive controller from [75].

The observables for which state-dependent uncertainty estimates are desired are $\boldsymbol{y}_t = [U_t,\ T_t,\ \boldsymbol{\tau}_t^{\text{grf}}]^\top$, with each component defined as follows:

- **$\mathbb{G}$-invariant Kinetic Energy.** $T(\boldsymbol{q}, \dot{\boldsymbol{q}}) = \frac{1}{2}\, \dot{\boldsymbol{q}}^\top\, M(\boldsymbol{q})\, \dot{\boldsymbol{q}} \in \mathbb{R}$, where $M(\boldsymbol{q})$ is the configuration-dependent inertia matrix. Noise is introduced through sensor measurement errors on the robot's degree of freedom (DoF) position $\boldsymbol{q} \in \mathbb{R}^{12}$ and velocity $\dot{\boldsymbol{q}} \in \mathbb{R}^{12}$.
- **$\mathbb{G}$-invariant Instantaneous Mechanical Work.** $U(\boldsymbol{q}, \dot{\boldsymbol{q}}, \boldsymbol{\tau}) \in \mathbb{R}$, representing the instantaneous mechanical work exerted or absorbed by the robot. This quantity depends on the actuator torques (typically measured with noisy, biased sensors) as well as the external forces (e.g. gravity, contact forces) that are not reliably measurable due to unobserved terrain parameters.
- **$\mathbb{G}$-equivariant Ground-Reaction Forces $\boldsymbol{\tau}_{\text{grf}} \in \mathbb{R}^{12}$**, a fundamental quantity in quadruped control, whose reliable estimation and uncertainty quantification are critical for downstream tasks in robotics [41, 76].

The observables of interest are predicted using a suit of onboard proprioceptive sensory signals available at time t:

$$\boldsymbol{x}_t \;=\; \Big[\boldsymbol{q}_t,\ \dot{\boldsymbol{q}}_t,\ \boldsymbol{a}_t,\ \boldsymbol{v}_t,\ \boldsymbol{v}_{t,\text{err}},\ \boldsymbol{\omega}_t,\ \boldsymbol{\omega}_{t,\text{err}},\ \boldsymbol{g}_t,\ \dot{\boldsymbol{p}}_{t,\text{feet}},\ \boldsymbol{\tau}_t^{\text{cmd}}\Big]^\top.$$

Specifically, $\boldsymbol{q}_t \in \mathbb{R}^{n_q}$ and $\dot{\boldsymbol{q}}_t \in \mathbb{R}^{n_q}$ are the joint positions and velocities, respectively; $\boldsymbol{a}_t \in \mathbb{R}^3$ is the linear acceleration of the robot's base frame measured by the IMU; $\boldsymbol{v}_t \in \mathbb{R}^3$ is the base linear velocity, while $\boldsymbol{v}_{t,\text{err}} \in \mathbb{R}^3$ the command error base linear velocity; $\boldsymbol{\omega}_t \in \mathbb{R}^3$ and $\boldsymbol{\omega}_{t,\text{err}} \in \mathbb{R}^3$ are the base angular velocity and its command error; $\boldsymbol{g}_t \in \mathbb{R}^3$ is the gravity vector expressed in the base frame; $\dot{\boldsymbol{p}}_{t,\text{feet}} \in \mathbb{R}^{12}$ stacks the linear velocities of the four feet (three components each); and $\boldsymbol{\tau}_t^{\text{cmd}} \in \mathbb{R}^{n_q}$ contains the commanded joint torques.

Hence we design the experiments to compare models of similar footprint in number of parameters, while the loss used for training differs between the NCP and eNCP models w.r.t to the CQR and eCQR models.

NN architectures We configure all models considered eNCP, NCP, eCQR, and CQR to have an inference-time NN architecture of the similar footprint. The backbone of all architectures is a standard multilayer perceptron consisting of three hidden layers, each with 512 units, followed by a final hidden layer containing 128 units. This final layer serves to encode the feature vector r for the NCP and eNCP models. Crucially, since $\mathbb{G}$-equivariance enforces weight sharing in the NN architecture the encoder NN for eNCP, eCQR have $\times 2$ less parameters than their symmetry-agnostic counterparts.

Results. Given sensory input $\mathbf{x}$, the model predicts a set $\mathbb{C}_{0.95}(\mathbf{x}) \subseteq \mathbb{R}^{14}$ satisfying $\mathbb{P}(\mathbf{y} \in \mathbb{C}_{0.95}(\mathbf{x}) \mid \mathbf{x}) \approx 0.95$, while minimizing its volume $\widehat{\mathbb{E}}_{\mathbf{x}}[\mathrm{vol}(\mathbb{C}_{0.95}(\mathbf{x}))]$. Empirically high coverage implies that the true $\mathbb{G}$-invariant kinetic energy, instantaneous mechanical work, and the $\mathbb{G}$-equivariant 12-dimensional ground-reaction forces lie within the predicted confidence set. In contrast, relaxed coverage (r-Coverage) quantifies the reliability of the estimates on a per-dimension basis. Tab. 4 summarizes the validation and test results for the eNCP, NCP, CQR, and eCQR models, and Fig. 12 illustrates a trajectory of GRF and their respective 90% confidence intervals for each model. Both CQR and eCQR tend to produce confidence intervals of smaller volume but fail to achieve the desired coverage on the testing set, implying that the models' confidence intervals are not reliable and require further calibration through retraining or conformal calibration [43]. In contrast, the eNCP model achieves the desired coverage on the test set while producing confidence intervals of larger volume, hence yielding reliable confidence intervals.

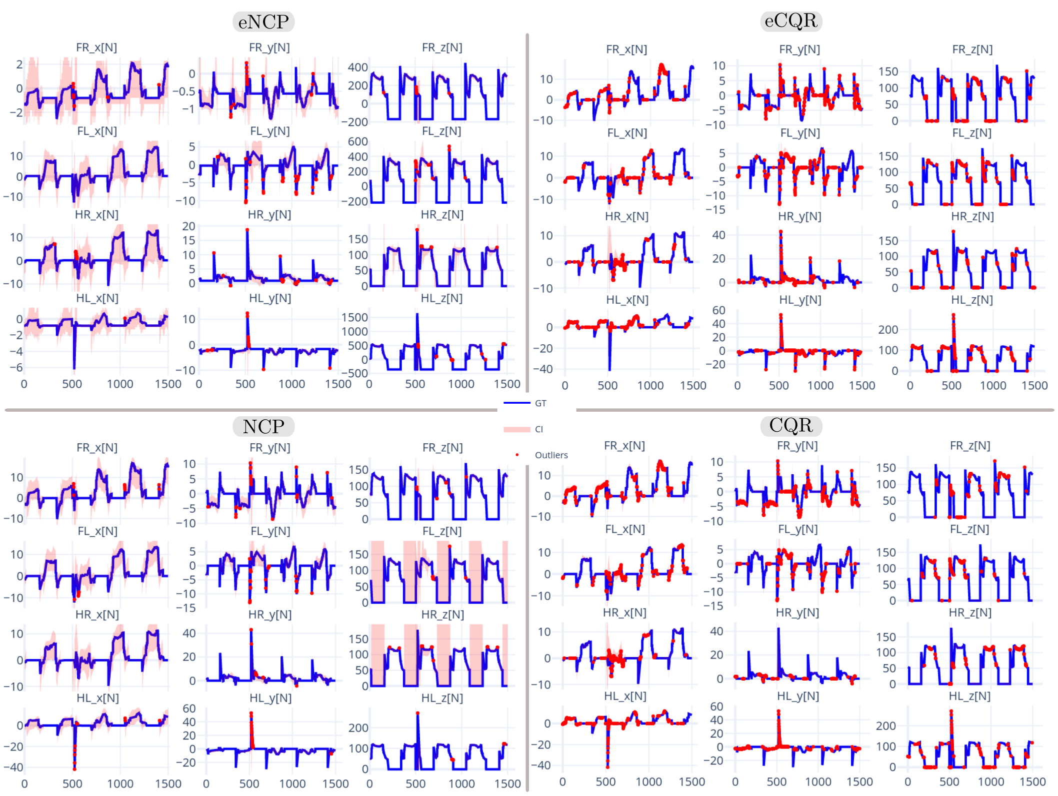

Figure 12: Prediction of 90% confidence intervals (CI) for the ground-reaction forces $\boldsymbol{\tau}_{\text{grf}} \in \mathbb{R}^{12}$ of a quadruped robot on rough terrain with varying friction. We compare the eNCP, NCP, eCQR, and CQR models based on relaxed coverage and set size (see Tab. 4). CIs are computed for each leg—front-right (FR), front-left (FL), hind-right (HR), and hind-left—along the x, y, and z axes. Forces outside the CI are highlighted in red, while those within appear in blue. Terrain variations cause significant variability in the x and y components due to differences in surface orientation and friction, whereas the z component is mainly influenced by local height changes that alter contact timing and produce short-duration high-impact forces.

H Conditional probability modeling via the conditional expectation operator

This section introduces the modelling of conditional probabilities for two random variables via the **conditional expectation operator**. Our goal is to understand conditional expectation from an operator-theoretic perspective. We begin by describing the marginal, joint, and conditional probabilities of the random variables within a measure-theoretic framework. This discussion extends the exposition of Kostic et al. [37].

Given two random variables $(\mathbf{x}, \mathbf{y})$ taking values in the measure spaces $(\mathcal{X}, \Sigma_{\mathcal{X}}, P_{\mathbf{x}})$ and $(\mathcal{Y}, \Sigma_{\mathcal{Y}}, P_{\mathbf{y}})$, we have that the marginal probability of any set $\mathbb{A} \in \Sigma_{\mathcal{X}}$ and $\mathbb{B} \in \Sigma_{\mathcal{Y}}$ are given by

$$\mathbb{P}(\mathbf{x} \in \mathbb{A}) = \int_{\mathcal{X}} \mathbb{1}_{\mathbb{A}}(\boldsymbol{x}) P_{\mathbf{x}}(d\boldsymbol{x}) = \int_{\mathbb{A}} P_{\mathbf{x}}(d\boldsymbol{x}) \quad \text{and} \quad \mathbb{P}(\mathbf{y} \in \mathbb{B}) = \int_{\mathcal{Y}} \mathbb{1}_{\mathbb{B}}(\boldsymbol{y}) P_{\mathbf{y}}(d\boldsymbol{y}) = \int_{\mathbb{B}} P_{\mathbf{y}}(d\boldsymbol{y}), \tag{33}$$

where $\mathbb{1}_{\mathbb{A}} \in \mathcal{L}^2_{\mathbf{x}}$ and $\mathbb{1}_{\mathbb{B}} \in \mathcal{L}^2_{\mathbf{y}}$ denote the characteristic functions of sets $\mathbb{A}$ and $\mathbb{B}$, respectively.

Furthermore, under the reasonable assumption that the joint probability measure is absolutely continuous w.r.t to the product of the marginals $P_{\mathbf{xy}} \ll P_{\mathbf{x}} \times P_{\mathbf{y}}$, we have that there exist a Radon-Nikodym derivative $\kappa : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ such that $P_{\mathbf{xy}}(d\boldsymbol{x}, d\boldsymbol{y}) = \kappa(\boldsymbol{x}, \boldsymbol{y}) P_{\mathbf{x}}(d\boldsymbol{x}) P_{\mathbf{y}}(d\boldsymbol{y})$. Note that κ is a kernel function that pointwise deforms the product of the marginals to produce the joint distribution [29] (see Fig. 3). This kernel function enable us to express the joint probability by:

$$\mathbb{P}(\mathbf{x} \in \mathbb{A}, \mathbf{y} \in \mathbb{B}) = \int_{\mathcal{X} \times \mathcal{Y}} \mathbb{1}_{\mathbb{A}}(\boldsymbol{x}) \mathbb{1}_{\mathbb{B}}(\boldsymbol{y}) \underbrace{\kappa(\boldsymbol{x}, \boldsymbol{y}) P_{\mathbf{y}}(d\boldsymbol{y}) P_{\mathbf{x}}(d\boldsymbol{x})}_{P_{\mathbf{xy}}(d\boldsymbol{x}, d\boldsymbol{y})} = \int_{\mathbb{A} \times \mathbb{B}} k(\boldsymbol{x}, \boldsymbol{y}) P_{\mathbf{x}}(d\boldsymbol{x}) P_{\mathbf{y}}(d\boldsymbol{y}). \tag{34}$$

Furthermore, given that $\mathbb{P}(\mathbf{y} \in \mathbb{B} | \mathbf{x} \in \mathbb{A}) = {}^{\mathbb{P}(\mathbf{x} \in \mathbb{A},\, \mathbf{y} \in \mathbb{B})} / {}_{\mathbb{P}(\mathbf{x} \in \mathbb{A})}$, the conditional probability of any set $\mathbb{B} \in \Sigma_{\mathcal{Y}}$ given a value of the random variable $\mathbf{x} = \boldsymbol{x}$ is given by:

$$\mathbb{P}(\mathbf{y} \in \mathbb{B} | \mathbf{x} = \boldsymbol{x}) = \int_{\mathcal{Y}} \mathbb{1}_{\mathbb{B}}(\boldsymbol{y}) P_{\mathbf{y}|\mathbf{x}}(d\boldsymbol{y} | \boldsymbol{x}) = \int_{\mathcal{Y}} \mathbb{1}_{\mathbb{B}}(\boldsymbol{y}) \kappa(\boldsymbol{x}, \boldsymbol{y}) P_{\mathbf{y}}(d\boldsymbol{y}) = \int_{\mathbb{B}} \kappa(\boldsymbol{x}, \boldsymbol{y}) P_{\mathbf{y}}(d\boldsymbol{y}), \tag{35}$$

where $P_{\mathbf{y}|\mathbf{x}} : \Sigma_{\mathcal{Y}} \times \mathcal{X} \mapsto [0, 1]$ denotes the **conditional probability measure**. This is a well-defined probability measure considering that:

$$\mathbb{P}(\mathbf{x} \in \mathbb{A}) := \mathbb{P}(\mathbf{x} \in \mathbb{A}, \mathbf{y} \in \mathcal{Y}) = \int_{\mathbb{A}} \underbrace{\left(\int_{\mathcal{Y}} \kappa(\boldsymbol{x}, \boldsymbol{y}) P_{\mathbf{y}}(d\boldsymbol{y}) \right)}_{\mathbb{E} P_{\mathbf{y}|\mathbf{x}}(d\mathbf{y} | \mathbf{x} = \boldsymbol{x}) = 1 \ \ \forall \boldsymbol{x} \in \mathcal{X}} P_{\mathbf{x}}(d\boldsymbol{x}) = \int_{\mathbb{A}} P_{\mathbf{x}}(d\boldsymbol{x}).$$

The operator perspective Every measurable function $h \in \mathcal{L}^2_{\mathbf{y}}$ can be approximated by simple functions—that is, as a combination of characteristic functions on measurable sets: $h(\cdot) \approx \sum_{i \in \mathbb{N}} \beta_i \mathbb{1}_{\mathbb{A}_i}(\cdot)$. Thus, Eq. (35) is a special case of the more general problem of approximating the conditional expectation of any function $h \in \mathcal{L}^2_{\mathbf{y}}$ given $\mathbf{x}$. This conditional expectation is captured by the action of a linear integral operator:

Definition H.1 (Conditional expectation operator)**.** *Let $(\mathbf{x}, \mathbf{y})$ be two random variables defined on the measure spaces $(\mathcal{X}, \Sigma_{\mathcal{X}}, P_{\mathbf{x}})$ and $(\mathcal{Y}, \Sigma_{\mathcal{Y}}, P_{\mathbf{y}})$, respectively, and let $\mathcal{L}^2_{\mathbf{x}}$ and $\mathcal{L}^2_{\mathbf{y}}$ denote the corresponding spaces of square-integrable functions. The conditional expectation operator $\mathsf{E}_{\mathbf{y}|\mathbf{x}} : \mathcal{L}^2_{\mathbf{y}} \to \mathcal{L}^2_{\mathbf{x}}$ is the linear integral operator—defined via the PMD Radon–Nikodym derivative $\kappa(\boldsymbol{x}, \boldsymbol{y}) = {}^{P_{\mathbf{xy}}(d\boldsymbol{x}, d\boldsymbol{y})} / {}_{P_{\mathbf{x}}(d\boldsymbol{x})\, P_{\mathbf{y}}(d\boldsymbol{y})}$ —which acts on any function $h \in \mathcal{L}^2_{\mathbf{y}}$ by computing its conditional expectation:*

$$[\mathsf{E}_{\mathbf{y}|\mathbf{x}} h](\boldsymbol{x}) = \mathbb{E}[h(\mathbf{y}) | \mathbf{x} = \boldsymbol{x}] := \int_{\mathcal{Y}} h(\boldsymbol{y}) P_{\mathbf{y}|\mathbf{x}}(d\boldsymbol{y} | \boldsymbol{x}) = \int_{\mathcal{Y}} h(\boldsymbol{y}) \frac{P_{\mathbf{xy}}(d\boldsymbol{y}, \boldsymbol{x})}{P_{\mathbf{x}}(d\boldsymbol{x})} = \int_{\mathcal{Y}} h(\boldsymbol{y}) \kappa(\boldsymbol{x}, \boldsymbol{y}) P_{\mathbf{y}}(d\boldsymbol{y}).$$

From a learning perspective, approximating the conditional expectation operator sufficiently well for a relevant set of functions in $\mathcal{L}^2_{\mathbf{y}}$ implies that we can approximate the conditional probability measure of any set $\mathbb{A} \in \Sigma_{\mathcal{Y}}$. This enables both regression *and* uncertainty quantification applications with a single model (see Eq. (2)).

I Background on group and representation theory

Group actions and representations This section provides a concise overview of the fundamental concepts in group and representation theory, which are used to define the symmetries of the random variables we consider in this work. For a comprehensive background on these topics in finite-dimensional vector spaces, see Weiler et al. [17]; for the infinite-dimensional case, consult Knapp [77]. These concepts will be referenced as needed in the main text. To begin, we define a group as an abstract mathematical object.

Definition I.1 (Group)**.** *A group is a set $\mathbb{G}$, endowed with a binary composition operator defined as:*

$$\begin{array}{rccc} (\circ): & \mathbb{G}\times\mathbb{G} & \longrightarrow & \mathbb{G} \\ & (g_1,g_2) & \longrightarrow & g_1\circ g_2, \end{array} \tag{36a}$$

such that the following axioms hold:

$$\begin{aligned} \textit{Associativity:} &\quad (g_1\circ g_2)\circ g_3 = g_1\circ(g_2\circ g_3), \quad \forall\, g_1,g_2,g_3\in\mathbb{G}, && \text{(36b)}\\ \textit{Identity:} &\quad \exists\, e\in\mathbb{G} \text{ such that } e\circ g = g = g\circ e, \quad \forall\, g\in\mathbb{G}, && \text{(36c)}\\ \textit{Inverses:} &\quad \forall\, g\in\mathbb{G},\ \exists\, g^{-1}\in\mathbb{G} \text{ such that } g\circ g^{-1} = e = g^{-1}\circ g. && \text{(36d)} \end{aligned}$$

We are primarily interested in symmetry groups, i.e., groups of transformations acting on a set $\mathcal{X}$. Each transformation is a bijection that leaves a fundamental property invariant. For example, if $\mathcal{X}$ represents states of a dynamical system, the invariant property is the state energy (see Fig. 7); if $\mathcal{X}$ is a data space, the preserved quantity is typically the probability density/distribution (see Fig. 3).

Definition I.2 (Group action on a set [17])**.** *Let $\mathcal{X}$ be a set endowed with symmetry group $\mathbb{G}$. The (left) group action of the group $\mathbb{G}$ on the set $\mathcal{X}$ is a map:*

$$\begin{array}{rccc} (\triangleright): & \mathbb{G}\times\mathcal{X} & \longrightarrow & \mathcal{X} \\ & (g,\boldsymbol{x}) & \longrightarrow & g\triangleright\boldsymbol{x} \end{array} \tag{37a}$$

that is compatible with the group composition and identity element $e\in\mathbb{G}$, in the sense that:

$$\begin{aligned} \textit{Identity:} &\quad e\triangleright\boldsymbol{x} = \boldsymbol{x}, && \forall\, \boldsymbol{x}\in\mathcal{X} && \text{(37b)}\\ \textit{Associativity:} &\quad (g_1\circ g_2)\triangleright\boldsymbol{x} = g_1\triangleright(g_2\triangleright\boldsymbol{x}), && \forall\, g_1,g_2\in\mathbb{G},\forall\,\boldsymbol{x}\in\mathcal{X}. && \text{(37c)} \end{aligned}$$

We are primarily interested in studying symmetry transformations on sets with a vector space structure. In most practical cases, the group action on a vector space is linear, allowing symmetry transformations to be represented as linear invertible maps. These maps can be expressed in matrix form once a basis for the space is chosen.

Definition I.3 (Linear group representation)**.** *Let $\mathcal{X}$ be a vector space endowed with symmetry group $\mathbb{G}$. A* linear representation *of $\mathbb{G}$ on $\mathcal{X}$ is a map, denoted by $\boldsymbol{\rho}_{\mathcal{X}}$, between symmetry transformation and invertible linear maps on $\mathcal{X}$ (i.e., elements of the general linear group $\mathbb{GL}(\mathcal{X})$):*

$$\begin{array}{rccc} \boldsymbol{\rho}_{\mathcal{X}}: & \mathbb{G} & \longrightarrow & \mathbb{GL}(\mathcal{X}) \\ & g & \longrightarrow & \boldsymbol{\rho}_{\mathcal{X}}(g), \end{array} \tag{38a}$$

such that the following properties hold:

$$\begin{aligned} \textit{composition}: \boldsymbol{\rho}_{\mathcal{X}}(g_1\circ g_2) &= \boldsymbol{\rho}_{\mathcal{X}}(g_1)\boldsymbol{\rho}_{\mathcal{X}}(g_2), && \forall\, g_1,g_2\in\mathbb{G}, && \text{(38b)}\\ \textit{inversion}: \boldsymbol{\rho}_{\mathcal{X}}(g^{-1}) &= \boldsymbol{\rho}_{\mathcal{X}}(g)^{-1}, && \forall\, g\in\mathbb{G}. && \text{(38c)}\\ \textit{identity}: \boldsymbol{\rho}_{\mathcal{X}}(g\circ g^{-1}) &= \boldsymbol{\rho}_{\mathcal{X}}(e) = \boldsymbol{I}, && && \text{(38d)} \end{aligned}$$

Whenever the vector space is of finite dimension $n<\infty$, linear maps admit a matrix form $\boldsymbol{\rho}_{\mathcal{X}}(g)\in\mathbb{R}^{n\times n}$, once a basis set $\mathbb{I}_{\mathcal{X}}$ for the vector space $\mathcal{X}$ is chosen. In this case, Eqs. (38b) to (38d) show how the composition and inversion of symmetry transformations translate to matrix multiplication and inversion, respectively. Moreover, $\boldsymbol{\rho}_{\mathcal{X}}$ allows to express a (linear) group action (Def. I.2) as a matrix-vector multiplication:

$$\begin{array}{rccc} (\triangleright): & \mathbb{G}\times\mathcal{X} & \longrightarrow & \mathcal{X} \\ & (g,\boldsymbol{x}) & \longrightarrow & g\triangleright\boldsymbol{x} := \boldsymbol{\rho}_{\mathcal{X}}(g)\boldsymbol{x}. \end{array} \tag{38e}$$

Since the matrix form of linear maps depends on the choice of basis, we can relate different matrix representations of the same linear map through changes of basis. This leads us to the concept of equivalent group representations.

Definition I.4 (Equivalent group representations)**.** *Let $\mathcal{X}$ be a vector space endowed with symmetry group $\mathbb{G}$, and let $\boldsymbol{\rho}'_{\mathcal{X}}$ and $\boldsymbol{\rho}_{\mathcal{X}}$ be two group representations of $\mathbb{G}$ on $\mathcal{X}$. They are said to be* equivalent*, denoted by $\boldsymbol{\rho}'_{\mathcal{X}} \sim \boldsymbol{\rho}_{\mathcal{X}}$, if there exists a change of basis $\boldsymbol{Q} : \mathcal{X} \to \mathcal{X}$ such that*

$$\boldsymbol{\rho}'_{\mathcal{X}}(g) = \boldsymbol{Q}\boldsymbol{\rho}_{\mathcal{X}}(g)\boldsymbol{Q}^{-1}, \quad \forall\, g \in \mathbb{G}. \tag{39}$$

Equivalent representations arise when the same group action $(\triangleright) : \mathbb{G} \times \mathcal{X} \to \mathcal{X}$ is expressed in different coordinate frames or bases. For instance, let $\mathbb{A}_{\mathcal{X}}$ and $\mathbb{B}_{\mathcal{X}}$ be two bases for $\mathcal{X} = span(\mathbb{A}_{\mathcal{X}}) = span(\mathbb{B}_{\mathcal{X}})$, and let $\boldsymbol{Q}^{\mathbb{B}}_{\mathbb{A}} : \mathcal{X} \to \mathcal{X}$ denote the change of basis from $\mathbb{A}_{\mathcal{X}}$ to $\mathbb{B}_{\mathcal{X}}$, so that $\boldsymbol{x}^{\mathbb{B}} = \boldsymbol{Q}^{\mathbb{B}}_{\mathbb{A}}\boldsymbol{x}^{\mathbb{A}}$ for all $\boldsymbol{x}^{\mathbb{A}} \in \mathcal{X}$. Then the group action admits equivalent representations, $\boldsymbol{\rho}^{\mathbb{A}}_{\mathcal{X}} \sim \boldsymbol{\rho}^{\mathbb{B}}_{\mathcal{X}}$, since

$$\begin{aligned}
g \triangleright \boldsymbol{x}^{\mathbb{B}} &:= \boldsymbol{Q}^{\mathbb{B}}_{\mathbb{A}}(g \triangleright \boldsymbol{x}^{\mathbb{A}}), \qquad \forall g \in \mathbb{G}, \\
\boldsymbol{\rho}^{\mathbb{B}}_{\mathcal{X}}(g)\boldsymbol{x}^{\mathbb{B}} &= \boldsymbol{Q}^{\mathbb{B}}_{\mathbb{A}}\big(\boldsymbol{\rho}^{\mathbb{A}}_{\mathcal{X}}(g)\boldsymbol{x}^{\mathbb{A}}\big) = \Big(\boldsymbol{Q}^{\mathbb{B}}_{\mathbb{A}}\boldsymbol{\rho}^{\mathbb{A}}_{\mathcal{X}}(g){\boldsymbol{Q}^{\mathbb{B}}_{\mathbb{A}}}^{-1}\Big)\boldsymbol{x}^{\mathbb{B}}, \\
\boldsymbol{\rho}^{\mathbb{B}}_{\mathcal{X}}(g) &= \boldsymbol{Q}^{\mathbb{B}}_{\mathbb{A}}\boldsymbol{\rho}^{\mathbb{A}}_{\mathcal{X}}(g){\boldsymbol{Q}^{\mathbb{B}}_{\mathbb{A}}}^{-1}.
\end{aligned} \tag{40}$$

To reveal the modular structure of symmetric vector spaces, we often change bases to decompose them into subspaces stable under the action of the group $\mathbb{G}$, termed $\mathbb{G}$-stable subspaces. This decomposition mirrors how a symmetry group breaks down into subgroups and is essential for analyzing and simplifying group representations. We introduce the following definition.

Definition I.5 ($\mathbb{G}$-stable and irreducible subspaces)**.** *Let $\mathcal{X}$ be a vector space endowed with a group action $(\triangleright)$ of the symmetry group $\mathbb{G}$. A subspace $\mathcal{X}' \subseteq \mathcal{X}$ is said to be* $\mathbb{G}$-stable *if the action of any group element on any vector in the subspace remains within the subspace, that is,*

$$g \triangleright \boldsymbol{x} \in \mathcal{X}', \quad \forall\, \boldsymbol{x} \in \mathcal{X}' \subseteq \mathcal{X}, \forall\, g \in \mathbb{G}.$$

If the only $\mathbb{G}$-stable subspaces of $\mathcal{X}$ are $\{\boldsymbol{0}\}$ and $\mathcal{X}$ itself, then $\mathcal{X}$ is a irreducible $\mathbb{G}$-stable space*.*

Decomposing symmetric vector spaces into $\mathbb{G}$-stable subspaces corresponds to decomposing the group representation associated with $\triangleright$ into smaller representations acting on these $\mathbb{G}$-stable subspaces:

Definition I.6 (Decomposable representation)**.** *Let $\mathcal{X}$ be a vector space with a group action $(\triangleright)$ defined by the representation $\boldsymbol{\rho}_{\mathcal{X}}$ in a chosen basis $\mathbb{A}_{\mathcal{X}}$. The representation is* decomposable *if it is equivalent to a direct sum of two lower-dimensional representations, $\boldsymbol{\rho}_{\mathcal{X}} \sim \boldsymbol{\rho}_{\mathcal{X}_1} \oplus \boldsymbol{\rho}_{\mathcal{X}_2}$, where $\mathcal{X}_1$ and $\mathcal{X}_2$ are $\mathbb{G}$-stable subspaces of $\mathcal{X}$. Equivalently, there exists a change of basis $\boldsymbol{Q}^{\mathbb{B}}_{\mathbb{A}} : \mathcal{X} \to \mathcal{X}$ such that*

$$\boldsymbol{\rho}^{\mathbb{B}}_{\mathcal{X}} = \begin{bmatrix} \boldsymbol{\rho}_{\mathcal{X}_1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\rho}_{\mathcal{X}_2} \end{bmatrix} = \boldsymbol{Q}^{\mathbb{B}}_{\mathbb{A}}\boldsymbol{\rho}_{\mathcal{X}}{\boldsymbol{Q}^{\mathbb{B}}_{\mathbb{A}}}^{-1}, \ \textit{and} \quad g \triangleright \boldsymbol{x}^{\mathbb{B}} := \boldsymbol{\rho}^{\mathbb{B}}_{\mathcal{X}}(g)\boldsymbol{x}^{\mathbb{B}} = \begin{bmatrix} \boldsymbol{\rho}_{\mathcal{X}_1}(g)\boldsymbol{x}^{\mathbb{B}}_1 \\ \boldsymbol{\rho}_{\mathcal{X}_2}(g)\boldsymbol{x}^{\mathbb{B}}_2 \end{bmatrix}, \ \textit{where } \boldsymbol{Q}^{\mathbb{B}}_{\mathbb{A}}\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}^{\mathbb{B}}_1 \in \mathcal{X}_1 \\ \boldsymbol{x}^{\mathbb{B}}_2 \in \mathcal{X}_2 \end{bmatrix}$$

This shows that the decomposition $\boldsymbol{\rho}_{\mathcal{X}} \sim \boldsymbol{\rho}_{\mathcal{X}_1} \oplus \boldsymbol{\rho}_{\mathcal{X}_2}$ corresponds to splitting the vector space into $\mathbb{G}$-stable subspaces, $\mathcal{X} = \mathcal{X}_1 \oplus \mathcal{X}_2$. Moreover, if the representation is block-diagonal in some basis set $\mathbb{B}_{\mathcal{X}}$, then $\mathbb{B}_{\mathcal{X}}$ is the union of disjoint basis sets $\mathbb{B}_{\mathcal{X}_1}$ and $\mathbb{B}_{\mathcal{X}_2}$ for $\mathcal{X}_1$ and $\mathcal{X}_2$, respectively.

Definition I.7 (Irreducible representation)**.** *Let $\mathcal{X}$ be a vector space endowed with a group action $(\triangleright)$ of a symmetry group $\mathbb{G}$. A representation $\boldsymbol{\rho}_{\mathcal{X}}$ of $\mathbb{G}$ on $\mathcal{X}$ is said to be* irreducible *if it cannot be decomposed into smaller representations acting on proper $\mathbb{G}$-stable subspaces (Def. I.5). That is, the only $\mathbb{G}$-stable subspaces $\mathcal{X}' \subseteq \mathcal{X}$ are $\mathcal{X}' = \{\boldsymbol{0}\}$ and $\mathcal{X}' = \mathcal{X}$ itself.*

We have now equipped all the necessary tools to decompose symmetric vector spaces into their smallest building blocks: irreducible $\mathbb{G}$-stable subspaces.

Irreducible representations are the fundamental building blocks for all representations of the group $\mathbb{G}$. Any unitary representation can be decomposed into a direct sum of irreducible representations, analogous to the prime factorization of integers. In terms of the vector spaces on which the group acts, this decomposition of the representation corresponds to decomposing the space into $\mathbb{G}$-irreducible subspaces (Def. I.5):

Theorem I.8 (Isotypic decomposition of symmetric Hilbert spaces [77])**.** *Let $\mathbb{G}$ be a compact group and $\mathcal{H}$ a separable Hilbert space with a unitary group representation $\boldsymbol{\rho}_{\mathcal{H}} : \mathbb{G} \to \mathbb{U}(\mathcal{H})$. Then we can identify $n_{\text{iso}} \leq |\mathbb{G}|$ irreducible representations $\bar{\boldsymbol{\rho}}_k : \mathbb{G} \to \mathbb{U}(\bar{\mathcal{H}}_k)$ that allow us to decompose $\mathcal{H}$ into a sum of orthogonal subspaces, denoted* isotypic subspaces*: $\mathcal{H} = \bigoplus^{\perp}_{1 \leq k \leq n_{\text{iso}}} \mathcal{H}_k$ where each $\mathcal{H}_k = \bigoplus_{j=1}^{m_k} \mathcal{H}_{k,j}$ is the sum of at most $m_k \leq \infty$ countably many subspaces isometrically isomorphic to $\bar{\mathcal{H}}_k$.*

Isotypic decomposition and disentangled representations Whenever the symmetric vector space of interest defines a vector valued representation of some data, the isotypic decomposition of the representation space is intricately linked with the concept of *disentangled representations*

Definition I.9 (Disentangled representation (Higgins et al. [18]))**.** *A vector representation is called a disentangled representation with respect to a particular decomposition of a symmetry group into subgroups, if it decomposes into independent subspaces, where each subspace is affected by the action of a single subgroup, and the actions of all other subgroups leave the subspace unaffected.*

The *subspaces* of Def. I.9 reefer to each of the isotypic subspaces $\mathcal{H}_i$, and the symmetry subgroups refer to the effective (matrix) group encoded by each irreducible representation $\bar{\boldsymbol{\rho}}_k : \mathbb{G} \mapsto \mathbb{U}(\bar{\mathcal{H}}_k)$. Which we denote in the main body as $\mathbb{G}^{(k)}$.

