Efficient Learning and Symmetry Discovery under Exact Invariances

Ashkan Soleymani[†,‡] Behrooz Tahmasebi[†,§] Patrick Jaillet[‡] Stefanie Jegelka[¶]

Abstract

Learning with group invariances is central to many scientific and geometric learning problems, yet its computational foundations remain poorly understood. Even for classical supervised regression settings, it has been unclear whether one can efficiently compute a regression function that is *exactly invariant* to a given group action. Recent work (Soleymani et al., 2025b) showed that exact invariance can be enforced in polynomial time when the underlying group is finite and known, but left open the cases of infinite groups and unknown symmetries. In this paper, we resolve both challenges. First, we present the first polynomial-time algorithm for learning with exact group invariances that applies uniformly to finite and infinite groups. The runtime is polynomial in the data dimension and sample size, and independent of the group, while achieving strong generalization guarantees. This provides a computational explanation for the empirical success of invariant and equivariant methods in geometric machine learning and partially answers a recent open question in the literature (Díaz et al., 2025). Second, we study learning in the *symmetry discovery* setting, where the invariance group is unknown. Focusing on the subgroup lattice of a finite group, we show that exact symmetries can be identified from data and exploited for learning in polynomial time. For regression over finite-dimensional feature spaces, our algorithm provably recovers the underlying symmetry, matches the minimax-optimal sample complexity of the known-symmetry setting, and runs in time polynomial in the data dimension and sample size. Our analysis relies on tools from random Cayley graphs and expander theory, which may be of independent interest.

Keywords: group invariances, symmetry, learning, computational statistics

[†]These authors contributed equally to this work.

[‡]MIT Laboratory for Information and Decision Systems (LIDS). Emails: {ashkanso,jaillet}@mit.edu

[§]Harvard John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA. Email: behrooz_tahmasebi@seas.harvard.edu

[¶]Technical University of Munich (CIT, MCML, MDSI) and MIT Computer Science and Artificial Intelligence Laboratory (CSAIL). Email: stefanie.jegelka@tum.de

Contents

1 Introduction

Invariances have been central to learning theory since the earliest days of machine learning, and their importance has only deepened with modern advances (Hinton, 1987; Kondor, 2008). Models that explicitly respect underlying symmetries consistently deliver remarkable gains in practice, combining efficiency with strong generalization (Bronstein et al., 2017; Weber, 2025). This empirical success has fueled an active line of research, though much of the theory remains focused on classic questions of expressivity, sample complexity, and testing (Elesedy, 2021; Bietti et al., 2021; Behboodi et al., 2022; Tahmasebi and Jegelka, 2023; Mei et al., 2021; Kiani et al., 2024; Soleymani et al., 2025a; Chen et al., 2023; Tahmasebi and Jegelka, 2025a; Díaz et al., 2025; Petrache and Trivedi, 2023; Tahmasebi and Jegelka, 2024). What is still missing is a systematic understanding of the computational price of incorporating invariances, even in fundamental settings such as the classical formulation of supervised regression.

How do we actually build invariances into algorithms? The most direct answer is brute force: expand the dataset through *augmentation* or sum over symmetries via *group averaging*. Unfortunately, both become infeasible once the symmetry group is large, sometimes even growing super-exponentially with input dimension. More "clever" alternatives, like *canonicalization* or *frame averaging*, often replace intractability with discontinuities, poor scalability, or the need for group-specific designs (Dym et al., 2024; Tahmasebi and Jegelka, 2025b).

1.1 Our Contributions

In this paper, we advance the foundational understanding of learning under exact group invariances by establishing the computational tractability of invariant learning beyond the finite-group regime and by developing provably efficient methods for symmetry discovery from data.

Settling the computational complexity for infinite groups. Given the potentially prohibitive size of the symmetry group, one might expect learning with exact invariances to be computationally intractable, even in classical settings such as kernel regression. Surprisingly, recent work (Soleymani et al., 2025b) introduced *spectral averaging*, an exactly invariant algorithm that achieves favorable population risk with runtime $\mathrm{poly}(n, d, \log|G|)$, where n is the number of samples, d is the input dimension, and $|G|$ is the group cardinality. For *finite groups*, this yields an effectively polynomial-time procedure, since $\log|G|$ is typically polynomial in d.

More precisely, spectral averaging decomposes functions into eigenfunction components, estimates the corresponding coefficients from data, and finally applies a projection onto the invariant subspace. Crucially, the projection operator is constructed using only $\mathcal{O}(\log|G|)$ group elements, leveraging the existence of small generating sets for finite groups.

This result naturally raises the question of whether similar guarantees can be achieved for *infinite groups*. In particular, one would hope for algorithms whose runtime is independent of the group cardinality. However, a key ingredient in the analysis of Soleymani et al. (2025b) is that any finite group admits a generating set of size $\mathcal{O}(\log|G|)$, a property that fails dramatically for infinite groups, which may admit arbitrarily large generating sets or no finite generating set at all. As a result, existing techniques do not extend to this setting.

Our first contribution is to resolve this gap by proposing a polynomial-time algorithm for learning with exact group invariances that applies uniformly to *both finite and infinite groups*. The runtime of our algorithm is *independent of the group size*, depending only polynomially on the sample size n and the input dimension d. In this sense, we provide an affirmative answer to the question of whether exact invariance can be enforced efficiently in the infinite-group regime. Our algorithm consists of a randomized preprocessing step that selects a specific small subset S of the group, followed by a deterministic polynomial-time procedure. Despite its group-independent runtime, it achieves generalization guarantees comparable to those of spectral averaging.

Theorem 1.1 (Informal version of Theorem 4.2)**.** *Consider a supervised regression problem in a kernel setting with n i.i.d. labeled samples from a d-dimensional domain, where the target function is invariant under a known group G, which may be finite or infinite. Then, there exists a randomized algorithm with runtime* $\mathrm{poly}(n, d)$ *such that, with high probability, it returns a G-invariant estimator with provable excess population risk guarantees. The algorithm uses randomness only in a data-independent preprocessing step, where it samples a small subset $S \subseteq G$; conditioned on S, the procedure is deterministic.*

The existence of such an algorithm provides a computational explanation for the empirical success of invariant and equivariant methods in geometric machine learning. Moreover, a recent open question in the literature (Díaz et al., 2025) concerns the design of computationally efficient algorithms for finite-dimensional invariant regression over general groups. While prior work proposes scalable solutions for specific group families, the existence of a polynomial-time algorithm applicable to *arbitrary* groups has remained open. This work resolves that question in the affirmative.

Symmetry discovery for learning under invariances. While group invariances arise naturally in many applications, in a range of recent problems, particularly in scientific discovery and the identification of governing laws from data, the underlying symmetry group is not known a priori and must itself be inferred from data. This setting gives rise to the problem of *symmetry discovery*.

In this setting, the learner must simultaneously identify the symmetry group and enforce the corresponding invariance constraints when learning the target function. This task is fundamentally more challenging than learning under a known symmetry, as it requires searching over a potentially

large class of candidate groups. Since finite groups may contain exponentially many subgroups, naive approaches based on enumerating candidate subgroups can be computationally infeasible.

To advance this direction, we study symmetry discovery for learning under exact invariances within the *subgroup lattice* of a given finite group, under a bounded-index assumption. Specifically, we consider a known finite reference group G and assume that the true symmetry group $H \subseteq G$ is an unknown subgroup governing the invariances of the task. Our goal is to recover H by returning a generating set for it and to learn a regression function that is exactly invariant with respect to H while achieving strong generalization guarantees.

Surprisingly, we show that exact symmetries can be identified from data and exploited for learning in polynomial time. For supervised regression problems with finite-dimensional kernels, we prove a stronger result: our algorithm provably recovers the underlying symmetry group, achieves the same minimax-optimal sample complexity as in the known-symmetry setting, and runs in time polynomial in the sample size and data dimension.

Theorem 1.2 (Informal version of Theorem 4.4)**.** *Consider a supervised regression problem in a kernel setting with n i.i.d. labeled samples from a d-dimensional domain, where the target function is invariant under an unknown group H. Assume that H is a subgroup of a known finite group G and satisfies $|H| \geq \kappa|G|$. Then, there exists a randomized algorithm with runtime $\mathrm{poly}\left(n, d, \frac{1}{\kappa}, \log|G|\right)$ such that, with high probability, it returns a generating set of size $\mathcal{O}(\log|H|)$ for H and an exactly H-invariant estimator with provable excess population risk guarantees. The algorithm uses randomness only in a data-independent preprocessing step, where it samples a small subset $S \subseteq G$; conditioned on S, the procedure is deterministic.*

Tools and ideas. Our analysis relies on tools from the theory of random Cayley graphs and expanders (Alon and Roichman, 1994), which we believe may be of independent interest for the theoretical study of learning under group invariances.

For settling the computational complexity of learning under exact invariances for infinite groups, our approach is conceptually simple: we sample a small number of group elements uniformly at random and show that, with high probability, these samples suffice to construct an accurate projection onto the space of invariant functions. This idea is formalized in Algorithm 1.

For the symmetry discovery setting, we again leverage random sampling over the reference group G. We design a rejection-sampling–type procedure that probabilistically "hits" the unknown subgroup H, enabling the recovery of a generating set of size $\mathcal{O}(\log|H|)$. The correctness and efficiency of this procedure follow from expansion properties of random Cayley graphs, which allow us to control both the runtime and the probability of success.

2 Related Work

Invariance has long been recognized as a powerful inductive bias in statistical learning. Incorporating symmetry into models can reduce sample complexity and improve generalization by restricting hypotheses to symmetry-respecting subsets (Hinton, 1987; Poggio and Vetter, 1992; Haussler, 1999; Kondor, 2008; Sokolic et al., 2017). A classical way of achieving this is through invariant kernels, constructed by *group averaging* (Schölkopf and Smola, 2002; Haasdonk and Burkhardt, 2007). Beyond this idea, alternative strategies include *frame averaging* (Puny et al., 2022), *canonicalization* (Kaba et al., 2023; Ma et al., 2024), *random projections* (Dym and Gortler, 2024), and *parameter sharing* (Ravanbakhsh et al., 2017). Each of these methods has distinct strengths, but also drawbacks. In particular, canonicalization and frame averaging can introduce discontinuities or violate

smoothness assumptions, a limitation highlighted in recent work on equivariant frames (Dym et al., 2024).

Beyond kernel methods, symmetries have played a central role in the design of specialized learning architectures. Graph Neural Networks (GNNs) exploit permutation symmetries in graphs (Scarselli et al., 2008; Xu et al., 2019), Convolutional Neural Networks (CNNs) leverage translation invariance in image data (Krizhevsky et al., 2012; Li et al., 2021), and PointNet architectures encode permutation invariance for point clouds (Qi et al., 2017a,b). Symmetry principles have also been integrated into generative models, including permutation-invariant normalizing flows and equivariant flows (Biloš and Günnemann, 2021; Niu et al., 2020; Köhler et al., 2020). For a broad discussion on geometric invariances across modalities, we refer to the survey of Bronstein et al. (2017).

