From Finite to Infinite Groups: A Polynomial-Time Algorithm for Learning with Exact Invariances

Ashkan Soleymani* ASHKANSO@MIT.EDU
MIT

Behrooz Tahmasebi* BZT@MIT.EDU
MIT

Patrick Jaillet JAILLET@MIT.EDU
MIT

Stefanie Jegelka STEFANIE.JEGELKA@TUM.DE
MIT and TUM

Abstract

Despite its broad range of applications in science, the theoretical foundations of learning with invariances have been only sparsely explored. Even in the case of polynomial regression, it has remained unclear whether one can efficiently compute an *exactly invariant* regression function, as traditional methods such as data augmentation, group averaging, and canonicalization fail to provably solve the task in polynomial time. Recent work (Soleymani et al., 2025b) has examined the statistical–computational trade-off of learning with invariances, demonstrating that for finite groups there exists a polynomial-time algorithm (in the data dimension, sample size, and logarithm of the group size) that yields functions which both generalize well and are exactly invariant. However, this approach is intrinsically limited to finite groups, leaving the tractability for learning with *infinite groups* unresolved. In this paper, we design and analyze a polynomial-time algorithm that applies to any group, including infinite ones, and learns functions that generalize well in polynomial time with respect to data dimension and sample size, independent of the group. This closes the gap and provides strong theoretical evidence that computationally efficient algorithms for learning under invariances can indeed generalize effectively, a phenomenon consistently supported by the empirical success of geometric machine learning.

Keywords: invariance, symmetry, groups, computational complexity

1. Introduction

Invariances have been central to learning 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). 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 detection (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

. * denotes equal contribution.

is still missing is a systematic understanding of the computational price of incorporating invariances—particularly in fundamental settings such as kernel methods.

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). Given these observations, especially the prohibitively large size of the group, one might expect that learning with exact invariances is not computationally tractable even in the basic setting of kernel regression. Surprisingly, recent work (Soleymani et al., 2025b) introduced *spectral-averaging*, an exact invariant algorithm that achieves desirable population risk in $\mathrm{poly}(n, d, \log|G|)$ time, that is, polynomial in the number of samples n, the input dimension d, and the logarithm of the group size $|G|$. For *finite groups*, this amounts to an effectively polynomial-time procedure since $\log|G|$ is polynomial in the input dimension d.

This result naturally begs the same question for *infinite groups*. In particular, one would hope for algorithms that achieve desirable population risk with time complexity independent of the group cardinality $|G|$. Remarkably, *we provide an affirmative answer to this problem* by designing a randomized algorithm. We begin by formalizing the setting with a precise problem statement.

1.1. Problem statement

We consider supervised learning on a smooth, compact, boundaryless Riemannian manifold $\mathcal{M}$ of dimension d. Given n independent samples $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^n$ with $x_i \in \mathcal{M}$ drawn uniformly[1], the labels follow $y_i = f^\star(x_i) + \epsilon_i$, where $f^\star \in C(\mathcal{M})$ is the unknown regression function and ϵ_i are independent, zero-mean, variance-σ^2 noise. The performance of an estimator $\widehat{f}$ is measured by its population risk

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

When $f^\star$ lies in an RKHS $\mathcal{H}$, kernel ridge regression (KRR) provides efficient estimators. For $\mathcal{H} = H^s(\mathcal{M})$, the KRR estimator achieves risk $\mathcal{O}(n^{-s/(s+d/2)})$ with $\mathcal{O}(n^3)$ computational cost (Bach, 2024).

Suppose that we are given a group G that acts smoothly and isometrically on $\mathcal{M}$, denoted by gx for any $g \in G$ and $x \in \mathcal{M}$. A regression function is G-invariant if $f^\star(gx) = f^\star(x)$ for all $g \in G, x \in \mathcal{M}$. In this case, the learner must find an estimator $\widehat{f}$ that is both accurate and invariant. For learning with Sobolev kernels, the KRR estimator $\widehat{f}_{\mathrm{KRR}}$ is *not* G-invariant. As a result, $\widehat{f}_{\mathrm{KRR}}$ cannot serve as a solution for learning under invariances, and additional care is required to construct estimators that explicitly respect the group structure. With these notations in place, we now formally state the problem.

