Learning with Exact Invariances in Polynomial Time

Ashkan Soleymani* MIT
Behrooz Tahmasebi* MIT
Stefanie Jegelka TUM, MIT
Patrick Jaillet MIT

Abstract

We study the statistical-computational trade-offs for learning with exact invariances (or symmetries) using kernel regression over manifold input spaces. Traditional methods, such as data augmentation, group averaging, canonicalization, and frame-averaging, either fail to provide a polynomial-time solution or are not applicable in the kernel setting. However, with oracle access to the geometric properties of the input space, we propose a polynomial-time algorithm that learns a classifier with *exact* invariances. Moreover, our approach achieves the same excess population risk (or generalization error) as the original kernel regression problem. To the best of our knowledge, this is the first polynomial-time algorithm to achieve exact (not approximate) invariances in this context. Our proof leverages tools from differential geometry, spectral theory, and optimization. A key result in our development is a new reformulation of the problem of learning under invariances, as optimizing an infinite number of linearly constrained convex quadratic programs, which may be of independent interest.

*These authors contributed equally to this work. Emails: {ashkanso, bzt}@mit.edu.

1 Introduction

While humans can readily observe symmetries or invariances in systems, it is generally challenging for machines to detect and exploit these properties from data. The objective of machine learning with invariances is to develop approaches that enable models to be trained and utilized under the symmetries inherent in the data. This framework is broadly applicable across various domains in the natural sciences and physics, including atomistic systems [21], molecular wavefunctions and electronic densities [51], interatomic potentials [2], and beyond [3]. While many applications involve Euclidean symmetries [46], the scope of such methods extends well beyond them to other geometries [8].

Learning with invariances has a longstanding history in machine learning [22, 27]. In recent years, there has been significant interest in the development and analysis of learning methods that account for various types of invariances. This surge in interest is strongly motivated by many models showing considerable success in practice. Empirical evidence suggests the existence of algorithms that can effectively learn under invariances while exhibiting strong generalization and computational efficiency. However, from a theoretical perspective, much of the focus has been on the expressive power of models, generalization bounds, and sample complexity. There remains a relative lack of understanding regarding the statistical-computational trade-offs in learning under invariances, even in foundational settings such as kernel regression.

Kernels, which have been among popular learning approaches, offer both statistical and computational efficiency [45]. Symmetries can be included in kernel learning in various ways. An immediate solution for learning with invariances seems to be *data augmentation* over the elements of the group. Moreover, most kernel-based approaches to learning with invariances rely on *group averaging*, a technique that involves summing over group elements. However, the typically large size of the group can make both of these approaches computationally prohibitive, even super-exponential in the dimension of input data. Alternative approaches, such as *canonicalization* and *frame averaging*, also suffer from issues like discontinuities and scalability challenges [16].

In light of these challenges, this paper seeks to address the following question:

> Is it possible to obtain an invariant estimator for the kernel regression problem that exhibits both strong generalization capabilities and computational efficiency?

The first contribution of this work is a detailed study of the problem of learning with invariances in the context of kernel methods. We argue that, while group averaging fails to produce exactly invariant estimators within a computationally efficient timeframe, alternative algorithms can generate invariant estimators for the kernel regression problem in time that is polylogarithmic in the size of the group. In other words, we demonstrate that it is possible to achieve an invariant estimator that is both computationally efficient and exhibits strong generalization. At first glance, this result may seem counterintuitive and even impossible, since it implies that enumerating all possible invariances is not required to design statistically efficient learning algorithms. This provides theoretical support for the empirical observation that computational efficiency and strong generalization are attainable in learning with invariances. To the best of our knowledge, this is the first algorithm that is both statistically and computationally efficient for learning with invariances in the kernel setting,

Learning with invariances can be formulated as a *nonconvex optimization* problem, which is not tractable directly. To design an efficient algorithm, we leverage the spectral theory of the Laplace-Beltrami operator on manifolds. Notably, since this operator commutes with all (isometric) group

actions on the manifold, it is possible to find an orthonormal basis of Laplacian eigenfunctions such that each group action on the manifold acts on the eigenspaces of the Laplacian via orthogonal matrices. This theoretical framework allows us to reformulate the original problem of learning with invariances on manifolds as a *infinite collection of finite-dimensional convex quadratic programs*—one for each eigenspace—each *constrained by linear conditions*. By truncating the number of quadratic programs solved, we can efficiently approximate solutions to the primary nonconvex optimization problem, thereby approximating kernel solutions to the problem of learning with invariances. This reformulation not only enables us to derive a polynomial-time algorithm for kernel regression under invariances, but it may also have broader applications, the exploration of which we defer to future research.

Finally, we emphasize again that this work is centered on achieving *exact* invariance, as many applications—especially neural networks with strong empirical performance—are explicitly designed to incorporate exact invariances by construction. In summary, this paper makes the following contributions:

- We initiate the exploration of statistical-computational trade-offs in the context of learning with exact invariances, focusing specifically on kernel regression over manifold-structured input spaces.
- We reformulate the problem of learning under invariances in kernel methods, leveraging differential geometry and spectral theory, and cast it as infinitely many convex quadratic program with linear constraints, for which we derive an efficient solution in terms of time complexity.
- We introduce the first polynomial algorithm for learning with invariances in the general setting of kernel regression over manifolds.

2 Related Work

Generalization bounds and sample complexity for learning with invariances have been extensively studied, particularly in the context of invariant kernels. Works such as Elesedy [17], Bietti et al. [6], Tahmasebi and Jegelka [48], and Mei et al. [33] provide insights into this area. Additionally, studies on equivariant kernels [18, 38] further our understanding of how equivariances affect learning. PAC-Bayesian methods have also been applied to derive generalization bounds under equivariances [4]. More recently, Kiani et al. [25] explored the complexity of learning under symmetry constraints for gradient-based algorithms. For studies on the optimization of kernels under invariances, see Teo et al. [50].

A variety of methods have been proposed to enhance the performance of kernel-based learning models. One prominent approach is the use of random feature models [42], which approximate kernels using randomly selected features. Low-rank kernel approximation techniques, such as the Nyström method [52, 14], have also been proposed to reduce the computational complexity of kernel methods; see also Bach [1], Cesa-Bianchi et al. [10]. Divide-and-conquer algorithms offer another pontential avenue for kernel approximation [54]. Additionally, the impact of kernel approximation on learning accuracy is well-documented in Cortes et al. [13].

Our work focuses on learning with invariances, which differs significantly from the tasks of learning invariances or measuring them in neural networks. For example, Benton et al. [5] address how neural networks can learn invariances, while Goodfellow et al. [20] study methods to measure the degree of invariance in network architectures.

Invariance in kernel methods is not limited to group averaging. Other approaches such as frame averaging [39], canonicalization [24, 32], random projections [15], and parameter sharing

[43] have also been proposed to construct invariant function classes. However, canonicalization and frame averaging face challenges, particularly concerning continuity, which has been addressed in recent works like Dym et al. [16].

In specialized tasks such as graphs, image, and pointcloud data, Graph Neural Networks (GNNs) [44, 53], Convolutional Neural Networks (CNNs) [28, 31], and Pointnet [40, 41] have demonstrated the effectiveness of leveraging symmetries. Symmetries have also been successfully integrated into generative models [7, 35, 26]. For a broader discussion on various types of invariances and their applications across machine learning tasks, see Bronstein et al. [8].