Compared with these approaches, our contribution can be seen as a *post-hoc invariantization* procedure: starting from (spectral) estimates of regression coefficients, we project onto the fixed-point subspaces determined by a (randomized) set of group elements. This yields estimators that are *exactly* invariant, with statistical guarantees matching standard bounds. Prior work on finite groups established this using generating sets of size at most $\log |G|$ (Soleymani et al., 2025b), while our randomized subset selection algorithm removes the dependence on $|G|$ altogether, extending polynomial-time learning with exact invariances to infinite groups (Soleymani et al., 2025b).

Learning under unknown (but existing) symmetries, often referred to as *symmetry discovery*, has recently attracted significant attention. A wide range of methods have been proposed, primarily based on deep learning architectures and Lie-algebraic techniques, including approaches for discovering continuous symmetries (Desai et al., 2022; Yang et al., 2023; Perin and Deny, 2025; Dehmamy et al., 2021; Romero and Lohit, 2022; Ko et al., 2024; Hu et al., 2025b) as well as methods tailored to finite groups (Huh, 2025). Related work explores symmetry identification via gradients of neural network layers (van der Ouderaa et al., 2023), flow-matching techniques for Lie group symmetries (Park et al., 2025), or joint procedures that simultaneously discover and enforce symmetries in learned models (Otto et al., 2025), or enforce them approximately (Tahmasebi and Weber, 2026a). While these approaches demonstrate strong empirical performance, they typically lack rigorous statistical or computational guarantees.

Symmetry discovery has also been studied beyond the data domain, for example in latent representations learned by neural networks (Yang et al., 2024a). In this setting, discovered symmetries may extend beyond affine transformations and relate to underlying manifold structure (Shaw et al., 2024; Bhat et al., 2025). Other approaches rely on data augmentation strategies (Santos-Escriche and Jegelka, 2025) or heuristic procedures (Karjol et al., 2025). We note that symmetry discovery is distinct from the problem of hypothesis testing for the presence of a *given* symmetry in data, which has been studied separately (Soleymani et al., 2025a).

Symmetry discovery is also relevant to dynamical systems arising in scientific applications (Tahmasebi and Weber, 2026b). For instance, Calvo-Barlés et al. (2025a,b) propose methods for identifying finite group invariances in dynamical systems, while approaches for discovering infinitesimal Lie group generators have been developed in (Hu et al., 2025c). Additional related work studies symmetry discovery in latent or continuous-time dynamical models (Li et al., 2025; Shaw et al., 2025; Gabel et al., 2024). For further connections between symmetry discovery and differential equations or partial differential equations, see (Kreider et al., 2025; Hu et al., 2025a; Yang et al., 2025, 2024b).

3 Problem Statement

We begin by formalizing the learning setting and introducing the problem studied in this paper.

Data generation and function space. We consider a well-specified supervised learning problem on a smooth, compact, boundaryless Riemannian manifold $\mathcal{M}$ of dimension d. Given n independent samples $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^n$, the inputs $x_i \in \mathcal{M}$ are drawn uniformly (i.e., with respect to the canonical Riemannian volume measure), and the labels are generated as

$$y_i = f^\star(x_i) + \epsilon_i,$$

where $f^\star \in L^2(\mathcal{M})$ is an unknown continuous regression function and the noise variables ϵ_i are independent, zero mean, and have variance at most σ^2. The uniformity assumption is made for simplicity; our results extend to distributions with bounded densities.

The performance of an estimator $\widehat{f}$ is measured by its population risk

$$\mathcal{R}(\widehat{f}) := \mathbb{E}\Big[\|\widehat{f} - f^\star\|^2_{L^2(\mathcal{M})}\Big],$$

where the expectation is over the randomness of the data.

Group actions and symmetries. Suppose that a known compact group G acts smoothly and isometrically on $\mathcal{M}$, with the action denoted by gx for any $g \in G$ and any $x \in \mathcal{M}$. A function $f^\star$ is said to be G-invariant if

$$f^\star(gx) = f^\star(x) \qquad \text{for all } g \in G,\ x \in \mathcal{M}.$$

In this setting, the goal is to construct an estimator $\widehat{f}$ that is both accurate and exactly invariant under the action of G.

Throughout the paper, we work with a Hilbert feature space $\mathcal{H} \subseteq L^2(\mathcal{M})$ that is closed under the group action, and we assume that this action is isometric (unitary) on $\mathcal{H}$. This is a natural compatibility condition between the feature space and the transformations: in finite dimensions, it can be enforced by choosing an invariant inner product, while in geometric settings it is canonically satisfied by many spectral constructions. In particular, on manifold domains, Laplace–Beltrami eigenspaces and their finite-dimensional truncations provide natural feature spaces that are compatible with both the geometry of $\mathcal{M}$ and Sobolev regularity. For example, when $\mathcal{M}$ is the sphere, these eigenspaces are the spaces of spherical harmonics. We denote by $\mathcal{H}^G$ the subspace of G-invariant functions in $\mathcal{H}$.

We consider two representative regimes. First, in the finite-dimensional regime, $\mathcal{H}$ is any finite-dimensional Hilbert feature space for which the kernel, or equivalently the associated feature map, can be evaluated, for instance through oracle access. In this setting, Empirical Risk Minimization (ERM) over $\mathcal{H}$ attains the minimax-optimal risk $\mathcal{O}(\dim(\mathcal{H})/n)$. Second, in the Sobolev regime, we take $\mathcal{H} = \mathcal{H}^s(\mathcal{M})$ with $s > d/2$. In this case, Kernel Ridge Regression (KRR) achieves the minimax-optimal risk rate $\mathcal{O}(n^{-s/(s+d/2)})$, albeit with a naive computational cost of $\mathcal{O}(n^3)$ when implemented using kernel evaluations (Bach, 2024). These two regimes are chosen for clarity and cover the main examples of interest; the framework itself applies more broadly to kernels and feature spaces compatible with the group action.

Note that even when $f^\star$ is G-invariant, standard learning procedures such as KRR with Sobolev kernels or ERM do not, in general, return invariant estimators (Soleymani et al., 2025b). As a result, additional structure must be imposed to enforce invariance. A classical approach is *group*

averaging, either by averaging the kernel or the learned estimator over the group G. However, such methods require $\mathcal{O}(|G|)$ operations even to evaluate the estimator, rendering them computationally intractable for large or infinite groups.

A recent breakthrough by Soleymani et al. (2025b) introduced *spectral averaging*, which achieves the same minimax-optimal risk rate (that is $\mathcal{O}\big(n^{-s/(s+d/2)}\big)$ in Sobolev spaces and $\mathcal{O}(\dim(\mathcal{H})/n)$ in finite-dimensional settings) while running in $\mathrm{poly}(n, d, \log|G|)$ time. This yields an exponential improvement over naive group averaging for finite groups. However, the dependence on $\log|G|$ fundamentally limits the applicability of spectral averaging to infinite groups, such as continuous rotation groups.

Consequently, any polynomial-time algorithm for learning with exact invariances under infinite groups must have a runtime that is independent of the group cardinality. The central goal of this paper is to address this challenge. In the following sections, we provide an affirmative answer by developing a novel algorithmic framework that enforces exact invariance while maintaining polynomial-time complexity independent of $|G|$.

Compactness and sampling from the group Our assumption that G is compact should be understood as an assumption on the effective transformation group acting on the data domain. Since our results only depend on the induced action on $\mathcal{M}$, one may always quotient out the kernel of the action and work with the corresponding faithful action. In this paper, we assume that this effective group is a compact group acting smoothly on the compact Riemannian manifold $\mathcal{M}$. In particular, this covers the standard setting of compact Lie group actions, for which there is a unique Haar probability measure. This measure provides the canonical notion of uniform sampling from the group G.

We emphasize, however, that sampling from G can be a separate algorithmic problem, depending on how the group is represented. For example, if $\mathcal{M} = S^1$ (the unit circle in two dimensions) and G is a finite subgroup of equally spaced rotations, then even specifying a uniformly random element requires $\Theta(\log|G|)$ random bits. More generally, for groups specified implicitly, such as automorphism groups of graphs, uniform generation can itself be computationally nontrivial. We therefore separate this issue from the statistical and approximation questions studied here, and assume oracle access to independent Haar-uniform samples from G. This assumption is standard in many settings of interest and is often mild; for instance, for several explicit matrix groups, finite cyclic groups, and permutation groups given by suitable generators, uniform or nearly uniform sampling can be implemented efficiently. For example, for permutation groups given by generators, this oracle can be implemented in polynomial time using standard Schreier–Sims methods, which also provide polynomial-time membership testing.

Unknown invariances and symmetry discovery. In many applications arising in scientific discovery, the symmetry group respected by the target function $f^\star \in \mathcal{H}$ is not known a priori and must be inferred from data. To model this setting, we consider a known finite reference group G and assume that the true invariance group is an unknown subgroup $H \subseteq G$. The collection of all subgroups of G forms a *subgroup lattice* under set inclusion.

The goal of the symmetry discovery problem is to use a labeled dataset $\mathcal{S}$ to identify the underlying symmetry structure and to construct an estimator that is exactly invariant with respect to H, while achieving provable generalization guarantees and efficient runtime. Note that a naive brute-force search over all subgroups of G is computationally infeasible, even when $|G|$ is only moderately large, due to the combinatorial size of the subgroup lattice.

To make the problem tractable, we focus on symmetry discovery within the class of *bounded-*

index subgroups. Specifically, we assume that the unknown subgroup H satisfies

$$[G : H] \coloneqq \frac{|G|}{|H|} \leq \frac{1}{\kappa} \iff |H| \geq \kappa |G|,$$

for some parameter $\kappa \in [0, 1]$. The index bound controls the size of the search space within the subgroup lattice: smaller values of κ correspond to broader exploration over candidate subgroups, while larger values restrict attention to subgroups closer to G itself.

4 Main Results

In this section, we present our main results. We begin with the setting of learning under *exact* invariances, where the symmetry group is known, and then turn to the more challenging problem of *symmetry discovery*, in which the invariances must be inferred from data.

4.1 Learning with Exact Invariances in Polynomial Time

We start by considering the case where the underlying symmetry group is known and acts through a given representation. A key algorithmic component underlying our results is a randomized subset selection procedure, which adaptively constructs a small subset of group elements sufficient to characterize the invariant subspace.

To state our theoretical guarantees, we take a closer look at the spectral averaging framework of Soleymani et al. (2025b). Throughout, we consider hypothesis classes $\mathcal{H}$ that are either finite-dimensional, or Sobolev spaces of order $s > \frac{d}{2}$.

The spectral-averaging algorithm of Soleymani et al. (2025b) shows that, for Sobolev spaces, the problem of learning with exact invariances can be reduced to solving an *infinite collection of finite-dimensional convex quadratic programs with linear constraints*. Each program corresponds to an eigenspace of the Laplace–Beltrami operator associated with the data manifold $\mathcal{M}$, and the reduction relies on tools from differential geometry and spectral theory.