1. Uniformity is assumed for simplicity; results extend to bounded densities.

Can we design an exactly G-invariant estimator that achieves the same excess population risk, $\mathcal{O}\big(n^{-s/(s+d/2)}\big)$, as in learning without invariances, while running in $\mathrm{poly}(n, d)$ time?

The recent breakthrough work of Soleymani et al. (2025b) introduced *spectral-averaging*, which achieves $\mathcal{O}\big(n^{-s/(s+d/2)}\big)$ risk in $\mathrm{poly}(n, d, \log|G|)$ time, thereby reducing the complexity exponentially compared to the $\mathrm{poly}(n, d, |G|)$ time required by canonical approaches to learning with invariances such as *group averaging*. While this algorithm indeed runs in polynomial time for finite groups G, *it falls short when extended to infinite groups*, such as rotation invariances. In effect, any polynomial-time algorithm for learning with exact invariance over infinite groups must be independent of the group cardinality $|G|$, which is the focus of this work. We address this problem affirmatively by designing a novel algorithm based on spectral-averaging.

2. Main Result

The spectral-averaging algorithm of Soleymani et al. (2025b) reduces the *nonconvex optimization* problem of learning with exact invariances to an *infinite collection of finite-dimensional quadratic convex programs with linear constraints*—one for each eigenspace of the Laplace–Beltrami operator on the data manifold—using machinery of differential geometry and spectral theory. Subsequently, by truncating the number of quadratic programs to be solved separately, they derive efficient approximations to the original nonconvex optimization problem, yielding approximate kernel solutions for learning with invariances. This *deterministic* procedure achieves an estimator with population risk of $\mathcal{O}\big(n^{-s/(s+d/2)}\big)$ in $\mathrm{poly}(n, d, \log|G|)$ time for finite groups.

The $\mathrm{poly}(\log|G|)$ dependence on the group cardinality arises from the number of linear constraints imposed in the quadratic convex programs associated with the eigenspaces of the spectral-averaging algorithm. These constraints are determined by the minimal generating set S of the group G, with one constraint required for each generator. Since the size of a minimal generating set of a finite group is bounded by $|S| \leq \log|G|$, this fully accounts for the polylogarithmic dependence of computational complexity on $|G|$. For infinite groups, the group cardinality $|G|$ is unbounded, and no efficient reduction of the nonconvex optimization problem is known.

In Algorithm 1, we introduce a new randomized procedure that, with high probability, identifies a set S governing the sufficient linear constraints in the quadratic convex programs associated with each eigenspace for general groups. This procedure returns a suitable subset $S \subseteq G$ in $\mathcal{O}\big(n \log \frac{1}{\delta}\big)$ time with probability at least $1 - \delta$ as formalized in Proposition 1, and its complexity is independent of the group cardinality $|G|$, holding even when G is infinite. The key idea is that, perhaps surprisingly, random group elements are sufficient to impose the necessary linear constraints with high probability. This stands in contrast to alternative approaches, such as generating sets, which are either inapplicable for infinite groups or require preprocessing the entire group, making them computationally inefficient.

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

Algorithm 1 Randomized Subset Selection

Input: Query access to a group representation $D(g) \in \mathbb{R}^{r\times r}$ for any $g \in G$, a parameter T
Output: A subset $S \subseteq G$

Initialization: $S \leftarrow \{\mathrm{id}_G\}$ and $\mathcal{B}_S = \{e_i : i \in [r]\}$ where e_i, $i \in [r]$, is the unit vector in i-th coordinate.
for $t = 1, \ldots, T$ **do**
  Sample $g \in G$ uniformly at random
  Sample $x = \sum_{v\in\mathcal{B}_S} N_v v$, such that $N_v \sim \mathcal{N}(0, \frac{1}{|\mathcal{B}_S|})$ independently for all $v \in \mathcal{B}_S$
  if $\|(D(g) - I_r)x\|_2^2 > 0$ **then**
    $S \leftarrow S \cup \{g\}$
    $\mathcal{B}_S \leftarrow$ orthonormal basis for $\mathrm{span}(\mathcal{B}_S) \bigcap \ker(I_r - D(g))$
  end if
end for
return The final subset $S \subseteq G$.

that $\bigcap_{g\in S} V_g = \bigcap_{g\in G} V_g$. *For the group representations considered in this paper, we have* $r = \mathcal{O}(\sqrt{n})$. *Therefore, the overall iteration complexity of Algorithm 1 is* $T = \mathcal{O}\big(n \log \frac{1}{\delta}\big)$.