3 Background and Problem Statement

Notation. We begin by establishing some frequently used notation. Let $\mathcal{M}$ be a smooth, compact, and boundaryless Riemannian manifold. The uniform distribution over the manifold is the normalized volume element corresponding to its metric. We denote the space of square-integrable functions over $\mathcal{M}$ by $L^2(\mathcal{M})$ and the space of continuous functions by $C(\mathcal{M})$. Furthermore, $H^s(\mathcal{M})$ represents the Sobolev space of functions on $\mathcal{M}$ with parameter s, defined as the set of functions with square-integrable derivatives up to order s. Larger values of s correspond to greater smoothness, and it holds that $H^s(\mathcal{M}) \subseteq C(\mathcal{M})$ if and only if $s > d/2$, a condition we will assume throughout this paper. For each $n \in \mathbb{N}$, we define $[n] \coloneqq \{1, 2, \ldots, n\}$. We use $\log$ to denote logarithm with base 2. We refer to Appendices A.1 and A.2 for a quick review on Riemannian manifolds.

Problem statement. We consider a general learning setup on a smooth, compact, and boundaryless Riemannian manifold $\mathcal{M}$ of dimension d. Our objective is to identify an estimator $\widehat{f} \in \mathcal{F}$ from a feasible space of estimators $\mathcal{F} \subseteq L^2(\mathcal{M})$, based on n independent and uniformly distributed labeled samples $\mathcal{S} = \{(x_i, y_i) : i \in [n]\} \subseteq (\mathcal{M} \times \mathbb{R})^n$ drawn from the manifold. Here, the labels y_i for $i \in [n]$ are produced based on the (unknown) ground truth regression function $f^\star \in C(\mathcal{M})$, meaning that $y_i = f^\star(x_i) + \epsilon_i$, for each $i \in [n]$, where ϵ_i, $i \in [n]$, is a sequence of independent zero-mean random variables with variance bounded by σ^2. The population risk (or generalization error) of an estimator $\widehat{f} \in L^2(\mathcal{M})$ is defined as:

$$\mathcal{R}(\widehat{f}) \coloneqq \mathbb{E}_{\mathcal{S}} \left[\|\widehat{f} - f^\star\|^2_{L^2(\mathcal{M})} \right],$$

which quantifies the quality of the estimation.

Given a dataset of size n, finding estimators with minimal population risk can be quite complex, often requiring the resolution of non-convex optimization objectives. However, in scenarios where $f^\star \in \mathcal{H}$, with $\mathcal{H} \subseteq L^2(\mathcal{M})$ being a Reproducing Kernel Hilbert Space (RKHS), it is feasible to compute kernel-based estimators with low risk efficiently. Specifically, the Kernel Ridge Regression (KRR) estimator for the RKHS $\mathcal{H} = H^s(\mathcal{M})$, denoted as $\widehat{f}_{\mathrm{KRR}}$, achieves a population risk of $\mathcal{R}(\widehat{f}_{\mathrm{KRR}}) = \mathcal{O}\left(n^{-s/(s+d/2)}\right)$ while being computable in time $\mathcal{O}(n^3)$, assuming access to an oracle that computes the kernel associated with the space. Note that Sobolev spaces $H^s(\mathcal{M})$ with $s > d/2$ are RKHS. We refer the reader to Appendices A.8 and A.9 for a detailed review of the KRR estimator and related topics on Sobolev spaces.

Learning with invariances. We assume that a finite group G acts smoothly and isometrically[1] on the manifold $\mathcal{M}$, represented by a smooth function $\theta : G \times \mathcal{M} \to \mathcal{M}$ mapping the product manifold $G \times \mathcal{M}$ to $\mathcal{M}$. We employ the notation $\theta(g, x)$ as gx for any $g \in G$ and $x \in \mathcal{M}$. In a scenario of

[1]The assumption of isometric action is made for simplicity; the proof can be extended to non-isometric actions using standard techniques in the literature, as discussed in [48].

learning under invariances, the regression function $f^\star$ is invariant under the action of the group G, satisfying $f^\star(gx) = f^\star(x)$ for each $g \in G$ and $x \in \mathcal{M}$. Thus, learning under invariances introduces an additional requirement: not only must we compute an estimator with minimal population risk efficiently, but $\widehat{f}$ must also be invariant with respect to G. This additional condition is often satisfied in neural network applications by constructing networks that are invariant *by design*, such as graph neural networks.

In the context of learning with Sobolev kernels, the KRR estimator $\widehat{f}_{\mathrm{KRR}}$ is *not* G-invariant (see Appendix A for more details). Consequently, the KRR estimator cannot provide a solution for learning under invariances. However, with a shift-invariant Positive Definite Symmetric (PDS) kernel[2] $K : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$, one can utilize *group averaging* to derive a new kernel and a new RKHS holding only G-invariant functions:

$$K_{\mathrm{inv}}(x_1, x_2) \coloneqq \frac{1}{|G|} \sum_{g \in G} K(gx_1, x_2).$$

Given that the Sobolev space $H^s(\mathcal{M})$ adopts a shift-invariant PDS kernel (Appendix A.8), one can apply the above method to construct and compute a G-invariant kernel (assuming access to evaluating its original kernel). This indicates that the KRR estimator on K_{inv} yields an invariant estimator for $f^\star$ with a desirable population risk (see [48] for a comprehensive study).

However, in terms of computational complexity, this method requires $\Omega(n^2|G|)$ time to compute the new kernel between pairs of input data. In many practical scenarios, $|G|$ can be intolerably large. For instance, for the permutation group P_d, we have $|G| = d! \sim \sqrt{d}(\frac{d}{e})^d$ which is super-exponential in d. Consequently, the group averaging method cannot provide an efficient algorithm for learning with exact invariances. We emphasize "exact invariance" here, as the sum involved in K_{inv} can be approximated by summing over a number of random group transformations. However, this does not guarantee exact invariance, which is the primary goal of this paper.

Other traditional approaches to achieve learning under invariances include data augmentation, canonicalization, and frame averaging. For data augmentation, we need to increase our dataset size by a multiplicative factor of $|G|$, which is often impractical within efficient time constraints. This is because for any datapoint $x_i \in \mathcal{S}$, new datapoint gx_i for any group element $g \in G$ should be added to dataset to ensure invariance of the underlying learning procedure in a blackbox way, leading to $\Omega(n|G|)$ complexity. Canonicalization involves mapping data onto the quotient space of the group action and subsequently finding an estimator (e.g., a KRR estimator) on the reduced input space. However, this method is also infeasible for kernels due to the unavoidable discontinuities and non-smoothness of the canonicalization maps, which violate RKHS requirements [16]. Finally, frame averaging is analogous to canonicalization, but it remains unclear how to address continuity issues for efficient frame sizes. Moreover, it requires careful design of frames tailored to the specific problem at hand, making it unsuitable for a general-purpose algorithm. Thus, motivated by these observations, we pose the following question:

> Is it possible to obtain a G-invariant estimator for $f^\star \in H^s(\mathcal{M})$ with a desirable population risk (similar to the case without invariances) in $\mathrm{poly}(n, d, \log(|G|))$ time?