To obtain a computationally efficient procedure, Soleymani et al. (2025b) truncate this infinite collection and solve only finitely many quadratic programs. This yields an approximation to the original optimization problem while incurring a negligible approximation error. Concretely, with a slight abuse of notation, let $\mathcal{H}$ denote the subspace of the Sobolev space $\mathcal{H}^s(\mathcal{M})$ spanned by the first $r \coloneqq n^{1/(1+\alpha)} = \mathcal{O}(\sqrt{n})$ eigenfunctions, $\alpha \coloneqq \frac{2s}{d}$. This truncation is canonical and reduces learning in $\mathcal{H}^s(\mathcal{M})$ to a finite-dimensional problem in $\mathcal{H} \subseteq L^2(\mathcal{M})$, with approximation error that vanishes at a polynomial rate in n. Consequently, we can focus on finite-dimensional hypothesis spaces with $r \coloneqq \dim(\mathcal{H})$.

Fix an orthonormal basis $\{\phi_\ell\}_{\ell=1}^r$ of $\mathcal{H}$. Any function $f \in \mathcal{H}$ can then be identified with its coefficient vector $f \in \mathbb{R}^r$. Let $D(g) \in \mathbb{R}^{r \times r}$ denote the matrix representation of the action of $g \in G$ on $\mathcal{H}$, defined by $(gf)(x) \coloneqq f(g^{-1}x), \forall f \in \mathcal{H}$. Under this identification, enforcing invariance of f under the group action amounts to the linear constraints $D(g)f = f$ for all $g \in G$.

The spectral-averaging framework, therefore, leads to the following optimization problem:

$$\min_{f \in \mathbb{R}^r} \quad \sum_{\ell=1}^{r} (f_\ell - \widetilde{f}_\ell)^2 \quad \text{s.t.} \qquad D(g)f = f, \quad \forall g \in S, \quad \widetilde{f}_\ell \coloneqq \frac{1}{n} \sum_{i=1}^{n} y_i \phi_\ell(x_i), \quad \forall \ell \in [r],$$

and $S \subseteq G$ is a *generating set* of the finite group G, meaning that every element of G can be expressed as a finite word over S.

Algorithm 1 Randomized Subset Selection

Input: Query access to $D(g) \in \mathbb{R}^{r\times r}$, $\forall g \in G$; number of iterations $T = \mathcal{O}\big(r^2 + r\log\frac{1}{\delta}\big)$
Output: A subset $S \subseteq G$

Initialization: $S \leftarrow \{\mathrm{id}_G\}$ and $\mathcal{B}_S \leftarrow \{e_i : i \in [r]\} \subseteq \mathbb{R}^r$, where e_i denotes the i-th standard basis vector.
for $t = 1, \ldots, T$ **do**
  Sample $g \sim \mathrm{Unif}(G)$
  Sample $x = \sum_{v\in\mathcal{B}_S} N_v v$, where $N_v \sim \mathcal{N}\Big(0, \frac{1}{|\mathcal{B}_S|}\Big)$ independently for all $v \in \mathcal{B}_S$
  if $\|(D(g) - I_r)x\|_2^2 > 0$ **then**
    $S \leftarrow S \cup \{g\}$
    $\mathcal{B}_S \leftarrow$ an orthonormal basis of $\mathrm{span}(\mathcal{B}_S) \cap \ker(I_r - D(g))$
  end if
end for
return S

This deterministic procedure yields an estimator achieving population risk $\mathcal{O}\big(n^{-s/(s+d/2)}\big)$ for Sobolev spaces, or $\mathcal{O}(r/n)$ in finite-dimensional settings, and can be implemented in $\mathrm{poly}(n, d, \log|G|)$ time for finite groups.

The $\mathrm{poly}(\log|G|)$ dependence on the group size arises from the number of linear constraints imposed by the invariance conditions. Specifically, one constraint is required for each generator in S. Since any finite group admits a generating set of size at most $\log_2|G|$, this accounts for the polylogarithmic dependence of the computational complexity on $|G|$. In contrast, for infinite groups, the cardinality $|G|$ is unbounded, and thus such a method fails to reduce the number of constraints.

In Algorithm 1, we introduce a new randomized procedure that, with high probability, identifies a subset $S \subseteq G$ imposing a sufficient collection of linear constraints in the quadratic programs associated with each eigenspace of the spectral-averaging framework. Crucially, this procedure applies to both finite and infinite groups.

Specifically, the algorithm returns a subset S of size $\mathcal{O}\big(r^2\log\frac{1}{\delta}\big)$ with probability at least $1-\delta$, as formalized in Proposition 4.1. The runtime and the size of the subset are entirely independent of the group cardinality $|G|$, and therefore remain valid even when G is infinite. The key insight is that random group elements are sufficient, with high probability, to enforce all necessary invariance constraints. This stands in sharp contrast to approaches based on generating sets, which either do not extend to infinite groups or require preprocessing the entire group, rendering them computationally infeasible.

Proposition 4.1. *For each $g \in G$, define $V_g := \ker(I_r - D(g))$. If $T = \mathcal{O}\big(r^2 + r\log\frac{1}{\delta}\big)$, then with probability at least $1-\delta$, Algorithm 1 returns a subset $S \subseteq G$ with $|S| \le r$ such that*

$$\bigcap_{g\in S} V_g \;=\; \bigcap_{g\in G} V_g.$$

With the randomized subset selection procedure in place, we are now ready to state our main algorithmic and statistical guarantee. The resulting method builds upon the spectral-averaging framework of Soleymani et al. (2025b), while replacing generating sets with random subsets.

Theorem 4.2. *Let G be a group, and let $\{(x_i, y_i)\}_{i=1}^n$ be a labeled dataset sampled from a d-dimensional manifold $\mathcal{M}$. Suppose the target regression function satisfies $f^\star \in \mathcal{H}^s(\mathcal{M})$ for some $s > d/2$, and define $\alpha := 2s/d$ (Sobolev regression), or alternatively that $\mathcal{H}$ is finite-dimensional.*

Then, Algorithm 2, achieved via spectral averaging with the randomized subset selection of Algorithm 1, running in $\mathrm{poly}(n, d, \log(1/\delta))$ *time, returns, with probability at least* $1-\delta$*, an exactly* G*-invariant estimator* $\widehat{f}$*. The estimator achieves excess population risk* $\mathcal{R}(\widehat{f}) = \mathcal{O}\big(n^{-s/(s+d/2)}\big)$ *for Sobolev regression, or* $\mathcal{O}(\dim(\mathcal{H}^G)/n)$ *in finite-dimensional settings. Thus, in the finite-dimensional case, this rate is minimax optimal over the class of* G*-invariant functions in* $\mathcal{H}$*.*

The above theorem establishes that exact invariance can be enforced in polynomial time without prior knowledge of the group structure, while simultaneously achieving statistically optimal sample complexity in finite-dimensional settings.

Proof sketch for Proposition 4.1 A central technical contribution of the paper is Proposition 4.1, which enables learning with exact invariances for arbitrary groups. The proposition shows that by randomly sampling group elements, one can construct a subset $S \subseteq G$, whose size is independent of $|G|$, such that

$$\bigcap_{g\in S} V_g \;=\; \bigcap_{g\in G} V_g.$$

At first glance, this may seem surprising. The space $\bigcap_{g\in G} V_g$ corresponds exactly to the subspace of invariant functions, whereas $\bigcap_{g\in S} V_g$ consists of functions invariant under the subgroup generated by S. While any generating set S of G trivially satisfies this equality, a key observation is that generation of the group is not necessary: since we only observe the action of G through its representation on $\mathbb{R}^r$, distinct group elements may induce redundant linear constraints.

The proof exploits this representation-theoretic viewpoint. Starting from the full space $\mathbb{R}^r$, each sampled group element g induces a constraint $D(g)f = f$, shrinking the feasible subspace to $V_g = \ker(I_r - D(g))$. Whenever the intersection $\bigcap_{g\in S} V_g$ has dimension strictly larger than that of $\bigcap_{g\in G} V_g$, a randomly sampled group element has a nontrivial probability of further reducing this dimension. Algorithm 1 implements this idea by iteratively sampling group elements and updating the intersection whenever a new constraint is informative.

Since the dimension of the invariant subspace is at most r, and each successful iteration reduces the dimension by at least one, after at most r effective reductions, occurring with high probability, the algorithm identifies a subset S for which

$$\bigcap_{g\in S} V_g = \bigcap_{g\in G} V_g.$$

Importantly, the total number of iterations required depends only on r, and not on the size or structure of the group G. This establishes the main technical ingredient behind our first contribution.

Remark 4.3. When G is the trivial group, Algorithm 2 reduces to the standard projection estimator in the chosen feature space. In particular, under the uniform distribution and an orthonormal feature basis, the population feature covariance is the identity, so the estimator can be viewed as the ordinary least-squares/projection estimator with the covariance plug-in replaced by its population value. Thus, in this special case, the algorithm agrees with ordinary linear estimation in the feature space up to the usual distinction between empirical OLS and projection using the identity covariance.

Algorithm 2 Spectral Averaging with Randomized Subset Selection

Input: Dataset $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^n$ and $\alpha = 2s/d > 1$ for Sobolev regression

Output: G-invariant estimator $\widehat{f}$

1: Set $r \leftarrow n^{1/(1+\alpha)}$ for Sobolev regression; otherwise set $r \leftarrow \dim(\mathcal{H})$

2: **for** each $\ell \in [r]$ **do**

3: $\quad \widetilde{f}_\ell \leftarrow \frac{1}{n}\sum_{i=1}^n y_i \phi_\ell(x_i)$

4: **end for**

5: $S \leftarrow$ Randomized Subset Selection for G and $\mathcal{H}$ (Algorithm 1)

6: Solve the linearly constrained quadratic program:

$$\widehat{f} \leftarrow \arg\min_{f \in \mathbb{R}^r} \sum_{\ell=1}^r (f_\ell - \widetilde{f}_\ell)^2, \qquad \text{s.t.} \quad D(g)f = f \ \ \forall g \in S$$

7: **Return** $\widehat{f}(x) = \sum_{\ell=1}^r \widehat{f}_\ell \phi_\ell(x)$

4.2 Sobolev Rates and Computational–Statistical Trade-offs

In the Sobolev setting, our guarantees should be interpreted relative to the computational cost of enforcing invariance. When G is trivial, our rate reduces to the classical minimax-optimal Sobolev regression rate. For nontrivial group actions, sharper invariant rates are known information-theoretically (Tahmasebi and Jegelka, 2023, Theorem 4.1); however, achieving these rates relies on a group-adapted regularization scheme (Tahmasebi and Jegelka, 2023, Eq. 122) and, in its kernel implementation, on averaging over the entire group. In our setting, this corresponds to a potentially large spectral truncation and a computational procedure that is infeasible when full group averaging is expensive.