I.1 Maps between symmetric vector spaces

We will frequently study and use linear and non-linear maps between symmetric vector spaces. Our focus is on maps that preserve entirely or partially the group structure of the vector spaces. These types of maps can be classified as $\mathbb{G}$-equivariant, $\mathbb{G}$-invariant maps:

Definition I.10 ($\mathbb{G}$-equivariant and $\mathbb{G}$-invariant maps)**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be two vector spaces endowed with the same symmetry group $\mathbb{G}$, with the respective group actions $\triangleright_{\mathcal{X}}$ and $\triangleright_{\mathcal{Y}}$. A map $f : \mathcal{X} \mapsto \mathcal{Y}$ is said to be $\mathbb{G}$-equivariant if it commutes with the group action, such that:*

$$\begin{aligned} g \triangleright_{\mathcal{Y}} \boldsymbol{y} = g \triangleright_{\mathcal{Y}} f(\boldsymbol{x}) = f(g \triangleright_{\mathcal{X}} \boldsymbol{x}), \quad \forall \boldsymbol{x} \in \mathcal{X}, g \in \mathbb{G}. \\ \boldsymbol{\rho}_{\mathcal{Y}}(g) f(\boldsymbol{x}) = f(\boldsymbol{\rho}_{\mathcal{X}}(g)\boldsymbol{x}) \end{aligned} \quad \iff \quad \begin{tikzcd} \mathcal{X} \arrow[r, "\triangleright_{\mathcal{X}}"] \arrow[d, "f"] & \mathcal{X} \arrow[d, "f"] \\ \mathcal{Y} \arrow[r, "\triangleright_{\mathcal{Y}}"] & \mathcal{Y} \end{tikzcd} \tag{41a}$$

A specific case of $\mathbb{G}$-equivariant maps are the $\mathbb{G}$-invariant ones, which are maps that commute with the group action and have trivial output group actions $\triangleright_{\mathcal{Y}}$ such that $\boldsymbol{\rho}_{\mathcal{Y}}(g) = \boldsymbol{I}$ for all $g \in \mathbb{G}$. That is:

$$\begin{aligned} \boldsymbol{y} = g \triangleright_{\mathcal{Y}} f(\boldsymbol{x}) = f(g \triangleright_{\mathcal{X}} \boldsymbol{x}), \quad \forall \boldsymbol{x} \in \mathcal{X}, g \in \mathbb{G}. \\ \boldsymbol{y} = \boldsymbol{\rho}_{\mathcal{Y}}(g) f(\boldsymbol{x}) = f(\boldsymbol{\rho}_{\mathcal{X}}(g)\boldsymbol{x}) \end{aligned} \quad \iff \quad \begin{tikzcd} \mathcal{X} \arrow[r, "\triangleright_{\mathcal{X}}"] \arrow[dr, "f"'] & \mathcal{X} \arrow[d, "f"] \\ & \mathcal{Y} \arrow[loop below, "\triangleright_{\mathcal{Y}}"] \end{tikzcd} \tag{41b}$$

I.2 Structure of $\mathbb{G}$-equivariant linear maps

Definition I.11 (Homomorphism, Isomorphism, and $\mathbb{G}$-equivariant linear maps)**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be two vector spaces endowed with the same symmetry group $\mathbb{G}$, with the respective group actions $\triangleright_{\mathcal{X}}: \mathbb{G} \times \mathcal{X} \mapsto \mathcal{X}$ and $\triangleright_{\mathcal{Y}}: \mathbb{G} \times \mathcal{Y} \mapsto \mathcal{Y}$. The spaces are said to be $\mathbb{G}$-homomorphic if there exists a linear map $\mathbb{A} : \mathcal{X} \mapsto \mathcal{Y}$ that commutes with the group action, such that $g \triangleright_{\mathcal{Y}} (\boldsymbol{A}\boldsymbol{x}) = \boldsymbol{A}(g \triangleright_{\mathcal{X}} \boldsymbol{x})$ for all $\boldsymbol{x} \in \mathcal{X}$. They are said to be $\mathbb{G}$-isomorphic if the linear map is invertible. Graphically, $\mathcal{X}$ and $\mathcal{Y}$*

are $\mathbb{G}$*-homomorphic or* $\mathbb{G}$*-isomorphic if the following diagrams commute:*

$$\underbrace{\begin{tikzcd} \mathcal{X} \arrow[r, "\triangleright_{\mathcal{X}}"] \arrow[d, "\boldsymbol{A}"] & \mathcal{X} \arrow[d, "\boldsymbol{A}"] \\ \mathcal{Y} \arrow[r, "\triangleright_{\mathcal{Y}}"] & \mathcal{Y} \end{tikzcd}}_{\textit{Homomorphism}} \quad \boldsymbol{A} \in \textit{Homo}_{\mathbb{G}}(\mathcal{X},\mathcal{Y}) \qquad \textit{or} \qquad \underbrace{\begin{tikzcd} \mathcal{X} \arrow[r, "\triangleright_{\mathcal{X}}"] \arrow[d, shift left, "\boldsymbol{A}"] & \mathcal{X} \arrow[d, shift left, "\boldsymbol{A}"] \\ \mathcal{Y} \arrow[r, "\triangleright_{\mathcal{Y}}"] \arrow[u, shift left, "\boldsymbol{A}^{-1}"] & \mathcal{Y} \arrow[u, shift left, "\boldsymbol{A}^{-1}"] \end{tikzcd}}_{\textit{Isomorphism}} \quad \boldsymbol{A} \in \textit{Iso}_{\mathbb{G}}(\mathcal{X},\mathcal{Y}). \tag{42}$$

Here, $\textit{Homo}_{\mathbb{G}}(\mathcal{X},\mathcal{Y})$ *denotes the space of* $\mathbb{G}$*-equivariant linear maps between* $\mathcal{X}$ *and* $\mathcal{Y}$*, and* $\textit{Iso}_{\mathbb{G}}(\mathcal{X},\mathcal{Y})$ *denotes the space of* $\mathbb{G}$*-equivariant invertible linear maps between* $\mathcal{X}$ *and* $\mathcal{Y}$*.*

Lemma I.12 (Schur's Lemma for unitary representations [77, Prop 1.5])**.** *Consider two Hilbert spaces,* $\mathcal{H}$ *and* $\mathcal{H}'$*, endowed with the irreducible unitary representations* $\bar{\boldsymbol{\rho}}_{\mathcal{H}} : \mathbb{G} \mapsto \mathbb{U}(\mathcal{H})$ *and* $\bar{\boldsymbol{\rho}}_{\mathcal{H}'} : \mathbb{G} \mapsto \mathbb{U}(\mathcal{H}')$*, respectively. Let* $\mathsf{T} : \mathcal{H} \mapsto \mathcal{H}'$ *be a linear* $\mathbb{G}$*-equivariant operator such that* $\bar{\boldsymbol{\rho}}_{\mathcal{H}'}\mathsf{T} = \mathsf{T}\bar{\boldsymbol{\rho}}_{\mathcal{H}}$*. If the irreducible representations are not equivalent, i.e.,* $\bar{\boldsymbol{\rho}}_{\mathcal{H}} \nsim \bar{\boldsymbol{\rho}}_{\mathcal{H}'}$*, then* T *is the trivial (or zero) map. Conversely, if* $\bar{\boldsymbol{\rho}}_{\mathcal{H}} \sim \bar{\boldsymbol{\rho}}_{\mathcal{H}'}$*, then* T *is a constant multiple of an isomorphism (Def. I.11). Denoting* I *as the identity operator, this can be expressed as:*

$$\bar{\boldsymbol{\rho}}_{\mathcal{H}} \nsim \bar{\boldsymbol{\rho}}_{\mathcal{H}'} \iff \qquad \boldsymbol{0}_{\mathcal{H}'} = \mathsf{T}\boldsymbol{h} \mid \forall\, \boldsymbol{h} \in \mathcal{H} \tag{43a}$$

$$\bar{\boldsymbol{\rho}}_{\mathcal{H}} \sim \bar{\boldsymbol{\rho}}_{\mathcal{H}'} \iff \qquad \mathsf{T} = \alpha\mathsf{U}, \alpha \in \mathbb{C}, \mathsf{U}\cdot\mathsf{U}^{H} = \mathsf{I} \tag{43b}$$

$$\bar{\boldsymbol{\rho}}_{\mathcal{H}} = \bar{\boldsymbol{\rho}}_{\mathcal{H}'} \iff \qquad \mathsf{T} = \alpha\mathsf{I} \tag{43c}$$

For intiution refeer to the following blog post

J Representation theory of symmetric function spaces

In this section, we study symmetry group actions on infinite-dimensional function spaces and specify the conditions needed to approximate these spaces in finite dimensions. Specifically, given a set $\mathcal{X}$ with a compact symmetry group $\mathbb{G}$ acting via $(\triangleright)$ (Def. I.2), the space of scalar-valued functions

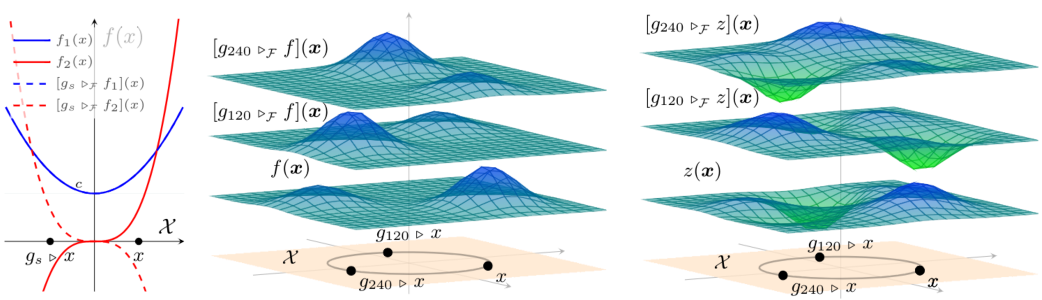

Figure 13: **Left:** Diagram of the group action $\triangleright_{\mathcal{F}}$ on functions $f_1(x) = x^2 + c$ and $f_2(x) = x^3$ defined on the domain $\mathcal{X} := \mathbb{R}$ endowed with the reflectional symmetry group $\mathbb{G} := \mathbb{C}_2 = \{e, g_s\}$, with the reflection action acting on the domain by $g_s \triangleright x = -x$ and on the function space $\mathcal{F} := \{f \mid f : \mathcal{X} \mapsto \mathbb{R}\}$ by $[g \triangleright_{\mathcal{F}} f](\boldsymbol{x}) = f(g \triangleright_{\mathcal{X}} \boldsymbol{x}) = f(-\boldsymbol{x})$. Hence we have that f_1 is a $\mathbb{G}$-invariant function, $g_s \triangleright_{\mathcal{F}} f_1(x) = f_1(x)$ and f_2 a $\mathbb{G}$-equivariant function $g_s \triangleright_{\mathcal{F}} f_2(x) = -x^3$. **Center:** Diagram representing the action $\triangleright_{\mathcal{F}}$ on the (arbitrarily chosen) function $f(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{c}_1, 2) + \mathcal{N}(\boldsymbol{x}; \boldsymbol{c}_2, 1)$ defined over the symmetric domain $\mathcal{X} = \mathbb{R}^2$ with the cyclic symmetry group $\mathbb{G} = \mathbb{C}_3 = \{e, g_{120}, g_{240}\}$ and group action $g \triangleright \boldsymbol{x} = \boldsymbol{\rho}_{\mathcal{X}}(g)\boldsymbol{x} = \boldsymbol{R}_g\boldsymbol{x}$, where $\boldsymbol{R}_g$ is a rotation matrix in 2D. Here, $g_{120} \triangleright_{\mathcal{F}} f$ is equivalent to evaluating f on a domain rotated by $-120°$. The same holds for $g_{240} \triangleright_{\mathcal{F}} f$. Note that the z-offsets are added for visualization purposes. **Right:** Diagram representing the action $\triangleright_{\mathcal{F}}$ on the function $z \in \widehat{\mathcal{F}}$, defined to be a member of the finite-dimensional symmetric function space $\widehat{\mathcal{F}} := \operatorname{span}(\mathbb{I}_{\widehat{\mathcal{F}}})$, constructed from a basis set composed of the group orbit of the (arbitrarily chosen) function $f \in \mathcal{F}$, that is $\mathbb{I}_{\widehat{\mathcal{F}}} := \mathbb{G}f = \{f, g_{120} \triangleright_{\mathcal{F}} f, g_{240} \triangleright_{\mathcal{F}} f\}$. This function space is $\mathbb{G}$-stable by construction, since $\mathbb{G}\mathbb{I}_{\widehat{\mathcal{F}}} = \mathbb{I}_{\widehat{\mathcal{F}}}$. Note that the z-offsets are added for visualization purposes.

on $\mathcal{X}$, $\mathcal{F} = \{f \mid f : \mathcal{X} \mapsto \mathbb{R}\}$, becomes a symmetric function space. The action of a symmetry transformation on a function is defined as:

Definition J.1 (Group action on a function space)**.** *Let $\mathcal{X}$ be a set endowed with the symmetry group $\mathbb{G}$, and let $\mathcal{F}$ be the space of scalar-valued functions on $\mathcal{X}$. The (left) action of $\mathbb{G}$ on a function $f \in \mathcal{F}$ is defined as the composition of f with the inverse of the group element g^{-1}:*

$$\begin{array}{rccl} (\triangleright_{\mathcal{F}}): & \mathbb{G} \times \mathcal{F} & \longrightarrow & \mathcal{F} \\ & (g, f) & \longrightarrow & [g \triangleright_{\mathcal{F}} f](\boldsymbol{x}) := [f \circ g^{-1}](\boldsymbol{x}) = f(g^{-1} \triangleright \boldsymbol{x}), \quad \forall\, \boldsymbol{x} \in \mathcal{X}. \end{array} \tag{44a}$$

In other words, the point-wise evaluation of f on a g^{-1}-transformed set $\mathcal{X}$ is equivalent to the evaluation of the transformed function $g \triangleright_{\mathcal{F}} f \in \mathcal{F}$ on the original set $\mathcal{X}$ (see simple examples in Fig. 13). Any function space that is stable under the group action Eq. (44a) is refereed to as a symmetric function space. Note that this action is compatible with the group composition and identity element $e \in \mathbb{G}$, such that the following properties hold:

$$\textit{Identity:} \quad e \triangleright_{\mathcal{F}} f(\cdot) = f(\cdot), \tag{44b}$$

$$\textit{Associativity:} \quad [(g_2 \circ g_1) \triangleright_{\mathcal{F}} f](\cdot) = [g_2 \triangleright_{\mathcal{F}} [g_1 \triangleright_{\mathcal{F}} f]](\cdot), \qquad \forall\, g_1, g_2 \in \mathbb{G}. \tag{44c}$$

Remark J.2. From an algebraic perspective, the inversion g^{-1} (*contragredient representation*) emerges to ensure that the associativity property of the group action (Eq. (44c)) holds:

$$\begin{aligned} &[(g_2 \circ g_1) \triangleright_{\mathcal{F}} f](\boldsymbol{x}) = [g_2 \triangleright_{\mathcal{F}} [g_1 \triangleright_{\mathcal{F}} f]](\boldsymbol{x}), \qquad \forall\, \boldsymbol{x} \in \mathcal{X} \\ &f((g_2 \circ g_1)^{-1} \triangleright \boldsymbol{x}) = [g_1 \triangleright_{\mathcal{F}} f](g_2^{-1} \triangleright \boldsymbol{x}) = f(g_1^{-1} \triangleright (g_2^{-1} \triangleright \boldsymbol{x})) \\ &f((g_2 \circ g_1)^{-1} \triangleright \boldsymbol{x}) = f((g_1 \circ g_2)^{-1} \triangleright \boldsymbol{x}). \end{aligned}$$

In the context of this work, we will study the scenario where the function space $\mathcal{F}$ is a separable Hilbert space and the group action of $\mathbb{G}$ on $\mathcal{F}$ is unitary, i.e., it preserves the inner product of the function space. This setup is crucial to enable us to approximate $\mathcal{F}$ and the group action on $\mathcal{F}$ in finite dimensions.

J.1 Unitary group representation on function spaces

Assume our symmetric set $\mathcal{X}$ is endowed with a measure space structure $(\mathcal{X}, \Sigma_{\mathcal{X}}, P_{\mathbf{x}})$, where $P_{\mathbf{x}} : \Sigma_{\mathcal{X}} \mapsto \mathbb{R}$ is the space measure. Then, consider a function space with a separable Hilbert space structure $\mathcal{F} := \mathcal{L}^2_{P_{\mathbf{x}}}\mathcal{X}, \mathbb{R}$, and inner product $\langle f_1, f_2 \rangle_{P_{\mathbf{x}}} = \int_{\mathcal{X}} f_1(\boldsymbol{x}) f_2(\boldsymbol{x}) P_{\mathbf{x}}(d\boldsymbol{x})$ for all $f_1, f_2 \in \mathcal{F}$. Then, the action $\triangleright_{\mathcal{F}}$ of the group $\mathbb{G}$ on the function space $\mathcal{F}$ is termed unitary if it preserves the inner product of the function space:

$$\begin{aligned} \langle f_1, f_2 \rangle_{P_{\mathbf{x}}} &= \langle g \triangleright_{\mathcal{F}} f_1, g \triangleright_{\mathcal{F}} f_2 \rangle_{P_{\mathbf{x}}} \qquad \forall\, f_1, f_2 \in \mathcal{F}, g \in \mathbb{G} \\ \int_{\mathcal{X}} f_1(\boldsymbol{x}) f_2(\boldsymbol{x}) P_{\mathbf{x}}(d\boldsymbol{x}) &= \int_{\mathcal{X}} (g \triangleright_{\mathcal{F}} f_1)(\boldsymbol{x}) (g \triangleright_{\mathcal{F}} f_2)(\boldsymbol{x}) P_{\mathbf{x}}(d\boldsymbol{x}) \\ &= \int_{\mathcal{X}} f_1(g^{-1} \triangleright \boldsymbol{x}) f_2(g^{-1} \triangleright \boldsymbol{x}) P_{\mathbf{x}}(d\boldsymbol{x}) \\ &= \int_{g \triangleright \mathcal{X} = \mathcal{X}} f_1(\boldsymbol{x}) f_2(\boldsymbol{x}) P_{\mathbf{x}}(g \triangleright d\boldsymbol{x}). \end{aligned} \tag{45}$$

That is, the group action is unitary if $P_{\mathbf{x}}$ is a $\mathbb{G}$-invariant measure $P_{\mathbf{x}}(g \triangleright d\boldsymbol{x}) = P_{\mathbf{x}}(d\boldsymbol{x}),\ \forall\, g \in \mathbb{G}, d\boldsymbol{x} \subseteq \mathcal{X}$. Note that an $\mathbb{G}$-invariant measure (and inner product) exists whenever $\mathbb{G}$ is finite, because for any measure $\eta : \Sigma_{\Omega} \mapsto \mathbb{R}$, we can use the group-average trick to obtain one, given by $P_{\mathbf{x}}(\mathbb{X}) = \Sigma_{g \in \mathbb{G}} \eta(g \triangleright \mathbb{X})$.[5]

The importance of the Hilbert space structure is that it enables the definition of a unitary group representation. Unitary representations have a well-studied modular structure that allows their decomposition (Thm. I.8) into $\mathbb{G}$-stable subspaces (Def. I.5), which is crucial for approximating symmetric function spaces using a finite set of basis elements. Let $\mathbb{I}_{\mathcal{F}} = \{\phi_i \mid \phi_i \in \mathcal{L}^2_{\mathbf{x}}\}_{i \in \mathbb{N}}$ be an orthogonal basis for the function space $\mathcal{F} = \text{span}(\mathbb{I}_{\mathcal{F}})$, so that any function $f \in \mathcal{F}$ can be represented

[5]Such a $\mathbb{G}$-invariant measure exists for any (finite or continuous) compact group. See discussion.

by its basis expansion coefficients $\boldsymbol{\alpha} = [\langle \phi_i \rangle_{P_{\mathbf{x}}} f]_{i\in\mathbb{N}}$, since $f_{\boldsymbol{\alpha}}(\boldsymbol{x}) = \sum_{i\in\mathbb{N}} \langle \phi_i, f \rangle_{P_{\mathbf{x}}} \phi_i(\boldsymbol{x})$. In this basis, the group action of $\mathbb{G}$ on $\mathcal{F}$ defines a unitary group representation mapping group elements to unitary linear integral operators on $\mathcal{F}$, which can be expressed in matrix form.

Definition J.3 (Unitary group representation on a function space)**.** *Let $\mathcal{F} = \mathcal{L}^2_{P_{\mathbf{x}}}\mathcal{X}, \mathbb{R}$ be a separable Hilbert space of scalar-valued functions on a set $\mathcal{X}$ endowed with the symmetry group $\mathbb{G}$. Let $\mathbb{I}_{\mathcal{F}}$ be an orthogonal basis set spanning $\mathcal{F}$. Then, the group action of $\mathbb{G}$ on $\mathcal{F}$ (Def. J.1) defines a unitary group representation mapping group elements to unitary linear integral operators on $\mathcal{F}$:*

$$\boldsymbol{\rho}_{\mathcal{F}}: \begin{array}{ccc} \mathbb{G} & \longrightarrow & \mathbb{U}(\mathcal{F}) \\ g & \longrightarrow & \boldsymbol{\rho}_{\mathcal{F}}(g) \end{array}, \qquad \textit{s.t.} \quad \boldsymbol{\rho}_{\mathcal{F}}(g)^* = \boldsymbol{\rho}_{\mathcal{F}}(g^{-1}). \tag{46}$$

Each unitary operator $\boldsymbol{\rho}_{\mathcal{F}}(g) : \mathcal{F} \mapsto \mathcal{F}$ admits an infinite-dimensional matrix representation with entries $[\boldsymbol{\rho}_{\mathcal{F}}(g)]_{i,j} := \langle \hat{f}_i, g \triangleright_{\mathcal{F}} \hat{f}_j \rangle_{P_{\mathbf{x}}}$, which characterize how the group action transforms the chosen basis functions. Consequently, once the group representation for a chosen basis set is defined, the group action on a function $f_{\boldsymbol{\alpha}} \in \mathcal{F}$ can be expressed as an (infinite-dimensional) matrix transformation of its basis expansion coefficients $\boldsymbol{\alpha}$, given by:

$$[g \triangleright_{\mathcal{F}} f_{\boldsymbol{\alpha}}](\cdot) := \sum_{i\in\mathbb{N}} \langle \hat{f}_i, g \triangleright_{\mathcal{F}} f_{\boldsymbol{\alpha}} \rangle_{P_{\mathbf{x}}} \hat{f}_i(\cdot) = \sum_{i\in\mathbb{N}} \left(\sum_{j\in\mathbb{N}} \langle \hat{f}_i, g \triangleright_{\mathcal{F}} \hat{f}_j \rangle_{P_{\mathbf{x}}} \underbrace{\langle \hat{f}_j, f \rangle_{P_{\mathbf{x}}}}_{\alpha_j} \right) \hat{f}_i(\cdot). \tag{47}$$

Example J.4 (Isotypic decomposition of symmetric function space). Let $(\mathcal{X}, \Sigma_{\mathcal{X}}, P_{\mathbf{x}})$ be a symmetric 2D measure space with domain $\mathcal{X} \sim \mathbb{R}^2$ and cyclic symmetry group $\mathbb{G} := \mathbb{C}_3 = \{e, g_{120}, g_{240}\}$, acting on the 2D plane by $120°$ rotations (Fig. 14). Define the finite-dimensional function space $\mathcal{F}_{\mathbf{x}} \subset \mathcal{L}^2_{\mathbf{x}}$ with basis $\mathbb{I}_{\mathcal{F}_{\mathbf{x}}} = \{\phi, g_{120} \triangleright \phi, g_{240} \triangleright \phi\}$, where $\phi \in \mathcal{F}_{\mathbf{x}}$ is an arbitrary measurable function (Fig. 14-left). In this basis, the group action $\triangleright_{\mathcal{F}_{\mathbf{x}}}$ for any function $z_{\boldsymbol{\alpha}} \in \mathcal{F}_{\mathbf{x}}$ is given by the regular representation $\boldsymbol{\rho}_{\mathcal{F}_{\mathbf{x}}} = \boldsymbol{\rho}_{\text{reg}}$ acting on the coefficient vector $\boldsymbol{\alpha} \in \mathbb{R}^3$ (Fig. 7-right).

$$[g \triangleright_{\mathcal{F}_{\mathbf{x}}} z_{\boldsymbol{\alpha}}](\cdot) = \sum_{i=1}^{3} \langle \phi_i, g \triangleright_{\mathcal{F}_{\mathbf{x}}} z_{\boldsymbol{\alpha}} \rangle_{P_{\mathbf{x}}} \phi_i(\cdot) \equiv (\boldsymbol{\rho}_{\text{reg}}(g)\boldsymbol{\alpha})^\top \begin{bmatrix} \phi(\cdot) \\ g_{120} \triangleright \phi(\cdot) \\ g_{240} \triangleright \phi(\cdot) \end{bmatrix}, \quad \boldsymbol{\rho}_{\text{reg}}(g) = \begin{cases} \boldsymbol{I}_3, & \text{if } g = e \\ \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}, & \text{if } g = g_{120} \\ \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, & \text{if } g = g_{240} \end{cases} \tag{48}$$

The group $\mathbb{C}_3$ possesses two types of (real-valued) irreducible representations, $n_{\text{iso}} = 2$: the trivial irreducible representation $\bar{\boldsymbol{\rho}}_{\text{inv}}$ and a 2D rotation representation $\bar{\boldsymbol{\rho}}_{2\pi/3}$, defined by:

$$\bar{\boldsymbol{\rho}}_{\text{inv}}(g) = \boldsymbol{I}_1, \forall\, g \in \mathbb{C}_3, \quad \text{and} \quad \bar{\boldsymbol{\rho}}_{2\pi/3}(g) = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}, \quad \text{s.t. } \theta = \begin{cases} 0°, & \text{if } g = e \\ 120° & \text{if } g = g_{120} \\ 240° & \text{if } g = g_{240} \end{cases} \tag{49}$$

Applying the appropriate change of basis, we decompose the regular representation into a direct sum of the group's irreducible representations: $\boldsymbol{\rho}_{\text{reg}} = \boldsymbol{Q}(\bar{\boldsymbol{\rho}}_{\text{inv}} \oplus \bar{\boldsymbol{\rho}}_{2\pi/3})\boldsymbol{Q}^{-1}$, where $\boldsymbol{Q}$ transitions from the regular basis to the isotypic basis of $\mathcal{F}_{\mathbf{x}}$. Since $\mathbb{C}_3$ is abelian, $\boldsymbol{Q}$ corresponds to the linear map defining the Fourier transform.

By Thm. I.8, this results in the orthogonal decomposition of the finite-dimensional function space into two orthogonal subspaces; $\mathcal{F}_{\mathbf{x}} = \mathcal{F}^{\text{inv}}_{\mathbf{x}} \oplus^{\perp} \mathcal{F}^{(2)}_{\mathbf{x}}$, where $\mathcal{F}^{\text{inv}}_{\mathbf{x}}$ denotes the 1-dimensional subspace of $\mathbb{G}$-invariant functions, and $\mathcal{F}^{(2)}_{\mathbf{x}}$ is the 2-dimensional subspace with group actions defined by the 2D irreducible representation $\bar{\boldsymbol{\rho}}_{2\pi/3}$. We can construct the basis set in the isotypic basis given:

$$\mathbb{I}^{\text{iso}}_{\mathcal{F}_{\mathbf{x}}} = \boldsymbol{Q} \begin{bmatrix} \phi(\cdot) \\ g_{120} \triangleright \phi(\cdot) \\ g_{240} \triangleright \phi(\cdot) \end{bmatrix} = \begin{bmatrix} u^{\text{inv}}(\cdot) \\ u^{(2)}_1(\cdot) \\ u^{(2)}_2(\cdot) \end{bmatrix} \qquad \text{s.t. } \boldsymbol{Q} = \begin{bmatrix} 1/\sqrt{3} & 1/\sqrt{3} & 1/\sqrt{3} \\ 2/\sqrt{6} & -1/\sqrt{6} & -1/\sqrt{6} \\ 0 & 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix} \tag{50}$$

The new basis functions in the isotypic basis are depicted in Fig. 14-right, and elucidate that the symmetry constraints on this 3-dimensional function space, result in $m = 2$ unique functions, each associated with a unique irreducible representation.

Assuming $P_{\mathbf{x}}$ is a $\mathbb{G}$-invariant probability measure, we compute the expected value of each basis function. In the regular basis, functions related by a symmetry transformation share the same expected value, i.e., $\mathbb{E}_{\mathbf{x}}\phi = \mathbb{E}_{\mathbf{x}} g \triangleright \phi$ for all $g \in \mathbb{C}_3$. In the isotypic basis, functions lacking a $\mathbb{G}$-invariant component (i.e., $u^{(2)}_1, u^{(2)}_2$) are centered: $\mathbb{E}_{\mathbf{x}} u^{(2)}_1 = \mathbb{E}_{\mathbf{x}} u^{(2)}_2 = 0$. In our example this constraint becomes clear from the nature of the change of basis $\boldsymbol{Q}$. Eq. (50).

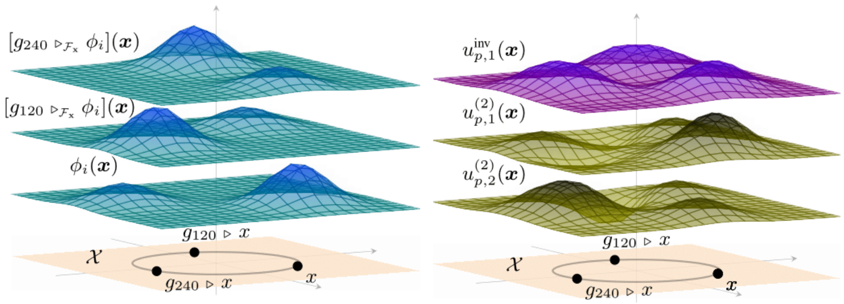

Figure 14: Visualization of the basis functions in the finite-dimensional symmetric function space $\mathcal{F}_{\mathbf{x}}$ from Example J.4. **Left**: Depiction of the basis functions in the regular basis $\mathbb{I}_{\mathcal{F}_{\mathbf{x}}} = \{\phi, g_{120} \triangleright \phi, g_{240} \triangleright \phi\}$, generated by the action of the cyclic group $\mathbb{C}_3$ on an arbitrary function $\phi \in \mathcal{F}_{\mathbf{x}}$. **Right**: Depiction of the basis functions in the isotypic basis $\mathbb{I}^{\text{iso}}_{\mathcal{F}_{\mathbf{x}}} = \{u^{\text{inv}}, u^{(2)}_1, u^{(2)}_2\}$, obtained via the change of basis matrix $\boldsymbol{Q}$. The first basis function u^{inv} corresponds to the $\mathbb{G}$-invariant subspace $\mathcal{F}^{\text{inv}}_{\mathbf{x}}$ and is visually invariant under the action of $\mathbb{C}_3$ on $\mathcal{X}$. The other two basis functions $u^{(2)}_1$, $u^{(2)}_2$ are constrained to span a $\mathbb{G}$-stable subspace of $\mathcal{L}^2_{\mathbf{x}}$, denoted by $\mathcal{F}^{(2)}_{\mathbf{x}}$ that transform according to the irreducible representation $\bar{\boldsymbol{\rho}}_{2\pi/3}$. Meaning for any function $f \in \mathcal{F}^{(2)}_{\mathbf{x}}$, the group action $g \triangleright_{\mathcal{F}_{\mathbf{x}}} f$ can be computed by a linear transformation of its basis expansion coefficients.

K $\mathbb{G}$-equivariant linear integral operators

This section gives an overview of $\mathbb{G}$-equivariant linear integral operators between symmetric function spaces. We define these operators, discuss their properties, and specify conditions under which they commute with group actions. In App. K.1 we examine their infinite-dimensional matrix form and the resulting algebraic constraints from $\mathbb{G}$-equivariance. In App. K.2 we then show how to exploit these constraints in a finite-rank approximation.