[2] A kernel $K : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$ is termed shift-invariant with respect to group G if and only if $K(gx_1, gx_2) = K(x_1, x_2)$ for each $g \in G$ and $x_1, x_2 \in \mathcal{M}$. Shift-invariant kernels are *not* necessarily G-invariant. For example, one can show that $H^s(\mathcal{M})$ adopts a shift-invariant kernel while still producing non-invariant functions in its RKHS. We cover the details in Appendix A.9.

We aim to answer this question affirmatively in the following section. This is surprising, as it suggests that even enumerating the set G is not required to find statistically efficient G-invariant estimators.

Oracles. To characterize computational complexity, first we need to specify the type of oracle access provided for the estimation. Before doing so, we briefly review the spectral theory of the Laplace-Beltrami operator on manifolds. For further details, we refer the reader to Appendix A.

The Laplace-Beltrami operator generalizes the Laplacian operator to Riemannian manifolds. It has a basis of smooth eigenfunctions $\phi_{\lambda,\ell} \in L^2(\mathcal{M})$, which serve as an orthonormal basis for $L^2(\mathcal{M})$. The index λ represents the eigenvalue corresponding to the eigenfunction $\phi_{\lambda,\ell}$, and $\ell \in [m_\lambda]$ runs over the multiplicity of λ, denoted by m_λ. The eigenvalues can be ordered such that $0 = \lambda_0 < \lambda_1 \leq \lambda_2 \leq \cdots \to \infty$. For example, in the case of the sphere S^{d-1}, the spherical harmonics, which are homogeneous harmonic polynomials, are a natural choice of eigenfunctions.

The sequence of eigenfunctions and their corresponding eigenvalues provide critical information about the geometry of the manifold. In this work, we make use of the following two types of oracles:

- The ability to evaluate any eigenfunction $\phi_{\lambda,\ell}(x)$ at a given point $x \in \mathcal{M}$.
- The ability to compute the $L^2(\mathcal{M})$ inner product between a shifted eigenfunction $\phi_{\lambda,\ell}(gx)$ and another eigenfunction $\phi_{\lambda,\ell'}(x)$ for any group element $g \in G$.

For both oracles, we assume free access as long as $D_\lambda := \sum_{\lambda' \leq \lambda} m_{\lambda'} = \mathrm{poly}(n, d)$, where D_λ denotes the number of eigenfunctions with eigenvalues less than or equal to λ. This assumption is motivated by the case of S^{d-1}, where spherical harmonics can be efficiently evaluated in low dimensions (in such cases, only a few monomials need to be processed, making the task simple). The first oracle handles the geometric structure of the manifold, while the second oracle captures the relationship between the group action and the manifold's spectrum. Both are crucial for obtaining our results.

4 Main Result

In this section, we address the question raised in the previous section by presenting the primary result of the paper, which is encapsulated in the following theorem.

Theorem 1 (Learning with exact invariances in polynomial time)**.** *Consider the problem of learning with invariances with respect to a finite group G using a labeled dataset of size n sampled from a manifold of dimension d. Assume that the optimal regression function belongs to the Sobolev space of functions of order s, i.e., $f^\star \in H^s(\mathcal{M})$ for some $s > d/2$ and let $\alpha := 2s/d$. Then, there exists an algorithm that, given the data, produces an estimator $\widehat{f}$ such that:*

- *It operates in time $\mathcal{O}\big(\log^3(|G|)n^{3/(1+\alpha)} + n^{(2+\alpha)/(1+\alpha)}\big)$;*
- *It achieves an excess population risk (or generalization error) of $\mathcal{R}(\widehat{f}) = \mathcal{O}\big(n^{-s/(s+d/2)}\big)$;*
- *It requires $\mathcal{O}\big(\log(|G|)n^{2/(1+\alpha)} + n^{(2+\alpha)/(1+\alpha)}\big)$ oracle calls to construct the estimator;*
- *For any $x \in \mathcal{M}$, the estimated label $\widehat{f}(x)$ can be computed in time $\mathcal{O}\big(n^{1/(1+\alpha)}\big)$ using $\mathcal{O}\big(n^{1/(1+\alpha)}\big)$ oracle calls.*

The full proof of Theorem 1 is presented in Appendix B.3, while a detailed proof sketch is provided in Section 5, and the algorithm is outlined in Algorithm 1.

Let us interpret the above theorem. Note that without any invariances, the Kernel Ridge Regression (KRR) estimator (details are given in Appendix A.9) provides an estimator $\widehat{f}_{\mathrm{KRR}}$ for the Sobolev space $H^s(\mathcal{M})$ that is computed in time $\mathcal{O}\big(n^3\big)$ and achieves the risk $\mathcal{R}(\widehat{f}_{\mathrm{KRR}}) = \mathcal{O}\big(n^{-s/(s+d/2)}\big)$. Here, while KRR cannot guarantee an exactly invariant estimator, we propose another estimator which is both exactly invariant and also converges with the same rate $\mathcal{O}\big(n^{-s/(s+d/2)}\big)$. As a result, we achieve exact invariances with statistically desirable risk (or sample complexity). In other words, the population risk is the same as the optimal case without invariances, which shows that the algorithm introduces no loss in statistical performance while enforcing group invariances.

We thus came to the following conclusion:

> The problem of learning with exact invariances can be efficiently solved in time $\mathrm{poly}(n, d, \log(|G|))$ and with excess population risk (or generalization error) $\mathcal{O}\big(n^{-s/(s+d/2)}\big)$ which is the same statistical performance as for learning without invariances.

It is worth mentioning that, according to the theorem, the proposed estimator $\widehat{f}$ is not only efficiently achievable but also efficiently computes new predictions on unlabeled data.

Remark 1. We notice that in the proof of Theorem 1, the actual time and sample complexity depends only on the size of minimum generating set of the group G, denoted by $\rho(G)$, instead of $\log(|G|)$. We use logarithm in the theorem just to give better insights to the reader about the improvement from the naive approach. Thus, the actual proof allows to even achieve a tighter result, since $\rho(G) \leq \log(|G|)$ for any finite group (see Proposition 5 in Appendix B.1). Note that for some cases (such as cyclic groups) $\rho(G) \ll \log(|G|)$.

5 Algorithm and Proof Sketch

In this section, we provide a proof sketch for Theorem 1, introducing several new notations and concepts necessary for achieving the reduction in time complexity.

We begin with the most natural optimization program for obtaining an estimator: the Empirical Risk Minimization (ERM), which proposes the following estimator:

$$\widehat{f}_{\mathrm{ERM}} \coloneqq \underset{f \in H^s(\mathcal{M})}{\arg\min} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(f(x_i) - y_i\right)^2 \right\},$$

where $\mathcal{S} = \{(x_i, y_i) : i \in [n]\} \subseteq (\mathcal{M} \times \mathbb{R})^n$ denotes the sampled (labeled) dataset.