Our contribution is different and complementary: we give a polynomial-time procedure that enforces exact invariance while attaining the full convergence rate of the corresponding no-invariance Sobolev problem. Thus, even though our Sobolev guarantee is not claimed to be the sharp invariant minimax rate, it achieves a statistically meaningful rate under exact invariance without requiring full group averaging. Determining the optimal computational–statistical trade-off for Sobolev spaces with invariances remains open. Resolving this question likely requires a finer understanding of how the group action interacts with low-frequency Laplace–Beltrami eigenspaces. In particular, it is currently unclear whether the sharp invariant minimax rate can be achieved in polynomial time without full group averaging, or whether an inherent computational gap is present. In contrast, in the finite-dimensional setting, our method achieves the optimal sample complexity of learning under invariances in polynomial time.

4.3 Symmetry Discovery for Learning with Exact Invariances

We now extend the previous setting to the case where the invariance group is *unknown*. Specifically, we assume that there exists an unknown subgroup $H \subseteq G$ such that the target function $f^\star \in \mathcal{H}$ is invariant under H and under no larger subgroup. Here, G denotes a known *reference finite group*.

More precisely, to ensure that information about the subgroup H is sufficiently present in the target function, we assume that $f^\star$ is drawn *isotropically* from the subspace of H-invariant functions within $\mathcal{H}$. In the Sobolev setting, isotropy is defined with respect to the truncated function class of dimension $r := \mathcal{O}(\sqrt{n})$. This assumption rules out pathological target functions for which the underlying symmetry would be statistically undetectable. Moreover, to avoid identifiability issues,

we identify any two subgroups that induce the same invariant function class.

A key observation is that, once a generating set $S \subseteq G$ for H is identified, learning reduces immediately to the exact-invariance setting studied in the previous section by simply running Algorithm 2 with this set of constraints. The central challenge in symmetry discovery is therefore to identify a generating set for H using only labeled data.

To make this problem tractable, we assume that H has bounded index in G, namely that

$$|H| \geq \kappa |G| \quad \text{for some } \kappa \in (0, 1].$$

Under this assumption, a uniformly random element $g \sim \mathrm{Unif}(G)$ lies in H with probability κ. Consequently, by sampling

$$N = \Omega\left(\frac{\log |G| + \log(1/\delta)}{\kappa}\right)$$

independent elements from G, we obtain, with probability at least $1 - \delta$, a subset of size $\mathcal{O}(\log |H|)$ contained in H. Classical results on random Cayley graphs (Alon and Roichman, 1994) imply that such a random subset generates H with high probability.

The remaining question is therefore algorithmic: given a fixed $g \in G$, can we test from data whether $f^\star$ is invariant under g, i.e., whether $f^\star(gx) = f^\star(x)$ holds almost surely?

Our approach answers this question via a data-driven hypothesis test. For each candidate element $g \in G$, we compare two estimators obtained via spectral averaging: one unconstrained estimator, and one estimator constrained to be invariant under g. Using a held-out test set, we evaluate whether imposing the constraint $D(g)f = f$ reduces the empirical test error. Elements for which this constraint leads to improved generalization performance are retained. Collecting such elements yields a candidate generating set for H.

This procedure leads to the following guarantee for symmetry discovery.

Theorem 4.4. *Let G be a finite reference group, and consider the family of subgroups $H \leq G$ satisfying $|H| \geq \kappa |G|$ for some $\kappa \in (0, 1]$. Let $\{(x_i, y_i)\}_{i=1}^n$ be a labeled dataset sampled from a d-dimensional manifold $\mathcal{M}$. Assume the unknown H-invariant target function satisfies $f^\star \in \mathcal{H}^s(\mathcal{M})$ for some $s > d/2$, and define $\alpha := 2s/d$ (Sobolev regression), or alternatively that $\mathcal{H}$ is finite-dimensional. Moreover, $f^\star \in \mathcal{H}$ is drawn* isotropically, *according to the definition above.*
Then Algorithm 3 returns, with probability at least $1 - \delta$, an exactly H-invariant estimator $\widehat{f}$ together with a generating set for H, in poly $\left(n, d, \frac{1}{\kappa}, \log |G|, \log \frac{1}{\delta}\right)$ *time. The estimator achieves excess population risk $\mathcal{R}(\widehat{f}) = \mathcal{O}\big(n^{-s/(s+d/2)}\big)$ for Sobolev regression, or $\mathcal{O}(\dim(\mathcal{H}^G)/n)$ in finite-dimensional settings. Thus, in the finite-dimensional case, it achieves the minimax optimal rate over the class of H-invariant functions in $\mathcal{H}$.*

Theorem 4.4 shows that even when the symmetry group is unknown, exact invariances can be discovered and enforced in polynomial time while retaining statistically optimal rates. This establishes a sharp statistical–computational trade-off for learning under exact invariances and provides theoretical justification for the empirical success of symmetry discovery methods.

Scope of the results. Our symmetry-discovery guarantee is stated for finite groups. The present approach relies on finite-group tools, including random Cayley graph constructions, and therefore does not directly extend to infinite groups. An extension to compact Lie groups would likely require different techniques, possibly exploiting the local and representation-theoretic structure of Lie groups; we leave this direction for future work.

Algorithm 3 Symmetry Discovery via Spectral Averaging

Input: Training set $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^n$; test set $\mathcal{S}' = \{(x_i', y_i')\}_{i=1}^n$; parameter $\alpha = 2s/d > 1$ for Sobolev regression; number of trials T; threshold η

Output: H-invariant estimator $\widehat{f}$ and a generating set S for H

Set $r \leftarrow n^{1/(1+\alpha)}$ for Sobolev regression; otherwise set $r \leftarrow \dim(\mathcal{H})$
for each $\ell \in [r]$ **do**
  $\widetilde{f}_\ell \leftarrow \frac{1}{n}\sum_{i=1}^n y_i \phi_\ell(x_i)$
end for
Compute the unconstrained estimator: $\widehat{f}_\emptyset \leftarrow \widetilde{f} \in \mathbb{R}^r$
Initialize $S \leftarrow \emptyset$
for $t = 1, \ldots, T$ **do**
  Sample $g \sim \mathrm{Unif}(G)$
  Compute the g-invariant estimator: $\widehat{f}_g \leftarrow \arg\min_{f \in \mathbb{R}^r} \sum_{\ell=1}^r (f_\ell - \widetilde{f}_\ell)^2 \quad \text{s.t.} \quad D(g)f = f$
  If $\frac{1}{n}\sum_{i=1}^n \big(\widehat{f}_g(x_i') - y_i'\big)^2 \le \frac{1}{n}\sum_{i=1}^n \big(\widehat{f}_\emptyset(x_i') - y_i'\big)^2 + \eta$ **then** $S \leftarrow S \cup \{g\}$.
end for
Compute the final estimator: $\widehat{f} \leftarrow \arg\min_{f \in \mathbb{R}^r} \sum_{\ell=1}^r (f_\ell - \widetilde{f}_\ell)^2 \quad \text{s.t.} \quad D(g)f = f \ \ \forall g \in S$
Return $\widehat{f}(x) = \sum_{\ell=1}^r \widehat{f}_\ell \phi_\ell(x)$

The isotropy assumption on $f^\star$ is also necessary for discovery. Without a separation condition, the target function may be invariant, or nearly invariant, under several distinct subgroups, making it impossible to identify a unique underlying group from data. Our assumption rules out this ambiguity and ensures that the relevant group is statistically identifiable.

Importantly, this assumption is needed only for symmetry discovery, not for invariant regression itself. Once a candidate subgroup H is specified, even if several subgroups are compatible with the data, Algorithm 2 can be applied to enforce H-invariance and obtain the convergence guarantee of Theorem 4.2 for that subgroup.

5 Conclusion

We presented the first *randomized* polynomial-time algorithm for learning with *exact invariances* under general group actions, accommodating both finite and infinite groups. This result represents a substantial step toward understanding the computational complexity of learning with symmetries, and clarifies the algorithmic boundary between tractable and intractable instances of invariant learning. For finite groups, prior work showed that spectral averaging combined with group generators yields a deterministic polynomial-time algorithm (Soleymani et al., 2025b). In contrast, for infinite groups no such deterministic procedure is currently known. Our results demonstrate that randomization suffices to overcome this barrier: spectral averaging with randomized subset selection enforces exact invariance in polynomial time without any dependence on the group cardinality. Whether a fully *deterministic* polynomial-time algorithm exists for learning with exact invariances under infinite groups remains an intriguing open question.

We further addressed the problem of *symmetry discovery*, where the invariance group is unknown. Under a bounded-index assumption, we showed that a simple sampling strategy, grounded in the theory of random Cayley graphs, enables efficient identification of the underlying symmetry group directly from data. Combining symmetry discovery with invariant learning yields estimators that are both computationally efficient and statistically optimal.

Taken together, our results provide the first rigorous statistical–computational guarantees for learning with exact invariances and for discovering unknown symmetries, offering a theoretical foundation for the growing empirical success of symmetry-aware learning methods.

Acknowledgments

AS and PJ were partially supported by the ONR grant N00014-24-1-2470. BT and SJ were partially supported by the Office of Naval Research under grant N00014-20-1-2023 (MURI ML-SCOPE) and by the National Science Foundation under award CCF-2112665 (TILOS AI Institute). SJ also acknowledges support from an Alexander von Humboldt Professorship.

References

Noga Alon and Yuval Roichman. Random cayley graphs and expanders. *Random Structures & Algorithms*, 5(2):271–284, 1994. 4, 12, 27

Francis Bach. *Learning theory from first principles*. MIT press, 2024. 6

Arash Behboodi, Gabriele Cesa, and Taco S Cohen. A pac-bayesian generalization bound for equivariant networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2

Manu Bhat, Jonghyun Park, Jianke Yang, Nima Dehmamy, Robin Walters, and Rose Yu. Atlasd: Automatic local symmetry discovery. In *Int. Conference on Machine Learning (ICML)*, 2025. 5

Alberto Bietti, Luca Venturi, and Joan Bruna. On the sample complexity of learning under geometric stability. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2

Marin Biloš and Stephan Günnemann. Scalable normalizing flows for permutation invariant densities. In *Int. Conference on Machine Learning (ICML)*, 2021. 5

Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4): 18–42, 2017. 2, 5

Pablo Calvo-Barlés, Sergio G Rodrigo, and Luis Martín-Moreno. Learning finite symmetry groups of dynamical systems via equivariance detection. *arXiv preprint arXiv:2503.03014*, 2025a. 5

Pablo Calvo-Barlés, Sergio G Rodrigo, and Luis Martín-Moreno. Machine learning for detection of equivariant finite symmetry groups in dynamical systems. *Machine Learning: Science and Technology*, 6(2):025058, 2025b. 5

Ziyu Chen, Markos Katsoulakis, Luc Rey-Bellet, and Wei Zhu. Sample complexity of probability divergences under group symmetry. In *Int. Conference on Machine Learning (ICML)*, 2023. 2

Nima Dehmamy, Robin Walters, Yanchen Liu, Dashun Wang, and Rose Yu. Automatic symmetry discovery with lie algebra convolutional network. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 5

Krish Desai, Benjamin Nachman, and Jesse Thaler. Symmetry discovery with deep learning. *Physical Review D*, 105(9):096031, 2022. 5