Let $\mathbb{G}$ be a compact group acting on two measure spaces $(\mathcal{X}, \Sigma_{\mathcal{X}}, P_{\mathbf{x}})$ and $(\mathcal{Y}, \Sigma_{\mathcal{Y}}, P_{\mathbf{y}})$ via the group actions $\triangleright_{\mathcal{X}}$ and $\triangleright_{\mathcal{Y}}$ (see Def. I.2). Assume that the measures $P_{\mathbf{x}}$ and $P_{\mathbf{y}}$ are $\mathbb{G}$-invariant, i.e., $P_{\mathbf{x}}(g \triangleright_{\mathcal{X}} \mathbb{B}) = P_{\mathbf{x}}(\mathbb{B})$ and $P_{\mathbf{y}}(g \triangleright_{\mathcal{Y}} \mathbb{A}) = P_{\mathbf{y}}(\mathbb{A})$ for all $g \in \mathbb{G}, \mathbb{B} \in \Sigma_{\mathcal{X}}$, and $\mathbb{A} \in \Sigma_{\mathcal{Y}}$ (see Def. I.10).

Let $\mathcal{L}^2_{\mathbf{x}} = \{f : \mathcal{X} \mapsto \mathbb{R} \mid \|f\|_{P_{\mathbf{x}}} < +\infty\}$ and $\mathcal{L}^2_{\mathbf{y}} = \{h : \mathcal{Y} \mapsto \mathbb{R} \mid \|h\|_{P_{\mathbf{y}}} < +\infty\}$ be the Hilbert spaces of square-integrable functions with respect to $P_{\mathbf{x}}$ and $P_{\mathbf{y}}$, respectively. Since $\mathcal{X}$ and $\mathcal{Y}$ have a $\mathbb{G}$-action, the spaces $\mathcal{L}^2_{\mathbf{x}}$ and $\mathcal{L}^2_{\mathbf{y}}$ inherit group actions defined by $[g \triangleright_{\mathcal{L}^2_{\mathbf{x}}} f](\boldsymbol{x}) = f(g^{-1} \triangleright_{\mathcal{X}} \boldsymbol{x})$, $\quad [g \triangleright_{\mathcal{L}^2_{\mathbf{y}}} h](\boldsymbol{y}) = h(g^{-1} \triangleright_{\mathcal{Y}} \boldsymbol{y})$, for all $f \in \mathcal{L}^2_{\mathbf{x}}$ and $h \in \mathcal{L}^2_{\mathbf{y}}$ (see Def. J.1).

We consider linear integral operators $\mathsf{T} : \mathcal{L}^2_{\mathbf{x}} \mapsto \mathcal{L}^2_{\mathbf{y}}$ defined by

$$h(\boldsymbol{y}) = [\mathsf{T}f](\boldsymbol{y}) = \int_{\mathcal{X}} \kappa(\boldsymbol{x}, \boldsymbol{y}) f(\boldsymbol{x})\, P_{\mathbf{x}}(d\boldsymbol{x}), \tag{51}$$

where $k : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$ is the kernel function of T. In this work we focus on those operators whose kernels are $\mathbb{G}$-invariant such operators are called $\mathbb{G}$-equivariant.

Definition K.1 ($\mathbb{G}$-equivariant linear intergral operators)**.** *Let $(\mathcal{X}, \Sigma_{\mathcal{X}}, P_{\mathbf{x}})$ and $(\mathcal{Y}, \Sigma_{\mathcal{Y}}, P_{\mathbf{y}})$ be two measure spaces endowed with group actions $\triangleright_{\mathcal{X}}$ and $\triangleright_{\mathcal{Y}}$ and $\mathbb{G}$-invariant measures $P_{\mathbf{x}}$ and $P_{\mathbf{y}}$ for a given compact symmetry group $\mathbb{G}$. Let $\mathsf{T} : \mathcal{L}^2_{\mathbf{x}} \mapsto \mathcal{L}^2_{\mathbf{y}}$ be a linear integral operator between the spaces of square-integrable functions defined on the two measure spaces. The operator T is said to* *be $\mathbb{G}$-equivariant if it commutes with the group action, that is $\forall\ f \in \mathcal{L}^2_{\mathbf{x}}, g \in \mathbb{G}$ and $\boldsymbol{y} \in \mathcal{Y}$:*

$$
\begin{aligned}
[\mathsf{T}[g \triangleright_{\mathcal{L}^2_{\mathbf{x}}} f]](\boldsymbol{y}) &= [g \triangleright_{\mathcal{L}^2_{\mathbf{y}}} [\mathsf{T} f]](\boldsymbol{y}) &\quad (52a)\\
\int_{\mathcal{X}} \kappa(\boldsymbol{x}, \boldsymbol{y}) f(g^{-1} \triangleright_{\mathcal{X}} \boldsymbol{x})\, P_{\mathbf{x}}(d\boldsymbol{x}) &= g \triangleright_{\mathcal{L}^2_{\mathbf{y}}} \left(\int_{\mathcal{X}} \kappa(\boldsymbol{x}, \boldsymbol{y}) f(\boldsymbol{x})\, P_{\mathbf{x}}(d\boldsymbol{x}) \right) \\
\int_{\mathcal{X}} k(g \triangleright_{\mathcal{X}} \boldsymbol{x}, \boldsymbol{y}) f(\boldsymbol{x})\, P_{\mathbf{x}}(g \triangleright_{\mathcal{X}} d\boldsymbol{x}) &= \int_{\mathcal{X}} k(\boldsymbol{x}, g^{-1} \triangleright_{\mathcal{Y}} \boldsymbol{y}) f(\boldsymbol{x})\, P_{\mathbf{x}}(d\boldsymbol{x}) \qquad \textit{s.t. } g \triangleright_{\mathcal{X}} \mathcal{X} := \mathcal{X} \\
\int_{\mathcal{X}} k(g \triangleright_{\mathcal{X}} \boldsymbol{x}, \boldsymbol{y}) f(\boldsymbol{x})\, P_{\mathbf{x}}(d\boldsymbol{x}) &= \int_{\mathcal{X}} k(\boldsymbol{x}, g^{-1} \triangleright_{\mathcal{Y}} \boldsymbol{y}) f(\boldsymbol{x})\, P_{\mathbf{x}}(d\boldsymbol{x}) \qquad \textit{s.t. } P_{\mathbf{x}}(g \triangleright_{\mathcal{X}} d\boldsymbol{x}) = P_{\mathbf{x}}(d\boldsymbol{x}) \\
k(g \triangleright_{\mathcal{X}} \boldsymbol{x}, \boldsymbol{y}) = k(\boldsymbol{x}, g^{-1} \triangleright_{\mathcal{Y}} \boldsymbol{y}) &\iff k(g \triangleright_{\mathcal{X}} \boldsymbol{x}, g \triangleright_{\mathcal{Y}} \boldsymbol{y}) = \kappa(\boldsymbol{x}, \boldsymbol{y}). &\quad (52b)
\end{aligned}
$$

Notice that the $\mathbb{G}$-equivariance of the operator T is linked to the $\mathbb{G}$-invariance of its kernel function, which is required to satisfy Eq. (52b).

Multiple approaches exist to parameterize and approximate linear integral operators with finite resources [78, sec. 4]. Here, we assume that both the input and output function spaces are separable Hilbert spaces, so that the operator can be represented as an infinite-dimensional matrix once appropriate basis sets are chosen. Its finite-dimensional (truncated or finite-rank) approximation is then obtained by selecting a finite number of basis functions in each space.

K.1 Infinite-dimensional matrix form of the operator

Since $\mathcal{L}^2_{\mathbf{x}}$ and $\mathcal{L}^2_{\mathbf{y}}$ are Hilbert spaces with inner products $\langle\cdot,\cdot\rangle_{P_{\mathbf{x}}}$ and $\langle\cdot,\cdot\rangle_{P_{\mathbf{y}}}$ respectively, we can choose orthogonal bases for both spaces: $\mathbb{I}_{\mathcal{L}^2_{\mathbf{x}}} = \{\phi_i \mid \phi_i \in \mathcal{L}^2_{\mathbf{x}}\}_{i\in\mathbb{N}}$ and $\mathbb{I}_{\mathcal{L}^2_{\mathbf{y}}} = \{\psi_j \mid \psi_j \in \mathcal{L}^2_{\mathbf{y}}\}_{j\in\mathbb{N}}$. This choice allows any function $f \in \mathcal{L}^2_{\mathbf{x}}$ and $h \in \mathcal{L}^2_{\mathbf{y}}$ to be represented by their infinite-dimensional coefficient vectors $\boldsymbol{\alpha} = [\langle\phi_i, f\rangle_{P_{\mathbf{x}}}]_{i\in\mathbb{N}}$ and $\boldsymbol{\beta} = [\langle\psi_j, h\rangle_{P_{\mathbf{y}}}]_{j\in\mathbb{N}}$, so that:

$$
f(\boldsymbol{x}) := f_{\boldsymbol{\alpha}}(\boldsymbol{x}) = \sum_{i=1}^{\infty} \langle\phi_i, f\rangle_{P_{\mathbf{x}}} \phi_i(\boldsymbol{x}) \equiv \boldsymbol{\alpha}^T \boldsymbol{\phi}(\boldsymbol{x}) \qquad h(\boldsymbol{y}) := h_{\boldsymbol{\beta}}(\boldsymbol{y}) = \sum_{j=1}^{\infty} \langle\psi_j, h\rangle_{P_{\mathbf{y}}} \psi_j(\boldsymbol{y}) \equiv \boldsymbol{\beta}^T \boldsymbol{\psi}(\boldsymbol{y}) \tag{53}
$$

Here, $\boldsymbol{\alpha}^T\boldsymbol{\phi}(\boldsymbol{x})$ and $\boldsymbol{\beta}^T\boldsymbol{\psi}(\boldsymbol{y})$ represent the function as the dot product of its expansion coefficients with the basis evaluations $\boldsymbol{\phi}(\boldsymbol{x}) = [\phi_i(\boldsymbol{x})]_{i\in\mathbb{N}}$ and $\boldsymbol{\psi}(\boldsymbol{y}) = [\psi_j(\boldsymbol{y})]_{j\in\mathbb{N}}$. This notation is useful when we later select a finite number of basis functions to form a finite-dimensional approximation of T.

With the chosen bases, the action of a linear integral operator $\mathsf{T} : \mathcal{L}^2_{\mathbf{y}} \to \mathcal{L}^2_{\mathbf{x}}$ on any $f \in \mathcal{L}^2_{\mathbf{x}}$ is determined by its action on the basis functions:

$$
[\mathsf{T} f_{\boldsymbol{\alpha}}](\boldsymbol{y}) = \int_{\mathcal{X}} \kappa(\boldsymbol{x}, \boldsymbol{y}) \Big(\sum_{i\in\mathbb{N}} \alpha_i\, \phi_i(\boldsymbol{x}) \Big) P_{\mathbf{x}}(d\boldsymbol{x}) = \sum_{i\in\mathbb{N}} \alpha_i \int_{\mathcal{X}} \kappa(\boldsymbol{x}, \boldsymbol{y})\, \phi_i(\boldsymbol{x}) P_{\mathbf{x}}(d\boldsymbol{x}) = \sum_{i\in\mathbb{N}} \alpha_i\, [\mathsf{T}\, \phi_i](\boldsymbol{y}) \tag{54}
$$

Since $[\mathsf{T}\phi_i] \in \mathcal{L}^2_{\mathbf{y}}$, each $[\mathsf{T}\,\phi_i](\boldsymbol{y})$ can be expanded using the output basis as $[\mathsf{T}\,\phi_i](\boldsymbol{y}) = \sum_{j\in\mathbb{N}} \langle\psi_j,\, \mathsf{T}\,\phi_i\rangle_{P_{\mathbf{y}}} \psi_j(\boldsymbol{y})$. Thus, the operator T can be represented by the infinite-dimensional matrix $\boldsymbol{T}$ with entries $\boldsymbol{T}_{ij} = \langle\psi_i,\, \mathsf{T}\,\phi_j\rangle_{P_{\mathbf{y}}}$. Therefore, the action of T on any $f_{\boldsymbol{\alpha}} \in \mathcal{L}^2_{\mathbf{x}}$ is given by the matrix–vector product $\boldsymbol{\beta} = \boldsymbol{T}\,\boldsymbol{\alpha}$, i.e.,

$$
\begin{aligned}
[\mathsf{T} f_{\boldsymbol{\alpha}}](\boldsymbol{y}) &= \sum_{j\in\mathbb{N}} \alpha_j\, [\mathsf{T}\,\phi_j](\boldsymbol{y}) = \sum_{j\in\mathbb{N}} \alpha_j \sum_{i\in\mathbb{N}} \langle\psi_i,\, \mathsf{T}\,\phi_j\rangle_{P_{\mathbf{y}}} \psi_i(\boldsymbol{y}) \\
&= \sum_{i\in\mathbb{N}} \sum_{j\in\mathbb{N}} \boldsymbol{T}_{ij}\, \alpha_j\, \psi_i(\boldsymbol{y}) \equiv (\boldsymbol{T}\,\boldsymbol{\alpha})^T\, \boldsymbol{\psi}(\boldsymbol{y})
\end{aligned} \tag{55}
$$

Eq. (55) shows that knowing the action of T on the bases $\mathbb{I}_{\mathcal{L}^2_{\mathbf{x}}}$ and $\mathbb{I}_{\mathcal{L}^2_{\mathbf{y}}}$ determines its action on any function in $\mathcal{L}^2_{\mathbf{x}}$. In the sections that follow, we describe how symmetry constrains this action by requiring the bases to be $\mathbb{G}$-stable and by imposing $\mathbb{G}$-equivariance on $\boldsymbol{T}$, thereby introducing exploitable algebraic constraints for improved finite-rank approximations.

K.1.1 $\mathbb{G}$-equivariant matrix form of the operator

Whenever the function spaces carry a symmetry group $\mathbb{G}$, the group action on their bases $\mathbb{I}_{\mathcal{L}^2_{\mathbf{x}}}$ and $\mathbb{I}_{\mathcal{L}^2_{\mathbf{y}}}$ is defined by the unitary representations $\boldsymbol{\rho}_{\mathcal{L}^2_{\mathbf{x}}} : \mathbb{G} \to \mathbb{U}()\mathcal{L}^2_{\mathbf{x}}$ and $\boldsymbol{\rho}_{\mathcal{L}^2_{\mathbf{y}}} : \mathbb{G} \to \mathbb{U}()\mathcal{L}^2_{\mathbf{y}}$ (see Def. J.3). As in Eq. (55), these representations can be expressed in (infinite-dimensional) matrix form so that the group action is given by a matrix-vector product:

$$
\begin{aligned}
[g \triangleright_{\mathcal{L}^2_x} f_{\boldsymbol{\alpha}}](\cdot) &\equiv (\boldsymbol{\rho}_{\mathcal{L}^2_x}(g)\boldsymbol{\alpha})^T \boldsymbol{\phi}(\cdot), \quad \forall\, f_{\boldsymbol{\alpha}} \in \mathcal{L}^2_{\mathbf{x}}, g \in \mathbb{G} \\
[g \triangleright_{\mathcal{L}^2_y} h_{\boldsymbol{\beta}}](\cdot) &\equiv (\boldsymbol{\rho}_{\mathcal{L}^2_y}(g)\boldsymbol{\beta})^T \boldsymbol{\psi}(\cdot), \quad \forall\, h_{\boldsymbol{\beta}} \in \mathcal{L}^2_{\mathbf{y}}, g \in \mathbb{G}
\end{aligned} \tag{56}
$$

Since the operator T is $\mathbb{G}$-equivariant by construction (Eq. (52a)), the matrix form $\boldsymbol{T}$ of the operator must also be $\mathbb{G}$-equivariant with respect to the group representations $\boldsymbol{\rho}_{\mathcal{L}^2_x}$ and $\boldsymbol{\rho}_{\mathcal{L}^2_y}$:

$$
\begin{aligned}
[\mathsf{T}[g \triangleright_{\mathcal{L}^2_x} f_{\boldsymbol{\alpha}}]](\boldsymbol{y}) &= [g \triangleright_{\mathcal{L}^2_y} [\mathsf{T} f_{\boldsymbol{\alpha}}]](\boldsymbol{y}) && \forall\, f_{\boldsymbol{\alpha}} \in \mathcal{L}^2_{\mathbf{x}}, g \in \mathbb{G}, \boldsymbol{y} \in \mathcal{Y} \\
(\boldsymbol{T}\boldsymbol{\rho}_{\mathcal{L}^2_x}(g)\boldsymbol{\alpha})^\top \boldsymbol{\psi}(\boldsymbol{y}) &= (\boldsymbol{\rho}_{\mathcal{L}^2_y}(g)\boldsymbol{T}\boldsymbol{\alpha})^\top \boldsymbol{\psi}(\boldsymbol{y}) && \text{s.t. Eqs. (55) and (56)} \\
\boldsymbol{T}\boldsymbol{\rho}_{\mathcal{L}^2_x}(g) &= \boldsymbol{\rho}_{\mathcal{L}^2_y}(g)\boldsymbol{T}
\end{aligned} \tag{57}
$$

With bases $\mathbb{I}_{\mathcal{L}^2_{\mathbf{x}}}$ and $\mathbb{I}_{\mathcal{L}^2_{\mathbf{y}}}$ for $\mathcal{L}^2_{\mathbf{x}}$ and $\mathcal{L}^2_{\mathbf{y}}$, the kernel (Def. K.1) can be written as $\kappa(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i,j \in \mathbb{N}} \boldsymbol{T}_{i,j}\, \phi_j(\boldsymbol{x})\, \psi_i(\boldsymbol{y})$. Hence, the $\mathbb{G}$-invariance condition (Eq. (52b)) on the kernel directly implies that the matrix $\boldsymbol{T}$ is $\mathbb{G}$-equivariant, as stated in the following proposition:

Proposition K.2 ($\mathbb{G}$-invariant kernel implies $\mathbb{G}$-equivariant matrix form)**.** *Let $\mathsf{T} : \mathcal{L}^2_{\mathbf{x}} \mapsto \mathcal{L}^2_{\mathbf{y}}$ be a $\mathbb{G}$-equivariant operator between symmetric function spaces endowed with the group actions $\triangleright_{\mathcal{L}^2_x}$ and $\triangleright_{\mathcal{L}^2_y}$ of a compact symmetry group $\mathbb{G}$. Let $\boldsymbol{\rho}_{\mathcal{L}^2_x}$ and $\boldsymbol{\rho}_{\mathcal{L}^2_y}$ be the group representation of the on the input/output function spaces on the chosen basis sets $\mathbb{I}_{\mathcal{L}^2_{\mathbf{x}}}$ and $\mathbb{I}_{\mathcal{L}^2_{\mathbf{y}}}$. Then the G-invariance of the operator's kernel function (Eq. (52b)) implies that the matrix form of the operator, in the chosen basis sets, is $\mathbb{G}$-equivariant w.r.t the group representations $\boldsymbol{\rho}_{\mathcal{L}^2_x}$ and $\boldsymbol{\rho}_{\mathcal{L}^2_y}$ (Eq. (57)).*

Proof. The proof follows by choosing appropriate $\mathbb{G}$-stable basis sets $\{\phi_i\} \subset \mathcal{L}^2_{\mathbf{x}}$ and $\{\psi_j\} \subset \mathcal{L}^2_{\mathbf{y}}$, so that for all $g \in \mathbb{G}$ we have $g \triangleright_{\mathcal{L}^2_x} \phi_i = \phi_{g \triangleright i}$ and $g \triangleright_{\mathcal{L}^2_y} \psi_j = \psi_{g \triangleright j}$ with $g \triangleright i, g \triangleright j \in \mathbb{N}$. This basis sets the $\mathbb{G}$-invariance of the kernel translates into algebraic constraints on the matrix form $\boldsymbol{T}$.

$$
\begin{gathered}
k(\boldsymbol{x}, \boldsymbol{y}) = k(g^{-1} \triangleright_{\mathcal{X}} \boldsymbol{x}, g^{-1} \triangleright_{\mathcal{Y}} \boldsymbol{y}) \qquad \forall\, g \in \mathbb{G}, \boldsymbol{x} \in \mathcal{X}, \boldsymbol{y} \in \mathcal{Y} \\
\sum_{i \in \mathbb{N}} \sum_{j \in \mathbb{N}} \boldsymbol{T}_{i,j} \phi_i(\boldsymbol{x}) \psi_j(\boldsymbol{y}) = \sum_{i \in \mathbb{N}} \sum_{j \in \mathbb{N}} \boldsymbol{T}_{i,j} [g \triangleright_{\mathcal{L}^2_x} \phi_i](\boldsymbol{x}) [g \triangleright_{\mathcal{Y}} \psi_j](\boldsymbol{y}) = \sum_{i \in \mathbb{N}} \sum_{j \in \mathbb{N}} \boldsymbol{T}_{i,j} \phi_{g \triangleright i}(\boldsymbol{x}) \psi_{g \triangleright j}(\boldsymbol{y})
\end{gathered} \tag{58}
$$

That is, the kernel is $\mathbb{G}$-equivariant if the operator's matrix satisfies $\boldsymbol{T}_{i,j} = \boldsymbol{T}_{g \triangleright i,\, g \triangleright j}$ for all $g \in \mathbb{G},\ i, j \in \mathbb{N}$. This condition exactly characterizes the $\mathbb{G}$-equivariance of the matrix form.

$$
\begin{aligned}
\boldsymbol{T}_{i,j} = \langle \psi_i, \mathsf{T}\phi_j \rangle_{P_{\mathbf{y}}} &= \langle \psi_{g \triangleright i}, \mathsf{T}\phi_{g \triangleright j} \rangle_{P_{\mathbf{y}}} = \boldsymbol{T}_{g \triangleright i, g \triangleright j} && \forall g \in \mathbb{G}, i, j \in \mathbb{N} \\
&= \langle g \triangleright_{\mathcal{L}^2_y} \psi_i, \mathsf{T}[g \triangleright_{\mathcal{L}^2_x} \phi_j] \rangle_{P_{\mathbf{y}}} \\
&= \langle g \triangleright_{\mathcal{L}^2_y} \psi_i, g \triangleright_{\mathcal{L}^2_y} [\mathsf{T}\phi_j] \rangle_{P_{\mathbf{y}}} && \text{s.t. Eq. (52a)} \\
&= \langle \psi_i, \mathsf{T}\phi_j \rangle_{P_{\mathbf{y}}} = \boldsymbol{T}_{i,j} && \text{s.t. Eq. (45)}
\end{aligned} \tag{59}
$$

□

K.1.2 Block-diagonal structure of the operator matrix form

According to Thm. I.8, a Hilbert space with a compact symmetry group $\mathbb{G}$ decomposes into n_{iso} orthogonal subspaces—one for each irreducible representation type—yielding an orthogonal decomposition of the operator's input and output spaces:

$$
\mathcal{L}^2_{\mathbf{x}} := \oplus^{\perp}_{1 \le k \le n_{\text{iso}}} \mathcal{L}^{2(k)}_{\mathbf{x}}, \qquad \text{and} \qquad \mathcal{L}^2_{\mathbf{y}} := \oplus^{\perp}_{1 \le k \le n_{\text{iso}}} \mathcal{L}^{2(k)}_{\mathbf{y}}, \tag{60}
$$

where $\mathcal{L}^{2(k)}_{\mathbf{x}}$ and $\mathcal{L}^{2(k)}_{\mathbf{y}}$ denote the k-th isotypic subspaces of $\mathcal{L}^2_{\mathbf{x}}$ and $\mathcal{L}^2_{\mathbf{y}}$, respectively. Such that any function in these spaces can be decomposed into a sum of its projections onto the isotypic subspaces:

$$
f(\boldsymbol{x}) = \sum_{k=1}^{n_{\text{iso}}} f^{(k)}(\boldsymbol{x}), \quad h(\boldsymbol{y}) = \sum_{k=1}^{n_{\text{iso}}} h^{(k)}(\boldsymbol{y}) \quad \text{with} \quad f^{(k)} \in \mathcal{L}^{2(k)}_{\mathbf{x}}, h^{(k)} \in \mathcal{L}^{2(k)}_{\mathbf{y}}. \tag{61}
$$

The orthogonal decomposition of the function spaces implies there exist unitary operators $\mathsf{A} : \mathcal{L}^2_{\mathbf{x}} \to \mathcal{L}^2_{\mathbf{x}}$ and $\mathsf{B} : \mathcal{L}^2_{\mathbf{y}} \to \mathcal{L}^2_{\mathbf{y}}$ (with matrix forms $\boldsymbol{A}$ and $\boldsymbol{B}$), that describe a change of basis from the canonical basis to an *isotypic basis*, $\mathbb{I}^{\text{iso}}_{\mathcal{L}^2_{\mathbf{x}}} = \cup_{k=1}^{n_{\text{iso}}} \mathbb{I}_{\mathcal{L}^{2(k)}_{\mathbf{x}}} = \mathsf{A}\mathbb{I}_{\mathcal{L}^2_{\mathbf{x}}}$ and $\mathbb{I}^{\text{iso}}_{\mathcal{L}^2_{\mathbf{y}}} = \cup_{k=1}^{n_{\text{iso}}} \mathbb{I}_{\mathcal{L}^{2(k)}_{\mathbf{y}}} = \mathsf{B}\mathbb{I}_{\mathcal{L}^2_{\mathbf{y}}}$, where the group's representations decompose into a direct sum of representations per isotypic subspace (see Def. I.4):

$$
\boldsymbol{\rho}^{\text{iso}}_{\mathcal{L}^2_x}(\cdot) := \boldsymbol{A}\boldsymbol{\rho}_{\mathcal{L}^2_x}(\cdot)\boldsymbol{A}^* = \oplus_{k=1}^{n_{\text{iso}}} \boldsymbol{\rho}_{\mathcal{L}^{2(k)}_x}(\cdot) \quad \text{and} \quad \boldsymbol{\rho}^{\text{iso}}_{\mathcal{L}^2_y}(\cdot) := \boldsymbol{B}\boldsymbol{\rho}_{\mathcal{L}^2_y}(\cdot)\boldsymbol{B}^* = \oplus_{k=1}^{n_{\text{iso}}} \boldsymbol{\rho}_{\mathcal{L}^{2(k)}_y}(\cdot). \tag{62}
$$

Then, denoting the matrix form of T in the isotypic basis by $\boldsymbol{T}^{\text{iso}} = \boldsymbol{B}^{*}\boldsymbol{T}\boldsymbol{A}$, the $\mathbb{G}$-equivariance of T results in the matrix form of the operator in the isotypic basis being block-diagonal, with each block being $\mathbb{G}$-equivariant with respect to the group representations on the isotypic subspaces:

$$
\begin{aligned}
\boldsymbol{T}^{\text{iso}} &= \boldsymbol{\rho}^{\text{iso}}_{\mathcal{L}^2_{\mathbf{y}}}(g)\boldsymbol{T}^{\text{iso}}\boldsymbol{\rho}^{\text{iso}}_{\mathcal{L}^2_{\mathbf{x}}}(g^{-1}) \\
&= \oplus_{k=1}^{n_{\text{iso}}}\boldsymbol{\rho}_{\mathcal{L}^{2(k)}_{\mathbf{y}}}(g)\boldsymbol{T}^{\text{iso}} \oplus_{k=1}^{n_{\text{iso}}} \boldsymbol{\rho}_{\mathcal{L}^{2(k)}_{\mathbf{x}}}(g^{-1}) \qquad\qquad \text{s.t. Eqs. (57) and (62)} \\
\boldsymbol{T}^{(k)} &= \boldsymbol{\rho}_{\mathcal{L}^{2(k)}_{\mathbf{y}}}(g)\boldsymbol{T}^{(k)}\boldsymbol{\rho}_{\mathcal{L}^{2(k)}_{\mathbf{x}}}(g^{-1}), \quad \forall\, k = 1, \dots, n_{\text{iso}} \quad \boldsymbol{T}^{\text{iso}} = \oplus_{k=1}^{n_{\text{iso}}}\boldsymbol{T}^{(k)} = \begin{bmatrix} \boldsymbol{T}^{(1)} & & \\ & \ddots & \\ & & \boldsymbol{T}^{(n_{\text{iso}})} \end{bmatrix}.
\end{aligned}
\tag{63}
$$

Each $\boldsymbol{T}^{(k)}$ represents the matrix form of the operator $\mathsf{T}^{(k)} : \mathcal{L}^{2(k)}_{\mathbf{x}} \mapsto \mathcal{L}^{2(k)}_{\mathbf{y}}$ in the isotypic basis, acting on the isotypic subspaces of type k in the input and output spaces. This shows that $\mathbb{G}$-equivariant operators preserve the structure of isotypic subspaces without mixing functions from different types.

This property is crucial for the finite-rank approximation of the operator T, as it reduces the problem to approximating lower-rank operators $\mathsf{T}^{(k)} : \mathcal{L}^{2(k)}_{\mathbf{x}} \mapsto \mathcal{L}^{2(k)}_{\mathbf{y}}$, for $k \in [1, n_{\text{iso}}]$. Moreover, the block diagonal structure of $\boldsymbol{T}^{\text{iso}}$ allows us to rewrite Eq. (55) in the isotypic basis in terms of the action of each $\mathsf{T}^{(k)}$ on the projection $f^{(k)}$ of the function onto the k^{th} isotypic subspace, see (61), such that:

$$
[\mathsf{T}f_{\boldsymbol{\alpha}}](\boldsymbol{y}) = \sum_{k=1}^{n_{\text{iso}}}[\mathsf{T}^{(k)}f^{(k)}](\boldsymbol{y}) \equiv \sum_{k=1}^{n_{\text{iso}}}(\boldsymbol{T}^{(k)}\boldsymbol{\alpha}^{(k)})^{\top}\boldsymbol{\psi}^{(k)}(\boldsymbol{y}). \qquad \boldsymbol{\psi}^{(k)}(\cdot) = [\psi^{(k)}_j(\cdot)]_{j\in\mathbb{N}}, \forall\, \psi^{(k)}_j \in \mathbb{I}_{\mathcal{L}^{2(k)}_{\mathbf{y}}}. \tag{64}
$$

In the isotypic basis $\mathbb{I}^{\text{iso}}_{\mathcal{L}^2_{\mathbf{x}}} = \cup_{k=1}^{n_{\text{iso}}}\mathbb{I}_{\mathcal{L}^{2(k)}_{\mathbf{x}}}$, the expansion coefficient vector $\boldsymbol{\alpha} = \oplus_{k=1}^{n_{\text{iso}}}\boldsymbol{\alpha}^{(k)}$ is formed from the projections of f onto each isotypic subspace: $\boldsymbol{\alpha}^{(k)} = [\langle \phi^{(k)}_i, f\rangle_{P_{\mathbf{x}}}]_{i\in\mathbb{N}}$. The block-diagonal structure of $\boldsymbol{T}^{\text{iso}}$ is only one of the algebraic constraints imposed on the matrix form of T by the $\mathbb{G}$-equivariance condition. The next section describes the further structural constraints on each block.

K.1.3 Structure of operators between isotypic subspaces

In this section, we shift the focus from the input and output function spaces, $\mathcal{L}^2_{\mathbf{x}}$ and $\mathcal{L}^2_{\mathbf{y}}$; and the operator $\mathsf{T} : \mathcal{L}^2_{\mathbf{x}} \mapsto \mathcal{L}^2_{\mathbf{y}}$, to their individual isotypic subspaces, $\mathcal{L}^{2(k)}_{\mathbf{x}}$ and $\mathcal{L}^{2(k)}_{\mathbf{y}}$ for $k \in [1, n_{\text{iso}}]$, and the operators $\mathsf{T}^{(k)} : \mathcal{L}^{2(k)}_{\mathbf{x}} \mapsto \mathcal{L}^{2(k)}_{\mathbf{y}}$ (Eq. (60)).

Recall from Thm. I.8, that each isotypic subspace possesses unitary group representations that decompose into direct sums of (infinitely many) multiplicities of the irreducible representation of type k; that is:

$$
\boldsymbol{\rho}_{\mathcal{L}^{2(k)}_{\mathbf{x}}}(g) \sim \oplus_{p=1}^{\infty}\bar{\boldsymbol{\rho}}_k(g) \qquad \text{and} \qquad \boldsymbol{\rho}_{\mathcal{L}^{2(k)}_{\mathbf{y}}}(g) \sim \oplus_{p=1}^{\infty}\bar{\boldsymbol{\rho}}_k(g). \tag{65}
$$

This implies that each isotypic subspace further decomposes into (infinitely many) finite-dimensional $\mathbb{G}$-stable subspaces: $\mathcal{L}^{2(k)}_{\mathbf{x}} := \oplus_{p=1}^{\infty}\mathcal{L}^{2k,p}_{\mathbf{x}}$ and $\mathcal{L}^{2(k)}_{\mathbf{y}} := \oplus_{p=1}^{\infty}\mathcal{L}^{2k,p}_{\mathbf{y}}$. Each subspace $\mathcal{L}^{2k,p}_{\mathbf{x}}$ (and similarly $\mathcal{L}^{2k,p}_{\mathbf{y}}$) has finite dimension $d_k \leq \infty$ and its elements transform according to the irreducible representation $\bar{\boldsymbol{\rho}}_k$ of the group $\mathbb{G}$.