However, as discussed, this method does not necessarily produce an estimator that is exactly invariant. A natural idea is to introduce group invariances as constraints into the above optimization, leading to the following constrained ERM solution:

$$\begin{aligned} \widehat{f}_{\text{ERM-C}} \coloneqq \underset{f \in H^s(\mathcal{M})}{\arg\min} & \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(f(x_i) - y_i\right)^2 \right\} \\ \text{s.t.} \quad & \forall (g, x) \in G \times \mathcal{M} : f(gx) = f(x). \end{aligned}$$

While this formulation ensures exact invariance, it introduces $|G|$ functional equations. This is problematic for two reasons: first, $|G|$ constraints are prohibitively many, and second, these

constraints require solving functional equalities, which are not easily achievable. Moreover, the functional equations involve non-linear (pointwise) constraints on the estimator function, which at first glance appear intractable due to nonconvexity of the contraints $f(gx) = f(x)$ for general choice of g.

Therefore, it is necessary to reformulate the above optimization program. The goals of the reformulation are to reduce the number of constraints and encode the functional equations into more tractable constraints, ideally linear ones.

Reducing the number of constraints. We begin by using the following basic property (based on the group law):

$$\Big(\forall g \in \{g_1, g_2\}, \forall x \in \mathcal{M} : f(gx) = f(x)\Big) \implies \Big(\forall x \in \mathcal{M} : f(g_1 g_2 x) = f(x)\Big).$$

This observation allows us to eliminate many unnecessary constraints. Specifically, we only need constraints over a small subset of G if this subset can generate any group element through arbitrary group multiplications. To formalize this, we introduce the following definition:

Definition 1. A finite group G is said to be generated by a subset $S \subseteq G$ if for every $g \in G$, there exists a sequence of elements $s_1, s_2, \ldots, s_k$ such that for each $i \in [k]$, either $s_i \in S$ or $s_i^{-1} \in S$ and $g = s_1 s_2 \cdots s_k$. The minimum size of such a subset S is denoted by $\rho(G)$.

Clearly, $\rho(G) \leq |G|$. However, it can be shown (see Appendix B.1) that $\rho(G) \leq \log(|G|)$, which represents an exponential improvement over the trivial upper bound.

Thus, we can reformulate the constrained ERM optimization as:

$$\begin{aligned} \widehat{f}_{\text{ERM-C}} \coloneqq \underset{f \in H^s(\mathcal{M})}{\arg\min} & \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(f(x_i) - y_i\right)^2 \right\} \\ \text{s.t.} \quad & \forall (g, x) \in S \times \mathcal{M} : f(gx) = f(x), \end{aligned}$$

where $|S| \leq \log(|G|)$. In this way, we reduce the number of constraints from $|G|$ to $\log(|G|)$ by leveraging the concept of minimal group generators. Note that this fact cannot be directly used in data augmentation, group averaging, or canonicalization techniques.

Optimization in the spectral domain. The constrained ERM formulation presented above, while advantageous in terms of reducing the number of constraints, involves optimizing over the infinite-dimensional space $H^s(\mathcal{M})$, which is computationally intractable. One way to make this problem tractable is to parametrize the estimator and search for the optimal parameters. To achieve this, we utilize the spectral theory of the Laplace-Beltrami operator over manifolds. While a detailed discussion of spectral theory is provided in Appendix A, we summarize the relevant concepts here.

As mentioned earlier, the Laplace-Beltrami operator yields a sequence of orthonormal eigenfunctions $\phi_{\lambda,\ell} \in L^2(\mathcal{M})$, where $\lambda \in \{\lambda_0, \lambda_1, \ldots\} \subseteq [0, \infty)$ represents the eigenvalue corresponding to the eigenfunction $\phi_{\lambda,\ell}$, and $\ell \in [m_\lambda]$ indexes the multiplicity of λ, denoted by m_λ. Therefore, any estimator $f \in L^2(\mathcal{M})$ can be expressed as:

$$f(x) = \sum_{\lambda} \sum_{\ell=1}^{m_\lambda} f_{\lambda,\ell} \phi_{\lambda,\ell}(x), \quad f_{\lambda,\ell} \coloneqq \langle f, \phi_{\lambda,\ell} \rangle_{L^2(\mathcal{M})}.$$

The idea is to parametrize the problem by finding the best coefficients $f_{\lambda,\ell}$. However, since there are infinitely many eigenvalues, there are infinitely many parameters to estimate, which is not feasible

in finite time. Fortunately, we know that $f^\star \in H^s(\mathcal{M})$. From the definition of Sobolev spaces (see Appendix A.8), we have:

$$\|f^\star\|^2_{H^s(\mathcal{M})} := \sum_\lambda \sum_{\ell=1}^{m_\lambda} (f^\star_{\lambda,\ell})^2 D_\lambda^\alpha,$$

where $D_\lambda = \sum_{\lambda' \le \lambda} m_{\lambda'}$, and $\alpha := \frac{2s}{d} > 1$.

Thus, we conclude that:

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

for any $D > 0$. This shows that for Sobolev regression functions $f^\star \in H^s(\mathcal{M})$, we can truncate the estimation of coefficients at a certain cutoff frequency λ, which allows the problem to be parametrized with finitely many parameters. Although this introduces bias into the estimation (since higher-frequency eigenfunctions will not be captured), the bias is bounded by the above inequality for Sobolev spaces.

Interestingly, this spectral approach yields a more meaningful optimization problem when considering the population risk function rather than ERM. The population risk, which is the primary objective in regression, is given by:

$$\mathcal{R}(f) = \mathbb{E}_{\mathcal{S}} \left[\|f - f^\star\|^2_{L^2(\mathcal{M})} \right] = \sum_\lambda \sum_{\ell=1}^{m_\lambda} \mathbb{E}[(f_{\lambda,\ell} - f^\star_{\lambda,\ell})^2].$$

Constrained spectral method. To review, we introduced an efficient way to impose the constraints related to group invariances in the ERM objective and later presented spectral methods for obtaining estimators. The last step here is to combine these to achieve exact invariances via a constrained spectral method. We use an important property of the Laplace-Beltrami operator to introduce the algorithm.

Let $\Delta_\mathcal{M}$ denote the Laplace-Beltrami operator on the manifold $\mathcal{M}$, and let G be a group acting isometrically on $\mathcal{M}$. Define the linear operator $T_g : f(x) \mapsto f(gx)$ for each group element $g \in G$ and any smooth function f on the manifold. Then, we have

$$\Delta_\mathcal{M}(T_g \phi) = T_g(\Delta_\mathcal{M}(\phi)),$$

for any smooth function ϕ on the manifold (for a formal proof, please refer to Appendix A.5).

This identity tells us that the Laplace-Beltrami operator $\Delta_\mathcal{M}$ commutes with the operator T_g for each g. Since both operators are linear, spectral theory implies that the commutativity shows the eigenspaces of $\Delta_\mathcal{M}$ are *preserved* under the action of the group G, meaning the operators can be simultaneously diagonalized. Specifically, for any λ, ℓ, and any $g \in G$, the function $\phi_{\lambda,\ell}(gx)$ is a linear combination of eigenfunctions $\phi_{\lambda,\ell'}$, $\ell' \in [m_\lambda]$. In particular, the group G acts via orthogonal matrices on the eigenspace $V_\lambda := \mathrm{span}(\phi_{\lambda,\ell} : \ell \in [m_\lambda])$ for each λ.