Mateo Díaz, Dmitriy Drusvyatskiy, Jack Kendrick, and Rekha R Thomas. Invariant kernels: Rank stabilization and generalization across dimensions. *arXiv preprint arXiv:2502.01886*, 2025. 1, 2, 3

Nadav Dym and Steven J Gortler. Low-dimensional invariant embeddings for universal geometric learning. *Foundations of Computational Mathematics*, pages 1–41, 2024. 4

Nadav Dym, Hannah Lawrence, and Jonathan W. Siegel. Equivariant frames and the impossibility of continuous canonicalization. In *Int. Conference on Machine Learning (ICML)*, 2024. 2, 5

Bryn Elesedy. Provably strict generalisation benefit for invariance in kernel methods. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2

Alex Gabel, Rick Quax, and Efstratios Gavves. Data-driven lie point symmetry detection for continuous dynamical systems. *Machine Learning: Science and Technology*, 5(1):015037, 2024. 5

Bernard Haasdonk and Hans Burkhardt. Invariant kernel functions for pattern analysis and machine learning. *Machine learning*, 68(1):35–61, 2007. 4

David Haussler. Convolution kernels on discrete structures. Technical report, Department of Computer Science, University of California at Santa Cruz, 1999. 4

Geoffrey E Hinton. Learning translation invariant recognition in a massively parallel networks. In *International conference on parallel architectures and languages Europe*, pages 1–13. Springer, 1987. 2, 4

Lexiang Hu, Yikang Li, and Zhouchen Lin. Governing equation discovery from data based on differential invariants. *arXiv preprint arXiv:2505.18798*, 2025a. 5

Lexiang Hu, Yikang Li, and Zhouchen Lin. Symmetry discovery for different data types. *Neural Networks*, page 107481, 2025b. 5

Lexiang Hu, Yikang Li, and Zhouchen Lin. Explicit discovery of nonlinear symmetries from dynamic data. In *Int. Conference on Machine Learning (ICML)*, 2025c. 5

Dongsung Huh. Discovering group structures via unitary representation learning. In *Int. Conference on Learning Representations (ICLR)*, 2025. 5

Sékou-Oumar Kaba, Arnab Kumar Mondal, Yan Zhang, Yoshua Bengio, and Siamak Ravanbakhsh. Equivariance with learned canonicalization functions. In *Int. Conference on Machine Learning (ICML)*, 2023. 4

Pavan Karjol, Vivek V Kashyap, Rohan Kashyap, et al. Learning equivariant functions via quadratic forms. *arXiv preprint arXiv:2509.22184*, 2025. 5

Bobak Kiani, Thien Le, Hannah Lawrence, Stefanie Jegelka, and Melanie Weber. On the hardness of learning under symmetries. In *Int. Conference on Learning Representations (ICLR)*, 2024. 2

Gyeonghoon Ko, Hyunsu Kim, and Juho Lee. Learning infinitesimal generators of continuous symmetries from data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 5

Jonas Köhler, Leon Klein, and Frank Noé. Equivariant flows: exact likelihood generative learning for symmetric densities. In *Int. Conference on Machine Learning (ICML)*, 2020. 5

Imre Risi Kondor. *Group theoretical methods in machine learning*. Columbia University, 2008. 2, 4

Max Kreider, John Harlim, and Daning Huang. A model-free method for discovering symmetry in differential equations. *arXiv preprint arXiv:2511.09779*, 2025. 5

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2012. 5

Haoran Li, Chenhan Xiao, Muhao Guo, and Yang Weng. Latent mixture of symmetries for sample-efficient dynamic learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. 5

Zewen Li, Fan Liu, Wenjie Yang, Shouheng Peng, and Jun Zhou. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems*, 33(12):6999–7019, 2021. 5

George Ma, Yifei Wang, Derek Lim, Stefanie Jegelka, and Yisen Wang. A canonization perspective on invariant and equivariant learning. *arXiv preprint arXiv:2405.18378*, 2024. 4

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Learning with invariances in random features and kernel models. In *Conference on Learning Theory (COLT)*, 2021. 2

Chenhao Niu, Yang Song, Jiaming Song, Shengjia Zhao, Aditya Grover, and Stefano Ermon. Permutation invariant graph generation via score-based generative modeling. In *Int. Conference on Machine Learning (ICML)*, 2020. 5

Samuel E Otto, Nicholas Zolman, J Nathan Kutz, and Steven L Brunton. A unified framework to enforce, discover, and promote symmetry in machine learning. *Journal of Machine Learning Research*, 26(248):1–83, 2025. 5

Jung Yeon Park, Yuxuan Chen, Floor Eijkelboom, Jan-Willem van de Meent, Lawson LS Wong, and Robin Walters. Discovering lie groups with flow matching. *arXiv preprint arXiv:2512.20043*, 2025. 5

Andrea Perin and Stephane Deny. On the ability of deep networks to learn symmetries from data: A neural kernel theory. *Journal of Machine Learning Research*, 26(145):1–70, 2025. 5

Mircea Petrache and Shubhendu Trivedi. Approximation-generalization trade-offs under (approximate) group equivariance. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2

Tomaso Poggio and Thomas Vetter. Recognition and structure from one 2d model view: Observations on prototypes, object classes and symmetries. Technical report, 1992. 4

Omri Puny, Matan Atzmon, Heli Ben-Hamu, Ishan Misra, Aditya Grover, Edward J Smith, and Yaron Lipman. Frame averaging for invariant and equivariant network design. In *Int. Conference on Learning Representations (ICLR)*, 2022. 4

Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017a. 5

Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017b. 5

Siamak Ravanbakhsh, Jeff Schneider, and Barnabas Poczos. Equivariance through parameter-sharing. In *Int. Conference on Machine Learning (ICML)*, 2017. 4

David W Romero and Suhas Lohit. Learning partial equivariances from data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 5

Eduardo Santos-Escriche and Stefanie Jegelka. Learning equivariant models by discovering symmetries with learnable augmentations. *arXiv preprint arXiv:2506.03914*, 2025. 5

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008. 5

Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002. 4

Ben Shaw, Abram Magner, and Kevin Moon. Symmetry discovery beyond affine transformations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 5

Ben Shaw, Sasidhar Kunapuli, Abram Magner, and Kevin R. Moon. Continuous symmetry discovery and enforcement using infinitesimal generators of multi-parameter group actions. *arXiv preprint arXiv:2505.08219*, 2025. 5

Jure Sokolic, Raja Giryes, Guillermo Sapiro, and Miguel Rodrigues. Generalization error of invariant classifiers. In *Artificial Intelligence and Statistics*, pages 1094–1103. PMLR, 2017. 4

Ashkan Soleymani, Behrooz Tahmasebi, Stefanie Jegelka, and Patrick Jaillet. A robust kernel statistical test of invariance: Detecting subtle asymmetries. In *Int. Conference on Artificial Intelligence and Statistics (AISTATS)*, 2025a. 2, 5

Ashkan Soleymani, Behrooz Tahmasebi, Stefanie Jegelka, and Patrick Jaillet. Learning with exact invariances in polynomial time. In *Forty-second International Conference on Machine Learning*, 2025b. 1, 3, 5, 6, 7, 8, 9, 13, 21, 25, 27

Behrooz Tahmasebi and Stefanie Jegelka. The exact sample complexity gain from invariances for kernel regression. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2, 11

Behrooz Tahmasebi and Stefanie Jegelka. Sample complexity bounds for estimating probability divergences under invariances. In *Int. Conference on Machine Learning (ICML)*, 2024. 2

Behrooz Tahmasebi and Stefanie Jegelka. Generalization bounds for canonicalization: A comparative study with group averaging. In *The Thirteenth International Conference on Learning Representations*, 2025a. 2

Behrooz Tahmasebi and Stefanie Jegelka. Regularity in canonicalized models: A theoretical perspective. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025b. 2

Behrooz Tahmasebi and Melanie Weber. Achieving approximate symmetry is exponentially easier than exact symmetry. In *Int. Conference on Learning Representations (ICLR)*, 2026a. 5

Behrooz Tahmasebi and Melanie Weber. Adaptive symmetry discovery for dynamical system identification. In *ICLR 2026 Workshop on Geometry-grounded Representation Learning and Generative Modeling*, 2026b. 5

Tycho van der Ouderaa, Alexander Immer, and Mark van der Wilk. Learning layer-wise equivariances automatically using gradients. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 5

Melanie Weber. Geometric machine learning. *Wiley Online Library*, 2025. 2

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *Int. Conference on Learning Representations (ICLR)*, 2019. 5

Jianke Yang, Robin Walters, Nima Dehmamy, and Rose Yu. Generative adversarial symmetry discovery. In *Int. Conference on Machine Learning (ICML)*, 2023. 5

Jianke Yang, Nima Dehmamy, Robin Walters, and Rose Yu. Latent space symmetry discovery. In *Int. Conference on Machine Learning (ICML)*, 2024a. 5

Jianke Yang, Wang Rao, Nima Dehmamy, Robin Walters, and Rose Yu. Symmetry-informed governing equation discovery. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024b. 5

Jianke Yang, Manu Bhat, Bryan Hu, Yadi Cao, Nima Dehmamy, Robin Walters, and Rose Yu. Discovering symbolic differential equations with symmetry invariants. *arXiv preprint arXiv:2505.12083*, 2025. 5

A Preliminaries on Groups and Manifolds

We collect here the basic notions from group actions and Riemannian geometry used throughout the paper. Our data domain is a smooth, compact, boundaryless manifold $\mathcal{M}$. This assumption is made for clarity of presentation and to avoid technicalities related to boundary conditions, noncompactness, or singularities. The same ideas extend to other sufficiently regular data domains, provided that the corresponding function spaces, measures, and group actions are well defined.

Groups. A group is a set G equipped with a binary operation $(g, h) \mapsto gh$ satisfying the following three axioms: (I) associativity, $(gh)k = g(hk)$ for all $g, h, k \in G$; (II) existence of an identity element $e \in G$ such that $eg = ge = g$ for all $g \in G$; and (III) existence of inverses, meaning that for every $g \in G$ there exists $g^{-1} \in G$ with $gg^{-1} = g^{-1}g = e$. Standard examples include finite cyclic groups, permutation groups such as the symmetric group S_d, finite groups of rotations, matrix groups such as the orthogonal group $O(d)$ and the special orthogonal group $SO(d)$, and compact Lie groups acting by smooth transformations.

Group actions on manifolds. A left action of a group G on $\mathcal{M}$ is a map

$$G \times \mathcal{M} \to \mathcal{M}, \qquad (g, x) \mapsto gx,$$

such that $ex = x$ and $(gh)x = g(hx)$ for all $g, h \in G$ and $x \in \mathcal{M}$. Throughout the paper, we assume that the action is smooth, meaning that each map $x \mapsto gx$ is a smooth diffeomorphism of $\mathcal{M}$, and, when G is a Lie group, the joint map $(g, x) \mapsto gx$ is smooth.