The modular structure of the isotypic subspaces implies that the $\mathbb{G}$-equivariant operator $\mathsf{T}^{(k)}$ further decomposes into $\mathbb{G}$-equivariant components acting between finite-dimensional, $\mathbb{G}$-stable subspaces: $\mathsf{T}^{(k,i,j)} : \mathcal{L}^{2k,i}_{\mathbf{x}} \mapsto \mathcal{L}^{2k,j}_{\mathbf{y}}$ for $i, j \in \mathbb{N}$. This is advantageous since—by Schur's lemma (Lem. I.12)—the space of $\mathbb{G}$-equivariant maps between irreducible subspaces is one-dimensional, i.e., $\dim(\text{Homo}_{\mathbb{G}}(\mathcal{L}^{2k,i}_{\mathbf{x}}, \mathcal{L}^{2k,j}_{\mathbf{y}})) = 1$ for all $i, j \in \mathbb{N}$.

To reveal the modular structure of $\mathsf{T}^{(k)}$ in matrix form, we select bases for the isotypic subspaces $\mathcal{L}^{2(k)}_{\mathbf{x}}$ and $\mathcal{L}^{2(k)}_{\mathbf{y}}$ that separate the basis functions by irreducible subspace, i.e., $\mathbb{I}_{\mathcal{L}^{2(k)}_{\mathbf{x}}} = \cup_{p=1}^{\infty}\mathbb{I}_{\mathcal{L}^{2k,p}_{\mathbf{x}}}$ and $\mathbb{I}_{\mathcal{L}^{2(k)}_{\mathbf{y}}} = \cup_{p=1}^{\infty}\mathbb{I}_{\mathcal{L}^{2k,p}_{\mathbf{y}}}$, so that $\boldsymbol{\rho}_{\mathcal{L}^{2(k)}_{\mathbf{x}}}(g) = \oplus_{p=1}^{\infty}\bar{\boldsymbol{\rho}}_k(g)$ and $\boldsymbol{\rho}_{\mathcal{L}^{2(k)}_{\mathbf{y}}}(g) = \oplus_{p=1}^{\infty}\bar{\boldsymbol{\rho}}_k(g)$. In these bases, each map $\mathsf{T}^{(k,i,j)}$ reduces to a scalar multiple of the identity, namely, $\mathsf{T}^{(k,i,j)} = \theta^{(k)}_{i,j}\mathsf{I}_k$, where $\theta^{(k)}_{i,j} \in \mathbb{R}$ captures the only degree of freedom (see (43c)). Consequently, the matrix representation of $\mathsf{T}^{(k)}$ consists of blocks that are scalar multiples of the identity.

$$
\boldsymbol{T}^{(k)} = \begin{bmatrix} \theta^{(k)}_{1,1}\boldsymbol{I}_k & \theta^{(k)}_{1,2}\boldsymbol{I}_k & \cdots \\ \theta^{(k)}_{2,1}\boldsymbol{I}_k & \ddots & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix} = \Theta^{(k)} \otimes \boldsymbol{I}_k, \quad \text{s.t.} \quad
\begin{aligned}
&\text{rank}(\boldsymbol{I}_k) = d_k, \\
&\boldsymbol{\rho}_{\mathcal{L}^{2(k)}_{\mathbf{y}}}(g)\boldsymbol{T}^{(k)} = \boldsymbol{T}^{(k)}\boldsymbol{\rho}_{\mathcal{L}^{2(k)}_{\mathbf{x}}}(g), \quad \forall\, g \in \mathbb{G} \\
&(\oplus_{p=1}^{\infty}\bar{\boldsymbol{\rho}}_k(g))\boldsymbol{T}^{(k)} = \boldsymbol{T}^{(k)}(\oplus_{p=1}^{\infty}\bar{\boldsymbol{\rho}}_k(g)),
\end{aligned}
\tag{66}
$$

where the (infinite-dimensional) matrix $\Theta^{(k)}$ parameterizes the degrees of freedom of $\boldsymbol{T}^{(k)}$.

Eq. (66) reveals the Kronecker product structure of $\mathbb{G}$-equivariant operators between isotypic subspaces when the appropriate input and output basis sets are chosen. To illustrate, consider a function $f^{(k)}_{\boldsymbol{\alpha}} \in \mathcal{L}^{2(k)}_{\mathbf{x}}$ with basis coefficients given by $\boldsymbol{\alpha} = \oplus_{p=1}^{\infty} \boldsymbol{\alpha}_p$, where each $\boldsymbol{\alpha}_p = [\langle \phi_i^{(k,p)}, f^{(k)} \rangle_{P_{\mathbf{x}}}]_{i=1}^{d_k} \in \mathbb{R}^{d_k}$ represents the projection of $f^{(k)}$ onto the p^{th} irreducible subspace $\mathcal{L}^{2k,p}_{\mathbf{x}}$. Then, if $h^{(k)}_{\boldsymbol{\beta}} = \mathsf{T}^{(k)} f^{(k)}_{\boldsymbol{\alpha}}$, the coefficients are computed as $\boldsymbol{\beta} = \Theta^{(k)} \boldsymbol{\alpha}$.

$$
\overbrace{\begin{bmatrix}
\langle \psi_1^{(k,1)}, \mathsf{T}^{(k)} f^{(k)} \rangle_{P_{\mathbf{y}}} \\
\langle \psi_2^{(k,1)}, \mathsf{T}^{(k)} f^{(k)} \rangle_{P_{\mathbf{y}}} \\
\vdots \\
\langle \psi_{d_k}^{(k,1)} \rangle_{P_{\mathbf{y}}} \mathsf{T}^{(k)} f^{(k)} \\
\hline
\langle \psi_1^{(k,2)}, \mathsf{T}^{(k)} f^{(k)} \rangle_{P_{\mathbf{y}}} \\
\langle \psi_2^{(k,2)}, \mathsf{T}^{(k)} f^{(k)} \rangle_{P_{\mathbf{y}}} \\
\vdots \\
\langle \psi_{d_k}^{(k,2)}, \mathsf{T}^{(k)} f^{(k)} \rangle_{P_{\mathbf{y}}} \\
\hline
\vdots
\end{bmatrix}}^{\boldsymbol{\beta}}
=
\overbrace{\left[\begin{array}{cccc|cccc|c}
\theta^{(k)}_{1,1} & 0 & \cdots & 0 & \theta^{(k)}_{1,2} & 0 & \cdots & 0 & \\
0 & \theta^{(k)}_{1,1} & \cdots & 0 & 0 & \theta^{(k)}_{1,2} & \cdots & 0 & \cdots \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \\
0 & 0 & \cdots & \theta^{(k)}_{1,1} & 0 & 0 & \cdots & \theta^{(k)}_{1,2} & \\
\hline
\theta^{(k)}_{2,1} & 0 & \cdots & 0 & & & & & \\
0 & \theta^{(k)}_{2,1} & \cdots & 0 & & & \ddots & & \cdots \\
\vdots & \vdots & \ddots & \vdots & & & & & \\
0 & 0 & \cdots & \theta^{(k)}_{2,1} & & & & & \\
\hline
 & & \vdots & & & & \vdots & & \ddots
\end{array}\right]}^{\boldsymbol{T}^{(k)} = \Theta^{(k)} \otimes \boldsymbol{I}_k}
\overbrace{\begin{bmatrix}
\langle \phi_1^{(k,1)}, f^{(k)} \rangle_{P_{\mathbf{x}}} \\
\langle \phi_2^{(k,1)}, f^{(k)} \rangle_{P_{\mathbf{x}}} \\
\vdots \\
\langle \phi_{d_k}^{(k,1)}, f^{(k)} \rangle_{P_{\mathbf{x}}} \\
\hline
\langle \phi_1^{(k,2)}, f^{(k)} \rangle_{P_{\mathbf{x}}} \\
\langle \phi_2^{(k,2)}, f^{(k)} \rangle_{P_{\mathbf{x}}} \\
\vdots \\
\langle \phi_{d_k}^{(k,2)}, f^{(k)} \rangle_{P_{\mathbf{x}}} \\
\hline
\vdots
\end{bmatrix}}^{\boldsymbol{\alpha}}
\tag{67}
$$

This structure can be interpreted as a constraint on the dimensionality of the singular spaces of the operator T to be of dimension larger than d_k, as summarized in the following proposition:

Proposition K.3 (Minimum dimensionality of singular space of $\mathbb{G}$-equivariant operators between isotypic subspaces)**.** *Let $\mathsf{T}^{(k)} : \mathcal{L}^{2(k)}_{\mathbf{x}} \mapsto \mathcal{L}^{2(k)}_{\mathbf{y}}$ be a $\mathbb{G}$-equivariant operator between isotypic subspaces $\mathcal{L}^{2(k)}_{\mathbf{x}}$ and $\mathcal{L}^{2(k)}_{\mathbf{y}}$ of type k. Then, the minimum dimension of a singular space of the operator is d_k.*

Proof. Let $\mathbb{I}_{\mathcal{L}^{2(k)}_{\mathbf{x}}} = \cup_{p=1}^{\infty} \mathbb{I}_{\mathcal{L}^{2k,p}_{\mathbf{x}}}$ and $\mathbb{I}_{\mathcal{L}^{2(k)}_{\mathbf{y}}} = \cup_{p=1}^{\infty} \mathbb{I}_{\mathcal{L}^{2k,p}_{\mathbf{x}}}$ be the basis sets that expose the Kronecker structure of the matrix form $\boldsymbol{T}^{(k)} = \Theta^{(k)} \otimes \boldsymbol{I}_k$, as per Eq. (66). Then the singular value decomposition of the matrix form inherits the Kronecker product structure such that $\boldsymbol{T}^{(k)} = \boldsymbol{U}^{(k)} \boldsymbol{\Sigma}^{(k)} (\boldsymbol{V}^{(k)})^* = (\boldsymbol{W}^{(k)} \otimes \boldsymbol{I}_k)(\boldsymbol{\Sigma}^{(k)} \otimes \boldsymbol{I}_k)((\boldsymbol{Q}^{(k)})^* \otimes \boldsymbol{I}_k)$, where $\Theta^{(k)} = \boldsymbol{W}^{(k)} \boldsymbol{\Sigma}^{(k)} (\boldsymbol{Q}^{(k)})^*$. The Kronecker structure of the diagonal singular value matrix $(\boldsymbol{\Sigma}^{(k)} \otimes \boldsymbol{I}_k)$ implies that each singular value has a minimum multiplicity of d_k. While the Kronecker strucuture of the change of bases $\boldsymbol{U}^{(k)}$ and $\boldsymbol{V}^{(k)}$ encodes the d_k orthogonal basis vectors of the singular spaces. $\square$

K.2 Finite-rank approximation of $\mathbb{G}$-equivariant operators

In practical applications, infinite-dimensional operators are approximated by finite-dimensional ones to enable computation. For any linear integral operator $\mathsf{T} : \mathcal{L}^2_{\mathbf{x}} \mapsto \mathcal{L}^2_{\mathbf{y}}$, the optimal rank-$r$ approximation in the Hilbert-Schmidt norm is obtained by truncating its SVD to the top r singular values and associated left/right singular functions. Let $\{\sigma_i\}_{i=1}^{\infty}$ be the singular values of T in decreasing order and let $\{u_i\}_{i=1}^{\infty} \subset \mathcal{L}^2_{\mathbf{x}}$, $\{v_i\}_{i=1}^{\infty} \subset \mathcal{L}^2_{\mathbf{y}}$ be the corresponding singular functions satisfying $\langle v_i, \mathsf{T} u_i \rangle_{P_{\mathbf{y}}} = \sigma_i$ for each $i \in \mathbb{N}$ and $\langle v_i, \mathsf{T} u_j \rangle_{P_{\mathbf{y}}} = 0$ when $i \neq j$. The best rank-r approximation of T is then given by [31]:

$$
\mathsf{T}_r f = \sum_{i=1}^{r} \sigma_i \langle u_i, f \rangle_{P_{\mathbf{x}}} v_i, \quad \forall f \in \mathcal{L}^2_{\mathbf{x}}, \qquad \iff \qquad \kappa(\boldsymbol{x}, \boldsymbol{y}) \approx \sum_{i=1}^{r} \sigma_i u_i(\boldsymbol{x}) v_i(\boldsymbol{y}). \tag{68}
$$

Since the left and right singular functions form orthonormal bases for $\mathcal{L}^2_{\mathbf{y}}$ and $\mathcal{L}^2_{\mathbf{x}}$, a rank-$r$ approximation reduces these infinite-dimensional spaces to the r-dimensional subspaces $\mathcal{F}_{\mathbf{x}} = \text{span}(\{u_i\}_{i=1}^{r})$ and $\mathcal{F}_{\mathbf{y}} = \text{span}(\{v_i\}_{i=1}^{r})$.

When $\mathcal{L}^2_{\mathbf{x}}$ and $\mathcal{L}^2_{\mathbf{y}}$ are symmetric function spaces with group actions $\triangleright_{\mathcal{L}^2_{\mathbf{x}}}$ and $\triangleright_{\mathcal{L}^2_{\mathbf{y}}}$ of a compact group $\mathbb{G}$, and T is $\mathbb{G}$-equivariant, the finite-rank approximation $\mathsf{T}_r : \mathcal{F}_{\mathbf{x}} \to \mathcal{F}_{\mathbf{y}}$ must satisfy that for all $f \in \mathcal{F}_{\mathbf{x}}$,

$h \in \mathcal{F}_{\mathbf{y}}$, and $g \in \mathbb{G}$, both $g \triangleright_{\mathcal{L}^2_{\mathbf{x}}} f \in \mathcal{F}_{\mathbf{x}}$ and $g \triangleright_{\mathcal{L}^2_{\mathbf{y}}} h \in \mathcal{F}_{\mathbf{y}}$. This ensures that $g \triangleright_{\mathcal{L}^2_{\mathbf{y}}} [\mathsf{T}_r f] = \mathsf{T}_r[g \triangleright_{\mathcal{L}^2_{\mathbf{x}}} f]$ (see App. J).

Moreover, since $\mathcal{L}^2_{\mathbf{x}}$ and $\mathcal{L}^2_{\mathbf{y}}$ decompose orthogonally into isotypic subspaces, $\mathcal{L}^2_{\mathbf{x}} = \bigoplus^{\perp}_{1\leq k\leq n_{\text{iso}}} \mathcal{L}^{2(k)}_{\mathbf{x}}$ and $\mathcal{L}^2_{\mathbf{y}} = \bigoplus^{\perp}_{1\leq k\leq n_{\text{iso}}} \mathcal{L}^{2(k)}_{\mathbf{y}}$, the operator T is completely determined by the n_{iso} operators $\mathsf{T}^{(k)} : \mathcal{L}^{2(k)}_{\mathbf{x}} \to \mathcal{L}^{2(k)}_{\mathbf{y}}$ (see App. K.1.2). Thus, the $\mathbb{G}$-equivariance of T_r depends on that of each finite-rank operator $\mathsf{T}^{(k)}_{r_k} : \mathcal{F}^{(k)}_{\mathbf{x}} \to \mathcal{F}^{(k)}_{\mathbf{y}}$, which requires the approximated subspaces $\mathcal{F}^{(k)}_{\mathbf{x}}$ and $\mathcal{F}^{(k)}_{\mathbf{y}}$ to be $\mathbb{G}$-stable. For simplicity, we assume $|\mathcal{F}^{(k)}_{\mathbf{x}}| = |\mathcal{F}^{(k)}_{\mathbf{y}}| = r_k$, although this equality need not hold in general.

K.2.1 Finite-rank approximation of $\mathbb{G}$-equivariant operators between isotypic subspaces

Each approximation of an isotypic subspace $\mathcal{L}^{2(k)}_{\mathbf{x}}$ (and similarly $\mathcal{L}^{2(k)}_{\mathbf{y}}$) is $\mathbb{G}$-stable if the group representation is defined using a truncated multiplicity $m_k < \infty$ for the k^{th} irreducible representation, i.e. $\boldsymbol{\rho}_{\mathcal{F}^{(k)}_{\mathbf{x}}} \sim \bigoplus^{m_k}_{p=1} \bar{\boldsymbol{\rho}}_k$ and $\boldsymbol{\rho}_{\mathcal{L}^{2(k)}_{\mathbf{y}}} \sim \bigoplus^{m_k}_{p=1} \bar{\boldsymbol{\rho}}_k$. Consequently, the dimension of the approximated subspaces is multiple of the irreducible representation's dimension: $r_k = d_k\, m_k$ (see App. K.1.3).

Given this structure, by Prop. K.3 the singular spaces of the finite-rank operators $\mathsf{T}^{(k)}_{r_k}$ have a minimum dimensionality of d_k. Consequently, the SVD of $\mathsf{T}^{(k)}_{r_k}$ exhibits a Kronecker structure:

$$
\begin{aligned}
\mathsf{T}^{(k)}_{r_k} &= \boldsymbol{U}^{(k)}\boldsymbol{\Sigma}^{(k)}(\boldsymbol{V}^{(k)})^* \in \mathbb{R}^{r_k\times r_k}, \\
&= (\boldsymbol{W}^{(k)} \otimes \boldsymbol{I}_k)\,(\boldsymbol{\Sigma}^{(k)} \otimes \boldsymbol{I}_k)\,((\boldsymbol{Q}^{(k)})^* \otimes \boldsymbol{I}_k)
\end{aligned}
\quad \text{s.t.} \quad
\begin{aligned}
\Theta^{(k)} &= \boldsymbol{W}^{(k)}\boldsymbol{\Sigma}^{(k)}(\boldsymbol{Q}^{(k)})^* \in \mathbb{R}^{m_k\times m_k}, \\
\operatorname{rank}(\boldsymbol{I}_k) &= d_k.
\end{aligned}
\tag{69}
$$

Here, $\Theta^{(k)}$ accounts for the m_k^2 degrees of freedom of $\mathsf{T}^{(k)}_{r_k}$, with each coefficient $\theta^{(k)}_{i,j}$ providing an isotropic scaling between the subspaces $\mathcal{L}^{2k,i}_{\mathbf{y}}$ and $\mathcal{L}^{2k,j}_{\mathbf{x}}$. Equation Eq. (69) constrains the finite-rank approximation of $\mathbb{G}$-equivariant operators between isotypic subspaces to approximate singular spaces of minimal dimensionality d_k.

This shows that the group representation on the isotypic basis also governs the singular (spectral) basis sets. As summarized in the following corollary:

Corollary K.4 (Group action on the spectral basis)**.** *The group representation on the spectral basis of each isotypic subspace $\mathcal{L}^{2(k)}_{\mathbf{x}}$ is given by its isotypic representation $\boldsymbol{\rho}_{\mathcal{L}^{2(k)}_{\mathbf{x}}} := \bigoplus^{m_k}_{p} \bar{\boldsymbol{\rho}}_k$. Similarly for $\mathcal{L}^{2(k)}_{\mathbf{y}}$.*

Proof. Lets consider a single isotypic subspace $\mathcal{L}^{2(k)}_{\mathbf{x}}$ and its group representation in the basis of singular functions:

$$
\begin{aligned}
\boldsymbol{\rho}^{\text{sng}}_{\mathcal{L}^{2(k)}_{\mathbf{x}}} := \boldsymbol{U}^{*(k)}\boldsymbol{\rho}_{\mathcal{L}^{2(k)}_{\mathbf{x}}}\boldsymbol{U}^{(k)} &= (\boldsymbol{W}^{*(k)} \otimes \boldsymbol{I}_{d_k})(\boldsymbol{I}_{m_k} \otimes \bar{\boldsymbol{\rho}}_k)(\boldsymbol{W}^{(k)} \otimes \boldsymbol{I}_{d_k}) \\
&= (\boldsymbol{W}^{*(k)} \otimes \bar{\boldsymbol{\rho}}_k)(\boldsymbol{W}^{(k)} \otimes \boldsymbol{I}_{d_k}) \\
&= (\boldsymbol{W}^{*(k)}\boldsymbol{W}^{(k)}) \otimes (\bar{\boldsymbol{\rho}}_k \boldsymbol{I}_{d_k}) = \bigoplus^{m_k}_{p} \bar{\boldsymbol{\rho}}_k = \boldsymbol{\rho}_{\mathcal{L}^{2(k)}_{\mathbf{x}}}.
\end{aligned}
\tag{70}
$$

□

K.2.2 Finite-rank approximation of a $\mathbb{G}$-equivariant operator

Given the block-diagonal structure of the operator T in the isotypic basis (Eq. (63)), the truncated SVD of T reduces to performing the truncated SVD of each per-isotypic operator $\mathsf{T}^{(k)}$ (Eq. (69)).

Let $\mathcal{F}_{\mathbf{x}} \subset \mathcal{L}^2_{\mathbf{x}}$ and $\mathcal{F}_{\mathbf{y}} \subset \mathcal{L}^2_{\mathbf{y}}$ be the $\mathbb{G}$-stable finite-dimensional approximations of the input/output spaces of T, endowed with group representations $\boldsymbol{\rho}_{\mathcal{F}_{\mathbf{x}}} = \bigoplus^{n_{\text{iso}}}_{k=1} \boldsymbol{\rho}_{\mathcal{F}^{(k)}_{\mathbf{x}}} = \bigoplus^{n_{\text{iso}}}_{k=1} \bigoplus^{m_k}_{p=1} \bar{\boldsymbol{\rho}}_k$ and $\boldsymbol{\rho}_{\mathcal{F}_{\mathbf{y}}} = \bigoplus^{n_{\text{iso}}}_{k=1} \boldsymbol{\rho}_{\mathcal{F}^{(k)}_{\mathbf{y}}} = \bigoplus^{n_{\text{iso}}}_{k=1} \bigoplus^{m_k}_{p=1} \bar{\boldsymbol{\rho}}_k$. Here, $m_k \in \mathbb{N}$ denotes the multiplicity of the irreducible representation of type k, and $d_k := |\bar{\boldsymbol{\rho}}_k|$ is its dimension. Then, the structural constraints on the SVD of the restriction of T to these spaces are summarized in the following theorem:

Theorem K.5 (Isotypic-spectral basis)**.** *Let T be a $\mathbb{G}$-equivariant operator and let $\mathsf{T}_\star : \mathcal{F}_{\mathbf{y}} \to \mathcal{F}_{\mathbf{x}}$ be its $\mathbb{G}$-equivariant restriction in finite dimensions. Then, the singular value decomposition of the restricted operator matrix representation $\boldsymbol{T}_\star$ reduces to:*

$$
\boldsymbol{T}_\star = \bigoplus^{n_{\text{iso}}}_{k=1} \boldsymbol{T}^{(k)}_\star = \bigoplus^{n_{\text{iso}}}_{k=1} \boldsymbol{W}^{(k)}_\star \boldsymbol{S}^{(k)}_\star \boldsymbol{M}^{(k)\top}_\star = \bigoplus^{n_{\text{iso}}}_{k=1} (\boldsymbol{U}^{(k)}_\star \boldsymbol{\Sigma}^{(k)}_\star \boldsymbol{V}^{(k)\top}_\star) \otimes \boldsymbol{I}_{d_k}
$$

Where $\boldsymbol{I}_{d_k}$ denotes the identity matrix in d_k-dimensions and $\boldsymbol{O}^{(k)} := \boldsymbol{U}^{(k)}_\star \boldsymbol{\Sigma}^{(k)}_\star \boldsymbol{V}^{(k)\top}_\star$ denotes the SVD of the free parameters of $\boldsymbol{T}^{(k)}_\star$.

Thm. K.5 shows that symmetries force each isotypic subspace's singular space to have dimension at least d_k, which is the minimum required for a faithful representation of $\mathbb{G}^{(k)}$ (see Def. I.7). Because in practice our goal is to approximate the top r singular spaces of T, this result precisely characterizes the constraints imposed by $\mathbb{G}$-equivariance on the optimal rank-r truncation's spectral basis and corresponding kernel function in Eq. (13), as summarized in the following corollary:

Corollary K.6 (Symmetry constraints on the spectral basis)**.** *Let T be a $\mathbb{G}$-equivariant operator and let $\mathsf{T}_\star : \mathcal{F}_{\mathbf{y}} \to \mathcal{F}_{\mathbf{x}}$ be its $\mathbb{G}$-equivariant restriction in r-dimensions. Then, the spectral basis of $\mathsf{T}_\star$ is given by:*

$$\kappa_\star(\boldsymbol{x}, \boldsymbol{y}) = \sum_{k=1}^{n_{\text{iso}}} \kappa_\star^{(k)}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{k=1}^{n_{\text{iso}}} \sum_{s=1}^{r_k} \sigma_s^{(k)} u_{s,i}^{(k)}(\boldsymbol{x}) v_{s,i}^{(k)}(\boldsymbol{y}), \tag{71}$$

where $\{u_{s,i}^{(k)}\}_{i\in[d_k]}$ and $\{v_{s,i}^{(k)}\}_{i\in[d_k]}$ are the left and right singular basis sets of the s^{th} singular space of $\mathsf{T}^{(k)}$. Note that the truncated dimension is restricted by the dimensionality and multiplicities of the individual irreducible representations $r = \sum_{k=1}^{d_{\text{iso}}} r_k = \sum_{k=1}^{d_{\text{iso}}} d_k m_k$.

L Relevant $\mathbb{G}$-equivariant operators in probability theory

In this section we study the properties of expectations and covariances of functions of symmetric random variables in the presence our assumed symmetry priors Eq. (6). In a nutshell, we characterize how expectations of observables of symmetric random variables are invariant to the group action, and that the covariance and cross-covariance matrices in these spaces are $\mathbb{G}$-equivariant and hence inherit rich structural constraints that can aid in empirical estimation.

Let $(\mathbf{x}, \mathbf{y})$ be two vector-valued random variables over the probability spaces $(\mathcal{X}, \Sigma_{\mathcal{X}}, P_{\mathbf{x}})$ and $(\mathcal{Y}, \Sigma_{\mathcal{Y}}, P_{\mathbf{y}})$, with $\mathcal{L}^2_{\mathbf{x}}$ and $\mathcal{L}^2_{\mathbf{y}}$ being the corresponding square-integrable function spaces and $\mathbb{1}_{P_{\mathbf{x}}} \in \mathcal{L}^2_{\mathbf{x}}$, $\mathbb{1}_{P_{\mathbf{y}}} \in \mathcal{L}^2_{\mathbf{y}}$ the characteristic functions of sets with nonzero probability.

When $\mathcal{L}^2_{\mathbf{x}}$ and $\mathcal{L}^2_{\mathbf{y}}$ are symmetric function spaces (see App. J), denote their orthogonal isotypic decompositions by $\mathcal{L}^2_{\mathbf{x}} := \bigoplus_{k=1}^{n_{\text{iso}}} \mathcal{L}^{2(k)}_{\mathbf{x}}$ and $\mathcal{L}^2_{\mathbf{y}} := \bigoplus_{k=1}^{n_{\text{iso}}} \mathcal{L}^{2(k)}_{\mathbf{y}}$ (cf. Thm. I.8). Any function $f \in \mathcal{L}^2_{\mathbf{x}}$ or $h \in \mathcal{L}^2_{\mathbf{y}}$ decomposes as $f = \sum_{k=1}^{n_{\text{iso}}} f^{(k)}$ and $h = \sum_{k=1}^{n_{\text{iso}}} h^{(k)}$ (see Eq. (61)). By convention, the first isotypic subspace corresponds to the trivial group action. Thus, we write $\mathcal{L}^{2\text{inv}}_{\mathbf{x}} := \mathcal{L}^{21}_{\mathbf{x}} \subset \mathcal{L}^2_{\mathbf{x}}$ and denote the $\mathbb{G}$-invariant component of f by $f^{\text{inv}} := f^{(1)}$ (and similarly for $\mathcal{L}^2_{\mathbf{y}}$).

L.1 The expectation operator

The expected value of a function $f \in \mathcal{F} := \mathcal{L}^2_{\mathbf{x}}$ can be interpreted as the result of applying a linear integral operator that projects each $f \in \mathcal{F}$ to a constant function evaluating to the function's expected value $\mathbb{E}_{P_{\mathbf{x}}} f$.

Definition L.1 (Expectation operator)**.** *Let $\mathcal{F} \subseteq \mathcal{L}^2_{\mathbf{x}}$ be a function space. The expectation operator $\mathsf{E}_{\mathbf{x}} : \mathcal{F} \mapsto \mathcal{F}$ is a linear integral operator defined by a constant kernel function $k_{\mathbb{E}}(\boldsymbol{x}, \boldsymbol{x}') = \mathbb{1}_{P_{\mathbf{x}}}(\boldsymbol{x})\mathbb{1}_{P_{\mathbf{x}}}(\boldsymbol{x}')$ for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$, such that this operator maps any function f to a constant function that evaluates to the function's expected value $\mathbb{1}_{P_{\mathbf{x}}}(\cdot)\mathbb{E}_{P_{\mathbf{x}}} f$, that is:*

$$[\mathsf{E}_{\mathbf{x}} f](\boldsymbol{x}') = \int_{\mathcal{X}} k_{\mathbb{E}}(\boldsymbol{x}, \boldsymbol{x}') f(\boldsymbol{x}) \mu(d\boldsymbol{x}) = \mathbb{1}_{P_{\mathbf{x}}}(\boldsymbol{x}') \int_{\mathcal{X}} f(\boldsymbol{x}) \mu(d\boldsymbol{x}) \equiv \mathbb{1}_{P_{\mathbf{x}}}(\boldsymbol{x}') \mathbb{E}_{P_{\mathbf{x}}} f. \tag{72}$$

Whenever $\mathcal{F}$ is a symmetric function space, the operator $\mathsf{E}_{\mathbf{x}}$ commutes with the group action and is $\mathbb{G}$-invariant (Def. I.10):

Proposition L.2 ($\mathbb{G}$-invariant expectation operator)**.** *Let $\mathcal{F}$ be a symmetric function space with the action $\triangleright_{\mathcal{F}}$ of a compact symmetry group $\mathbb{G}$. Then, the expectation operator commutes with the group action and is a $\mathbb{G}$-invariant operator $\mathsf{E}_{\mathbf{x}} : \mathcal{F} \mapsto \mathcal{F}^{inv} \subseteq \mathcal{F}$:*

$$\mathsf{E}_{\mathbf{x}}[g \triangleright_{\mathcal{F}} f] = g \triangleright_{\mathcal{F}} [\mathsf{E}_{\mathbf{x}} f] \qquad \textit{and} \qquad \mathsf{E}_{\mathbf{x}} f = \mathsf{E}_{\mathbf{x}}[g \triangleright_{\mathcal{F}} f] \in \mathcal{F}^{inv}, \qquad \forall\, f \in \mathcal{F}, g \in \mathbb{G}. \tag{73}$$

Proof. The operator $\mathsf{E}_{\mathbf{x}}$ commutes with the group action as its kernel function $k_{\mathbb{E}}$ is constant and therefore $\mathbb{G}$-invariant (Def. K.1). Furthermore since the image of the expectation operator are constant functions, these functions belong to the subspace of $\mathbb{G}$-invariant functions, $\mathcal{F}^{\text{inv}}$. □

As an operator that commutes with the group action, the expectation operator decomposes into $\mathsf{E}_{\mathbf{x}} := \bigoplus_{k=1}^{n_{\text{iso}}} \mathsf{E}_{\mathbf{x}}^{(k)}$, where $\mathsf{E}_{\mathbf{x}}^{(k)} : \mathcal{F}^{(k)} \mapsto \mathcal{F}^{(k)}$ denotes the restriction of $\mathsf{E}_{\mathbf{x}}$ to the isotypic subspace $\mathcal{F}^{(k)}$ (App. K.1.2). However, since the image of the operator lies in the subspace of $\mathbb{G}$-invariant functions, $\mathrm{Im}(\mathsf{E}_{\mathbf{x}}) \subset \mathcal{F}^{\text{inv}}$, it follows that $\mathsf{E}_{\mathbf{x}}^{(k)} = \mathbf{0}$ for every $k \neq$ inv. Consequently, we obtain the following:

Corollary L.3 (Expectation of a function depends only on its $\mathbb{G}$-invariant component)**.** *For any function $f \in \mathcal{F}$, the expectation depends only on its $\mathbb{G}$-invariant component:*

$$[\mathsf{E}_{\mathbf{x}} f](\cdot) = \sum_{k=1}^{n_{\text{iso}}} [\mathsf{E}_{\mathbf{x}}^{(k)} f^{(k)}](\cdot) = [\mathsf{E}_{\mathbf{x}}^{inv} f^{inv}](\cdot) := \mathbb{1}_{\mu}(\cdot)\mathbb{E}_{\mu} f^{inv}. \tag{74}$$

Corollary L.4 (Functions without a $\mathbb{G}$-invariant component are centered)**.** *Any function $f = \sum_{k=1}^{n_{\text{iso}}} f^{(k)} \in \mathcal{L}^2_{\mathbf{x}}$ without a $\mathbb{G}$-invariant compoment, i.e., $f^{inv} = 0$, is centered:*

$$[\mathsf{E}_{\mathbf{x}} f](\cdot) = \sum_{k=2}^{n_{\text{iso}}} [\mathsf{E}_{\mathbf{x}}^{(k)} f^{(k)}](\cdot) = \mathbb{1}_{\mu}(\cdot)0, \qquad \Longleftrightarrow \qquad \mathbb{E}_{\mu} f = 0, \quad \forall\, f \in \mathcal{L}^{2\,inv\,\perp}_{\mathbf{x}}. \tag{75}$$

To better comprehend these concepts we refer the reader to Example J.4.