With the randomized subset selection of Algorithm 2 in place to find sufficiently small representer sets S for each quadratic convex program, we are now ready to state our full algorithm, which builds upon the spectral-averaging framework of Soleymani et al. (2025b).

Theorem 2 *Let G be a general group, and let $\{x_i, y_i\}_{i=1}^n$ be a labeled dataset of size n sampled from a d-dimensional manifold $\mathcal{M}$. Suppose the optimal regression function satisfies $f^\star \in H^s(\mathcal{M})$ for some $s > d/2$, and set $\alpha := 2s/d$. Then, `Spec-Avg` with Randomized Subset Selection (Algorithm 2) returns, with probability at least $1-\delta$, an exactly G-invariant estimator $\widehat{f}$ in $\mathrm{poly}(n, d, \log(1/\delta))$ time that achieves excess population risk $\mathcal{R}(\widehat{f}) = \mathcal{O}\big(n^{-s/(s+d/2)}\big)$.*

3. Conclusion

We have designed the first *randomized* polynomial-time algorithm for learning with *exact invariances* under general groups, accommodating both finite and infinite cases. This marks a significant step toward settling the computational complexity of learning with invariances. For finite groups, spectral averaging with group generators (Soleymani et al., 2025b) is known to run in polynomial time. For infinite groups, however, our approach—spectral averaging with subset selection—is inherently randomized, and it remains open whether a *deterministic* polynomial-time algorithm exists or not. We leave this as an intriguing direction for future work.

Acknowledgments

BT and SJ acknowledge support from NSF AI Institute TILOS and the Alexander von Humboldt Foundation.

Algorithm 2 Spectral Averaging (`Spec-Avg`) with Randomized Subset Selection

Input: $\mathcal{S} = \{(x_i, y_i) : i \in [n]\}$ and $\alpha = 2s/d \in (1, \infty)$.
Output: $\widehat{f}(x)$.

Initialize $D \leftarrow n^{1/(1+\alpha)}$.
for each λ such that $D_\lambda \leq D$ **do**
  for each $\ell \in [m_\lambda]$ **do**
    $\widetilde{f}_{\lambda,\ell} \leftarrow \frac{1}{n} \sum_{i=1}^{n} y_i \phi_{\lambda,\ell}(x_i)$.
  end for
end for
Initialize $S \leftarrow$ Randomized Subset Selection (G) ▷ [Algorithm 1]
for each λ such that $D_\lambda \leq D$ **do**
  Solve the following linearly constrained quadratic program over m_λ variables:

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

end for
Return: $\widehat{f}(x) = \sum_{\lambda : D_\lambda \leq D} \sum_{\ell=1}^{m_\lambda} \widehat{f}_{\lambda,\ell} \phi_{\lambda,\ell}(x)$.

References

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

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.

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.

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

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.

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.

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

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

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

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

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

David Haussler et al. Convolution kernels on discrete structures. Technical report, Technical report, Department of Computer Science, University of California ..., 1999.

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.

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.

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.

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

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

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

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.

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.

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

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.

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

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

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.

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.

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.

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

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.

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

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

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.

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.

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

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

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

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

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

Appendix A. 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 et al., 1999; Kondor, 2008; Sokolic et al., 2017). A classical way of achieving this is through invariant kernels, constructed by averaging over transformation groups (Schölkopf and Smola, 2002; Haasdonk and Burkhardt, 2007).

Invariance in kernel methods is not limited to group averaging. 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). By contrast, our spectral projection approach preserves Sobolev regularity while enforcing exact invariances via linear constraints.

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 kernel regression without invariances. 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).