The kernel of the action is

$$\ker(G \curvearrowright \mathcal{M}) := \{g \in G : \ gx = x \text{ for all } x \in \mathcal{M}\}.$$

Elements in the kernel act trivially on all data points and hence have no statistical or computational effect in our setting. Therefore, without loss of generality, one may replace G by the effective quotient

$$G_{\text{eff}} := G/\ker(G \curvearrowright \mathcal{M}),$$

which acts faithfully on $\mathcal{M}$. Thus, throughout the paper, we identify G with its effective action and assume that the action is faithful (i.e., no non-identity element acts trivially on all of $\mathcal{M}$).

Compactness and Haar measure. Our results are stated for compact groups G. More precisely, after quotienting out the kernel of the action, we assume that the effective transformation group acting on $\mathcal{M}$ is compact. This assumption covers the main examples of interest, including finite groups, compact matrix groups, and compact Lie group actions. Compactness gives a canonical probability measure on the group: the normalized Haar measure, which is the unique probability measure invariant under left and right multiplication. This measure is the natural notion of the uniform distribution on G and is used whenever we sample random group elements.

Group representations. A representation of a group G on a finite-dimensional vector space V is a map $D : G \to \mathrm{GL}(V)$ such that

$$D(gh) = D(g)D(h), \qquad D(e) = I.$$

Thus, a representation realizes each abstract group element $g \in G$ as an invertible linear operator $D(g)$ acting on V, in a way that preserves the group law. If V is equipped with an inner product and each $D(g)$ preserves this inner product, then the representation is called orthogonal over $\mathbb{R}$ and unitary over $\mathbb{C}$; equivalently,

$$D(g)^\top D(g) = I \quad \text{or} \quad D(g)^* D(g) = I, \qquad \forall g \in G.$$

In this paper, such representations arise by restricting the action of G on functions to a finite-dimensional feature space $\mathcal{H}$: after choosing a basis of $\mathcal{H}$, the operator $T_g f(x) = f(g^{-1}x)$ is represented by a matrix $D(g)$, and the identity $T_{gh} = T_g T_h$ becomes $D(gh) = D(g)D(h)$.

Invariant Riemannian metrics and isometric actions. The compactness assumption also allows us to regard the action as isometric without loss of generality. Indeed, starting from any Riemannian metric $\mathfrak{g}_0$ on $\mathcal{M}$, we may average it over the group:

$$\mathfrak{g}_x(u, v) := \int_G \mathfrak{g}_{0,g^{-1}x}\big(Dg_x^{-1}u, Dg_x^{-1}v\big)\, d\mu_G(g).$$

The resulting metric $\mathfrak{g}$ is smooth, positive definite, and G-invariant. Consequently, every transformation $x \mapsto gx$ is an isometry of $(\mathcal{M}, \mathfrak{g})$. The corresponding Riemannian volume measure is therefore also G-invariant. In this sense, assuming that the group acts isometrically with respect to the canonical volume measure is not restrictive for compact group actions; it can always be enforced by choosing an invariant Riemannian metric.

Once such a metric is fixed, the isometry group

$$\mathrm{Iso}(\mathcal{M})$$

is the group of all diffeomorphisms of $\mathcal{M}$ preserving the metric. By the Myers–Steenrod theorem, $\mathrm{Iso}(\mathcal{M})$ is a Lie group, and for compact $\mathcal{M}$ it is compact. Thus, after choosing a G-invariant metric, the group G may be viewed as a compact subgroup of $\mathrm{Iso}(\mathcal{M})$. This provides the geometric setting used throughout the paper: a compact group acting smoothly, faithfully, and isometrically on a compact Riemannian manifold, equipped with the corresponding invariant volume measure.

Setting and notation for Sobolev-space regression. Let $(\mathcal{M}, \mathfrak{g})$ be a smooth, compact, connected, boundaryless d-dimensional Riemannian manifold, and let G act smoothly and isometrically on $\mathcal{M}$, so that $x \mapsto gx$ is an isometry for every $g \in G$. We observe i.i.d. samples $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^n$, where x_i is drawn uniformly from the Riemannian volume measure on $\mathcal{M}$ and

$$y_i = f^\star(x_i) + \varepsilon_i.$$

Here $f^\star \in \mathcal{H}^s(\mathcal{M})$ for some $s > d/2$, and the noise variables ε_i are independent, mean zero, and have variance σ^2.

Let $\Delta_{\mathcal{M}}$ denote the Laplace–Beltrami operator. By spectral theory, there exists an $L^2(\mathcal{M})$-orthonormal basis $\{\varphi_{\lambda,\ell}\}_{\lambda,\ell}$ of eigenfunctions, grouped by eigenvalues

$$0 = \lambda_0 < \lambda_1 \le \lambda_2 \le \cdots$$

with multiplicities m_λ, such that every $f \in L^2(\mathcal{M})$ admits the expansion

$$f(x) = \sum_{\lambda} \sum_{\ell=1}^{m_\lambda} f_{\lambda,\ell}\, \varphi_{\lambda,\ell}(x).$$

In these coordinates, the Sobolev space $\mathcal{H}^s(\mathcal{M})$ is defined spectrally as

$$\mathcal{H}^s(\mathcal{M}) \coloneqq \left\{ f = \sum_{\lambda} \sum_{\ell=1}^{m_\lambda} f_{\lambda,\ell} \varphi_{\lambda,\ell} : \|f\|^2_{\mathcal{H}^s(\mathcal{M})} \coloneqq \sum_{\lambda} \sum_{\ell=1}^{m_\lambda} D_\lambda^\alpha f_{\lambda,\ell}^2 < \infty \right\},$$

where $\alpha \coloneqq 2s/d$ where $\alpha \coloneqq 2s/d$ and D_λ denotes the cumulative spectral dimension up to eigenvalue λ, namely

$$D_\lambda \coloneqq \sum_{\lambda' \leq \lambda} m_{\lambda'}.$$

For each eigenvalue λ, let

$$V_\lambda \coloneqq \operatorname{span}\{\varphi_{\lambda,\ell}\}_{\ell=1}^{m_\lambda}.$$

Since the action is isometric, the induced operator

$$T_g f(x) \coloneqq f(g^{-1}x)$$

is unitary on $L^2(\mathcal{M})$ and commutes with $\Delta_{\mathcal{M}}$. Hence each eigenspace V_λ is G-invariant, and the action restricts to an orthogonal representation

$$D_\lambda(g) : G \to O(m_\lambda)$$

on the coefficient space of V_λ. This commutativity and the resulting blockwise orthogonal action are discussed in detail by Soleymani et al. (2025b). Writing

$$f_\lambda \coloneqq (f_{\lambda,\ell})_{\ell=1}^{m_\lambda},$$

the invariance condition on the λ-th eigenspace is

$$D_\lambda(g) f_\lambda = f_\lambda \qquad \text{for all } g \in G.$$

Invariant subspaces in finite-dimensional feature spaces. Let $\mathcal{H} \subseteq L^2(\mathcal{M})$ be a finite-dimensional Hilbert feature space that is closed under the action of G. The induced action on functions is given by

$$T_g f(x) \coloneqq f(g^{-1}x), \qquad g \in G,$$

so that $T_g \mathcal{H} \subseteq \mathcal{H}$. After choosing a basis of $\mathcal{H}$ with $\dim(\mathcal{H}) = r$, each T_g is represented by a matrix $D(g) \in \mathbb{R}^{r \times r}$. The subspace of G-invariant functions is

$$\mathcal{H}^G \coloneqq \{f \in \mathcal{H} : T_g f = f, \ \forall g \in G\}.$$

Equivalently, defining

$$V_g \coloneqq \{f \in \mathcal{H} : T_g f = f\},$$

we have

$$\mathcal{H}^G = \bigcap_{g \in G} V_g.$$

In coordinates, this is the common fixed subspace

$$\mathcal{H}^G \equiv \bigcap_{g\in G} \ker(D(g) - I).$$

Thus, in the finite-dimensional setting, enforcing invariance amounts to representing this intersection of fixed spaces. A central goal of our analysis is to identify small subsets $S \subseteq G$ for which

$$\bigcap_{g\in S} V_g = \bigcap_{g\in G} V_g,$$

or for which the corresponding approximation is sufficient for the desired statistical guarantees.

B Proofs

B.1 Proof of Proposition 4.1

For any subset $S \subseteq G$, let us define a probability measure μ_S as follows. Consider the set $\mathcal{B}_S$, which is an orthonormal basis for the subspace $\bigcap_{g\in S} V_g$. We construct μ_S as the distribution of random linear combinations of vectors in this basis:

$$\mu_S := \text{ law of } \sum_{v\in\mathcal{B}_S} N_v v,$$

where the coefficients N_v are independent Gaussian random variables, each drawn as $N_v \sim \mathcal{N}(0, \frac{1}{|\mathcal{B}_S|})$. This choice ensures that μ_S is isotropic in the span of $\mathcal{B}_S$, i.e., the covariance of the distribution is proportional to the identity restricted to $\text{span}(\mathcal{B}_S)$. Intuitively, this means that μ_S places uniform weights along all directions in the subspace $\bigcap_{g\in S} V_g$.

Now define the functional

$$A(S) := \mathbb{E}_g \, \mathbb{E}_{x\sim\mu_S} \|(D(g) - I_r)x\|_2^2, \tag{1}$$

where $g \in G$ is sampled uniformly at random. This quantity measures, in expectation, how much the action of $D(g)$ deviates from the identity transformation on vectors x sampled from μ_S. In other words, $A(S)$ introduces some kind of discrepancy between invariance under the full group G and invariance under the restricted subset S.

Step 1. The case $A(S) = 0$. If $A(S) = 0$, then for every $g \in G$ and every x in the support of μ_S we must have

$$(D(g) - I_r)x = 0,$$

i.e., $D(g)x = x$. This means that $x \in \bigcap_{g\in G} V_g$.

Since the support of μ_S is exactly $\text{span}(\mathcal{B}_S)$, it follows that the entire subspace $\text{span}(\mathcal{B}_S)$ is fixed by the group G. We claim that $\text{span}(\mathcal{B}_S) = \bigcap_{g\in S} V_g$, which follows from the way we update it in Algorithm 1. This implies

$$\bigcap_{g\in S} V_g \subseteq \bigcap_{g\in G} V_g \implies \bigcap_{g\in S} V_g = \bigcap_{g\in G} V_g,$$

so in this case the invariant subspace with respect to G is already fully captured by the smaller intersection over S.

Step 2. The case $A(S) > 0$. Suppose now that $A(S) > 0$. We will prove a quantitative lower bound on $A(S)$. Namely, we prove that

$$A(S) \geq \frac{2}{r}.$$

Expanding the square inside the definition of $A(S)$ gives

$$A(S) = \mathbb{E}_g \mathbb{E}_{x\sim\mu_S}\Big[\|x\|_2^2 + \|D(g)x\|_2^2 - x^\top D(g)x - x^\top D(g)^\top x\Big]. \tag{2}$$

Since the representation $D(g)$ is orthogonal, we have $\|D(g)x\|_2^2 = \|x\|_2^2$. This allows us to rewrite the expression as:

$$\begin{aligned} A(S) &= \mathbb{E}_g \mathbb{E}_{x\sim\mu_S}\big[\|x\|_2^2 + \|x\|_2^2 - x^\top D(g)x - x^\top D(g)^\top x\big] &(3)\\ &= 2 - \mathbb{E}_g \mathbb{E}_{x\sim\mu_S}\big[x^\top D(g)x + x^\top D(g)^\top x\big], &(4)\end{aligned}$$

where in above we used the fact that $\mathbb{E}_{x\sim\mu_S}[\|x\|_2^2] = 1$, according to the definition of μ_S.