L.2 The cross-covariance operator

Given two vector-valued random variables ($\mathbf{x} = [\mathrm{x}_1, \dots, \mathrm{x}_n], \mathbf{y} = [\mathrm{y}_1, \dots, \mathrm{y}_m]$) defined on the measure spaces $(\mathcal{X}, \Sigma_{\mathcal{X}}, P_{\mathbf{x}})$ and $(\mathcal{Y}, \Sigma_{\mathcal{Y}}, P_{\mathbf{y}})$, a key statistic assessing the linear relationship between scalar components is the covariance:

$$\mathrm{Cov}(\mathrm{x}_i, \mathrm{y}_j) = \mathbb{E}_{P_{\mathbf{xy}}}[(\mathrm{x}_i - \mathbb{E}_{\mathbf{x}}[\mathrm{x}_i])(\mathrm{y}_j - \mathbb{E}_{\mathbf{y}}[\mathrm{y}_j])] = \mathbb{E}_{P_{\mathbf{xy}}}[\mathrm{x}_i \mathrm{y}_j] - \mathbb{E}_{\mathbf{x}}[\mathrm{x}_i]\mathbb{E}_{\mathbf{y}}[\mathrm{y}_j].$$

For vector-valued random variables, the cross-covariance matrix $\mathrm{Cov}(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{n \times m}$ is defined entrywise by $\mathrm{Cov}(\mathbf{x}, \mathbf{y})_{i,j} := \mathrm{Cov}(\mathrm{x}_i, \mathrm{y}_j)$. The cross-covariance operator is the extension of this concept to the Hilbert spaces of functions $\mathcal{L}^2_{\mathbf{x}}$ and $\mathcal{L}^2_{\mathbf{y}}$.

Definition L.5 (Cross-covariance operator [28])**.** *Let $\mathcal{F}_{\mathbf{x}} \subseteq \mathcal{L}^2_{\mathbf{x}}$ and $\mathcal{L}^2_{\mathbf{y}} \subseteq \mathcal{L}^2_{\mathbf{y}}$ be two Hilbert spaces of functions defined on the random variables $\mathbf{x}$ and $\mathbf{y}$, which take values in the measure spaces $(\mathcal{X}, \Sigma_{\mathcal{X}}, P_{\mathbf{x}})$ and $(\mathcal{Y}, \Sigma_{\mathcal{Y}}, P_{\mathbf{y}})$, respectively. The cross-covariance operator $\mathsf{C}_{\mathbf{xy}} : \mathcal{L}^2_{\mathbf{y}} \mapsto \mathcal{L}^2_{\mathbf{x}}$ is a linear integral operator defined by*

$$\langle f, \mathsf{C}_{\mathbf{xy}} h \rangle_{P_{\mathbf{x}}} := \mathrm{Cov}(f, h) = \mathbb{E}_{P_{\mathbf{xy}}}[f(\mathbf{x})h(\mathbf{y})] - \mathbb{E}_{\mathbf{x}}[f(\mathbf{x})]\mathbb{E}_{\mathbf{y}}[h(\mathbf{y})], \qquad \forall\, f \in \mathcal{L}^2_{\mathbf{x}},\ h \in \mathcal{L}^2_{\mathbf{y}}. \tag{76}$$

Choosing separable basis sets for the two spaces, $\mathbb{I}_{\mathcal{L}^2_{\mathbf{x}}} = \{\phi_i\}_{i \in \mathbb{N}}$ and $\mathbb{I}_{\mathcal{L}^2_{\mathbf{y}}} = \{\psi_i\}_{i \in \mathbb{N}}$, the matrix representation of the cross-covariance operator has entries $[\boldsymbol{C}_{\mathbf{x},\mathbf{y}}]_{i,j} := \langle \phi_i, \mathsf{C}_{\mathbf{xy}} \psi_j \rangle_{P_{\mathbf{x}}} = \mathrm{Cov}(\phi_i, \psi_j)$, where the covariance is computed with respect to the joint measure $P_{\mathbf{xy}}$ and the marginals $P_{\mathbf{x}}$ and $P_{\mathbf{y}}$. Given a dataset of N samples from the joint distribution $(\boldsymbol{x}, \boldsymbol{y}) \sim P_{\mathbf{xy}}$, the empirical estimate of the matrix form of the cross-covariance operator is

$$\widehat{\boldsymbol{C}}_{\mathbf{xy}} = \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{\phi}(\boldsymbol{x}_n)\boldsymbol{\psi}(\boldsymbol{y}_n)^{\top} - \widehat{\mathbb{E}}_{\mathbf{x}}[\boldsymbol{\phi}(\boldsymbol{x}_n)]\,\widehat{\mathbb{E}}_{\mathbf{y}}[\boldsymbol{\psi}(\boldsymbol{y}_n)]^{\top}, \quad \boldsymbol{\phi}(\cdot) = [\phi(\cdot)]_{i \in \mathbb{N}},\ \boldsymbol{\psi}(\cdot) = [\psi(\cdot)]_{i \in \mathbb{N}}. \tag{77}$$

Note that the adjoint of the operator is defined by $\mathsf{C}^*_{\mathbf{xy}} = \mathsf{C}_{\mathbf{yx}} : \mathcal{L}^2_{\mathbf{x}} \mapsto \mathcal{L}^2_{\mathbf{y}}$. In the case $\mathcal{L}^2_{\mathbf{x}} = \mathcal{L}^2_{\mathbf{y}}$, the cross-covariance operator reduces to the covariance operator, and has an analog definition to Def. L.5.

Covariance and cross-covariance operators of symmetric Hilbert spaces of functions Whenever $\mathcal{L}^2_{\mathbf{x}}$ and $\mathcal{L}^2_{\mathbf{y}}$ are symmetric function spaces, and the joint probability measure is $\mathbb{G}$-invaraint, the cross-covariance operator $\mathsf{C}_{\mathbf{xy}}$ commute with the group action and is $\mathbb{G}$-equivariant (App. I.2):

Proposition L.6 ($\mathbb{G}$-equivariant cross-covariance operator)**.** *Let $\mathcal{L}^2_{\mathbf{x}} \subseteq \mathcal{L}^2_{\mathbf{x}}$ and $\mathcal{L}^2_{\mathbf{y}} \subseteq \mathcal{L}^2_{\mathbf{y}}$ be symmetric Hilbert spaces of functions endowed with the group actions $\triangleright_{\mathcal{L}^2_{\mathbf{x}}}$ and $\triangleright_{\mathcal{L}^2_{\mathbf{y}}}$ of a compact symmetry group $\mathbb{G}$. Then, whenever the joint probability measure is $\mathbb{G}$-invariant, i.e., $P_{\mathbf{xy}}(\mathbb{B}, \mathbb{A}) = P_{\mathbf{xy}}(g \triangleright_{\mathcal{X}} \mathbb{B}, g \triangleright_{\mathcal{Y}} \mathbb{A})$ for all $g \in \mathbb{G}, \mathbb{B} \in \Sigma_{\mathcal{X}}, \mathbb{A} \in \Sigma_{Y}$, the cross-covariance operator $\mathsf{C}_{\mathbf{xy}} : \mathcal{L}^2_{\mathbf{y}} \mapsto \mathcal{L}^2_{\mathbf{x}}$ (Def. L.5) commutes with the group actions and is a $\mathbb{G}$-equivariant operator (Def. K.1):*

$$g \triangleright_{\mathcal{L}^2_{\mathbf{x}}} [\mathsf{C}_{\mathbf{xy}} h] = \mathsf{C}_{\mathbf{xy}}[g \triangleright_{\mathcal{L}^2_{\mathbf{y}}} h], \qquad \forall\, h \in \mathcal{L}^2_{\mathbf{y}}, g \in \mathbb{G}. \tag{78}$$

Proof. To proof that the operator is $\mathbb{G}$-equivariant we must show its kernel function is $\mathbb{G}$-invariant (see Def. K.1). The proof follows naturally in any regular basis of the input and output functions spaces $\mathbb{I}_{\mathcal{L}^2_{\mathbf{x}}} = \{\phi_i\}_{i\in\mathbb{N}}$ and $\mathbb{I}_{\mathcal{L}^2_{\mathbf{y}}} = \{\psi_i\}_{i\in\mathbb{N}}$, in which the group action on basis functions acts by permutations of basis functions, such that, $g \triangleright_{\mathcal{L}^2_{\mathbf{x}}} \phi_i \equiv \phi_{g\triangleright i} \in \mathbb{I}_{\mathcal{L}^2_{\mathbf{x}}}$ and $g \triangleright_{\mathcal{L}^2_{\mathbf{y}}} \psi_j \equiv \psi_{g\triangleright j} \in \mathbb{I}_{\mathcal{L}^2_{\mathbf{y}}}$, where $g \triangleright i, g \triangleright j \in \mathbb{N}$. Then we must show that that:

$$\begin{aligned} k(\boldsymbol{x},\boldsymbol{y}) &= k(g^{-1} \triangleright_{\mathcal{X}} \boldsymbol{x}, g^{-1} \triangleright_{\mathcal{Y}} \boldsymbol{y}) && \forall\, g\in\mathbb{G}, \boldsymbol{x}\in\mathcal{X}, \boldsymbol{y}\in\mathcal{Y} \\ \sum_{i\in\mathbb{N}}\sum_{j\in\mathbb{N}} [\boldsymbol{C}_{\mathbf{x},\mathbf{y}}]_{i,j}\phi_i(\boldsymbol{x})\psi_j(\boldsymbol{y}) &= \sum_{i\in\mathbb{N}}\sum_{j\in\mathbb{N}} [\boldsymbol{C}_{\mathbf{x},\mathbf{y}}]_{i,j}[g \triangleright_{\mathcal{L}^2_{\mathbf{x}}} \phi_i](\boldsymbol{x})[g \triangleright_{\mathcal{Y}} \psi_j](\boldsymbol{y}) && \text{s.t. Defs. J.1 and L.5} \\ \sum_{i\in\mathbb{N}}\sum_{j\in\mathbb{N}} \mathrm{Cov}(\phi_i,\psi_j)\phi_i(\boldsymbol{x})\psi_j(\boldsymbol{y}) &= \sum_{i\in\mathbb{N}}\sum_{j\in\mathbb{N}} \mathrm{Cov}(\phi_i,\psi_j)\phi_{g\triangleright i}(\boldsymbol{x})\psi_{g\triangleright j}(\boldsymbol{y}). \end{aligned} \tag{79}$$

Hence, the cross-covariance operator's kernel function is $\mathbb{G}$-invariant only if the covariance is $\mathbb{G}$-invariant:

$$\begin{aligned} \mathrm{Cov}(\phi_i,\psi_j) &= \mathrm{Cov}(g \triangleright_{\mathcal{L}^2_{\mathbf{x}}} \phi_i, g \triangleright_{\mathcal{Y}} \psi_j) && \forall\, g\in\mathbb{G}, i,j\in\mathbb{N} \\ \mathbb{E}_{P_{\mathbf{xy}}}[\phi_i(\mathbf{x})\psi_j(\mathbf{y})] &= \mathbb{E}_{P_{\mathbf{xy}}}[\phi_i(g^{-1} \triangleright_{\mathcal{X}} \mathbf{x})\psi_j(g^{-1} \triangleright_{\mathcal{Y}} \mathbf{y})] && \mathbb{E}_\mu f = \mathbb{E}_\mu g \triangleright f \\ \int_{\mathcal{X}\times\mathcal{Y}} \phi_i(\mathbf{x})\psi_j(\mathbf{y})\, P_{\mathbf{xy}}(d\mathbf{x}, d\mathbf{y}) &= \int_{\mathcal{X}\times\mathcal{Y}} \phi_i(g^{-1} \triangleright_{\mathcal{X}} \mathbf{x})\psi_j(g^{-1} \triangleright_{\mathcal{Y}} \mathbf{y})\, P_{\mathbf{xy}}(d\mathbf{x}, d\mathbf{y}) \\ &= \int_{\mathcal{X}\times\mathcal{Y}} \phi_i(\mathbf{x})\psi_j(\mathbf{y})\, P_{\mathbf{xy}}(g \triangleright d\mathbf{x}, g \triangleright d\mathbf{y}) \\ &= \mathrm{Cov}(\phi_i,\psi_j). \end{aligned} \tag{80}$$

$\square$

An equivalent result follows for covariance operators of symmetric Hilbert spaces.

M Statistical Learning Theory

This section provides the development and proofs of the statistical learning guarantees in Thm. C.1 for regression and conditional probability estimation using our proposed model.

Recall that regression and conditional probabilities can be expressed in terms of the conditional expectation operator $\mathsf{E}_{\mathbf{y}|\mathbf{x}} \colon \mathcal{L}^2_{\mathbf{y}} \to \mathcal{L}^2_{\mathbf{x}}$ (see Eqs. (1) and (2)). Given that the operator is compact [37], it admits a singular value decomposition. Hence, the kernel function defining the operator Eq. (1) can be expanded in terms of the operator spectral basis:

$$\kappa(\boldsymbol{x},\boldsymbol{y}) := \frac{dP_{\mathbf{xy}}(\boldsymbol{x},\boldsymbol{y})}{d(P_{\mathbf{x}}(\boldsymbol{x}) \times P_{\mathbf{y}}(\boldsymbol{y}))} = \sum_{i=0}^{\infty} \sigma_i u_i(\boldsymbol{x}) v_i(\boldsymbol{y}). \tag{81}$$

Where $(\sigma_i)_{i\in\mathbb{N}}$ denotes the operator's singular values, and $(u_i)_{i\in\mathbb{N}}$ and $(v_i)_{i\in\mathbb{N}}$ denote the left and right singular functions, which form complete orthonormal basis sets for $\mathcal{L}^2_{\mathbf{x}}$ and $\mathcal{L}^2_{\mathbf{y}}$, respectively. Given that the operator's first singular value is $\sigma_0 = 1$, associated with the constant functions $u_0 = \mathbb{1}_{\mathcal{X}}$, $v_0 = \mathbb{1}_{\mathcal{Y}}$, the conditional expectation operator can be defined as:

$$\mathsf{E}_{\mathbf{y}|\mathbf{x}} = \sum_{i=1}^{\infty} \sigma_i u_i \langle v_i, \cdot\rangle_{P_{\mathbf{y}}} = \mathbb{1}_{\mathcal{X}}\langle \mathbb{1}_{\mathcal{Y}}, \cdot\rangle_{P_{\mathbf{y}}} + \underbrace{\sum_{i=1}^{\infty} \sigma_i u_i \langle v_i, \cdot\rangle_{P_{\mathbf{y}}}}_{\mathsf{D}_{\mathbf{y}|\mathbf{x}}}. \tag{82}$$

Where $\mathsf{D}_{\mathbf{y}|\mathbf{x}}$ denotes the *deflated* operator, excluding the first eigen triplet (σ_0, u_0, v_0). Leveraging the SVD of $\mathsf{E}_{\mathbf{y}|\mathbf{x}}$, we approximate the operator's action for any $h \in \mathcal{L}^2_{\mathbf{y}}$ using a rank-r $(1 < r < \infty)$ operator given by:

$$\begin{gathered} \mathbb{E}[h(\mathbf{y})|\mathbf{x}{=}\boldsymbol{x}] = [\mathsf{E}_{\mathbf{y}|\mathbf{x}} h](\boldsymbol{x}) \approx \mathbb{E}[h(\mathbf{y})] + \sum_{i=1}^{r} \sigma_i u_i^{\boldsymbol{\theta}}(\boldsymbol{x}) \mathbb{E}[v_i^{\boldsymbol{\theta}}(\mathbf{y}) h(\mathbf{y})], \\ \text{s.t. } \mathbb{E}[u_i^{\boldsymbol{\theta}}(\mathbf{x})] = \mathbb{E}[v_i^{\boldsymbol{\theta}}(\mathbf{y})] = 0, \forall\, i \geq 1. \end{gathered} \tag{83}$$

Where $(u_i^{\boldsymbol{\theta}})_{i=1}^r$ and $(v_i^{\boldsymbol{\theta}})_{i=1}^r$ denote parametrizations of the top-r left and right singular functions. Given that the operator's kernel Eq. (81) preserves the probability mass, that is $\int_{\mathcal{X}\times\mathcal{Y}} \kappa(\boldsymbol{x},\boldsymbol{y})dP_{\mathbf{x}}(\boldsymbol{x})dP_{\mathbf{y}}(\boldsymbol{y}) = 1$, every non-constant singular function is constrained to be centered, as described in the r.h.s of Eq. (83).

In the context of symmetries, we note that $\mathsf{D}_{\mathbf{y}|\mathbf{x}}$ admits a block-diagonal structure w.r.t. to isotypic basis of associated $\mathcal{L}^2$ spaces. Indeed we have the following from Thm. K.5.

$$\mathsf{Q}_{\mathbf{x}}^*\mathsf{D}_{\mathbf{y}|\mathbf{x}}\mathsf{Q}_{\mathbf{y}} = \oplus_{k=1}^{n_{\text{iso}}}\mathsf{Q}_{\mathbf{x}}^{(k)*}\mathsf{D}_{\mathbf{y}|\mathbf{x}}^{(k)}\mathsf{Q}_{\mathbf{y}}^{(k)} = \oplus_{k=1}^{n_{\text{iso}}}\big[(\mathsf{U}^{(k)}\mathsf{S}^{(k)}\mathsf{V}^{(k)*})\otimes\boldsymbol{I}_{d_k}\big]. \tag{84}$$

Where the unitary operators $\mathsf{Q}_{\mathbf{x}}\colon \mathcal{L}^2_{\mathbf{x}} \to \mathcal{L}^2_{\mathbf{x}}$ and $\mathsf{Q}_{\mathbf{y}}\colon \mathcal{L}^2_{\mathbf{y}} \to \mathcal{L}^2_{\mathbf{y}}$ change the basis to the isotypic decompositions $\mathbb{I}_{\mathcal{L}^2_{\mathbf{x}}} = \{\phi_{i,j}^{(k)}\}_{k\in[n_{\text{iso}}],\, i\in[m_k],\, j\in[d_k]}$ and $\mathbb{I}_{\mathcal{L}^2_{\mathbf{y}}} = \{\psi_{i,j}^{(k)}\}_{k\in[n_{\text{iso}}],\, i\in[m_k],\, j\in[d_k]}$, with i indexing each irreducible $\mathbb{G}$-stable subspace and j indexing the dimensions within that subspace (see App. K.2.2).

Further, by Thm. K.5, the SVD of $\mathsf{D}_{\mathbf{y}|\mathbf{x}}$ forces each isotypic subspace to have dimension at least $d_k = \bar{\boldsymbol{\rho}}_k$ for every $k\in[n_{\text{iso}}]$.

$$\mathsf{Q}_{\mathbf{x}}^{(k)*}\mathsf{D}_{\mathbf{y}|\mathbf{x}}^{(k)}\mathsf{Q}_{\mathbf{y}}^{(k)} = \big[\mathsf{U}^{(k)}\otimes\boldsymbol{I}_{d_k}\big]\big[\mathsf{S}^{(k)}\otimes\boldsymbol{I}_{d_k}\big]\big[\mathsf{V}^{(k)}\otimes\boldsymbol{I}_{d_k}\big]^*,\ k\in[n_{\text{iso}}], \tag{85}$$

where $\mathsf{Q}_{\mathbf{x}}^{(k)}\mathsf{Q}_{\mathbf{x}}^{(k)*}$ and $\mathsf{Q}_{\mathbf{y}}^{(k)}\mathsf{Q}_{\mathbf{y}}^{(k)*}$ are orthogonal projectors on k-th isotypic subspace, and

$$\mathsf{Q}_{\mathbf{x}}^*\mathsf{D}_{\mathbf{y}|\mathbf{x}}\mathsf{Q}_{\mathbf{y}} = \big[\boldsymbol{I}_{n_{\text{iso}}}\otimes\mathsf{U}^{(k)}\otimes\boldsymbol{I}_{d_k}\big]\big[I_{n_{\text{iso}}}\otimes\mathsf{S}^{(k)}\otimes\boldsymbol{I}_{d_k}\big]\big[I_{n_{\text{iso}}}\otimes\mathsf{V}^{(k)}\otimes\boldsymbol{I}_{d_k}\big]^*. \tag{86}$$

Further, observe that the singular values of $\mathsf{D}_{\mathbf{y}|\mathbf{x}}$ are elements of positive diagonal operators $\mathsf{S}^{(k)}$, denoted as $(\mathsf{S}^{(k)})_i = \sigma_i^{(k)}$, while the left and right singular functions are $u_i^{(k)}\otimes\boldsymbol{e}_j^{d_k}$ and $v_i^{(k)}\otimes\boldsymbol{e}_j^{d_k}$, respectively, for $i\in\mathbb{N}$, $j\in[d_k]$ and $k\in[n_{\text{iso}}]$, where $\boldsymbol{e}_j^d$ is j-th vector of standard basis of $\mathbb{R}^d$.

Given the constraints on the spectral basis of $\mathbb{G}$-equivariant operators (see Cor. K.6), our representation learning procedure approach results in feature maps:

$$\begin{aligned}
\boldsymbol{u}_{\boldsymbol{\theta}}(\cdot) &= \sum_{k\in[n_{\text{iso}}],i\in[m],j\in[d_k]}[\boldsymbol{e}_k^{n_{\text{iso}}}\otimes\boldsymbol{e}_i^m\otimes\boldsymbol{e}_j^{d_k}]\,u_{i,j}^{\boldsymbol{\theta}(k)}(\cdot)\colon\mathcal{X}\to\mathbb{R}^{r_m}\\
\boldsymbol{v}_{\boldsymbol{\theta}}(\cdot) &= \sum_{k\in[n_{\text{iso}}],i\in[m],j\in[d_k]}[\boldsymbol{e}_k^{n_{\text{iso}}}\otimes\boldsymbol{e}_i^m\otimes\boldsymbol{e}_j^{d_k}]\,v_{i,j}^{\boldsymbol{\theta}(k)}(\cdot)\colon\mathcal{X}\to\mathbb{R}^{r_m},
\end{aligned} \tag{87}$$

which can further be separated into n_{iso} orthogonal blocks $\boldsymbol{u}_{\boldsymbol{\theta}}^{(k)} = \sum_{i\in[m],j\in[d_k]}\phi_{i,j}^{\boldsymbol{\theta}(k)}$ and $\boldsymbol{\psi}_{\boldsymbol{\theta}}^{(k)} = \sum_{i\in[m],j\in[d_k]}\psi_{i,j}^{\boldsymbol{\theta}(k)}$ as

$$\boldsymbol{u}_{\boldsymbol{\theta}}^{(k)} = \sum_{i\in[m],j\in[d_k]}[\boldsymbol{e}_i^m\otimes\boldsymbol{e}_j^{d_k}]\,u_{i,j}^{\boldsymbol{\theta}(k)}(\cdot)\quad\text{and}\quad\boldsymbol{v}_{\boldsymbol{\theta}}^{(k)} = \sum_{i\in[m],j\in[d_k]}[\boldsymbol{e}_i^m\otimes\boldsymbol{e}_j^{d_k}]\,v_{i,j}^{\boldsymbol{\theta}(k)}(\cdot). \tag{88}$$

In addition, the singular value matrices have a tensor form $\boldsymbol{S}_{\boldsymbol{\theta}} = diag(\boldsymbol{S}_{\boldsymbol{\theta}}^{(1)},\ldots,\boldsymbol{S}_{\boldsymbol{\theta}}^{(n_{\text{iso}})})$, where $\boldsymbol{S}_{\boldsymbol{\theta}}^{(k)} = diag(\sigma_1^{\boldsymbol{\theta}(k)},\ldots,\sigma_m^{\boldsymbol{\theta}(k)})\otimes\boldsymbol{I}_{d_k}$ and $\sigma_i^{\boldsymbol{\theta}(k)}\in[0,1]$, $i\in[m], k\in[n_{\text{iso}}]$. Thus, we obtain the operator $\mathsf{D}_{\boldsymbol{\theta}} = \mathsf{E}_{\boldsymbol{\theta}} - \mathbb{1}_{P_{\mathbf{x}}}\otimes\mathbb{1}_{P_{\mathbf{y}}}$ in block form, $\mathsf{D}_{\boldsymbol{\theta}} = \oplus_{k\in[n_{\text{iso}}]}\mathsf{D}_{\boldsymbol{\theta}}^{(k)}$, where each $\mathsf{D}_{\boldsymbol{\theta}}^{(k)}$ acts on the k-th isotypic subspace as

$$[\mathsf{D}_{\boldsymbol{\theta}}^{(k)}f](\boldsymbol{x}) := \boldsymbol{u}_{\boldsymbol{\theta}}^{(k)}(\boldsymbol{x})^\top\boldsymbol{S}_{\boldsymbol{\theta}}^{(k)}\,\mathbb{E}_{\mathbf{y}}[\boldsymbol{v}_{\boldsymbol{\theta}}^{(k)}(\mathbf{y})f^{(k)}(\mathbf{y})],\quad f\in\mathcal{L}^2_{\mathbf{y}}, \tag{89}$$

and hence

$$[\mathsf{D}_{\boldsymbol{\theta}}f](\boldsymbol{x}) := \boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x})^\top\boldsymbol{S}_{\boldsymbol{\theta}}\,\mathbb{E}_{\mathbf{y}}[\boldsymbol{v}_{\boldsymbol{\theta}}(\mathbf{y})f(\mathbf{y})],\quad f\in\mathcal{L}^2_{\mathbf{y}}. \tag{90}$$

Finally, we extend the definition of $\mathsf{D}_{\boldsymbol{\theta}}$ to vector-valued observables $\boldsymbol{h}\colon\mathcal{Y}\to\mathcal{Z}$ via basis expansions.

$$[\mathsf{D}_{\boldsymbol{\theta}}\boldsymbol{h}](\boldsymbol{x}) := \sum_{\ell}\boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x})^\top\boldsymbol{S}_{\boldsymbol{\theta}}\,\mathbb{E}_{\mathbf{y}}[\boldsymbol{v}_{\boldsymbol{\theta}}(\mathbf{y})(\langle\boldsymbol{h}(\mathbf{y}),\boldsymbol{z}_\ell\rangle_{\mathcal{Z}}\boldsymbol{z}_\ell)],\quad\boldsymbol{h}\in\mathcal{L}^2_{\mathbf{y}}(\mathcal{Y},\mathcal{Z}) \tag{91}$$

where $(\boldsymbol{z}_i)_{i\in[n_{\mathcal{Z}}]}$ is the orthonormal basis of $\mathcal{Z}$.

By doing so, we ensure that $\mathsf{D}_{\boldsymbol{\theta}}$ and, consequently, $\mathsf{E}_{\boldsymbol{\theta}}$ are $\mathbb{G}$-equivariant operators for both the scalar map $\mathcal{L}^2_{\mathbf{y}}\to\mathcal{L}^2_{\mathbf{x}}$ and the vector-valued map $\mathcal{L}^2_{\mathbf{y}}(\mathcal{Y},\mathcal{Z})\to\mathcal{L}^2_{\mathbf{x}}(\mathcal{X},\mathcal{Z})$. Moreover, a direct consequence of (91) is as follows.

Proposition M.1. *Let with $\mathcal{Z}$ being a real Euclidean space endowed with symmetry group $\mathbb{G}$, and let $\mathsf{E}_{\boldsymbol{\theta}} \colon \mathcal{L}^2_{P_{\mathbf{y}}}(\mathcal{Y},\mathcal{Z}) \mapsto \mathcal{L}^2_{P_{\mathbf{x}}}(\mathcal{X},\mathcal{Z})$ be given by $\mathsf{E}_{\boldsymbol{\theta}}\boldsymbol{f} = \mathbb{E}_{\mathbf{y}}[\boldsymbol{f}(\mathbf{y})] + \mathsf{D}_{\boldsymbol{\theta}}\boldsymbol{f}$. Then for every $\mathbb{G}$-equivariant $\boldsymbol{f} \in \mathcal{L}^2_{P_{\mathbf{y}}}(\mathcal{Y},\mathcal{Z})$ and every $\boldsymbol{x} \in \mathcal{X}$*

$$[\mathsf{E}_{\boldsymbol{\theta}}\boldsymbol{f}](g \triangleright_{\mathcal{X}} \boldsymbol{x}) = \mathbb{E}_{\mathbf{y}}[\boldsymbol{f}(\mathbf{y})] + [\mathsf{D}_{\boldsymbol{\theta}}\boldsymbol{f}](g \triangleright_{\mathcal{X}} \boldsymbol{x}) = \mathbb{E}_{\mathbf{y}}[\boldsymbol{f}(\mathbf{y})] + g \triangleright_{\mathcal{Z}} [\mathsf{D}_{\boldsymbol{\theta}}\boldsymbol{f}](\boldsymbol{x}) = g \triangleright_{\mathcal{Z}} [\mathsf{E}_{\boldsymbol{\theta}}\boldsymbol{f}](\boldsymbol{x}). \tag{92}$$

Proof. Since $\mathsf{D}_{\boldsymbol{\theta}}$ is $\mathbb{G}$-equivaraint, for every $g \in \mathbb{G}$ we have that

$$[\mathsf{D}_{\boldsymbol{\theta}}\boldsymbol{h}](g^{-1} \triangleright_{\mathcal{X}} \boldsymbol{x}) = [\mathsf{D}_{\boldsymbol{\theta}}[\boldsymbol{h}(g^{-1} \triangleright_{\mathcal{Y}} \cdot)]](\boldsymbol{x}) = \sum_i \boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x})^\top \boldsymbol{S}_{\boldsymbol{\theta}}\, \mathbb{E}_{\mathbf{y}}[\boldsymbol{v}_{\boldsymbol{\theta}}(\mathbf{y}) \langle \boldsymbol{h}(g^{-1} \triangleright_{\mathcal{Y}} \mathbf{y}), \boldsymbol{z}_i\rangle_{\mathcal{Z}} \boldsymbol{z}_i],$$

which, using g instead of g^{-1} and the assumption that f is $\mathbb{G}$-equivariant, implies

$$\begin{aligned}[\mathsf{D}_{\boldsymbol{\theta}}\boldsymbol{h}](g \triangleright_{\mathcal{X}} \boldsymbol{x}) &= \sum_i (\boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x})^\top \boldsymbol{S}_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{y}}[\boldsymbol{v}_{\boldsymbol{\theta}}(\mathbf{y}) \langle g \triangleright_{\mathcal{Z}} \boldsymbol{h}(\mathbf{y}), \boldsymbol{z}_i\rangle_{\mathcal{Z}} \boldsymbol{z}_i] \\ &= \sum_i (\boldsymbol{u}_{\boldsymbol{\theta}}(\mathrm{x})^\top \boldsymbol{S}_{\boldsymbol{\theta}}\, \mathbb{E}_{\mathbf{y}}[\boldsymbol{v}_{\boldsymbol{\theta}}(\mathbf{y}) \langle \boldsymbol{h}(\mathbf{y}), g^{-1} \triangleright_{\mathcal{Z}} \boldsymbol{z}_i\rangle_{\mathcal{Z}}) \boldsymbol{z}_i.\end{aligned}$$

Thus, changing the basis to $(g^{-1} \triangleright_{\mathcal{Z}} \boldsymbol{z}_i)_{i \in [n_{\mathcal{Z}}]}$ we obtain the result when $\mathbb{E}_{\mathbf{y}}[\boldsymbol{h}(\mathbf{y})] = 0$. But since $\mathbb{1}_{\mathcal{X}}(g \triangleright_{\mathcal{X}} \boldsymbol{x}) = 1$ for every $\boldsymbol{x} \in \mathcal{X}$ and $g \in \mathbb{G}$, the same holds for $\mathsf{E}_{\boldsymbol{\theta}}$. $\square$

Recall that for the effective latent dimension m the true latent dimension is constrained by the dimensionality of the singular spaces, i.e., $r_m = \sum_{k \in [n_{\mathrm{iso}}]} r_k = \sum_{k \in [n_{\mathrm{iso}}]} m d_k$. Further, given a measurable set $\mathbb{A} \subseteq \mathcal{X}$ and collection of group elements $\mathbb{G}' \subseteq \mathbb{G}$, let us define the following symmetry index of a set $\mathbb{A}$ w.r.t. probability distribution of random variable $\mathbf{x}$

$$\gamma_{\mathbb{G}'}(\mathbb{A}) = \frac{1}{|\mathbb{G}'|(|\mathbb{G}'|-1)} \sum_{\substack{g_1, g_2 \in \mathbb{G}' \\ g_1 \neq g_2}} \frac{\mathbb{P}[\mathbf{x} \in g_1 \triangleright \mathbb{A} \cap g_2 \triangleright \mathbb{A}]}{\mathbb{P}[\mathbf{x} \in \mathbb{A}]}, \tag{93}$$

which in the case when $\mathbb{G}'$ is a subgroup of $\mathbb{G}$ simplifies as

$$\gamma_{\mathbb{G}'}(\mathbb{A}) = \frac{1}{|\mathbb{G}'|-1} \sum_{\substack{g \in \mathbb{G}' \\ g \neq e}} \frac{\mathbb{P}[\mathbf{x} \in \mathbb{A} \cap g \triangleright \mathbb{A}]}{\mathbb{P}[\mathbf{x} \in \mathbb{A}]}. \tag{94}$$

Observe that always $\gamma_{\mathbb{G}'}(\mathbb{A}) \in [0,1]$, where extremes correspond to the cases $\gamma_{\mathbb{G}'}(\mathbb{A}) = 1$ when set $\mathbb{A}$ is $\mathbb{G}'$ invariant, and $\gamma_{\mathbb{G}'}(\mathbb{A}) = 0$ when $\mathbb{A}$ equals its coset w.r.t. $\mathbb{G}'$, that is $g \triangleright \mathbb{A} \cap \mathbb{A} = \emptyset$ for all $g \in \mathbb{G}'$, meaning that the set is fully asymmetric w.r.t transformations $g \in \mathbb{G}'$.