Let $D^\lambda(g)$ denote the $m_\lambda \times m_\lambda$ orthogonal matrix corresponding to the action of an element $g \in G$ on V_λ for each λ. Then, a function

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

is G-invariant if and only if

$$D^\lambda(g) f_\lambda = f_\lambda, \quad \forall g \in G \, \forall \lambda \in \{\lambda_0, \lambda_1, \ldots\},$$

where $f_\lambda \coloneqq (f_{\lambda,\ell})_{\ell \in [m_\lambda]} \in \mathbb{R}^{m_\lambda}$ for each λ. We can further reduce the number of conditions by passing G to a generator set, which gives only $\log(|G|)$ conditions.

Thus, the commutativity of the Laplace-Beltrami operator and any isometric group action allows us to introduce only linear constraints on the spectral method to achieve exact invariances. This leads to the following optimization program:

$$\begin{aligned} \min_{f_{\lambda,\ell}} \; & \sum_\lambda \sum_{\ell=1}^{m_\lambda} \mathbb{E}[(f_{\lambda,\ell} - f^\star_{\lambda,\ell})^2], \\ \text{s.t.} \quad & \forall g \in S \, \forall \lambda \in \{\lambda_0, \lambda_1, \ldots\} : D^\lambda(g) f_\lambda = f_\lambda. \end{aligned}$$

Here, $f^\star_{\lambda,\ell} = \mathbb{E}_x[f^\star(x)\phi_{\lambda,\ell}(x)] = \mathbb{E}_{x,y}[y\phi_{\lambda,\ell}(x)]$ is not known a priori; only n samples $(x_i, y_i) \in \mathcal{M} \times \mathbb{R}$, $i \in [n]$, are given. Furthermore, the constraints are independent for different eigenspaces (i.e., different λ), and the objective is a sum over eigenspaces. This means we can decompose the problem into a set of linearly constrained optimization programs, one for each eigenspace V_λ:

$$\begin{aligned} \min_{f_{\lambda,\ell}} \; & \sum_{\ell=1}^{m_\lambda} \mathbb{E}[(f_{\lambda,\ell} - f^\star_{\lambda,\ell})^2], \\ \text{s.t.} \quad & \forall g \in S : D^\lambda(g) f_\lambda = f_\lambda. \end{aligned}$$

This reformulation allows us to propose efficient estimators for the problem.

Empirical estimator. In this paper, we suggest the following auxiliary empirical mean estimator from the data for the above optimization program on V_λ:

$$\widetilde{f}_{\lambda,\ell} = \frac{1}{n} \sum_{i=1}^{n} y_i \phi_{\lambda,\ell}(x_i), \quad \forall \ell \in [m_\lambda]. \tag{1}$$

Moreover, we stop estimation and set $\widetilde{f}_{\lambda,\ell} = 0$ when $D_\lambda > D$, where D is a hyperparameter. To find a G-invariant using our primary estimator, we solve the following quadratic program to find a solution satisfying the constraints for each V_λ with $D_\lambda \leq D$:

$$\begin{aligned} \widehat{f}_{\lambda,\ell} \coloneqq \arg\min_{f_{\lambda,\ell}} \; & \sum_{\ell=1}^{m_\lambda} (f_{\lambda,\ell} - \widetilde{f}_{\lambda,\ell})^2, \\ \text{s.t.} \quad & \forall g \in S : D^\lambda(g) f_\lambda = f_\lambda. \end{aligned}$$

This optimization problem is a convex quadratic program with linear constraints that can be solved iteratively using the rich convex optimization machinery. Additionally, it has a closed-form solution as noted in Proposition 6 in Appendix B.2. Let $B^\lambda \in \mathbb{R}^{|S| m_\lambda \times m_\lambda}$ defined as the augmented matrix resulted by concatenating $D^\lambda(g) - I$ for all $g \in S$ on top of each other, i.e., $B^\lambda = [(D^\lambda(g_1) - I)^\top, (D^\lambda(g_2) - I)^\top, \ldots, (D^\lambda(g_{|S|}) - I)^\top]^\top$. Then,

$$\widehat{f}_{\lambda,\ell} = \widetilde{f}_{\lambda,\ell} - {B^\lambda}^\top (B^\lambda {B^\lambda}^\top)^\dagger (B^\lambda \widetilde{f}_\lambda)[\ell],$$

where $\dagger$ denotes Moore–Penrose inverse.

The final estimator of the algorithm is given by

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

This meta approach to design G-invariant estimator $\widehat{f}$ from any primary estimtor $\widetilde{f}$ is novel and can be of independent interest. A pseudocode for this method is presented in Algorithm 1.

Algorithm 1 Learning with Exact Invariances

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

Output: Output $\widehat{f}(x)$.

Initialize $D \leftarrow n^{1/(1+\alpha)}$.
for each λ such that $D_\lambda \le D$ **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
for each λ such that $D_\lambda \le D$ **do**
  Solve the following linearly constrained quadratic program over m_λ variables:

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

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

We conclude this section by reviewing how we apply the results from Algorithm 1 to the two following important examples.

Example 1. Consider the problem of learning under invariances over the unit sphere $S^{d-1} := \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$, where the group G is the group of all permutations of coordinates. Note that $|G| = d!$, which is prohibitively large for data augmentation or group averaging. However, this group is generated by only two elements: $\sigma_1 = (1\,2)$ and $\sigma_2 = (1\,2\,\ldots\,d)$. Here, σ_1 swaps the first and second coordinates, while σ_2 is a cycle that maps $1 \to 2$, $2 \to 3$, and so on, cyling with $d \to 1$.

The eigenspaces V_λ for the sphere are precisely the sets of homogeneous harmonic polynomials of degree k, where $\lambda = k(k + d - 2)$. The permutation group acts on V_λ by permuting the variables of the polynomials. This action is clearly linear, and the matrices $D^\lambda(g)$ can be efficiently computed (using tensor products) as long as k is small. Moreover, homogeneous polynomials of degree k can also be computed efficiently for small k. This shows that the oracles considered in this paper align perfectly with what we observe in the important case of spheres and polynomial regression. In Algorithm 1, we first compute the coefficients of each polynomial for degree k, up to a small k, and then solve a quadratic program with only two linear constraints to obtain an exactly invariant polynomial solution.

Example 2. Consider the same setup as the previous example but assume $d = 2$, i.e., the manifold is the unit circle. In this case, each eigenspace V_λ is spanned by $\sin(k\theta)$ and $\cos(k\theta)$, where $\lambda = k^2$. Let us assume our task is to find an estimator invariant with respect to rotations by integer multiples of

$\frac{2\pi}{|G|}$. This group is cyclic and is generated by only one element $g_0 = \frac{2\pi}{|G|}$. Thus, we have only one constraint for each eigenspace. Indeed, one can observe that $D^{\lambda}(g_0) = R(k\frac{2\pi}{|G|})$, where $R(.) \in \mathbb{R}^{2\times 2}$ is the two-dimensional rotation matrix. Thus, this example further illustrates how our oracles are defined to solve the problem.

6 Discussion and Future Directions

We initiated the study on computational-statistical trade-offs in learning with exact invariances. We designed an algorithm that shows achieving the desirable population risk (the same as kernel regression without invariances) in $\mathrm{poly}(n, d, \log(|G|))$ time for the task of kernel regression with invariances on general manifolds. We note that, for simplicity, we have focused on boundaryless manifolds and isometric group actions. However, using standard techniques, the theory can be extended to more general cases as well[3]. While the proposed spectral algorithm is computationally efficient, it does not offer any improvement in sample complexity over the baseline $\mathcal{R}(\widehat{f}) = \mathcal{O}\big(n^{-s/(s+d/2)}\big)$. It has been observed that without computational constraints, better convergence rates are possible for learning with invariances [48], which are minimax optimal. Thus, it remains open whether those improved rates are achievable in $\mathrm{poly}(n, d, \log(|G|))$ time.