But since $D(g)$ is orthogonal, averaging $D(g)$ or $D(g)^\top$ over $g \in G$ yields the same expectation. We therefore obtain

$$A(S) = 2 - 2\mathbb{E}_{x\sim\mu_S}[x^\top \bar{D}x],$$

where we define the average

$$\bar{D} := \mathbb{E}_g[D(g)] = \mathbb{E}_g[D(g)^\top].$$

Step 3. Structure of $\bar{D}$. From basic representation theory, the matrix $\bar{D}$ is the orthogonal projection onto the subspace of invariant vectors (Schur's lemma). Equivalently, there exists an orthogonal change of basis U such that

$$\bar{D} = UPU^\dagger,$$

where P is the projection matrix onto the first r_{inv} coordinates, with $r_{\text{inv}} = \dim\left(\bigcap_{g\in G} V_g\right)$. In other words,

$$\bar{D} = U \begin{bmatrix} I_{r_{\text{inv}}} & 0 \\ 0 & 0 \end{bmatrix} U^\dagger.$$

Thus, $\bar{D}$ corresponds to selecting exactly those elements invariant under the full group.

Moreover, since $S \subseteq G$, we automatically have

$$\bigcap_{g\in G} V_g \subseteq \bigcap_{g\in S} V_g.$$

That is, the invariant subspace for the full group is always contained within the invariant subspace defined by any subset of group elements. Therefore, $\text{span}(\mathcal{B}_S)$ contains the true invariant subspace, and the expectation $\mathbb{E}_{x\sim\mu_S}[x^\top \bar{D}x]$ reflects the fraction of L^2-norm of the elements in the support of μ_S lying in this smaller subspace.

Step 4. Dimension ratio interpretation. Since μ_S is isotropic over $\bigcap_{g\in S} V_g$, the expectation of the quadratic form $x^\top \bar{D}x$ is equal to the ratio

$$\frac{\dim(\bigcap_{g\in G} V_g)}{\dim(\bigcap_{g\in S} V_g)}.$$

Substituting this back, we find

$$A(S) = 2\Big(1 - \frac{\dim(\bigcap_{g\in G} V_g)}{\dim(\bigcap_{g\in S} V_g)}\Big). \tag{5}$$

Step 5. Lower bound on $A(S)$. If $A(S) > 0$, then necessarily

$$\dim(\bigcap_{g\in S} V_g) > \dim(\bigcap_{g\in G} V_g).$$

The smallest possible difference between the two dimensions is exactly one, so

$$\frac{\dim(\bigcap_{g\in G} V_g)}{\dim(\bigcap_{g\in S} V_g)} \leq \frac{r-1}{r}.$$

Therefore,

$$A(S) \geq 2\Big(1 - \frac{r-1}{r}\Big) = \frac{2}{r}. \tag{6}$$

This establishes the claimed lower bound.

Step 6. Probabilistic argument. It remains to control the number of random trials needed to detect all strict dimension decreases. By the previous step, whenever a strict discrepancy remains, namely

$$\bigcap_{g\in G} V_g \subsetneq \bigcap_{g\in S} V_g,$$

a fresh random draw succeeds in detecting it with probability at least $2/r$. Indeed, if $X \in [0,1]$ denotes the corresponding detection statistic, then $\mathbb{E}X \geq 2/r$, and hence

$$\mathbb{P}(X > 0) \geq \mathbb{E}X \geq \frac{2}{r}.$$

Let Z be the number of successful detections among T independent trials. Then Z stochastically dominates a binomial random variable $\mathrm{Bin}(T, 2/r)$. Therefore, by a standard Chernoff bound,

$$\mathbb{P}(Z < r) \leq \delta$$

provided that

$$T \geq Cr\left(r + \log\frac{1}{\delta}\right)$$

for a universal constant $C > 0$. Since the dimension can decrease at most r times, r successful detections suffice to certify that no further strict discrepancy remains. Thus, with probability at least $1 - \delta$, the procedure terminates succesfully after

$$T = \mathcal{O}\left(r^2 + r\log\frac{1}{\delta}\right)$$

independent trials.

Conclusion. Putting everything together, with T iterations we guarantee that, with probability at least $1 - \delta$, the subspace defined by S coincides with the true invariant subspace:

$$\bigcap_{g\in G} V_g = \bigcap_{g\in S} V_g,$$

while $|S| \leq r$ since at each iteration at most one group element is chosen. This completes the proof.

B.2 Proof of Theorem 4.2

The proof follows the structure of the finite-group proof (Soleymani et al., 2025b, Theorem 1) and differs only in how the constraint set S is chosen and analyzed; the statistical analysis is mostly unchanged.

Step 1: Spectral reduction and decoupled convex programs. Because $\Delta_{\mathcal{M}}$ commutes with all isometries (hence with the G-action), G preserves each eigenspace V_λ and acts on it through an orthogonal matrix $D_\lambda(g)$. Therefore f is G-invariant if and only if $D_\lambda(g) f_\lambda = f_\lambda$ for all λ and all $g \in G$. As in the finite-group analysis, minimizing the population risk $\mathbb{E}\|f - f^\star\|_{L^2}^2$ over G-invariant f reduces to *independent* quadratic problems (QP) on the retained eigenspaces V_λ with linear constraints $D_\lambda(g)u = u$. Replacing the intractable population coefficients by their empirical means $\widehat{f}_{\lambda,\ell}$ yields the *empirical* QPs mentioned above; their minimizers are the orthogonal projections of $\widehat{f}_\lambda$ onto the fixed-point subspaces (Soleymani et al., 2025b).

Step 2: A single random subset S suffices for *all* retained eigenspaces. Define the block-diagonal representation

$$R(g) \;:=\; \bigoplus_{\lambda:\, D_\lambda \le D} D_\lambda(g) \in O(r), \qquad r := \sum_{\lambda:\, D_\lambda \le D} m_\lambda = D.$$

Let $V_g := \ker(I_r - R(g))$ be its fixed-point subspace. Run Algorithm 1 (Randomized Subset Selection) *once* on the representation $R(\cdot)$ with $T = \Theta(r^2 + r\log(1/\delta))$ iterations to obtain a set $S \subseteq G$ of size $|S| \le r$ such that, with probability at least $1-\delta$,

$$\bigcap_{g\in S} V_g \;=\; \bigcap_{g\in G} V_g.$$

This result follows from Proposition 4.1 and its proof is discussed in Section B.1: the statistic $A(S) := \mathbb{E}_g\, \mathbb{E}_{x\sim\mu_S}\|(R(g) - I_r)x\|_2^2$ either vanishes (in which case the intersections coincide) or is bounded below by a positive constant depending on r; a standard concentration argument then shows that each time $A(S) > 0$ one detects it in r trials and reduces the candidate basis dimension by 1, hence $\mathcal{O}(r^2 + r\log(1/\delta))$ trials suffice.

Because $R(g)$ is block-diagonal, $V_g = \bigoplus_{\lambda:\, D_\lambda \le D} \ker\big(I_{m_\lambda} - D_\lambda(g)\big)$ and therefore

$$\bigcap_{g\in S} V_g = \bigoplus_{\lambda:\, D_\lambda \le D} \bigcap_{g\in S} \ker\big(I_{m_\lambda} - D_\lambda(g)\big), \qquad \bigcap_{g\in G} V_g = \bigoplus_{\lambda:\, D_\lambda \le D} \bigcap_{g\in G} \ker\big(I_{m_\lambda} - D_\lambda(g)\big).$$

Thus the equality of intersections at the block level implies, *for every retained* λ, that

$$\bigcap_{g\in S} \ker\big(I_{m_\lambda} - D_\lambda(g)\big) \;=\; \bigcap_{g\in G} \ker\big(I_{m_\lambda} - D_\lambda(g)\big).$$

Consequently, projecting $\widehat{f}_\lambda$ onto the fixed-point subspace defined by S is the same as projecting onto the G-invariant subspace of V_λ. Hence the resulting estimator $\widetilde{f}$ is *exactly G-invariant* with probability at least $1-\delta$.

Finally, note that by construction $r = D$. Since $\alpha > 1$, we have $D = n^{1/(1+\alpha)} \le n^{1/2}$, hence $r = \mathcal{O}(\sqrt{n})$, matching the choice of r in Proposition 4.1.

Step 3: Risk bound (classic bias-variance analysis). Write $f^\star = f^\star_{\le D} + f^\star_{>D}$ for the orthogonal decomposition into the retained and discarded spectral parts. Exactly as the proof of Soleymani et al. (2025b, Theorem 1), we decompose

$$\mathbb{E}\Big[\|\widetilde{f} - f^\star\|_{L^2}^2\Big] \;\le\; 2\,\mathbb{E}\Big[\|\widetilde{f} - f^\star_{\le D}\|_{L^2}^2\Big] \;+\; 2\,\|f^\star_{>D}\|_{L^2}^2.$$

The *bias term* obeys $\|f^\star_{>D}\|^2_{L^2} \le D^{-\alpha}\|f^\star\|^2_{H^s(M)}$ by $f^\star \in H^s(\mathcal{M})$ and the spectral definition of the Sobolev norm. This is because,

$$\begin{aligned}
\|f^\star_{>D}\|^2_{L^2} &= \sum_{\lambda: D_\lambda > D} \sum_{\ell=1}^{m_\lambda} (f^\star_{\lambda,\ell})^2 \\
&= \sum_{\lambda: D_\lambda > D} \sum_{\ell=1}^{m_\lambda} D_\lambda^{-\alpha} D_\lambda^{\alpha} (f^\star_{\lambda,\ell})^2 \\
&\le D^{-\alpha} \sum_{\lambda: D_\lambda > D} \sum_{\ell=1}^{m_\lambda} D_\lambda^{\alpha} (f^\star_{\lambda,\ell})^2 \\
&\le D^{-\alpha} \sum_{\lambda} \sum_{\ell=1}^{m_\lambda} D_\lambda^{\alpha} (f^\star_{\lambda,\ell})^2 \\
&= D^{-\alpha} \|f^\star\|^2_{H^s(\mathcal{M})}.
\end{aligned}$$

In turn, we focus on the *variance term*,

$$\mathbb{E}\Big[\|\widehat{f}_{\le D} - f^\star_{\le D}\|^2_{L^2}\Big] = \sum_{D_\lambda \le D} \sum_{\ell=1}^{m_\lambda} \mathbb{E}\Big[\big|\widehat{f}_{\lambda,\ell} - f^\star_{\lambda,\ell}\big|^2\Big].$$