We first generalize the approximation error bound in Lemma 1 from [37] to the case of vector valued functions in the presence of symmetries.

Theorem M.2 (Approximation error)**.** *Given a group of symmetries $\mathbb{G}$, let $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$ be Hilbert spaces endowed with symmetry group $\mathbb{G}$, and let $P_{\mathbf{x}}$, $P_{\mathbf{y}}$ and $P_{\mathbf{xy}}$ be $\mathbb{G}$-invariant probability distributions on $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{X} \times \mathcal{Y}$. Then, for every $\boldsymbol{h} \in \mathcal{L}^2_{\mathbf{y}}(\mathcal{Y},\mathcal{Z})$ it holds that*

$$\|\mathbb{E}_{\mathbf{y}}[\boldsymbol{h}(\mathbf{y}) \,|\, \mathbf{x} = \cdot] - \mathsf{E}_{\boldsymbol{\theta}}\boldsymbol{h}\|_{\mathcal{L}^2_{P_{\mathbf{x}}}(\mathcal{X},\mathcal{Z})} \leq \left(\sigma^\star_{r_m+1} + \left\| [\![\mathsf{D}_{\mathbf{y}|\mathbf{x}}]\!]_{r_m} - \mathsf{D}_{\boldsymbol{\theta}} \right\|\right) \|\boldsymbol{h}\|_{\mathcal{L}^2_{\mathbf{y}}(\mathcal{Y},\mathcal{Z})}. \tag{95}$$

Moreover, denoting

$$\mathsf{E}_{\boldsymbol{\theta}}[f(\mathbf{y}) \,|\, \mathbf{x} \in \mathbb{A}] = \mathbb{E}_{\mathbf{y}}[f] + \frac{\mathbb{E}_{\mathbf{x}}[\mathbb{1}_{\mathbb{A}}(\mathbf{x})[\mathsf{D}_{\boldsymbol{\theta}}f](\mathbf{x})]}{\mathbb{P}[\mathbf{x} \in \mathbb{A}]}, \tag{96}$$

if $\boldsymbol{h}$ is either $\mathbb{G}'$-invariant or $\mathbb{G}'$-equivariant for some $\mathbb{G}' \subseteq \mathbb{G}$, then for every measurable set $\mathbb{A}$

$$\|\mathbb{E}[\boldsymbol{h}(\mathbf{y}) \,|\, \mathbf{x} \in \mathbb{A}] - \mathsf{E}_{\boldsymbol{\theta}}[\boldsymbol{h}(\mathbf{y}) \,|\, \mathbf{x} \in \mathbb{A}]\|_{\mathcal{Z}} \leq \left(\sigma^\star_{r_m+1} + \left\| [\![\mathsf{D}_{\mathbf{y}|\mathbf{x}}]\!]_{r_m} - \mathsf{D}_{\boldsymbol{\theta}} \right\|\right) \frac{\|f\|_{\mathcal{L}^2_{\mathbf{y}}(\mathcal{Y},\mathcal{Z})}}{\sqrt{\mathbb{P}[\mathbf{x} \in \mathbb{A}]}} \sqrt{\frac{1 + (|\mathbb{G}'|-1)\gamma_{\mathbb{G}'}(\mathbb{A})}{|\mathbb{G}'|}}. \tag{97}$$

Proof. Start by observing that

$$\begin{aligned}\|\mathbb{E}[\boldsymbol{h}(\mathbf{y}) \,|\, \mathbf{x} = \cdot] - \mathsf{E}_{\boldsymbol{\theta}}\boldsymbol{h}\|_{\mathcal{L}^2_{P_{\mathbf{x}}}(\mathcal{X},\mathcal{Z})} &\leq \left\|\mathsf{D}_{\mathbf{y}|\mathbf{x}} - \mathsf{D}_{\boldsymbol{\theta}}\right\|_{\mathcal{L}^2_{\mathbf{y}}(\mathcal{Y},\mathcal{Z}) \to \mathcal{L}^2_{P_{\mathbf{x}}}(\mathcal{X},\mathcal{Z})} \|\boldsymbol{h}\|_{\mathcal{L}^2_{\mathbf{y}}(\mathcal{Y},\mathcal{Z})} \\ &= \left\|\mathsf{D}_{\mathbf{y}|\mathbf{x}} - \mathsf{D}_{\boldsymbol{\theta}}\right\|_{\mathcal{L}^2_{P_{\mathbf{y}}}(\mathcal{Y}) \to \mathcal{L}^2_{P_{\mathbf{x}}}(\mathcal{X})} \|\boldsymbol{h}\|_{\mathcal{L}^2_{\mathbf{y}}(\mathcal{Y},\mathcal{Z})},\end{aligned}$$

where the equality holds since we extended operators $\mathsf{D}_{\mathbf{y}|\mathbf{x}}$ and $\mathsf{D}_{\boldsymbol{\theta}}$ to vector valued setting as integral operators with the same scalar kernel. Hence, (95) readily follows.

To prove (97), start with noting

$$\mathbb{E}[\boldsymbol{h}(\mathbf{y}) \,|\, \mathbf{x} \in \mathbb{A}] - \mathsf{E}_{\boldsymbol{\theta}}[\boldsymbol{h}(\mathbf{y}) \,|\, \mathbf{x} \in \mathbb{A}] = \frac{\mathbb{E}_{\mathbf{x}}[\mathbb{1}_{\mathbb{A}}(\mathbf{x})[(\mathsf{D}_{\mathbf{y}|\mathbf{x}} - \mathsf{D}_{\boldsymbol{\theta}})\boldsymbol{h}](\mathbf{x})]}{\mathbb{P}[\mathbf{x} \in \mathbb{A}]}.$$

Then, if $\boldsymbol{h}$ is $\mathbb{G}$-equivariant, then, using that invariance of the probability distribution $P_{\mathbf{x}}$, $\mathbb{G}$-equivariance of $\mathsf{D}_{\mathbf{y}|\mathbf{x}}$ and, due to Proposition M.1, of $\mathsf{D}_{\boldsymbol{\theta}}$, we have that for every $g \in \mathbb{G}' \subseteq \mathbb{G}$

$$\begin{aligned}
\mathbb{E}_{\mathbf{x}}[\mathbb{1}_{\mathbb{A}}(\mathbf{x})[(\mathsf{D}_{\mathbf{y}|\mathbf{x}} - \mathsf{D}_{\boldsymbol{\theta}})\boldsymbol{h}](\mathbf{x})] &= \mathbb{E}_{\mathbf{x}}[\mathbb{1}_{\mathbb{A}}(g \triangleright_{\mathcal{X}} \mathbf{x})[(\mathsf{D}_{\mathbf{y}|\mathbf{x}} - \mathsf{D}_{\boldsymbol{\theta}})\boldsymbol{h}](g \triangleright_{\mathcal{X}} \mathbf{x})] \\
&= \mathbb{E}_{\mathbf{x}}[\mathbb{1}_{g^{-1} \triangleright_{\mathcal{X}} \mathbb{A}}(\mathbf{x}) \, g \triangleright_{\mathcal{Z}} [(\mathsf{D}_{\mathbf{y}|\mathbf{x}} - \mathsf{D}_{\boldsymbol{\theta}})\boldsymbol{h}](\mathbf{x})] \\
&= \mathbb{E}_{\mathbf{x}}[\mathbb{1}_{g^{-1} \triangleright_{\mathcal{X}} \mathbb{A}}(\mathbf{x}) \bar{\boldsymbol{\rho}}_{\mathcal{Z}}(g) \, [(\mathsf{D}_{\mathbf{y}|\mathbf{x}} - \mathsf{D}_{\boldsymbol{\theta}})\boldsymbol{h}](\mathbf{x})].
\end{aligned}$$

Hence, averaging over $\mathbb{G}'$ we obtain

$$\mathbb{E}[\boldsymbol{h}(\mathbf{y}) \,|\, \mathbf{x} \in \mathbb{A}] - \mathsf{E}_{\boldsymbol{\theta}}[\boldsymbol{h}(\mathbf{y}) \,|\, \mathbf{x} \in \mathbb{A}] = \mathbb{E}_{\mathbf{x}}[\mathsf{H}(\mathbf{x})\overline{\boldsymbol{z}}(\mathbf{x})],$$

where

$$\mathsf{H}(\boldsymbol{x}) = \frac{1}{|\mathbb{G}'|\mathbb{P}[\mathbf{x} \in \mathbb{A}]} \sum_{g \in \mathbb{G}'} \mathbb{1}_{g^{-1} \triangleright_{\mathcal{X}} \mathbb{A}}(\boldsymbol{x}) \bar{\boldsymbol{\rho}}_{\mathcal{Z}}(g) \quad \text{and} \quad \boldsymbol{z}(\boldsymbol{x}) = [(\mathsf{D}_{\mathbf{y}|\mathbf{x}} - \mathsf{D}_{\boldsymbol{\theta}})\boldsymbol{h}](\boldsymbol{x}).$$

Since due to Cauchy-Schwartz inequality we have

$$\|\mathbb{E}_{\mathbf{x}}[\mathsf{H}(\mathbf{x})\boldsymbol{z}(\mathbf{x})]\|_{\mathcal{Z}}^2 \leq [\mathbb{E}_{\mathbf{x}}\|\mathsf{H}(\mathbf{x})\|_{\mathcal{Z}\to\mathcal{Z}}^2][\mathbb{E}_{\mathbf{x}}\|\boldsymbol{z}(\mathbf{x})\|_{\mathcal{Z}}^2] = \|\boldsymbol{z}\|_{\mathcal{L}^2_{P_{\mathbf{x}}}(\mathcal{X},\mathcal{Z})}^2 \, [\mathbb{E}_{\mathbf{x}}\|\mathsf{H}(\mathbf{x})\|_{\mathcal{Z}\to\mathcal{Z}}^2]$$

and $\|\boldsymbol{z}\|_{\mathcal{L}^2_{P_{\mathbf{x}}}(\mathcal{X},\mathcal{Z})} \leq \left\|\mathsf{D}_{\mathbf{y}|\mathbf{x}} - \mathsf{D}_{\boldsymbol{\theta}}\right\|_{\mathcal{L}^2_{\mathbf{y}}(\mathcal{Y},\mathcal{Z}) \to \mathcal{L}^2_{P_{\mathbf{x}}}(\mathcal{X},\mathcal{Z})} \|\boldsymbol{h}\|_{\mathcal{L}^2_{\mathbf{y}}(\mathcal{Y},\mathcal{Z})}^2$, it remains to bound $\mathbb{E}_{\mathbf{x}}\|\mathsf{H}(\mathbf{x})\|_{\mathcal{Z}\to\mathcal{Z}}^2$. But, the group actions in the vector spaces are unitary, so using the $\mathbb{G}$-invariance of the distribution of $\mathbf{x}$ we obtain

$$\begin{aligned}
\mathbb{E}_{\mathbf{x}}\|\mathsf{H}(\mathbf{x})\|_{\mathcal{Z}\to\mathcal{Z}}^2 &\leq \mathbb{E}_{\mathbf{x}}\Big[\frac{1}{|\mathbb{G}'|\mathbb{P}[\mathbf{x} \in \mathbb{A}]} \sum_{g \in \mathbb{G}'} \mathbb{1}_{g^{-1} \triangleright_{\mathcal{X}} \mathbb{A}}(\boldsymbol{x})\Big]^2 \\
&= \frac{1}{|\mathbb{G}'|^2\mathbb{P}[\mathbf{x} \in \mathbb{A}]^2} \sum_{g,g' \in \mathbb{G}'} \mathbb{E}_{\mathbf{x}}[\mathbb{1}_{g^{-1} \triangleright_{\mathcal{X}} \mathbb{A}}(\boldsymbol{x})\mathbb{1}_{g'^{-1} \triangleright_{\mathcal{X}} \mathbb{A}}(\boldsymbol{x})] \\
&= \frac{1}{|\mathbb{G}'|^2\mathbb{P}[\mathbf{x} \in \mathbb{A}]^2} \sum_{g,g' \in \mathbb{G}'} \mathbb{E}_{\mathbf{x}}[\mathbb{1}_{g \triangleright_{\mathcal{X}} \mathbb{A} \cap g' \triangleright_{\mathcal{X}} \mathbb{A}}(\boldsymbol{x})] \\
&= \frac{1}{|\mathbb{G}'|^2\mathbb{P}[\mathbf{x} \in \mathbb{A}]} \sum_{g,g' \in \mathbb{G}'} \frac{\mathbb{P}[\mathbf{x} \in g \triangleright_{\mathcal{X}} \mathbb{A} \cap g' \triangleright_{\mathcal{X}} \mathbb{A}]}{\mathbb{P}[\mathbf{x} \in \mathbb{A}]} \\
&= \frac{1}{\mathbb{P}[\mathbf{x} \in \mathbb{A}]} \frac{1 + (|\mathbb{G}'| - 1)\gamma_{\mathbb{G}'}(\mathbb{A})}{|\mathbb{G}'|},
\end{aligned}$$

which completes the proof of (97) for $\mathbb{G}'$-equivariant functions. Finally, if f is $\mathbb{G}'$-invariant, the proof follows the same lines by replacing group actions ($\triangleright_{\mathcal{Z}}$) by their respective group representation $\boldsymbol{\rho}_{\mathcal{Z}}$ (see Def. I.3) with identity. $\square$

Next we analyze the errors when, instead of applying learned operators $\mathsf{E}_{\boldsymbol{\theta}}$, we apply their empirical counterparts in inference tasks. To that end, we define now estimators of $\mathbb{E}[\boldsymbol{h}(\mathbf{x})]$ and $\mathbb{E}[\boldsymbol{z}(\mathbf{y})]$ exploiting the $\mathbb{G}$-invariance of the distributions of $\mathbf{x}$ and $\mathbf{y}$. First, define the empirical $\mathbb{G}$-invariant distributions

$$\widehat{\mathbb{P}}_{\mathbf{x}} := \frac{1}{|\mathbb{G}|N} \sum_{i=1}^{N} \sum_{g \in g} \delta_{g \triangleright \mathbf{x}_i}(\cdot), \quad \widehat{\mathbb{P}}_{\mathbf{y}} := \frac{1}{|\mathbb{G}|N} \sum_{i=1}^{N} \sum_{g \in g} \delta_{g \triangleright \mathbf{y}_i}(\cdot).$$

Hence we can define the equivariant empirical mean of any function $f \in \mathcal{L}^2_{\mathbf{x}}$, $h \in \mathcal{L}^2_{\mathbf{y}}$ as

$$\widehat{\mathbb{E}}_{\mathbf{x}}[f] = \frac{1}{|\mathbb{G}|N} \sum_{i=1}^{N} \sum_{g \in \mathbb{G}} f(g \triangleright_{\mathcal{X}} \mathbf{x}_i), \quad \widehat{\mathbb{E}}_{\mathbf{y}}[h] = \frac{1}{|\mathbb{G}|N} \sum_{i=1}^{N} \sum_{g \in \mathbb{G}} h(g \triangleright_{\mathcal{Y}} \mathbf{y}_i). \tag{98}$$

This extends naturally to operator on a function space $\mathcal{L}^2_{\mathbf{y}}(\mathcal{Y},\mathcal{Z})$ where $\mathcal{Z}$ is endowed with an inner product $\langle\cdot,\cdot\rangle = \langle\cdot,\cdot\rangle_{\mathcal{Z}}$. If the distribution of $\mathbf{y}$ is $\mathbb{G}'$-invariant, then for any $\boldsymbol{h}\in\mathcal{L}^2_{\mathbf{y}}(\mathcal{Y},\mathcal{Z})$, we use the estimator $\widehat{\mathbb{E}}_{\mathbf{y}}[\boldsymbol{h}(\mathbf{y})]$ in (98) as an estimator of $\mathbb{E}[\boldsymbol{h}(\mathbf{y})]$:

$$\widehat{\mathbb{E}}_{\mathbf{y}}[\boldsymbol{h}] = \frac{1}{|\mathbb{G}|N}\sum_{i=1}^{N}\sum_{g\in\mathbb{G}} \boldsymbol{h}(g \triangleright_{\mathcal{Y}} \mathbf{y}_i). \tag{99}$$

In this notation, we define our empirical estimators

$$[\mathsf{E}_{\theta}\boldsymbol{h}](\boldsymbol{x}) \approx [\widehat{\mathsf{E}}_{\theta}\boldsymbol{h}](\boldsymbol{x}) = \widehat{\mathbb{E}}_{\mathbf{y}}[\boldsymbol{h}(\mathbf{y})] + \sum_{k\in[n_{\mathrm{iso}}]}\sum_{i\in[m]}\sum_{j\in[d_k]} \sigma_i^{\boldsymbol{\theta}(k)} u_{i,j}^{\boldsymbol{\theta}(k)}(\boldsymbol{x})\widehat{\mathbb{E}}_{\mathbf{y}}[v_{i,j}^{\boldsymbol{\theta}(k)}\boldsymbol{h}]$$

and

$$\mathsf{E}_{\theta}[\boldsymbol{h}(\mathbf{y}) \,|\, \mathbf{x}\in\mathbb{A}] \approx \widehat{\mathsf{E}}_{\theta}[\boldsymbol{h}(\mathbf{y}) \,|\, \mathbf{x}\in\mathbb{A}] = \widehat{\mathbb{E}}_{\mathbf{y}}[\boldsymbol{h}] + \sum_{k\in[n_{\mathrm{iso}}]}\sum_{i\in[m]}\sum_{j\in[d_k]} \sigma_i^{\boldsymbol{\theta}(k)} \frac{\widehat{\mathbb{E}}_{\mathbf{x}}[u_{i,j}^{\boldsymbol{\theta}(k)}\mathbb{1}_{\mathbb{A}}]}{\widehat{\mathbb{E}}_{\mathbf{x}}[\mathbb{1}_{\mathbb{A}}]}\widehat{\mathbb{E}}_{\mathbf{y}}[v_{i,j}^{\boldsymbol{\theta}(k)}\boldsymbol{h}].$$

and, by choosing $\boldsymbol{h} = \mathbb{1}_{\mathbb{B}}$,

$$P[\mathbf{y}\in\mathbb{B} \mid \mathbf{x}\in\mathbb{A}] \approx \widehat{P}_{\boldsymbol{\theta}}[\mathbf{y}\in\mathbb{B} \mid \mathbf{x}\in\mathbb{A}] = \widehat{\mathbb{E}}_{\mathbf{y}}[\mathbb{1}_{\mathbb{B}}] + \sum_{k\in[n_{\mathrm{iso}}]}\sum_{i\in[m]}\sum_{j\in[d_k]} \sigma_i^{\boldsymbol{\theta}(k)} \frac{\widehat{\mathbb{E}}_{\mathbf{x}}[u_{i,j}^{\boldsymbol{\theta}(k)}\mathbb{1}_{\mathbb{A}}]}{\widehat{\mathbb{E}}_{\mathbf{x}}[\mathbb{1}_{\mathbb{A}}]}\widehat{\mathbb{E}}_{\mathbf{y}}[v_{i,j}^{\boldsymbol{\theta}(k)}\mathbb{1}_{\mathbb{B}}].$$

Direct consequence of the above construction which ensures that $\widehat{P}_{\mathbf{x}}$ and $\widehat{P}_{\mathbf{y}}$ are $\mathbb{G}$-invariant is the following result.

Proposition M.3. *Let $P_{\mathbf{x}}$ and $P_{\mathbf{y}}$ are $\mathbb{G}$-invariant, and $\mathsf{D}_{\boldsymbol{\theta}}$ from (90) is $\mathbb{G}$-equivariant model, and let $z\in\mathcal{L}^2_{\mathbf{x}}(\mathcal{X},\mathbb{R})$ and $\boldsymbol{h}\in\mathcal{L}^2_{P_{\mathbf{x}}}(\mathcal{Y},\mathcal{Z})$ be arbitrary. If for every $k\in[n_{\mathrm{iso}}]$*

$$\Big\{\big\|\mathsf{D}^{(k)}_{\mathbf{y}|\mathbf{x}} - \mathsf{D}^{(k)}_{\boldsymbol{\theta}}\big\|, \Big\|\mathbb{E}_{\mathbf{x}}[\boldsymbol{u}^{(k)}_{\boldsymbol{\theta}\,(1)}(\mathbf{x})\boldsymbol{u}^{(k)}_{\boldsymbol{\theta}\,(1)}(\mathbf{x})^{\top}] - \boldsymbol{I}_m\Big\|, \Big\|\mathbb{E}_{\mathbf{y}}[\boldsymbol{v}^{(k)}_{\boldsymbol{\theta}\,(1)}(\mathbf{y})\boldsymbol{v}^{(k)}_{\boldsymbol{\theta}\,(1)}(\mathbf{y})^{\top}] - \boldsymbol{I}_m\Big\|\Big\} \leq \mathcal{E}^{(k)}_{\boldsymbol{\theta}}$$

holds with $\boldsymbol{u}^{(k)}_{\boldsymbol{\theta}\,(1)} = [u_{1,1}^{\boldsymbol{\theta}(k)}|\ldots|u_{m,1}^{\boldsymbol{\theta}(k)}]^{\top}\in\mathbb{R}^m$ and $\boldsymbol{v}^{(k)}_{\boldsymbol{\theta}\,(1)} = [v_{1,1}^{\boldsymbol{\theta}(k)}|\ldots|v_{m,1}^{\boldsymbol{\theta}(k)}]^{\top}\in\mathbb{R}^m$, and if

$$\begin{aligned}\frac{\Big\|\widehat{\mathbb{E}}_{\mathbf{x}}[\boldsymbol{u}^{(k)}_{\boldsymbol{\theta}\,(1)} z_1^{(k)}] - \mathbb{E}_{\mathbf{x}}[\boldsymbol{u}^{(k)}_{\boldsymbol{\theta}\,(1)}(\mathbf{x}) z_1^{(k)}(\mathbf{x})]\Big\|}{\big\|z_1^{(k)}\big\|_{\mathcal{L}^2_{\mathbf{x}}}} &\leq A(\boldsymbol{u}_{\boldsymbol{\theta}}, z),\\ \frac{\Big\|\widehat{\mathbb{E}}_{\mathbf{y}}[\boldsymbol{v}^{(k)}_{\boldsymbol{\theta}\,(1)} \otimes \boldsymbol{h}_1^{(k)}] - \mathbb{E}_{\mathbf{y}}[\boldsymbol{v}^{(k)}_{\boldsymbol{\theta}\,(1)}(\mathbf{y}) \otimes \boldsymbol{h}_1^{(k)}(\mathbf{y})]\Big\|}{\big\|\boldsymbol{h}_1^{(k)}\big\|_{\mathcal{L}^2_{\mathbf{y}}}} &\leq A(\boldsymbol{v}_{\boldsymbol{\theta}}, \boldsymbol{h}),\end{aligned} \tag{100}$$

where $z = \sum_{k\in[n_{\mathrm{iso}}]}\sum_{j\in[d_k]} z_j^{(k)}$ and $\boldsymbol{h} = \sum_{k\in[n_{\mathrm{iso}}]}\sum_{j\in[d_k]} \boldsymbol{h}_j^{(k)}$ are isospectral decompositions, then

$$\Big\|\mathsf{E}_{\boldsymbol{\theta}}\boldsymbol{h} - \widehat{\mathsf{E}}_{\boldsymbol{\theta}}\boldsymbol{h}\Big\|^2_{\mathcal{L}^2_{P_{\mathbf{x}}}(\mathcal{X},\mathcal{Z})} \leq \Big\|\mathbb{E}_{\mathbf{y}}[\boldsymbol{h}(\mathbf{y}) - \widehat{\mathbb{E}}_{\mathbf{y}}[\boldsymbol{h}]]\Big\|^2_{\mathcal{Z}} + \Big[1 + \max_{k\in[n_{\mathrm{iso}}]}\mathcal{E}^{(k)}_{\boldsymbol{\theta}}\Big]^3 \|\boldsymbol{h}\|^2_{\mathcal{L}^2_{\mathbf{y}}(\mathcal{Y},\mathcal{Z})}\,[A(\boldsymbol{v}_{\boldsymbol{\theta}},\boldsymbol{h})]^2. \tag{101}$$

Moreover, the empirical estimation error is upper bounded by

$$\begin{aligned}&\Big\|\mathbb{E}_{\mathbf{x}}[z(\mathbf{x})[\mathsf{D}_{\boldsymbol{\theta}}\boldsymbol{h}](\mathbf{x})] - \widehat{\mathbb{E}}_{\mathbf{x}}[z[\widehat{\mathsf{D}}^{(k)}_{\boldsymbol{\theta}}\boldsymbol{h}]]\Big\|^2_{\mathcal{Z}} \leq\\ &(1+\mathcal{E}_{\boldsymbol{\theta}})^3\Big[A(\boldsymbol{u}_{\boldsymbol{\theta}},z) + A(\boldsymbol{v}_{\boldsymbol{\theta}},\boldsymbol{h}) + A(\boldsymbol{u}_{\boldsymbol{\theta}},z)A(\boldsymbol{v}_{\boldsymbol{\theta}},\boldsymbol{h})\Big]^2 \|z\|^2_{\mathcal{L}^2_{P_{\mathbf{x}}}(\mathcal{X})}\|\boldsymbol{h}\|^2_{\mathcal{L}^2_{\mathbf{y}}(\mathcal{Y},\mathcal{Z})}.\end{aligned} \tag{102}$$

Proof. First, observe that due to $\mathbb{G}$-invariance of distribution $P_{\mathbf{xy}}$ and $\mathbb{G}$-equivaraince of $\mathsf{E}_{\boldsymbol{\theta}}$ and $\mathsf{D}_{\boldsymbol{\theta}}$ we have that

$$\mathsf{E}_{\boldsymbol{\theta}}\boldsymbol{h} = \mathbb{E}_{\mathbf{y}}[\boldsymbol{h}^{(1)}(\mathbf{y})] + \sum_{k\in[n_{\mathrm{iso}}]} \mathsf{D}^{(k)}_{\boldsymbol{\theta}}\boldsymbol{h}^{(k)}, \tag{103}$$

and

$$\mathbb{E}_{\mathbf{x}}[z(\mathbf{x})[\mathsf{E}_{\boldsymbol{\theta}}\boldsymbol{h}](\mathbf{x})] = \mathbb{E}_{\mathbf{x}}[z^{(1)}(\mathbf{x})]\mathbb{E}_{\mathbf{y}}[\boldsymbol{h}^{(1)}(\mathbf{y})] + \sum_{k\in[n_{\mathrm{iso}}]} \mathbb{E}_{\mathbf{x}}[z^{(k)}(\mathbf{x})[\mathsf{D}_{\boldsymbol{\theta}}^{(k)}\boldsymbol{h}^{(k)}](\mathbf{x})]. \tag{104}$$

In the same way, since the empirical distributions $\widehat{P}_{\mathbf{x}}$ and $\widehat{P}_{\mathbf{y}}$ are $\mathbb{G}$-invariant, we have that

$$\widehat{\mathsf{E}}_{\boldsymbol{\theta}}\boldsymbol{h} = \widehat{\mathbb{E}}_{\mathbf{y}}[\boldsymbol{h}^{(1)}] + \sum_{k\in[n_{\mathrm{iso}}]} \widehat{\mathsf{D}}_{\boldsymbol{\theta}}^{(k)}\boldsymbol{h}^{(k)}, \tag{105}$$

and

$$\widehat{\mathbb{E}}_{\mathbf{x}}[z[\widehat{\mathsf{E}}_{\boldsymbol{\theta}}\boldsymbol{h}]] = \widehat{\mathbb{E}}_{\mathbf{x}}[z^{(1)}]\widehat{\mathbb{E}}_{\mathbf{y}}[\boldsymbol{h}^{(1)}] + \sum_{k\in[n_{\mathrm{iso}}]} \widehat{\mathbb{E}}_{\mathbf{x}}[z^{(k)}[\widehat{\mathsf{D}}_{\boldsymbol{\theta}}^{(k)}\boldsymbol{h}^{(k)}]], \tag{106}$$

where

$$[\widehat{\mathsf{D}}_{\boldsymbol{\theta}}^{(k)}\boldsymbol{h}^{(k)}](\boldsymbol{x}) = \boldsymbol{u}_{\boldsymbol{\theta}}^{(k)}(\boldsymbol{x})^{\top}\boldsymbol{S}_{\boldsymbol{\theta}}^{(k)}\widehat{\mathbb{E}}_{\mathbf{y}}[\boldsymbol{v}_{\boldsymbol{\theta}}^{(k)} \otimes \boldsymbol{h}^{(k)}]. \tag{107}$$

Therefore, combining (103) and (105), we obtain that

$$\begin{aligned}[\mathsf{E}_{\boldsymbol{\theta}}\boldsymbol{h}](\boldsymbol{x}) - [\widehat{\mathsf{E}}_{\boldsymbol{\theta}}\boldsymbol{h}](\boldsymbol{x}) = &\Big(\mathbb{E}_{\mathbf{y}}[\boldsymbol{h}^{(1)}(\mathbf{y})] - \widehat{\mathbb{E}}_{\mathbf{y}}[\boldsymbol{h}^{(1)}]\Big)\mathbb{1}_{\mathcal{X}}(\boldsymbol{x}) \\ &+ \sum_{k\in[n_{\mathrm{iso}}]}\Big([\mathsf{D}_{\boldsymbol{\theta}}^{(k)}\boldsymbol{h}^{(k)}](\boldsymbol{x}) - [\widehat{\mathsf{D}}_{\boldsymbol{\theta}}^{(k)}\boldsymbol{h}^{(k)}](\boldsymbol{x})\Big),\end{aligned}$$

which after taking the norm in $\mathcal{L}^2_{P_{\mathbf{x}}}(\mathcal{X},\mathcal{Z})$, due to orthonormality of isotypic subspaces gives

$$\begin{aligned}\Big\|\mathsf{E}_{\boldsymbol{\theta}}\boldsymbol{h} - \widehat{\mathsf{E}}_{\boldsymbol{\theta}}[\boldsymbol{h}]\Big\|^2_{\mathcal{L}^2_{P_{\mathbf{x}}}(\mathcal{X},\mathcal{Z})} = &\Big\|\mathbb{E}_{\mathbf{y}}[\boldsymbol{h}^{(1)}(\mathbf{y})] - \widehat{\mathbb{E}}_{\mathbf{y}}[\boldsymbol{h}^{(1)}]\Big\|^2_{\mathcal{Z}} \\ &+ \sum_{k\in[n_{\mathrm{iso}}]}\Big\|[\mathsf{D}_{\boldsymbol{\theta}}^{(k)}\boldsymbol{h}^{(k)}] - [\widehat{\mathsf{D}}_{\boldsymbol{\theta}}^{(k)}\boldsymbol{h}^{(k)}]\Big\|^2_{\mathcal{L}^2_{P_{\mathbf{x}}}(\mathcal{X},\mathcal{Z})}.\end{aligned}$$

Now, observe that, since

$$[\mathsf{D}_{\boldsymbol{\theta}}^{(k)}\boldsymbol{h}^{(k)}](\boldsymbol{x}) - [\widehat{\mathsf{D}}_{\boldsymbol{\theta}}^{(k)}\boldsymbol{h}^{(k)}](\boldsymbol{x}) = \boldsymbol{u}_{\boldsymbol{\theta}}^{(k)}(\boldsymbol{x})^{\top}\boldsymbol{S}_{\boldsymbol{\theta}}^{(k)}\Big(\mathbb{E}_{\mathbf{y}}[\boldsymbol{v}_{\boldsymbol{\theta}}^{(k)} \otimes \boldsymbol{h}^{(k)}] - \widehat{\mathbb{E}}_{\mathbf{y}}[\boldsymbol{v}_{\boldsymbol{\theta}}^{(k)} \otimes \boldsymbol{h}^{(k)}]\Big)$$

applying the norm we have that $\Big\|[\mathsf{D}_{\boldsymbol{\theta}}^{(k)}\boldsymbol{h}^{(k)}] - [\widehat{\mathsf{D}}_{\boldsymbol{\theta}}^{(k)}\boldsymbol{h}^{(k)}]\Big\|^2_{\mathcal{L}^2_{P_{\mathbf{x}}}(\mathcal{X},\mathcal{Z})}$ equals

$$\Big(\mathbb{E}_{\mathbf{y}}[\boldsymbol{v}_{\boldsymbol{\theta}}^{(k)}\otimes\boldsymbol{h}^{(k)}] - \widehat{\mathbb{E}}_{\mathbf{y}}[\boldsymbol{v}_{\boldsymbol{\theta}}^{(k)}\otimes\boldsymbol{h}^{(k)}]\Big)^{\top}\boldsymbol{S}_{\boldsymbol{\theta}}^{(k)}\Big(\mathbb{E}_{\mathbf{x}}[\boldsymbol{u}_{\boldsymbol{\theta}}^{(k)}(\boldsymbol{x})\boldsymbol{u}_{\boldsymbol{\theta}}^{(k)}(\boldsymbol{x})^{\top}]\Big)\boldsymbol{S}_{\boldsymbol{\theta}}^{(k)}\Big(\mathbb{E}_{\mathbf{y}}[\boldsymbol{v}_{\boldsymbol{\theta}}^{(k)}\otimes\boldsymbol{h}^{(k)}] - \widehat{\mathbb{E}}_{\mathbf{y}}[\boldsymbol{v}_{\boldsymbol{\theta}}^{(k)}\otimes\boldsymbol{h}^{(k)}]\Big)$$

which using constraints within each isotypic block and

$$\mathbb{E}_{\mathbf{x}}[{\boldsymbol{u}_{\boldsymbol{\theta}}^{(k)}}_{(m)}(\boldsymbol{x}){\boldsymbol{u}_{\boldsymbol{\theta}}^{(k)}}_{(m)}(\boldsymbol{x})^{\top}] \preceq \Big\|\mathbb{E}_{\mathbf{x}}[{\boldsymbol{u}_{\boldsymbol{\theta}}^{(k)}}_{(m)}(\boldsymbol{x}){\boldsymbol{u}_{\boldsymbol{\theta}}^{(k)}}_{(m)}(\boldsymbol{x})^{\top}]\Big\|\boldsymbol{I}_m \leq (1+\mathcal{E}_{\boldsymbol{\theta}}^{(k)})\boldsymbol{I}_m,$$

implies, due to (100), that

$$\begin{aligned}\Big\|\mathsf{D}_{\boldsymbol{\theta}}^{(k)}\boldsymbol{h}^{(k)} - \widehat{\mathsf{D}}_{\boldsymbol{\theta}}^{(k)}\boldsymbol{h}^{(k)}\Big\|^2_{\mathcal{L}^2_{P_{\mathbf{x}}}(\mathcal{X},\mathcal{Z})} &\leq d_k\,(1+\mathcal{E}_{\boldsymbol{\theta}}^{(k)})\,(\sigma_1^{\boldsymbol{\theta}(k)})^2 \\ &\quad\cdot\Big\|\mathbb{E}_{\mathbf{y}}[{\boldsymbol{v}_{\boldsymbol{\theta}}^{(k)}}_{(1)}(\mathbf{y})\otimes\boldsymbol{h}_1^{(k)}(\mathbf{y})] - \widehat{\mathbb{E}}_{\mathbf{y}}[{\boldsymbol{v}_{\boldsymbol{\theta}}^{(k)}}_{(1)}\otimes\boldsymbol{h}_1^{(k)}]\Big\|^2_{\mathbb{R}^m\times\mathcal{Z}} \\ &\leq d_k\,(1+\mathcal{E}_{\boldsymbol{\theta}}^{(k)})\,(\sigma_1^{\boldsymbol{\theta}(k)})^2\,[A(\boldsymbol{v}_{\boldsymbol{\theta}},\boldsymbol{h})]^2\,\Big\|\boldsymbol{h}_1^{(k)}\Big\|^2_{\mathcal{L}^2_{P_{\mathbf{x}}}(\mathcal{Y},\mathcal{Z})}.\end{aligned}$$

Therefore, bounding $\sigma_1^{\boldsymbol{\theta}(k)} \leq \sigma_1^{(k)} + |\sigma_1^{(k)} - \sigma_1^{\boldsymbol{\theta}(k)}| \leq 1 + \Big\|\mathsf{D}_{\mathbf{y}|\mathbf{x}}^{(k)} - \mathsf{D}_{\boldsymbol{\theta}}^{(k)}\Big\|$ and summing over isotypic components, since $\|\boldsymbol{h}\|^2_{\mathcal{L}^2_{P_{\mathbf{x}}}(\mathcal{Y},\mathcal{Z})} = \sum_{k\in[n_{\mathrm{iso}}],j\in[d_k]}\Big\|\boldsymbol{h}_j^{(k)}\Big\|^2_{\mathcal{L}^2_{P_{\mathbf{x}}}(\mathcal{Y},\mathcal{Z})} = \sum_{k\in[n_{\mathrm{iso}}]} d_k\Big\|\boldsymbol{h}_1^{(k)}\Big\|^2_{\mathcal{L}^2_{P_{\mathbf{x}}}(\mathcal{Y},\mathcal{Z})}$, we complete the proof of (101).