We note that the oracle access we assumed is primarily motivated by the case of the sphere, where polynomials can be evaluated, multiplied, composed by group elements, and integrated efficiently when they are of relatively low degree. We believe this is the most natural oracle access for this problem, as it aligns well with applications involving polynomials. An interesting future work could be to investigate the statistical-computational trade-offs using alternative oracles, e.g., similar to the kernel trick, how to design computationally efficient algorithms that have only access to the inner product of the RKHS. Another interesting future direction is to find whether random feature models as approximation for kernels, can significantly improve the statistical-computational trade-off of learning with invariances? At present, our theory does not apply to random feature models.

We also observe that the spectral algorithm used in this paper does not employ the kernel trick, as it requires access to the entire set of features, rather than just their inner products. An interesting question is whether it is possible to utilize kernel tricks and find an alternative (polynomial-time) algorithm for learning under invariances. This approach could potentially improve the statistical efficiency of the spectral algorithm. In the end, we would like to note that capturing computational-statistical trade-offs in other estimation problems with invariances such as density estimation [12, 49] could serve as a compelling avenue for future research.

7 Conclusion

In this paper, we explore the statistical-computational trade-offs in learning with invariances, focusing specifically on kernel regression. We observe that while the Kernel Ridge Regression (KRR) estimator can address this problem, it is not necessarily invariant without group averaging. Furthermore, since performing group averaging can be costly for large groups, we ask whether it is possible to develop statistically sound estimators with efficient time complexity. Our findings show that by reformulating the problem and reducing the number of constraints using group laws, we can express it as solving an infinite series of quadratic optimization programs under

[3] See e.g., Tahmasebi and Jegelka [48].

linear constraints. We conclude with an algorithm that achieves an exactly invariant estimator with polynomial time complexity and highlight several additional questions for future research.

8 Reproducibility and Ethics Statement

The main focus of this work is the theoretical understanding of computational-statistical trade-offs in learning with invariances; therefore, there are not significant ethical concerns. To ensure reproducibility, we have provided a detailed proof sketch of the algorithm in the main text. Additionally, all proofs and relevant background information are discussed in detail in the appendix.

References

[1] Francis Bach. Sharp analysis of low-rank kernel matrix approximations. In *Conference on Learning Theory (COLT)*, 2013. 3

[2] Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E Smidt, and Boris Kozinsky. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature communications*, 13(1): 2453, 2022. 2

[3] Simon Batzner, Albert Musaelian, and Boris Kozinsky. Advancing molecular simulation with equivariant interatomic potentials. *Nature Reviews Physics*, 5(8):437–438, 2023. 2

[4] 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. 3

[5] Gregory Benton, Marc Finzi, Pavel Izmailov, and Andrew G Wilson. Learning invariances in neural networks from training data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 3

[6] 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. 3

[7] Marin Biloš and Stephan Günnemann. Scalable normalizing flows for permutation invariant densities. In *Int. Conference on Machine Learning (ICML)*, 2021. 4

[8] 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, 4

[9] Yaiza Canzani. Analysis on manifolds via the laplacian. *Lecture Notes available at: http://www. math. harvard. edu/canzani/docs/Laplacian. pdf*, pages 41–44, 2013. 21

[10] Nicolo Cesa-Bianchi, Yishay Mansour, and Ohad Shamir. On the complexity of learning with kernels. In *Conference on Learning Theory (COLT)*, 2015. 3

[11] Isaac Chavel. *Eigenvalues in Riemannian geometry*. Academic press, 1984. 19, 20

[12] 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. 12

[13] Corinna Cortes, Mehryar Mohri, and Ameet Talwalkar. On the impact of kernel approximation on learning accuracy. In *Int. Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010. 3

[14] Petros Drineas, Michael W Mahoney, and Nello Cristianini. On the nyström method for approximating a gram matrix for improved kernel-based learning. *Journal of Machine Learning Research*, 2005. 3

[15] Nadav Dym and Steven J Gortler. Low-dimensional invariant embeddings for universal geometric learning. *Foundations of Computational Mathematics*, pages 1–41, 2024. 3

[16] 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, 4, 5

[17] Bryn Elesedy. Provably strict generalisation benefit for invariance in kernel methods. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 3

[18] Bryn Elesedy and Sheheryar Zaidi. Provably strict generalisation benefit for equivariant models. In *Int. Conference on Machine Learning (ICML)*, 2021. 3

[19] Lawrence C Evans. *Partial differential equations*, volume 19. American Mathematical Society, 2022. 20

[20] Ian Goodfellow, Honglak Lee, Quoc Le, Andrew Saxe, and Andrew Ng. Measuring invariances in deep networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2009. 3

[21] Andrea Grisafi, David M Wilkins, Gábor Csányi, and Michele Ceriotti. Symmetry-adapted machine learning for tensorial properties of atomistic systems. *Physical review letters*, 120(3): 036002, 2018. 2

[22] 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

[23] Lars Hörmander. The spectral function of an elliptic operator. In *Mathematics Past and Present Fourier Integral Operators*, pages 217–242. Springer, 1968. 21

[24] 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. 3

[25] 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. 3

[26] Jonas Köhler, Leon Klein, and Frank Noé. Equivariant flows: exact likelihood generative learning for symmetric densities. In *Int. Conference on Machine Learning (ICML)*, 2020. 4

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

[28] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2012. 4

[29] John Lee. *Introduction to Smooth Manifolds*, volume 218. Springer Science & Business Media, 2012. 17

[30] John M Lee. *Riemannian manifolds: an introduction to curvature*, volume 176. Springer Science & Business Media, 2006. 17

[31] 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. 4

[32] 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. 3

[33] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Learning with invariances in random features and kernel models. In *Conference on Learning Theory (COLT)*, 2021. 3

[34] Sumner B Myers and Norman Earl Steenrod. The group of isometries of a riemannian manifold. *Annals of Mathematics*, 40(2):400–416, 1939. 19

[35] 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. 4

[36] Richard S Palais. On the differentiability of isometries. *Proceedings of the American Mathematical Society*, 8(4):805–807, 1957. 19

[37] Peter Petersen. Riemannian geometry. *Graduate Texts in Mathematics/Springer-Verlarg*, 2006. 17, 19

[38] Mircea Petrache and Shubhendu Trivedi. Approximation-generalization trade-offs under (approximate) group equivariance. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 3

[39] 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. 3

[40] 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)*, 2017. 4

[41] 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)*, 2017. 4

[42] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2007. 3

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

[44] 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. 4

[45] Bernhard Scholkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2018. 2

[46] Tess E Smidt. Euclidean symmetry and equivariance in machine learning. *Trends in Chemistry*, 3(2):82–85, 2021. 2

[47] Christopher D Sogge. Concerning the lp norm of spectral clusters for second-order elliptic operators on compact manifolds. *Journal of functional analysis*, 77(1):123–138, 1988. 21