By definition, we obtain

$$f^\star_{\lambda,\ell} = \mathbb{E}_x[f^\star(x)\phi_{\lambda,\ell}(x)] = \mathbb{E}_{x,y}[y\phi_{\lambda,\ell}(x)], \tag{7}$$

for every λ, ℓ. In addition, $\widetilde{f}_{\lambda,\ell}$ denotes the empirical estimate derived from the data:

$$\widetilde{f}_{\lambda,\ell} = \frac{1}{n}\sum_{i=1}^{n} y_i \phi_{\lambda,\ell}(x_i). \tag{8}$$

Thus, we get

$$\begin{aligned}
\mathbb{E}[|\widetilde{f}_{\lambda,\ell} - f^\star_{\lambda,\ell}|^2] &= \frac{1}{n}\mathbb{E}\left[|y\phi_{\lambda,\ell}(x) - \mathbb{E}[y\phi_{\lambda,\ell}(x)]|^2\right] \\
&= \frac{1}{n}\mathbb{E}\left[|\epsilon\phi_{\lambda,\ell}(x) + f^\star(x)\phi_{\lambda,\ell}(x) - \mathbb{E}[f^\star(x)\phi_{\lambda,\ell}(x)]|^2\right] \\
&= \frac{1}{n}\left(\sigma^2\mathbb{E}[\phi^2_{\lambda,\ell}] + \mathbb{E}\left[|f^\star(x)\phi_{\lambda,\ell}(x) - \mathbb{E}[f^\star(x)\phi_{\lambda,\ell}(x)]|^2\right]\right) \\
&\le \frac{1}{n}\left(\sigma^2 + \mathbb{E}[f^\star(x)^2\phi^2_{\lambda,\ell}(x)]\right) \\
&\le \frac{1}{n}\left(\sigma^2 + \|f^\star\|^2_{L^\infty(\mathcal{M})}\right),
\end{aligned}$$

since the $\varphi_{\lambda,\ell}$'s are orthonormal and $\widehat{f}_{\lambda,\ell}$ are empirical means. Summing over dimensions up to D, we obtain

$$\mathbb{E}[\|\widetilde{f} - f^\star_{\le D}\|^2_{L^2(\mathcal{M})}] \le \frac{D}{n}\left(\sigma^2 + \|f^\star\|^2_{L^\infty(\mathcal{M})}\right).$$

Because $\widetilde{f}$ is the *orthogonal projection* (in each V_λ) of $\widehat{f}$ onto a linear subspace, the projection can only reduce squared error, so

$$\mathbb{E}\Big[\|\widetilde{f} - f^\star_{\le D}\|^2_{L^2}\Big] \le \mathbb{E}\Big[\|\widehat{f}_{\le D} - f^\star_{\le D}\|^2_{L^2}\Big] \le \frac{D}{n}\left(\sigma^2 + \|f^\star\|^2_{L^\infty(\mathcal{M})}\right).$$

Putting the two parts (*bias* and *variance*) together and taking $D = n^{1/(1+\alpha)}$ yields

$$\mathbb{E}\|\widetilde{f} - f^\star\|_{L^2}^2 \;\le\; \Big(\sigma^2 + \|f^\star\|_{L^\infty(\mathcal{M})}^2\Big)\,\frac{D}{n} + \Big(\|f^\star\|_{H^s(\mathcal{M})}^2\Big)\,D^{-\alpha} \;=\; \mathcal{O}\Big(n^{-\frac{\alpha}{1+\alpha}}\Big),$$

exactly as for the finite group settings. *All the calculations of this step are borrowed verbatim from the proof of for finite groups in Soleymani et al. (2025b).*

Step 4: Running-time bound. Computing the primary coefficients $\widehat{f}_{\lambda,\ell}$ for $D_\lambda \le D$ takes $\mathcal{O}\big(nD\big) = \mathcal{O}\big(n^{\frac{2+\alpha}{1+\alpha}}\big)$ time. Forming the constraint matrices $\{D_\lambda(g)\}_{g\in S}$ costs $\mathcal{O}\big(|S|\sum_{D_\lambda\le D} m_\lambda^2\big) \le \mathcal{O}\big(|S|\,D^2\big)$ oracle calls, because $\sum m_\lambda^2 \le (\sum m_\lambda)^2 = D^2$. Solving the QPs by the closed form uses a pseudoinverse of a matrix of size $(|S|m_\lambda) \times (|S|m_\lambda)$ and hence time $\mathcal{O}\big(|S|^3 m_\lambda^3\big)$ per λ; summing gives $\mathcal{O}\big(|S|^3\sum m_\lambda^3\big) \le \mathcal{O}\big(|S|^3 D^3\big)$. By Step 2, with probability at least $1-\delta$ we have $|S| = \mathcal{O}(D)$, and since $D = n^{1/(1+\alpha)} \le n^{1/2}$, this is polynomial in n (and independent of $|G|$). Thus the total running time is $\mathrm{poly}\big(n, d, \log(1/\delta)\big)$.

Combining Steps 1–4 concludes the proof: the estimator $\widetilde{f}$ is exactly G-invariant (with probability $\ge 1-\delta$), achieves the same excess-risk rate as in the finite-group case, and the algorithm runs in time polynomial in $n, d, \log(1/\delta)$, *independent of the cardinality of* G.

Remark B.1. The proof for the finite-dimensional setting proceeds analogously (except that we do not need to cap the dimension and bound the tail), since the basis is projected onto the quotient space. From this point onward, the minimax optimality argument follows standard lines.

B.3 Proof of Theorem 4.4

In this section, we present a proof of the main result of the paper on symmetry discovery.

We begin by noting that, in light of the proof of Theorem 4.2, it suffices to show that Algorithm 3 returns a set $S \subseteq G$ that generates the true symmetry group H with high probability. Once such a generating set is obtained, exact H-invariance of the final estimator follows immediately from the construction of the algorithm, and the stated statistical guarantees then follow from standard results on invariant regression.

Sampling a generating set for H. We first argue that sampling

$$T = \Omega\left(\frac{1}{\kappa}\big(\log|G| + \log\tfrac{1}{\delta}\big)\right)$$

elements uniformly at random from G is sufficient to ensure that, with probability at least $1-\delta$, the sampled elements contain a generating set for H.

It is a classical result in the theory of random Cayley graphs that, for any finite group H, drawing

$$T_H = \Omega\big(\log|H| + \log\tfrac{1}{\delta}\big)$$

elements uniformly at random from H produces a generating set for H with probability at least $1-\delta$; see, e.g., (Alon and Roichman, 1994).

Since $|H| \ge \kappa|G|$, each draw from $\mathrm{Unif}(G)$ lands in H with probability at least κ. By standard concentration bounds for binomial random variables, after

$$T = \Omega\left(\frac{1}{\kappa}\big(\log|H| + \log\tfrac{1}{\delta}\big)\right)$$

samples, the number of draws falling inside H is at least T_H with probability at least $1-\delta$. Consequently, with high probability, the sampled elements contain a generating set for H.

Acceptance and rejection of sampled elements. Having ensured that a generating set for H appears among the sampled group elements, it remains to show that Algorithm 3 correctly distinguishes elements of H from elements in $G \setminus H$, where the latter denotes transformations $g \in G$ for which $T_g f^\star \neq f^\star$ (assuming that H denotes the largest subgroup under which $f^\star$ is invariant, to resolve any ambiguity arising from identifiability issues).

For each randomly sampled element $g \in G$, there are two types of errors to control:

- If $g \in H$, we must ensure that the algorithm accepts g with high probability.
- If $g \notin H$, we must ensure that the algorithm rejects g with high probability.

The remainder of the proof is devoted to establishing these two guarantees by comparing the population and empirical risks of the unconstrained estimator and the estimator constrained to be invariant under g. We show that invariance constraints corresponding to elements of H do not increase the population risk, while invariance constraints corresponding to elements outside H introduce a detectable bias with high probability under the isotropy assumption on $f^\star$.

For a candidate element $g \in G$, let

$$W \coloneqq \mathcal{H}^H, \qquad W_g \coloneqq \mathcal{H}^{\langle H, g\rangle} = W \cap \mathcal{H}^{\langle g\rangle},$$

and let P_g denote the orthogonal projector from W onto W_g. Since $f^\star \in W$, imposing invariance under g introduces the population-level bias

$$b_g \coloneqq \|(I - P_g) f^\star\|^2_{L^2(\mathcal{M})}.$$

If $g \in H$, then $W_g = W$ and hence $b_g = 0$. If $g \notin H$, then, under the identifiability convention that H is the maximal invariance subgroup of $f^\star$, we have $W_g \subsetneq W$, and therefore $b_g > 0$.

The key point is that this separation holds uniformly over the finitely many candidate elements inspected by the algorithm. Let $\mathcal{G}_{\text{test}} \subseteq G$ denote this candidate set, with $|\mathcal{G}_{\text{test}}| = T$. Conditional on $\mathcal{G}_{\text{test}}$, the subspaces $\{W_g : g \in \mathcal{G}_{\text{test}} \setminus H\}$ are fixed strict subspaces of W. Under the isotropy assumption, we may write $f^\star = Ru$, where u is uniformly distributed on the unit sphere in W and $R = \|f^\star\|_{L^2(\mathcal{M})}$. If $d_H \coloneqq \dim(W)$, then for any fixed strict subspace $U \subsetneq W$ and any $\tau \in (0, 1)$,

$$\mathbb{P}\big(\operatorname{dist}(u, U)^2 \leq \tau\big) \leq C\sqrt{d_H \tau},$$

for a universal constant $C > 0$; the worst case is codimension one. A union bound over the at most T tested elements gives

$$\mathbb{P}\big(\exists g \in \mathcal{G}_{\text{test}} \setminus H : b_g \leq R^2 \tau\big) \leq CT\sqrt{d_H \tau}.$$

Thus, choosing

$$\tau = \frac{c\delta^2}{T^2 d_H}$$

for a sufficiently small universal constant $c > 0$, we obtain, with probability at least $1 - \delta$, the simultaneous separation

$$b_g = 0 \quad \text{for all } g \in H, \qquad b_g \geq \gamma \quad \text{for all } g \in \mathcal{G}_{\text{test}} \setminus H,$$

where

$$\gamma \coloneqq R^2 \tau = \Omega\left(\frac{R^2 \delta^2}{T^2 d_H}\right).$$

In particular, when $T = \mathrm{poly}(d)$ and $d_H \le r = \mathrm{poly}(d)$, this gap is inverse-polynomial in d.

It remains to ensure that the empirical comparisons resolve this population gap. Using an independent validation sample, standard concentration bounds imply that, for all candidates in $\mathcal{G}_{\mathrm{test}}$ simultaneously, the empirical losses of the unconstrained and g-constrained estimators are within $o(\gamma)$ of their corresponding population losses, provided the validation size is polynomial in r, $1/\gamma$, and $\log(T/\delta)$. On this event, a thresholded comparison with tolerance, say $\eta = \gamma/4$, is correct: elements $g \in H$ introduce no bias and are accepted, whereas elements $g \notin H$ introduce bias at least γ and are rejected. Consequently, with high probability over the isotropic draw of $f^\star$, Algorithm 3 correctly distinguishes all sampled elements of the group H from all sampled elements outside H.