To show (102), we combine (104) and (106), and obtain that $\mathbb{E}_{\mathbf{x}}[z(\mathbf{x})[\mathsf{D}_{\boldsymbol{\theta}}\boldsymbol{h}](\mathbf{x})] - \widehat{\mathbb{E}}_{\mathbf{x}}[z[\widehat{\mathsf{D}}_{\boldsymbol{\theta}}\boldsymbol{h}]]$ can be written as

$$\sum_{k\in[n_{\mathrm{iso}}]} d_k\Bigg[\mathbb{E}_{\mathbf{x}}[{\boldsymbol{u}_{\boldsymbol{\theta}}^{(k)}}_{(1)}(\mathbf{x})z_1^{(k)}(\mathbf{x})]^{\top}\boldsymbol{S}_{\boldsymbol{\theta}}^{(k)}\mathbb{E}_{\mathbf{y}}[{\boldsymbol{v}_{\boldsymbol{\theta}}^{(k)}}_{(1)}(\mathbf{y})\otimes\boldsymbol{h}_1^{(k)}(\mathbf{y})] - \widehat{\mathbb{E}}_{\mathbf{x}}[\boldsymbol{u}_{\boldsymbol{\theta}}^{(k)}z_1^{(k)}]^{\top}\boldsymbol{S}_{\boldsymbol{\theta}}^{(k)}\widehat{\mathbb{E}}_{\mathbf{y}}[{\boldsymbol{v}_{\boldsymbol{\theta}}^{(k)}}_{(1)}\otimes\boldsymbol{h}_1^{(k)}]\Bigg].$$

Adding and subtracting mixed terms we then obtain for each isotypic component, $\frac{1}{d_k}\mathbb{E}_{\mathbf{x}}[z(\mathbf{x})[\mathsf{D}_{\boldsymbol{\theta}}^{(k)}\boldsymbol{h}](\mathbf{x})] - \widehat{\mathbb{E}}_{\mathbf{x}}[z[\widehat{\mathsf{D}}_{\boldsymbol{\theta}}^{(k)}\boldsymbol{h}]]$ can be expressed as

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{x}}[\boldsymbol{u}_{\boldsymbol{\theta}}^{(k)}{}_{(1)}(\mathbf{x})z_1^{(k)}(\mathbf{x})]^\top \boldsymbol{S}^{(k)}\left(\mathbb{E}_{\mathbf{y}}[\boldsymbol{v}_{\boldsymbol{\theta}}^{(k)}{}_{(1)}(\mathbf{y})\otimes\boldsymbol{h}_1^{(k)}(\mathbf{y})]-\widehat{\mathbb{E}}_{\mathbf{y}}[\boldsymbol{v}_{\boldsymbol{\theta}}^{(k)}{}_{(1)}\otimes\boldsymbol{h}_1^{(k)}]\right)\\
&+\left(\mathbb{E}_{\mathbf{x}}[\boldsymbol{u}_{\boldsymbol{\theta}}^{(k)}{}_{(1)}(\mathbf{x})z_1^{(k)}(\mathbf{x})]-\widehat{\mathbb{E}}_{\mathbf{x}}[\boldsymbol{u}_{\boldsymbol{\theta}}^{(k)}{}_{1}z_1^{(k)}]\right)^\top \boldsymbol{S}^{(k)}\mathbb{E}_{\mathbf{y}}[\boldsymbol{v}_{\boldsymbol{\theta}}^{(k)}{}_{(1)}(\mathbf{x})\otimes\boldsymbol{h}_1^{(k)}(\mathbf{y})]\\
&+\left(\mathbb{E}_{\mathbf{x}}[\boldsymbol{u}_{\boldsymbol{\theta}}^{(k)}{}_{(1)}(\mathbf{x})z_1^{(k)}(\mathbf{x})]-\widehat{\mathbb{E}}_{\mathbf{x}}[\boldsymbol{u}_{\boldsymbol{\theta}}^{(k)}{}_{1}z_1^{(k)}]\right)^\top \boldsymbol{S}^{(k)}\left(\mathbb{E}_{\mathbf{y}}[\boldsymbol{v}_{\boldsymbol{\theta}}^{(k)}{}_{(1)}(\mathbf{y})\otimes\boldsymbol{h}_1^{(k)}(\mathbf{y})]-\widehat{\mathbb{E}}_{\mathbf{y}}[\boldsymbol{v}_{\boldsymbol{\theta}}^{(k)}{}_{(1)}\otimes\boldsymbol{h}_1^{(k)}]\right),
\end{aligned}
$$

and consequently bounded using (100) as

$$
\begin{aligned}
\left\|\mathbb{E}_{\mathbf{x}}[z(\mathbf{x})[\mathsf{D}_{\boldsymbol{\theta}}^{(k)}\boldsymbol{h}](\mathbf{x})] - \widehat{\mathbb{E}}_{\mathbf{x}}[z[\widehat{\mathsf{D}}_{\boldsymbol{\theta}}^{(k)}\boldsymbol{h}]]\right\|_{\mathcal{Z}} \leq d_k\sigma_1^{\boldsymbol{\theta}(k)}\Big[&A(\boldsymbol{u}_{\boldsymbol{\theta}},z)+A(\boldsymbol{v}_{\boldsymbol{\theta}},\boldsymbol{h})\\
&+A(\boldsymbol{u}_{\boldsymbol{\theta}},z)A(\boldsymbol{v}_{\boldsymbol{\theta}},\boldsymbol{h})\Big]\left\|z_1^{(k)}\right\|_{\mathcal{L}^2_{P_{\mathbf{x}}}(\mathcal{X})}\left\|\boldsymbol{h}_1^{(k)}\right\|_{\mathcal{L}^2_{\mathbf{y}}(\mathcal{Y},\mathcal{Z})}.
\end{aligned}
$$

Summing across isotypic components and bounding $\sigma_1^{\boldsymbol{\theta}(k)}$ as before, we complete the proof. □

First note that coupling (101) with (95) ensures that we can prove regression bound via concentration result ensuring (100). To obtain similar result for set-wise regression, we set $z = \mathbb{1}_{\mathbb{A}}$ and use (102) to obtain the following.

Proposition M.4. *Under the assumptions of Proposition M.3, let* $A(\boldsymbol{u}_{\boldsymbol{\theta}},\mathbb{1}_{\mathbb{A}})A(\boldsymbol{v}_{\boldsymbol{\theta}},\boldsymbol{h}) \leq A(\boldsymbol{u}_{\boldsymbol{\theta}},\mathbb{1}_{\mathbb{A}})+A(\boldsymbol{v}_{\boldsymbol{\theta}},\boldsymbol{h})$. *If*

$$
|\mathbb{E}_{\mathbf{x}}[\mathbb{1}_{\mathbb{A}}(\mathbf{x})] - \widehat{\mathbb{E}}_{\mathbf{x}}[\mathbb{1}_{\mathbb{A}}]|/\mathbb{E}_{\mathbf{x}}[\mathbb{1}_{\mathbb{A}}(\mathbf{x})] \leq \eta_{\mathbb{A}} \tag{108}
$$

and $\eta_{\mathbb{A}} < 1/2$, *then*

$$
\begin{aligned}
\left\|\mathsf{E}_\theta[\boldsymbol{h}(\mathbf{y})|\mathbf{x}\in\mathbb{A}] - \widehat{\mathsf{E}}_\theta[\boldsymbol{h}(\mathbf{y})|\mathbf{x}\in\mathbb{A}]\right\|_{\mathcal{Z}} \leq \left\|\mathbb{E}_{\mathbf{y}}[\boldsymbol{h}]-\widehat{\mathbb{E}}_{\mathbf{y}}[\boldsymbol{h}]\right\|_{\mathcal{Z}} &+ \frac{2\|\boldsymbol{h}\|_{\mathcal{L}^2_{P_{\mathbf{x}}}(\mathcal{Y},\mathcal{Z})}}{\sqrt{P[\mathbf{x}\in\mathbb{A}]}}\\
&\times\Big[2(1+\mathcal{E}_{\boldsymbol{\theta}})\Big(A(\boldsymbol{u}_{\boldsymbol{\theta}},\mathbb{1}_{\mathbb{A}})+A(\boldsymbol{v}_{\boldsymbol{\theta}},\boldsymbol{h})\Big)+\eta_{\mathbb{A}}\Big],
\end{aligned} \tag{109}
$$

and for $\boldsymbol{h} = \mathbb{1}_{\mathbb{B}}$

$$
\begin{aligned}
|P[\mathbf{y}\in\mathbb{B}\mid\mathbf{x}\in\mathbb{A}]-\widehat{P}_{\boldsymbol{\theta}}[\mathbf{y}\in\mathbb{B}\mid\mathbf{x}\in\mathbb{A}]| \leq& \left\|\mathbb{E}_{\mathbf{y}}[\boldsymbol{h}]-\widehat{\mathbb{E}}_{\mathbf{y}}[\boldsymbol{h}]\right\|_{\mathcal{Z}}\\
&+\frac{2}{\widehat{\mathbb{E}}_{\mathbf{x}}[\boldsymbol{h}]}\sqrt{\frac{P[\mathbf{y}\in\mathbb{B}]}{P[\mathbf{x}\in\mathbb{A}]}}\Big[2(1+\mathcal{E}_{\boldsymbol{\theta}})[A(\boldsymbol{u}_{\boldsymbol{\theta}},\mathbb{1}_{\mathbb{A}})+A(\boldsymbol{v}_{\boldsymbol{\theta}},\mathbb{1}_{\mathbb{B}})]+\eta_{\mathbb{A}}\Big].
\end{aligned} \tag{110}
$$

Proof. Leveraging the representations in (104) and (106) with $z = \mathbb{1}_{\mathbb{A}}$, we get

$$
\begin{aligned}
&\mathsf{E}_\theta[\boldsymbol{h}(\mathbf{y})|\mathbf{x}\in\mathbb{A}] - \widehat{\mathsf{E}}_\theta[\boldsymbol{h}(\mathbf{y})|\mathbf{x}\in\mathbb{A}] = \mathbb{E}_{\mathbf{y}}[\boldsymbol{h}]-\widehat{\mathbb{E}}_{\mathbf{y}}[\boldsymbol{h}]+\tfrac{\mathbb{E}_{\mathbf{x}}[\mathbb{1}_{\mathbb{A}}(\mathbf{x})[\mathsf{D}_{\boldsymbol{\theta}}\boldsymbol{h}](\mathbf{x})]}{\mathbb{E}[\mathbb{1}_{\mathbb{A}}]} - \tfrac{\widehat{\mathbb{E}}_{\mathbf{x}}[\mathbb{1}_{\mathbb{A}}[\widehat{\mathsf{D}}_{\boldsymbol{\theta}}\boldsymbol{h}]]}{\widehat{\mathbb{E}}_{\mathbf{x}}[\mathbb{1}_{\mathbb{A}}]} =\\
&\mathbb{E}_{\mathbf{y}}[\boldsymbol{h}]-\widehat{\mathbb{E}}_{\mathbf{y}}[\boldsymbol{h}]+\mathbb{E}_{\mathbf{x}}[\mathbb{1}_{\mathbb{A}}(\mathbf{x})[\mathsf{D}_{\mathbf{y}|\mathbf{x}}\boldsymbol{h}](\mathbf{x})]\left(\tfrac{1}{\mathbb{E}[\mathbb{1}_{\mathbb{A}}(\mathbf{x})]}-\tfrac{1}{\widehat{\mathbb{E}}_{\mathbf{x}}[\mathbb{1}_{\mathbb{A}}]}\right)+\tfrac{\mathbb{E}_{\mathbf{x}}[\mathbb{1}_{\mathbb{A}}(\mathbf{x})[\mathsf{D}_{\boldsymbol{\theta}}\boldsymbol{h}](\mathbf{x})]-\widehat{\mathbb{E}}_{\mathbf{x}}[\mathbb{1}_{\mathbb{A}}[\widehat{\mathsf{D}}_{\boldsymbol{\theta}}\boldsymbol{h}]]}{\widehat{\mathbb{E}}_{\mathbf{x}}[\mathbb{1}_{\mathbb{A}}]}.
\end{aligned}
$$

By triangular inequality applied to the norm in $\mathcal{Z}$, we get

$$
\begin{aligned}
&\left\|\mathsf{E}_\theta[\boldsymbol{h}(\mathbf{y})|\mathbf{x}\in\mathbb{A}] - \widehat{\mathsf{E}}_\theta[\boldsymbol{h}(\mathbf{y})|\mathbf{x}\in\mathbb{A}]\right\|_{\mathcal{Z}}\\
&\leq \left\|\mathbb{E}_{\mathbf{y}}[\boldsymbol{h}]-\widehat{\mathbb{E}}_{\mathbf{y}}[\boldsymbol{h}]\right\|_{\mathcal{Z}}+\|\mathbb{E}[\mathbb{1}_{\mathbb{A}}(\mathbf{x})[\mathsf{D}_{\boldsymbol{\theta}}f(\mathbf{x})]]\|_{\mathcal{Z}}\left|\tfrac{1}{\mathbb{E}[\mathbb{1}_{\mathbb{A}}(\mathbf{x})]}-\tfrac{1}{\widehat{\mathbb{E}}_{\mathbf{x}}[\mathbb{1}_{\mathbb{A}}]}\right|+\tfrac{\left\|\mathbb{E}_{\mathbf{x}}[\mathbb{1}_{\mathbb{A}}(\mathbf{x})[\mathsf{D}_{\boldsymbol{\theta}}\boldsymbol{h}](\mathbf{x})]-\widehat{\mathbb{E}}_{\mathbf{x}}[\mathbb{1}_{\mathbb{A}}[\widehat{\mathsf{D}}_{\boldsymbol{\theta}}\boldsymbol{h}]]\right\|_{\mathcal{Z}}}{\widehat{\mathbb{E}}_{\mathbf{x}}[\mathbb{1}_{\mathbb{A}}]}\\
&\leq \left\|\mathbb{E}_{\mathbf{y}}[\boldsymbol{h}]-\widehat{\mathbb{E}}_{\mathbf{y}}[\boldsymbol{h}]\right\|_{\mathcal{Z}}+\|\mathbb{E}[\mathbb{1}_{\mathbb{A}}(\mathbf{x})[\mathsf{D}_{\boldsymbol{\theta}}\boldsymbol{h}](\mathbf{x})]\|_{\mathcal{Z}}\tfrac{2\eta_{\mathbb{A}}}{\mathbb{P}[\mathbf{x}\in\mathbb{A}]}+\tfrac{\left\|\mathbb{E}_{\mathbf{x}}[\mathbb{1}_{\mathbb{A}}(\mathbf{x})[\mathsf{D}_{\boldsymbol{\theta}}\boldsymbol{h}](\mathbf{x})]-\widehat{\mathbb{E}}_{\mathbf{x}}[\mathbb{1}_{\mathbb{A}}[\widehat{\mathsf{D}}_{\boldsymbol{\theta}}\boldsymbol{h}]]\right\|_{\mathcal{Z}}}{\widehat{\mathbb{E}}_{\mathbf{x}}[\mathbb{1}_{\mathbb{A}}]},
\end{aligned}
$$

where we have used Condition (108) in the last line to get that

$$\left|\frac{1}{\mathbb{P}[\mathbf{x}\in\mathbb{A}]}-\frac{1}{\widehat{\mathbb{E}}_{\mathbf{x}}[\mathbb{1}_{\mathbb{A}}]}\right|\le\frac{\eta_{\mathbb{A}}}{(1-\eta_{\mathbb{A}})\mathbb{P}[\mathbf{x}\in\mathbb{A}]}\le\frac{2\eta_{\mathbb{A}}}{\mathbb{P}[\mathbf{x}\in\mathbb{A}]}.$$

From Proposition M.3 and Condition (108) we get that

$$\tfrac{1}{\widehat{\mathbb{E}}_{\mathbf{x}}[\mathbb{1}_{\mathbb{A}}]}\left\|\mathbb{E}_{\mathbf{x}}[\mathbb{1}_{\mathbb{A}}(\mathbf{x})[\mathsf{D}_{\boldsymbol{\theta}}\boldsymbol{h}](\mathbf{x})]-\widehat{\mathbb{E}}_{\mathbf{x}}[\mathbb{1}_{\mathbb{A}}(\mathbf{x})[\widehat{\mathsf{D}}_{\boldsymbol{\theta}}\boldsymbol{h}]]\right\|_{\mathcal{Z}}\le\frac{2(1+\mathcal{E}_{\boldsymbol{\theta}})\left[A(\boldsymbol{u}_{\boldsymbol{\theta}},\mathbb{1}_{\mathbb{A}})A(\boldsymbol{v}_{\boldsymbol{\theta}},\boldsymbol{h})\right]}{\mathbb{P}(\mathbf{x}\in\mathbb{A})}\|\mathbb{1}_{\mathbb{A}}\|_{\mathcal{L}^2_{P_{\mathbf{x}}}}\|\boldsymbol{h}\|_{\mathcal{L}^2_{P_{\mathbf{y}}}(\mathcal{Y},\mathcal{Z})}$$

Cauchy's Schwarz's inequality again and $\|\mathsf{D}_{\boldsymbol{\theta}}\|\le1$ give

$$\|\mathbb{E}[\mathbb{1}_{\mathbb{A}}(\mathbf{x})[\mathsf{D}_{\boldsymbol{\theta}}\boldsymbol{h}](\mathbf{x})]\|_{\mathcal{Z}}\le\|\mathbb{1}_{\mathbb{A}}\|_{\mathcal{L}^2_{P_{\mathbf{x}}}}\|\mathsf{D}_{\boldsymbol{\theta}}\|\|\boldsymbol{h}\|_{\mathcal{L}^2_{P_{\mathbf{y}}}}\le\|\mathbb{1}_{\mathbb{A}}\|_{\mathcal{L}^2_{P_{\mathbf{x}}}}\|\boldsymbol{h}\|_{\mathcal{L}^2_{P_{\mathbf{y}}}}=\sqrt{\mathbb{P}[\mathbf{x}\in\mathbb{A}]}\|\boldsymbol{h}\|_{\mathcal{L}^2_{P_{\mathbf{y}}}(\mathcal{Y},\mathcal{Z})}.$$

Combining the last four displays give the first result. The second result follows immediately for $\boldsymbol{h}=\mathbb{1}_{\mathbb{B}}$. □

Consequence of this result is that we can bound the error in probability as we can derive concentration inequalities on the terms in (100) and (108). Then an union bound gives the estimation result for regression conditional on sets.

Next, we recall that $\mathsf{E}_{\mathbf{y}|\mathbf{x}}$ being $(1/\alpha)$-Schatten class operator, implies:

Assumption M.5. *Let there exist some constant $c>0$ such that for $\alpha>0$, any $i\ge1$ and any $k\in[n_{\text{iso}}]$, we have $\sigma_i^{(k)}\le c\,i^{-\alpha}$.*

Further, for any $\boldsymbol{h}\in\mathcal{L}^2_{\mathbf{y}}(\mathcal{Y},\mathcal{Z})$, we define $\overline{\boldsymbol{h}}(\mathbf{y})=\boldsymbol{h}(\mathbf{y})-\mathbb{E}[\boldsymbol{h}(\mathbf{y})]$ and

$$\gamma_{\mathbb{G}'}(\boldsymbol{h}):=\frac{1}{|\mathbb{G}'|-1}\sum_{\substack{g\in\mathbb{G}'\\ g\neq e}}\mathbb{E}[\langle\overline{\boldsymbol{h}}(\mathbf{y}),\overline{\boldsymbol{h}}(g\triangleright_{\mathcal{Y}}\mathbf{y})\rangle].\tag{111}$$

In the following, we consider observables h satisfying the following condition (that is clearly satisfied for an indicator of a set of positive measure)

Assumption M.6. *Let there exists an absolute constant $C_0\ge1$ such that $(|\mathbb{G}'|-1)\gamma_{\mathbb{G}'}(h)\le C_0\,\mathbb{E}[\|h(\mathbf{y})\|^2_{\mathcal{Z}}]$.*

Define

$$\eta_{\mathbb{A}}=\eta_{\mathbb{A}}(\delta):=\left(\frac{1-\mathbb{P}[\mathbf{x}\in\mathbb{G}\triangleright\mathbb{A}]}{\mathbb{P}[\mathbf{x}\in\mathbb{G}\triangleright\mathbb{A}]}\right)\frac{\log2\delta^{-1}}{N}+\sqrt{2\frac{\log2\delta^{-1}}{N}}\sqrt{\frac{1-\mathbb{P}[\mathbf{x}\in\mathbb{G}\triangleright\mathbb{A}]}{\mathbb{P}[\mathbf{x}\in\mathbb{G}\triangleright\mathbb{A}]}}.$$

Theorem M.7. *Let Assumptions M.6 and M.5 be satisfied. Let $P_{\mathbf{x}}$ and $P_{\mathbf{y}}$ are $\mathbb{G}$-invariant, and $\mathsf{D}_{\boldsymbol{\theta}}$ from (90) is $\mathbb{G}$-equivariant model, and let $\boldsymbol{h}\in\mathcal{L}^2_{\mathbf{y}}(\mathcal{Y},\mathcal{Z})$ and $\boldsymbol{f}\in\mathcal{L}^2_{\mathbf{x}}(\mathcal{X},\mathcal{Z})$ (with values in $\mathcal{Z}$) be subGaussian random variables. Assume in addition that the event $\mathbb{A}$ is anti-symmetric for $\mathbb{G}$ and that $m_k=m$ for all $k\in[n_{\text{iso}}]$. Assume that $N\ge|\mathbb{G}|$. Then for any $\delta\in(0,1)$, it holds w.p.a.l $1-\delta$*

$$\left\|\mathbb{E}[\boldsymbol{h}(\mathbf{y})\,|\,\mathbf{x}\in\mathbb{A}]-\widehat{\mathsf{E}}_{\boldsymbol{\theta}}[\boldsymbol{h}(\mathbf{y})|\mathbf{x}\in\mathbb{A}]\right\|_{\mathcal{Z}}\lesssim_{C_0}\frac{\|\boldsymbol{h}\|_{\mathcal{L}^2_{\mathbf{y}}(\mathcal{Y},\mathcal{Z})}}{\sqrt{\mathbb{P}[\mathbf{x}\in\mathbb{G}\triangleright_{\mathcal{X}}\mathbb{A}]}}\left(\mathcal{E}_{\boldsymbol{\theta}}+\frac{\log\left(2n_{\text{iso}}\delta^{-1}\right)}{(d_{\text{iso}}N)^{\frac{\alpha}{1+2\alpha}}}\right),$$

and

$$|\mathbb{P}(\mathbf{y}\in\mathbb{B}\mid\mathbf{x}\in\mathbb{A})-\widehat{\mathbb{P}}_{\boldsymbol{\theta}}(\mathbf{y}\in\mathbb{B}\mid\mathbf{x}\in\mathbb{A})|\lesssim_{C_0}\sqrt{\tfrac{\mathbb{P}[\mathbf{y}\in\mathbb{B}]}{\mathbb{P}[\mathbf{x}\in\mathbb{G}\triangleright_{\mathcal{X}}\mathbb{A}]}}\left(\mathcal{E}_{\boldsymbol{\theta}}+\frac{\log\left(2n_{\text{iso}}\delta^{-1}\right)}{(d_{\text{iso}}N)^{\frac{\alpha}{1+2\alpha}}}+\sqrt{|\mathbb{G}|}\eta_{\mathbb{A}}\right).$$

Proof. This result follows immediately from Propositions M.3 and M.4 combined with Lemmas M.9 and Lemma M.10. Set

$$A(\boldsymbol{u}_{\boldsymbol{\theta}},\boldsymbol{f}):=C\sqrt{\tfrac{1}{|\mathbb{G}'|\,N}}\sqrt{C_0\vee\tfrac{|\mathbb{G}'|}{N}}\,\log\left(2n_{\text{iso}}\delta^{-1}\right),$$

$$A(\boldsymbol{v}_{\boldsymbol{\theta}},\boldsymbol{h}):=C\sqrt{\tfrac{\max_{k\in[n_{\text{iso}}]}\{m_k\}}{|\mathbb{G}'|\,N}}\sqrt{C_0\vee\tfrac{|\mathbb{G}'|}{N}}\,\log\left(2n_{\text{iso}}\delta^{-1}\right),$$

for some large enough absolute constant $C > 0$.

Then an union bound based on Lemmas M.9 and M.10 guarantees that (109) is satisfied w.p.a.l. $1 - \delta$ (up to a rescaling of the constant C):

$$\left\|\mathsf{E}_{\boldsymbol{\theta}}[\boldsymbol{h}(\mathbf{y})|\mathbf{x}\in\mathbb{A}] - \widehat{\mathsf{E}}_{\boldsymbol{\theta}}[\boldsymbol{h}(\mathbf{y})|\mathbf{x}\in\mathbb{A}]\right\|_{\mathcal{Z}} \leq$$
$$C\frac{\|\boldsymbol{h}\|_{\mathcal{L}^2_{\mathbf{x}}(\mathcal{X},\mathcal{Z})}}{\sqrt{P[\mathbf{x}\in\mathbb{A}]}}\left[2(1+\mathcal{E}_{\boldsymbol{\theta}})\left(\sqrt{\frac{\max_{k\in[n_{\mathrm{iso}}]}\{m_k\}}{|\mathbb{G}'|\,N}}\sqrt{C_0\vee\frac{|\mathbb{G}'|}{N}}\,\log\big(2n_{\mathrm{iso}}\delta^{-1}\big)\right)\right].$$

Next we use our bound on the representation bias in (97)

$$\|\mathbb{E}[\boldsymbol{y}(\mathbf{y})\,|\,\mathbf{x}\in\mathbb{A}] - \mathsf{E}_{\boldsymbol{\theta}}[\boldsymbol{y}(\mathbf{y})\,|\,\mathbf{x}\in\mathbb{A}]\|_{\mathcal{Z}} \leq \left(\sigma^{\star}_{r_m+1} + \mathcal{E}_{\boldsymbol{\theta}}\right)\frac{\|\boldsymbol{h}\|_{\mathcal{L}^2_{\mathbf{y}}(\mathcal{Y},\mathcal{Z})}}{\sqrt{\mathbb{P}[\mathbf{x}\in\mathbb{A}]}}\sqrt{\frac{1+(|\mathbb{G}'|-1)\gamma_{\mathbb{G}}(\mathbb{A})}{|\mathbb{G}'|}}. \quad (112)$$

Recall that $\mathcal{E}_{\boldsymbol{\theta}} = \max_{k\in[n_{\mathrm{iso}}]}\{\mathcal{E}^{(k)}_{\boldsymbol{\theta}}\}$. Under Assumption M.5, we have $\left\|[\![\mathsf{D}_{\mathbf{y}|\mathbf{x}}]\!]_{r_m} - \mathsf{D}_{\boldsymbol{\theta}}\right\| \leq \frac{1}{(d_{\mathrm{iso}}m)^{\alpha}}$. In addition, $(|\mathbb{G}'|-1)\gamma_{\mathbb{G}}(\mathbb{A}) \leq C_0$ under Assumption M.6.

Combining the last two display gives w.p.a.l $1-\delta$

$$\left\|\mathbb{E}[\boldsymbol{h}(\mathbf{y})\,|\,\mathbf{x}\in\mathbb{A}] - \widehat{\mathsf{E}}_{\boldsymbol{\theta}}[\boldsymbol{h}(\mathbf{y})|\mathbf{x}\in\mathbb{A}]\right\|_{\mathcal{Z}} \lesssim_{C_0} \frac{\|\boldsymbol{h}\|_{\mathcal{L}^2_{\mathbf{y}}(\mathcal{Y},\mathcal{Z})}}{\sqrt{\mathbb{P}[\mathbf{x}\in\mathbb{G}\triangleright_{\mathcal{X}}\mathbb{A}]}}\left(\mathcal{E}_{\boldsymbol{\theta}} + \frac{1}{(d_{\mathrm{iso}}m)^{\alpha}} + \sqrt{\frac{m}{N}}\,\log\big(2n_{\mathrm{iso}}\delta^{-1}\big)\right).$$

Balancing the previous display w.r.t. dimension m, we get that $m \asymp (d_{\mathrm{iso}}^{-2\alpha}N)^{\frac{1}{1+2\alpha}}$ and the first result follows.

The bound for the conditional probability follows by picking $\boldsymbol{y} = \mathbb{1}_{\mathbb{B}}$. $\square$

M.1 Quadratic error regression bound

Our goal is to estimate the conditional expectation function

$$\boldsymbol{z}(\boldsymbol{x}) = \mathbb{E}[\boldsymbol{h}(\mathbf{y})|\mathbf{x}{=}\boldsymbol{x}] = \mathbb{E}[\boldsymbol{h}(\mathbf{y})] + [\mathsf{D}_{\mathbf{y}|\mathbf{x}}\boldsymbol{h}](\boldsymbol{x}).$$

Our estimator is

$$\widehat{\boldsymbol{z}}_{\boldsymbol{\theta}}(\cdot) = \widehat{\mathbb{E}}_{\mathbf{y}}[\mathbf{y}] + [\widehat{\mathsf{D}}_{\boldsymbol{\theta}}\boldsymbol{h}](\cdot).$$

Theorem M.8. *Assume that Y is a sub-Gaussian random vector. Let Assumption M.5 be satisfied. Assume in addition that $\mathcal{E}_{\boldsymbol{\theta}} \leq 1$, $m_k = m$ for all $k \in [n_{\mathrm{iso}}]$. Then for any $\delta \in (0,1)$ such that $N \geq (c_u \vee c_v)^2 m \log\big(e\delta^{-1}n_{\mathrm{iso}}\big) \vee |\mathbb{G}|$, it holds w.p.a.l. $1-\delta$*

$$\|\boldsymbol{z} - \widehat{\boldsymbol{z}}_{\boldsymbol{\theta}}\|^2_{\mathcal{L}^2_{\mathbf{x}}(\mathcal{X},\mathcal{Z})} \lesssim \mathrm{Tr}(\mathrm{Cov}(Y))\left(\mathcal{E}^2_{\boldsymbol{\theta}} + (d_{\mathrm{iso}}|\mathbb{G}|N)^{\frac{-2\alpha}{1+2\alpha}}\log^2(\delta^{-1}n_{\mathrm{iso}})\right). \quad (113)$$

Discussion When the training of the NN is successful, we expect the statistical rate to dominate the optimization error $\max_{k\in[n_{\mathrm{iso}}]}\{\mathcal{E}^{(k)}_{\boldsymbol{\theta}}\}$ for large enough sample size N. For distribution containing symmetry invariants with large isotopic components (m is large), we observe that exploiting this information in the construction of the NCP operator yields a substantial improvement in the statistical error rate as we go from a rate $N^{-\frac{\alpha}{1+2\alpha}}$ for standard NCP to $(N\,m)^{-\frac{\alpha}{1+2\alpha}}$ for eNCP.