[48] Behrooz Tahmasebi and Stefanie Jegelka. The exact sample complexity gain from invariances for kernel regression. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 3, 4, 5, 12, 21, 22

[49] Behrooz Tahmasebi and Stefanie Jegelka. Sample complexity bounds for estimating probability divergences under invariances. In *Int. Conference on Machine Learning (ICML)*, 2024. 12

[50] Choon Teo, Amir Globerson, Sam Roweis, and Alex Smola. Convex learning with invariances. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2007. 3

[51] Oliver Unke, Mihail Bogojeski, Michael Gastegger, Mario Geiger, Tess Smidt, and Klaus-Robert Müller. Se(3)-equivariant prediction of molecular wavefunctions and electronic densities. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2

[52] Christopher Williams and Matthias Seeger. Using the nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2000. 3

[53] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *Int. Conference on Learning Representations (ICLR)*, 2019. 4

[54] Yuchen Zhang, John Duchi, and Martin Wainwright. Divide and conquer kernel ridge regression. In *Conference on Learning Theory (COLT)*, 2013. 3

A Background

A.1 Riemannian Manifolds

In this section, we review some fundamental definitions from differential geometry and refer the reader to Lee [30], Petersen [37], Lee [29] for further details.

Definition 2 (Manifold)**.** A topological *manifold* $\mathcal{M}$ of dimension $\dim(\mathcal{M})$ is a completely separable Hausdorff space that is locally homeomorphic to an open subset of Euclidean space of the same dimension, specifically $\mathbb{R}^{\dim(\mathcal{M})}$. More formally, for each point $x \in \mathcal{M}$, there exists an open neighborhood $U \subseteq \mathcal{M}$ and a homeomorphism $\phi : U \to \widehat{U}$, where $\widehat{U} \subseteq \mathbb{R}^{\dim(\mathcal{M})}$.

The value $\dim(\mathcal{M})$ is referred to as the *dimension* of the manifold. Examples of manifolds include tori, spheres, $\mathbb{R}^d$, and graphs of continuous functions. *Manifolds with boundaries* differ from boundaryless manifolds in that they may have neighborhoods that locally resemble open subsets of closed $\dim(\mathcal{M})$*-dimensional upper half-spaces*, denoted as $\mathbb{H}^{\dim(\mathcal{M})} \subseteq \mathbb{R}^{\dim(\mathcal{M})}$, defined as follows:

$$\mathbb{H}^d = \{(x_1, x_2, \ldots, x_d) \in \mathbb{R}^d \mid x_d \geq 0\}.$$

Definition 3 (Local Coordinates)**.** Given a *chart* (U, ϕ)—a pair consisting of a local neighborhood U and the corresponding homeomorphism $\phi : U \to \widehat{U}$—on a manifold $\mathcal{M}$ with dimension d, we define *local coordinates* $(x^1, x^2, \ldots, x^d)$ such that

$$\phi(p) = (x^1(p), x^2(p), \ldots, x^d(p)),$$

for each point $p \in U$.

Definition 4 (Tangent Space)**.** At each point $x \in \mathcal{M}$, the *tangent space* $T_x\mathcal{M}$ is defined as the vector space formed by the tangent vectors to the manifold $\mathcal{M}$ at x. A tangent vector $v \in T_x\mathcal{M}$ can be represented as the derivative of a smooth curve $\gamma(t) : (-\epsilon, \epsilon) \to \mathcal{M}$ defined on the manifold with the property that $\gamma(0) = x$. It is expressed as

$$\nu = \frac{d}{dt}\gamma(t)\bigg|_{t=0}.$$

The tangent space $T_x\mathcal{M}$ is a real vector space with dimension $\dim(\mathcal{M})$.

Definition 5 (Riemannian Metric Tensor)**.** A *Riemannian metric tensor* g[4] on a manifold $\mathcal{M}$ is a smooth inner product defined on the tangent space $T_x\mathcal{M}$ at each point $x \in \mathcal{M}$. For any two tangent vectors $u, v \in T_x\mathcal{M}$, the metric assigns a real number $g_x(u, v) \in \mathbb{R}$.

Definition 6 (Riemannian Manifold)**.** A *Riemannian manifold* is defined as a pair $(\mathcal{M}, g)$, where $\mathcal{M}$ is a manifold and g is a *Riemannian metric tensor* defined on the tangent space $T_x\mathcal{M}$ at each point $x \in \mathcal{M}$.

A Riemannian metric tensor provides essential tools for the study of manifolds, which we formalize below. It enables the following:

- the definition of the geodesic distance $d(x, y)$ between any two points $x, y \in \mathcal{M}$ on the manifold,

[4]This notation differs from g, which denotes group elements.

- a volume element $d\,\mathrm{vol}_g(x)$ over the manifold, serving as the measure for the Borel sigma-algebra over open subsets of the manifold $\mathcal{M}$, and
- the measurement of the angle between any two tangent vectors $u, v \in T_x\mathcal{M}$, which in turn provides the size of tangent vectors.

Definition 7 (Geodesic Distance)**.** The *geodesic distance* $d_{\mathcal{M}}(x, y)$ between any two points $x, y \in \mathcal{M}$ on the manifold is defined as the infimum length among all smooth curves $\gamma : [0, 1] \to \mathcal{M}$ connecting x to y ($\gamma(0) = x, \gamma(1) = y$). The length of a curve γ is defined as

$$L(\gamma) = \int_0^1 \sqrt{g_{\gamma(t)}\left(\dot{\gamma}, \dot{\gamma}\right)}\, dt,$$

where $\dot{\gamma}$ denotes the derivative $\frac{d\gamma}{dt}$.

Definition 8 (Volume Element)**.** The *volume element* $d\mathrm{vol}_g(x)$ on a Riemannian manifold $(\mathcal{M}, g)$ is defined as

$$d\mathrm{vol}_g = \sqrt{\det(g_{ij})}\, dx^1 \wedge dx^2 \wedge \cdots \wedge dx^n,$$

where g_{ij} are the components of the Riemannian metric tensor, $(x^1, x^2, \ldots, x^n)$ are the local coordinates, and $\wedge$ denotes the exterior product.

The volume element provides a way to compute the volume of subsets of $\mathcal{M}$ by integrating functions over $\mathcal{M}$. Moreover, a Borel measure μ on open subsets of $\mathcal{M}$ can be derived form the volume element to form probability measure space, e.g., uniformly over the manifold.

Definition 9 (Smooth Map)**.** A maping $f : \mathcal{M} \to \mathcal{N}$ is a smooth map if for any charts (U, ϕ) on $\mathcal{M}$, and (V, ψ) on $\mathcal{N}$, the composition function $\psi \circ f \circ \phi^{-1} : \mathbb{R}^{\dim(\mathcal{M})} \to \mathbb{R}^{\dim(\mathcal{N})}$ is infinitely differentiable.

Definition 10 (Pullback of the metric tensor)**.** Given Riemannian manifolds $\mathcal{M}$, $(\mathcal{N}, g)$ and $\varphi : \mathcal{M} \to \mathcal{N}$ a smooth map between them. The *pullback of the metric tensor* g by ϕ, denoted by $\varphi^* g$ is the Riemannian metric tensor on manifold $\mathcal{M}$ defined by,

$$(\varphi^* g)_x(u, v) = g_{\varphi(x)}(d\varphi_x(u), d\varphi_x(v)), \text{ for all points } x \in \mathcal{M} \text{ and all } u, v \in T_x\mathcal{M},$$

where $d\varphi_x : T_x\mathcal{M} \to T_{\varphi(x)}\mathcal{N}$ is the differential of the map φ at point x.