Appendix B. Proofs

Setting and notation. Let $(\mathcal{M}, g)$ be a smooth, compact, connected, boundaryless d-dimensional Riemannian manifold, and let a (possibly infinite) group G act smoothly and isometrically on $\mathcal{M}$ (so $x \mapsto gx$ is an isometry for each $g \in G$). We observe i.i.d. samples $S = \{(x_i, y_i)\}_{i=1}^n$ with x_i uniform on $\mathcal{M}$ and

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

where $f^\star \in H^s(\mathcal{M})$ for some $s > d/2$ and ε_i are independent, mean 0, variance σ^2. $H^s(\mathcal{M})$ is the space of s-Sobolov functions on the Riemannian manifold $\mathcal{M}$, defined as,

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

where $\alpha \coloneqq 2s/d$.

Let $\Delta_{\mathcal{M}}$ denote the Laplace-Beltrami operator; by spectral theory there is an $L^2(\mathcal{M})$-orthonormal basis $\{\varphi_{\lambda,\ell}\}_{\lambda,\ell}$ of eigenfunctions, grouped by eigenvalues $\lambda \in \{0 = \lambda_0 < \lambda_1 \leq \lambda_2 \leq \cdots\}$ with multiplicities m_λ, such that

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

For each λ, the G-action induces an orthogonal representation $D_\lambda : G \to O(m_\lambda)$ on $V_\lambda := \mathrm{span}\{\varphi_{\lambda,\ell}\}_{\ell=1}^{m_\lambda}$, because Δ_M commutes with the isometries $T_g : f \mapsto (x \mapsto f(gx))$.[2] We will write $D_\lambda(g) f_\lambda = f_\lambda$ for the linear invariance constraints on the coefficient vector

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

B.1. Proof of Proposition 1

Proof 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 $\mathrm{span}(\mathcal{B}_S)$. 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

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

2. This commutativity, together with the induced blockwise orthogonal actions, is discussed in detail in Soleymani et al. (2025b).

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 $\mathrm{span}(\mathcal{B}_S)$, it follows that the entire subspace $\mathrm{span}(\mathcal{B}_S)$ is fixed by the group G. We claim that $\mathrm{span}(\mathcal{B}_S) = \bigcap_{g\in S} V_g$, 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 $r_{\mathrm{inv}} = \dim\left(\bigcap_{g\in G} V_g\right)$. In other words,

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

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, $\mathrm{span}(\mathcal{B}_S)$ 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. Thus, whenever $A(S) > 0$, the expectation is at least $\frac{2}{r}$. By Hoeffding's inequality, the probability of failing to detect the strict inequality

$$\bigcap_{g \in G} V_g \neq \bigcap_{g \in S} V_g$$

after T_1 independent trials is at most $\exp(-\Omega(T_1/r))$. Since detecting such discrepancies in dimensions requires at most $\mathcal{O}(r)$ successful events, the total number of iterations needed is upper-bounded as

$$T = \mathcal{O}\big(r^2 \log \tfrac{1}{\delta}\big)\,.$$

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 $|S| = \mathcal{O}\big(r^2 \log \frac{1}{\delta}\big)$ since at each iteration at most one group element is chosen. This completes the proof. ■

B.2. Proof of Theorem 2

Proof 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 iff $D_\lambda(g) f_\lambda = f_\lambda$ 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 $T = \Theta(r^2 \log(1/\delta))$ iterations to obtain a set $S \subseteq G$ of size $|S| = \Theta(r^2 \log(1/\delta))$ 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 1 and its proof is discussed in Appendix 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 $\mathcal{O}(r)$ trials and reduces the candidate basis dimension by 1, hence $\mathcal{O}(r^2 \log(1/\delta))$ 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 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 $\|f^\star_{>D}\|_{L^2}^2 \le D^{-\alpha}\|f^\star\|_{H^s(M)}^2$ by $f^\star \in H^s(\mathcal{M})$ and the spectral definition of the Sobolev norm. This is because,

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

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

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

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

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

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 $|S| = \Theta(r^2\log(1/\delta)) = \Theta(D^2\log(1/\delta))$, 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. ■