Proof. Combining (101) with Lemma M.9 gives w.p.a.l. $1-\delta$

$$\begin{aligned}\left\|\mathsf{E}_{\boldsymbol{\theta}}\,\mathbf{y} - \widehat{\mathsf{E}}_{\boldsymbol{\theta}}\,\mathbf{y}\right\|^2_{\mathcal{L}^2_{P_{\mathbf{x}}}(\mathcal{X})} &\lesssim (1+\mathcal{E}_{\boldsymbol{\theta}})^3\mathrm{Tr}(\mathrm{Cov}(\mathbf{y}))\frac{m}{|\mathbb{G}|N}\log^2(2n_{\mathrm{iso}}\delta^{-1})\\ &\lesssim \mathrm{Tr}(\mathrm{Cov}(\mathbf{y}))\frac{m}{|\mathbb{G}|N}\log^2(n_{\mathrm{iso}}\delta^{-1}),\end{aligned}$$

provide that $\mathcal{E}_{\boldsymbol{\theta}} \leq 1$. We derived in (95) an upper bound on the bias term

$$\left\|\mathbb{E}_{\mathbf{y}|\mathbf{x}}[\mathbf{y}\,|\,\mathbf{x}=\cdot] - \mathsf{E}_{\boldsymbol{\theta}}\mathbf{y}\right\|^2_{\mathcal{L}^2_{P_{\mathbf{x}}}(\mathcal{X},\mathcal{Z})} \leq \mathrm{Tr}(\mathrm{Cov}(\mathbf{y}))\left(\frac{1}{(d_{\mathrm{iso}}m)^{2\alpha}} + \mathcal{E}^2_{\boldsymbol{\theta}}\right). \quad (114)$$

Balancing the two bounds in the last two displays w.r.t. $m \asymp (|\mathbb{G}|d_{\mathrm{iso}}N)^{\frac{1}{1+2\alpha}}$, we get the result. $\square$

M.2 Auxiliary results.

Consider the function space $\mathcal{L}^2_{\mathbf{y}}(\mathcal{Y}, \mathcal{Z})$ where $\mathcal{Z}$ is endowed with an inner product $\langle \cdot, \cdot \rangle = \langle \cdot, \cdot \rangle_{\mathcal{Z}}$. If the distribution of $\mathbf{y}$ is $\mathbb{G}'$-invariant, then for any $h \in \mathcal{L}^2_{\mathbf{y}}(\mathcal{Y}, \mathcal{Z})$, we use the estimator $\widehat{\mathbb{E}}_{\mathbf{y}}[h]$ in (98) as an estimator of $\mathbb{E}[h(\mathbf{y})]$.

Lemma M.9. *Assume that the distribution $P_{\mathbf{y}}$ of $\mathbf{y}$ is $\mathbb{G}$-invariant and let $\mathbb{G}' \leq \mathbb{G}$. Let there exists a function $\boldsymbol{h} \in \mathcal{L}^2_{\mathbf{y}}(\mathcal{Y}, \mathcal{Z})$ such that $\boldsymbol{h}(\mathbf{y})$ is subGaussian. Then there exists an absolute constant $C > 0$ such that for any $\delta \in (0, 1)$, it holds w.p.a.l. $1 - \delta$*

$$\left\| \widehat{\mathbb{E}}_{\mathbf{y}}[\boldsymbol{h}] - \mathbb{E}[\boldsymbol{h}(\mathbf{y})] \right\|_{\mathcal{Z}} \leq C \sqrt{\frac{\log^2 2\delta^{-1}}{|\mathbb{G}'|\, N}} \sqrt{\mathbb{E}[\left\|\overline{\boldsymbol{h}}(\mathbf{y})\right\|^2_{\mathcal{Z}}] + (|\mathbb{G}'| - 1)\gamma_{\mathbb{G}'}(\boldsymbol{h}) + \frac{|\mathbb{G}'| \mathbb{E}[\left\|\overline{\boldsymbol{h}}(\mathbf{y})\right\|^2_{\mathcal{Z}}]}{N}}.$$

Assume in addition that there exists an absolute constant $C_0 \geq 1$ such that $(|\mathbb{G}'| - 1)\gamma_{\mathbb{G}'}(\overline{\boldsymbol{h}}) \leq C_0\, \mathbb{E}[\|\boldsymbol{h}(\mathbf{y})\|^2_{\mathcal{Z}}]$. Then for any $\delta \in (0, 1)$, it holds w.p.a.l. $1 - \delta$

$$\left\| \widehat{\mathbb{E}}_{\mathbf{y}}[\boldsymbol{h}] - \mathbb{E}[\boldsymbol{h}(\mathbf{y})] \right\|_{\mathcal{Z}} \leq C \sqrt{\frac{\mathbb{E}[\left\|\overline{\boldsymbol{h}}(\mathbf{y})\right\|^2_{\mathcal{Z}}]}{|\mathbb{G}'|\, N}} \sqrt{(1 + C_0) + \frac{|\mathbb{G}'|}{N}}\, \log 2\delta^{-1}.$$

Note that similar bounds hold valid for the $\mathbb{G}$-invariant distribution $P_{\mathbf{x}}$ and any function $\boldsymbol{f} \in \mathcal{L}^2_{P_{\mathbf{x}}}(\mathcal{X}, \mathcal{Z})$ such that $\boldsymbol{f}(\mathbf{x})$ is subGaussian.

Proof. We note that

$$\widehat{\mathbb{E}}_{\mathbf{y}}[\boldsymbol{h}] - \mathbb{E}[\boldsymbol{h}(\mathbf{y})] = \frac{1}{N} \sum_{i=1}^{N} Z_i \quad \text{with} \quad Z_i = \frac{1}{|\mathbb{G}'|} \sum_{g \in \mathbb{G}'} \boldsymbol{h}(g \triangleright_{\mathcal{Y}} \mathrm{y}_i) - \mathbb{E}_{\mathrm{y}_i}[\boldsymbol{h}(g \triangleright_{\mathcal{Y}} \mathrm{y}_i)],\ \forall i \in [N].$$

Define

$$Z := \frac{1}{|\mathbb{G}'|} \sum_{g \in \mathbb{G}'} \boldsymbol{h}(g \triangleright_{\mathcal{Y}} \mathbf{y}) - \mathbb{E}_{\mathbf{y}}[\boldsymbol{h}(g \triangleright_{\mathcal{Y}} \mathbf{y})], \tag{115}$$

and, for brevity, set $\|\boldsymbol{z}\| = \|\boldsymbol{z}\|_{\mathcal{Z}} = \sqrt{\langle \boldsymbol{z}, \boldsymbol{z} \rangle_{\mathcal{Z}}}$ for any $\boldsymbol{z} \in \mathcal{Z}$. We apply Proposition M.12, to get w.p.a.l. $1 - \delta$

$$\left\| \widehat{\mathbb{E}}_{\mathbf{y}}[\boldsymbol{h}] - \mathbb{E}_{\mathbf{y}}[\boldsymbol{h}(\mathbf{y})] \right\| \leq \frac{4\sqrt{2}}{\sqrt{N}} \sqrt{\mathrm{Var}_{\mathbf{y}}(\|Z\|) + \frac{\|Z\|^2_{\psi_2}}{N}} \log \frac{2}{\delta}.$$

Using the triangular inequality successively on $\|\cdot\|$ and $\|\cdot\|_{\psi_2}$ and the $\mathbb{G}'$-invariance of $P_{\mathbf{y}}$, $\left\|\overline{\boldsymbol{h}}(g \triangleright_{\mathcal{Y}} \mathbf{y})\right\|_{\psi_2} = \left\|\overline{\boldsymbol{h}}(\mathbf{y})\right\|_{\psi_2}$ for any $g \in \mathbb{G}'$, we get that

$$\| \|Z\| \|_{\psi_2} \lesssim \left\| \left\|\overline{\boldsymbol{h}}(\mathbf{y})\right\| \right\|_{\psi_2}.$$

We note next that $\left\|\overline{\boldsymbol{h}}(\mathbf{y})\right\|$ is subGaussian. Consequently the well-known property of equivalence of moments for subGaussian distributions gives $\|Z\|_{\psi_2} \lesssim \left\| \left\|\overline{\boldsymbol{h}}(\mathbf{y})\right\| \right\|_{\psi_2} \lesssim \mathbb{E}[\left\|\overline{\boldsymbol{h}}(\mathbf{y})\right\|^2]$. We derive now a control on $\mathrm{Var}_{\mathbf{y}}(\|Z\|) \leq \mathbb{E}[\|Z\|^2]$. Using the $\mathbb{G}'$-invariance of $P_{\mathbf{y}}$, we get

$$\begin{aligned}
\mathrm{Var}(\|Z\|) &\leq \frac{\mathbb{E}[\left\|\overline{\boldsymbol{h}}(\mathbf{y})\right\|^2]}{|\mathbb{G}'|} + \frac{1}{|\mathbb{G}'|} \sum_{\substack{g \in \mathbb{G}' \\ g \neq e}} \mathbb{E}[\langle \boldsymbol{h}(\mathbf{y}) - \mathbb{E}[\boldsymbol{h}(\mathbf{y})], \boldsymbol{h}(g \triangleright_{\mathcal{Y}} \mathbf{y}) - \mathbb{E}[\boldsymbol{h}(\mathbf{y})] \rangle] \\
&= \frac{\mathbb{E}[\left\|\overline{\boldsymbol{h}}(\mathbf{y})\right\|^2]}{|\mathbb{G}'|} + \frac{(|\mathbb{G}'| - 1)\gamma_{\mathbb{G}'}(\boldsymbol{h})}{|\mathbb{G}'|} \leq (1 + C_0) \frac{\mathbb{E}[\left\|\overline{\boldsymbol{h}}(\mathbf{y})\right\|^2]}{|\mathbb{G}'|}.
\end{aligned} \tag{116}$$

Hence we get the result.

$\square$

We focus now on a concentration bound for indicator functions $z = \mathbb{1}_{\mathbb{A}}$ for any event $\mathbb{A} \in \Sigma_{\mathcal{X}}$. We define

$$Z_A := \widehat{\mathbb{E}}_{\mathbf{x}}[\mathbb{1}_{\mathbb{A}}] - \mathbb{P}[\mathbf{x} \in \mathbb{A}] = \frac{1}{|\mathbb{G}'|} \sum_{g \in \mathbb{G}'} (\mathbb{1}_{g^{-1} \triangleright_{\mathcal{X}} \mathbb{A}}(\mathbf{x}) - \mathbb{E}[\mathbb{1}_{g^{-1} \triangleright_{\mathcal{X}} \mathbb{A}}(\mathbf{x})])$$
$$= \frac{1}{|\mathbb{G}'|} \sum_{g \in \mathbb{G}'} \left(\mathbb{1}_{g^{-1} \triangleright_{\mathcal{X}} \mathbb{A}}(\mathbf{x}) - \mathbb{P}[\mathbf{x} \in \mathbb{A}]\right) = \left(\frac{1}{|\mathbb{G}'|} \sum_{g \in \mathbb{G}'} \mathbb{1}_{g^{-1} \triangleright_{\mathcal{X}} \mathbb{A}}(\mathbf{x})\right) - \mathbb{P}[\mathbf{x} \in \mathbb{A}]. \tag{117}$$

Note that we always have $|Z_{\mathbb{A}}| \leq 1$ but this bound can be quite conservative as we could get a much sharper bound for some events $\mathbb{A}$. We denote by $\gamma_{\mathbb{G}',\infty}(\mathbb{A})$ the smallest deterministic upper-bound on $\frac{1}{|\mathbb{G}'|} \sum_{g \in \mathbb{G}'} \mathbb{1}_{g^{-1} \triangleright_{\mathcal{X}} \mathbb{A}}(\mathbf{x})$ (For instance when $\mathbb{A}$ is an antisymmetric event, then we have $\gamma_{\mathbb{G}',\infty}(\mathbb{A}) = 1/|\mathbb{G}'|$). Then we have

$$-\mathbb{P}[\mathbf{x} \in \mathbb{A}] \leq Z_{\mathbb{A}} \leq \gamma_{\mathbb{G}',\infty}(\mathbb{A}) - \mathbb{P}[\mathbf{x} \in \mathbb{A}]. \tag{118}$$

Define also

$$\Upsilon_{\mathbb{G}',X}(\mathbb{A}) := \mathbb{P}(\mathbf{x} \in \mathbb{A})(1 - \mathbb{P}(\mathbf{x} \in \mathbb{A})) + (|\mathbb{G}'| - 1)\left(\gamma_{\mathbb{G}'}(\mathbb{A}) - \mathbb{P}[\mathbf{x} \in \mathbb{A}]\right)\mathbb{P}[\mathbf{x} \in \mathbb{A}]. \tag{119}$$

Lemma M.10. *Let the distribution of* $\mathbf{x}$ *be* $\mathbb{G}'$*-invariant. Then for any* $\mathbb{A} \in \Sigma_{\mathcal{X}}$ *and any* $\delta \in (0, 1)$*, it holds w.p.a.l.* $1 - \delta$

$$\left|\widehat{\mathbb{E}}_{\mathbf{x}}[\mathbb{1}_{\mathbb{A}}] - \mathbb{E}[\mathbb{1}_{\mathbb{A}}(\mathbf{x})]\right| \leq |\gamma_{\mathbb{G}',\infty}(\mathbb{A}) - \mathbb{P}[\mathbf{x} \in \mathbb{A}]| \frac{\log 2\delta^{-1}}{N} + \sqrt{\frac{\Upsilon_{\mathbb{G}',\mathbf{x}}(\mathbb{A})}{|\mathbb{G}'|}} \sqrt{2\frac{\log 2\delta^{-1}}{N}}.$$

Assume in addition that $g \triangleright \mathbb{A} \cap \mathbb{A} = \emptyset$ *for all* $g \in \mathbb{G}' \setminus \{e\}$*. Then it holds w.p.a.l.* $1 - \delta$

$$\frac{\left|\widehat{\mathbb{E}}_{\mathbf{x}}[\mathbb{1}_{\mathbb{A}}] - \mathbb{E}[\mathbb{1}_{\mathbb{A}}(\mathbf{x})]\right|}{\mathbb{E}[\mathbb{1}_{\mathbb{A}}(\mathbf{x})]} \leq \left(\frac{1 - \mathbb{P}[\mathbf{x} \in \mathbb{G}' \triangleright \mathbb{A}]}{\mathbb{P}[\mathbf{x} \in \mathbb{G}' \triangleright \mathbb{A}]}\right) \frac{\log 2\delta^{-1}}{N} + \sqrt{2\frac{\log 2\delta^{-1}}{N}} \sqrt{\frac{1 - \mathbb{P}[\mathbf{x} \in \mathbb{G}' \triangleright \mathbb{A}]}{\mathbb{P}[\mathbf{x} \in \mathbb{G}' \triangleright \mathbb{A}]}}.$$

If the distribution of $\mathbf{y}$ *is* $\mathbb{G}'$*-invariant, then an identical result is immediately available for* $\mathbf{y}$ *by the same proof argument.*

Remark M.11. Using the standard empirical mean estimator that does not take advantage of $\mathbb{G}$-invariance, we obtain a concentration bound with a slower rate. For example, for an antisymmetric event A, we would achieve, w.p.a.l. $1 - \delta$, the following result:

$$\frac{\left|\widehat{\mathbb{E}}_{\mathbf{x}}[\mathbb{1}_{\mathbb{A}}] - \mathbb{E}[\mathbb{1}_{\mathbb{A}}(\mathbf{x})]\right|}{\mathbb{E}[\mathbb{1}_{\mathbb{A}}(\mathbf{x})]} \leq \left(\frac{1 - \mathbb{P}[\mathbf{x} \in \mathbb{A}]}{\mathbb{P}[\mathbf{x} \in \mathbb{A}]}\right) \frac{\log 2\delta^{-1}}{N} + \sqrt{2\frac{\log 2\delta^{-1}}{N}} \sqrt{\frac{1 - \mathbb{P}[\mathbf{x} \in \mathbb{A}]}{\mathbb{P}[\mathbf{x} \in \mathbb{A}]}}.$$

Specifically, leveraging $\mathbb{G}'$-invariance allows us to replace $\mathbb{P}[\mathbf{x} \in \mathbb{A}]$ with $\mathbb{P}[\mathbf{x} \in \mathbb{G}' \triangleright_{\mathcal{X}} \mathbb{A}]$, which represents the probability of the entire orbit of A under the action of $\mathbb{G}'$. This becomes particularly interesting when $\mathbb{P}[\mathbf{x} \in \mathbb{A}] \ll \mathbb{P}[\mathbf{x} \in \mathbb{G}' \triangleright_{\mathcal{X}} \mathbb{A}]$, especially in the case of rare events where $\mathbb{P}[\mathbf{x} \in \mathbb{A}] \approx 0$.

Proof. Since $P_{\mathbf{x}}$ is $\mathbb{G}'$-invariant, we have $\mathbb{E}[\mathbb{1}_{g^{-1} \triangleright \mathbb{A}}(\mathbf{x})] = \mathbb{P}[\mathbf{x} \in \mathbb{A}]$ and $\mathrm{Var}(\mathbb{1}_{g^{-1} \triangleright \mathbb{A}}(\mathbf{x})) = \mathrm{Var}(\mathbb{1}_{\mathbb{A}}(\mathbf{x})) = \mathbb{P}[\mathbf{x} \in \mathbb{A}](1 - \mathbb{P}[\mathbf{x} \in \mathbb{A}])$, for any $g \in \mathbb{G}'$. Hence

$$\widehat{\mathbb{E}}_{\mathbf{x}}[\mathbb{1}_{\mathbb{A}}] - \mathbb{E}[\mathbb{1}_{\mathbb{A}}(\mathbf{x})] = \frac{1}{N} \sum_{i=1}^{N} Z_i \quad \text{with} \quad Z_i = \frac{1}{|\mathbb{G}'|} \sum_{g \in \mathbb{G}'} \mathbb{1}_{g^{-1} \triangleright \mathbb{A}}(\mathbf{x}_i) - \mathbb{E}[\mathbb{1}_{g^{-1} \triangleright \mathbb{A}}(\mathbf{x}_i)], \ \forall i \in [N].$$

The Z_i's are i.i.d. copies of $Z = Z_{\mathbb{A}}$. In view of (118), we can apply Hoeffding's inequality Bercu et al. [79, Theorem 2.16]. We get for any $\delta \in (0, 1)$ w.p.a.l $1 - \delta$

$$\left|\widehat{\mathbb{E}}_{\mathbf{x}}[\mathbb{1}_{\mathbb{A}}] - \mathbb{E}[\mathbb{1}_{\mathbb{A}}(\mathbf{x})]\right| \leq \gamma_{\mathbb{G}',\infty}(\mathbb{A}) \sqrt{\frac{\log 2\delta^{-1}}{2N}}. \tag{120}$$

We propose to prove another bound based on application of Bernstein's inequality. We first prove an improved bound on $\mathrm{Var}(Z)$ as compared to the standard empirical mean estimator which does not exploit $\mathbb{G}$-invariance. Indeed we have

$$\begin{aligned}\mathrm{Var}(Z) &= \frac{1}{|\mathbb{G}'|^2}\left(\sum_{g\in\mathbb{G}'}\mathrm{Var}(\mathbb{1}_{g^{-1}\triangleright\mathbb{A}}(\mathbf{x})) + \sum_{g\neq g'}\mathrm{Cov}\Big(\mathbb{1}_{g^{-1}\triangleright\mathbb{A}}(\mathbf{x}), \mathbb{1}_{(g')^{-1}\triangleright\mathbb{A}}(\mathbf{x})\Big)\right)\\ &= \frac{\mathbb{P}(\mathbf{x}\in\mathbb{A})(1-\mathbb{P}(\mathbf{x}\in\mathbb{A}))}{|\mathbb{G}'|} + \frac{1}{|\mathbb{G}'|^2}\sum_{g\neq g'}\mathrm{Cov}\Big(\mathbb{1}_{g^{-1}\triangleright\mathbb{A}}(\mathbf{x}), \mathbb{1}_{(g')^{-1}\triangleright\mathbb{A}}(\mathbf{x})\Big).\end{aligned}$$

Next, using again that $\mathbb{P}_X$ is $\mathbb{G}$-invariant, we get for any $g, g' \in \mathbb{G}'$

$$\mathrm{Cov}\Big(\mathbb{1}_{g^{-1}\triangleright\mathbb{A}}(\mathbf{x}),\mathbb{1}_{(g')^{-1}\triangleright\mathbb{A}}(\mathbf{x})\Big) = \tag{121}$$

$$\begin{aligned}&\mathbb{P}[\mathbf{x}\in g^{-1}\triangleright\mathbb{A}\cap(g')^{-1}\triangleright\mathbb{A}] - \mathbb{P}[\mathbf{x}\in g^{-1}\triangleright\mathbb{A}]\,\mathbb{P}[\mathbf{x}\in(g')^{-1}\triangleright\mathbb{A}]\\ &= \mathbb{P}[\mathbf{x}\in g^{-1}\triangleright\mathbb{A}\cap(g')^{-1}\triangleright\mathbb{A}] - \mathbb{P}[\mathbf{x}\in\mathbb{A}]^2.\end{aligned}\tag{122}$$

Using again the invariance assumption, we note that

$$\sum_{g\neq g'}\mathrm{Cov}\big(\mathbb{1}_{g^{-1}\triangleright\mathbb{A}}(\mathbf{x}), \mathbb{1}_{(g')^{-1}\triangleright\mathbb{A}}(\mathbf{x})\big) = |\mathbb{G}'|\big(\sum_{g\in\mathbb{G}', g\neq e}\mathbb{P}[\mathbf{x}\in\mathbb{A}\cap g\triangleright\mathbb{A}]\big) - |\mathbb{G}'|(|\mathbb{G}'|-1)\mathbb{P}[\mathbf{x}\in\mathbb{A}]^2$$

Consequently by definition of $\gamma_{\mathbb{G}'}(A)$ in (93) and (94), we get

$$\sum_{g\in\mathbb{G}', g\neq e}\mathbb{P}[\mathbf{x}\in\mathbb{A}\cap g\triangleright A] = (|\mathbb{G}'|-1)\,\gamma_{\mathbb{G}'}(A)\,\mathbb{P}(\mathbf{x}\in\mathbb{A}).$$

Combining the last four displays, we get

$$\mathrm{Var}(Z) = \frac{\mathbb{P}(\mathbf{x}\in\mathbb{A})(1-\mathbb{P}(\mathbf{x}\in\mathbb{A})) + (|\mathbb{G}'|-1)\,(\gamma_{\mathbb{G}'}(A) - \mathbb{P}[\mathbf{x}\in\mathbb{A}])\mathbb{P}[\mathbf{x}\in\mathbb{A}]}{|\mathbb{G}'|} = \frac{\Upsilon_{\mathbb{G}',\mathbf{x}}(A)}{|\mathbb{G}'|}.\tag{123}$$

We note that for any $p \geq 3$

$$\sum_{i=1}^{N}\mathbb{E}[\big(\max(0, Z_i)\big)^p] \leq \frac{p!}{2}\max\big(0, \gamma_{\mathbb{G}',\infty}(\mathbb{A}) - \mathbb{P}[\mathbf{x}\in\mathbb{A}]\big)^{p-2}\,N\,\mathrm{Var}(Z).$$

Then Bercu et al. [79, Theorem 2.1] gives w.p.a.l. $1-\delta$

$$\widehat{\mathbb{E}}_{\mathbf{x}}[\mathbb{1}_{\mathbb{A}}] - \mathbb{E}[\mathbb{1}_{\mathbb{A}}(\mathbf{x})] \leq \max\big(0, \gamma_{\mathbb{G}',\infty}(\mathbb{A}) - \mathbb{P}[\mathbf{x}\in\mathbb{A}]\big)\frac{\log\delta^{-1}}{N} + \sqrt{\mathrm{Var}(Z)}\sqrt{2\frac{\log\delta^{-1}}{N}}.$$

Applying the same reasoning to variables $-Z_1, \ldots, -Z_N$ and an union bound gives gives w.p.a.l. $1-2\delta$

$$\Big|\widehat{\mathbb{E}}_{\mathbf{x}}[\mathbb{1}_{\mathbb{A}}] - \mathbb{E}[\mathbb{1}_{\mathbb{A}}(\mathbf{x})]\Big| \leq \big|\gamma_{\mathbb{G}',\infty}(\mathbb{A}) - \mathbb{P}[\mathbf{x}\in\mathbb{A}]\big|\frac{\log\delta^{-1}}{N} + \sqrt{\mathrm{Var}(Z)}\sqrt{2\frac{\log\delta^{-1}}{N}}.\tag{124}$$

Next, we note that when $g\triangleright\mathbb{A}\cap\mathbb{A} = \emptyset$ for all $g\in\mathbb{G}'\setminus\{e\}$, then $\gamma_{\mathbb{G}'}(\mathbb{A}) = 0$ and $\mathbb{P}[\mathbf{x}\in\mathbb{A}] = \mathbb{P}[\mathbf{x}\in\mathbb{G}'\triangleright\mathbb{A}]/|\mathbb{G}'|$. Consequently we get

$$\Upsilon_{\mathbb{G}',\mathbf{x}}(\mathbb{A}) = \frac{\mathbb{P}[\mathbf{x}\in\mathbb{G}'\triangleright\mathbb{A}](1-\mathbb{P}[\mathbf{x}\in\mathbb{G}'\triangleright\mathbb{A}])}{|\mathbb{G}'|} \quad\text{and}\quad \frac{\gamma_{\mathbb{G}',\infty}(\mathbb{A}) - \mathbb{P}[\mathbf{x}\in\mathbb{A}]}{\mathbb{P}[\mathbf{x}\in\mathbb{A}]} = \frac{1}{\mathbb{P}[\mathbf{x}\in\mathbb{G}'\triangleright\mathbb{A}]} - 1.$$

Hence under the additional assumptions, dividing by $\mathbb{E}[\mathbb{1}_{\mathbb{A}}(\mathbf{x})] = \mathbb{P}[\mathbf{x}\in\mathbb{A}]$ gives w.p.a.l. $1-2\delta$

$$\frac{\Big|\widehat{\mathbb{E}}_{\mathbf{x}}[\mathbb{1}_{\mathbb{A}}] - \mathbb{E}[\mathbb{1}_{\mathbb{A}}(\mathbf{x})]\Big|}{\mathbb{E}[\mathbb{1}_{\mathbb{A}}(\mathbf{x})]} \leq \left(\frac{1}{\mathbb{P}[\mathbf{x}\in\mathbb{G}'\triangleright\mathbb{A}]} - 1\right)\frac{\log\delta^{-1}}{N} + \sqrt{2\frac{\log\delta^{-1}}{N}}\sqrt{\frac{1-\mathbb{P}[\mathbf{x}\in\mathbb{G}'\triangleright\mathbb{A}]}{\mathbb{P}[\mathbf{x}\in\mathbb{G}'\triangleright\mathbb{A}]}}.$$

Replacing δ by $\delta/2$ gives w.p.a.l. $1-\delta$

$$\frac{\Big|\widehat{\mathbb{E}}_{\mathbf{x}}[\mathbb{1}_{\mathbb{A}}] - \mathbb{E}[\mathbb{1}_{\mathbb{A}}(\mathbf{x})]\Big|}{\mathbb{E}[\mathbb{1}_{\mathbb{A}}(\mathbf{x})]} \leq \left(\frac{1}{\mathbb{P}[\mathbf{x}\in\mathbb{G}'\triangleright\mathbb{A}]} - 1\right)\frac{\log 2\delta^{-1}}{N} + \sqrt{2\frac{\log 2\delta^{-1}}{N}}\sqrt{\frac{1-\mathbb{P}[\mathbf{x}\in\mathbb{G}'\triangleright\mathbb{A}]}{\mathbb{P}[\mathbf{x}\in\mathbb{G}'\triangleright\mathbb{A}]}}.\tag{125}$$

$\square$

Proposition M.12. *Let A_i, $i \in [N]$ be i.i.d copies of a random variable A in a separable Hilbert space with norm $\|\cdot\|$. If there exist constants $L > 0$ and $\sigma > 0$ such that for every $m \geq 2$, $\mathbb{E}\|A\|^m \leq \frac{1}{2} m! L^{m-2} \sigma^2$, then with probability at least $1-\delta$*

$$\left\| \frac{1}{N} \sum_{i \in [N]} A_i - \mathbb{E}A \right\| \leq \frac{4\sqrt{2}}{\sqrt{N}} \sqrt{\sigma^2 + \frac{L^2}{N}} \log \frac{2}{\delta}. \tag{126}$$

Lemma M.13 ((Sub-Gaussian random variable) Lemma 5.5. in [80])**.** *Let Z be a random variable. Then, the following assertions are equivalent with parameters $K_i > 0$ differing from each other by at most an absolute constant factor.*

1. *Tails: $\mathbb{P}\{|Z| > t\} \leq \exp\big(1 - t^2/K_1^2\big)$ for all $t \geq 0$;*
2. *Moments: $(\mathbb{E}|Z|^p)^{1/p} \leq K_2 \sqrt{p}$ for all $p \geq 1$;*
3. *Super-exponential moment: $\mathbb{E}\exp\big(Z^2/K_3^2\big) \leq 2$.*

A random variable Z satisfying any of the above assertions is called a sub-Gaussian random variable. We will denote by K_3 the sub-Gaussian norm.

Consequently, a sub-Gaussian random variable satisfies the following equivalence of moments property. There exists an absolute constant $c > 0$ such that for any $m \geq 2$,

$$\big(\mathbb{E}|Z|^m\big)^{1/m} \leq cK_3\sqrt{m}\big(\mathbb{E}|Z|^2\big)^{1/2}.$$

Lemma M.14. *Assume that Y is sub-Gaussian with sub-Gaussian norm K. We set $\sigma_\theta^2(Y) := \mathrm{Var}(\|Y - \mathbb{E}[\mathbf{y}]\|)$. Then there exists an absolute constant $C > 0$ such that for any $\delta \in (0,1)$, it holds w.p.a.l. $1-\delta$*

$$\left\| \widehat{\mathbb{E}}_{\mathbf{y}}[\mathbf{y}] - \mathbb{E}[\mathbf{y}] \right\| \leq \frac{C}{\sqrt{N}} \sqrt{\sigma^2(\mathbf{y}) + \frac{K^2}{N}} \log\big(2\delta^{-1}\big).$$

Proof. Set $Z := \|\mathbf{y} - \mathbb{E}\mathbf{y}\|$ and we recall that $\sigma^2(\mathbf{y}) := \mathrm{Var}(\|\mathbf{y} - \mathbb{E}[\mathbf{y}]\|)$. We check that the moment condition,

$$\mathbb{E}Z^m \leq \frac{1}{2} m! L^{m-2} \sigma^2(\mathbf{y})^2, \quad \forall m \geq 2,$$

for some constant $L > 0$ to be specified.

The condition is obviously satisfied for $m = 2$. Next for any $m \geq 3$, the Cauchy-Schwarz inequality and the equivalence of moment property give

$$\mathbb{E}Z^m \leq \left(\mathbb{E}Z^{2(m-2)}\right)^{1/2} \left(\mathbb{E}Z^4\right)^{1/2} \leq 4K_3^2 \sigma_\theta^2(Y)^2 \left(\mathbb{E}Z^{2(m-2)}\right)^{1/2}.$$

Next, by homogeneity, rescaling Z to Z/K_1 we can assume that $K_1 = 1$ in Lemma M.13. We recall that if Z is in addition non-negative random variable, then for every integer $p \geq 1$, we have

$$\mathbb{E}Z^p = \int_0^\infty \mathbb{P}\{Z \geq t\}\, pt^{p-1}\, dt \leq \int_0^\infty e^{1-t^2} pt^{p-1}\, dt = \big(\frac{ep}{2}\big)\Gamma\big(\frac{p}{2}\big).$$

With $p = 2(m-2)$, we get that $\mathbb{E}Z^p \leq e(m-2)\Gamma\big(m-2\big) = e(m-2)! = em!/2$. Using again Lemma M.13, we can take $L = cK$ for some large enough absolute constant $c > 0$. Then Proposition M.12 gives the result.

∎