Thus, the pullback metric $\varphi^* g$ on $\mathcal{M}$ captures the relation between tangent vectors of $\mathcal{M}$ in terms of how they are mapped to the manifold $\mathcal{N}$ via φ.

Definition 11 (Connected Manifold)**.** A manifold $\mathcal{M}$ is *connected* if for any two points $x, x' \in \mathcal{M}$, there is a smooth curve $\gamma : [0, 1] \to \mathcal{M}$ such that $\gamma(0) = x$ and $\gamma(1) = x'$.

Throughout this paper, we focus on smooth, connected, compact and boundaryless Riemannian manifolds (M, g) unless stated otherwise. For a Riemannian manifolds (M, g), we denoted the dot product induced by the metric tensor g as $\langle u, v\rangle_{g_x} = g_x(u, v)$ for all $u, v \in T_x$. We drop the subscript x whenever it is clear from the context.

A.2 Functional Spaces over Manifolds

Now equipped with probability measures on manifold discussed in Appendix A.1, we are ready to define functional spaces $L^p(\mathcal{M})$ and Sobolev spaces $\mathcal{H}^s(\mathcal{M})$ on manifold $\mathcal{M}$ analogously to their Euclidean counterparts in the following,

Definition 12 (Functional Spaces on Manifolds)**.** The Lebesgue functional spaces $L^p(\mathcal{M})$ for $p \in [1, \infty]$, and the Sobolev spaces $H^s(\mathcal{M})$ for $s \geq 0$ on a smooth Manifold $\mathcal{M}$, are defined as follows:

- The *Lebesgue space* $L^p(\mathcal{M})$ consists of measurable functions $f : \mathcal{M} \to \mathbb{R}$ such that $\|f\|_{L^p(\mathcal{M})} < \infty$ where,

$$\|f\|_{L^p(\mathcal{M})} = \begin{cases} \left(\int_{\mathcal{M}} |f(x)|^p \, d\mu(x)\right)^{1/p} & \text{if } p \in [1, \infty) \\ \operatorname{ess\,sup}_{x \in \mathcal{M}} |f(x)| < \infty. & \text{if } p = \infty \end{cases},$$

  where μ is the uniform measure over the manifold $\mathcal{M}$.

- The *Sobolev space* $H^s(\mathcal{M})$ consists of measurable functions whose derivatives up to order s are in $L^2(\mathcal{M})$, i.e.,

$$H^s(\mathcal{M}) = \left\{ f \in L^2(\mathcal{M}) \mid D^\alpha f \in L^2(\mathcal{M}) \text{ for all multi-indices } \alpha \text{ with } |\alpha| \leq s \right\}.$$

A.3 Lie Group of Isometry Maps

In this section, first we state basic definition of isometric mappings over manifolds and then wrap up by characterizing the isometry group over the manifold.

Definition 13 (Isometry Map)**.** A bijective mapping $\tau : \mathcal{M} \to \mathcal{M}$ is an *isometry* on the manifold ($\mathcal{G}$, g) if $d(\tau(x), \tau(x')) = d(x, x')$.

We also state a brief definition of Lie groups for completeness.

Definition 14 (Lie group)**.** A group G is a *Lie group* with smooth group operations (multiplication and inversion) if it is additionally a smooth manifold.

The space of bijective Riemannian isometries defined on the manifold $(\mathcal{M}, g)$, denoted by $\mathrm{ISO}(\mathcal{M}, g)$ constitutes a group with composition operation. The celebrated Myers–Steenrod theorem states that any isometry map $\tau \in \mathrm{ISO}(\mathcal{M}, g)$ between connected manifolds is an isometry [34, 36]. Myers and Steenrod [34] took it a step further and proved that isometry group of a Riemannian manifold $(\mathcal{M}, g)$ is a Lie group.

Alternatively, $\mathrm{ISO}(\mathcal{M}, g)$ can be charecterized by the pullback of the metric tensor. In terms, $\tau \in \mathrm{ISO}(\mathcal{M}, g)$ if and only if $g = \tau^* g$ [37].

A.4 Laplacian on Manifolds

In this section, we reiterate over definition of Laplace-Beltrami operator on manifolds (which is the generalization of the Laplacian operator $\Delta = \partial_1^2 + \partial_2^2 + \cdots + \partial_d^2$ defined on the Euclidean space $\mathbb{R}^d$) and state a several interesting properties that will utilize later. We refer to Chavel [11] for additional details.

Definition 15 (Laplace-Beltrami operator)**.** Given a Riemannian manifold $(\mathcal{M}, g)$, the *Laplace-Beltrami* operator $\Delta_g : \mathcal{H}^s(\mathcal{M}) \to \mathcal{H}^{s-2}(\mathcal{M})$ acts on a smooth function $f : \mathcal{M} \to \mathbb{R}$ by

$$\Delta_g f = \mathrm{div}_g(\mathrm{grad}_g(f)).$$

Moreover, $\Delta_g f$ has an equivalent weak formulation [19], as the unique continuous linear operator $\Delta_g : \mathcal{H}^s(\mathcal{M}) \to \mathcal{H}^{s-2}(\mathcal{M})$ which is a solution to the equation,

$$\int_{\mathcal{M}} \psi(x)\Delta_g\phi(x) d\,\mathrm{vol}_g(x) + \int_{\mathcal{M}} \langle \nabla_g\psi(x), \nabla_g\phi(x)\rangle_g d\,\mathrm{vol}_g(x) = 0, \forall \phi, \psi \in \mathcal{H}^s(\mathcal{M}). \tag{2}$$

The Laplace-Beltrami operator Δ_g is self-adjoint, eliptic and diagonalizable in $L^p(\mathcal{M})$ [11, 19], yielding a sequence of orthonormal eigenfunctions $\phi_{\lambda,\ell} \in L^2(\mathcal{M})$, where $\lambda \in \{\lambda_0, \lambda_1, \ldots\} \subseteq [0, \infty)$ represents the eigenvalue corresponding to the eigenfunction $\phi_{\lambda,\ell}$, and $\ell \in [m_\lambda]$ indexes the multiplicity of λ, denoted by m_λ such that $\Delta_g\phi_{\lambda_i,\ell} + \lambda_\ell\phi_{\lambda_i,\ell} = 0$ for all $\ell \in \{1, \ldots, m_{\lambda_i}\}$. Note that the basis starts with the constant function $\phi_0 \equiv 1$ and $\lambda_0 = 0$. Hence, one can write $\Delta_g f = -\sum_{i=0}^{\infty}\sum_{\ell=1}^{m_{\lambda_i}} \lambda_i\langle f, \phi_{\lambda_i,\ell}\rangle\phi_{\lambda_i,\ell}$.

Lemma 2. *For any function $f \in L^2(\mathcal{M})$, such that f is decomposed into the basis $\{\phi_{\lambda,\ell}\}_{\lambda=1}^{\infty}$ as $f = \sum_{i=0}^{\infty}\sum_{\ell=1}^{m_{\lambda_i}}\langle f, \phi_{\lambda_i,\ell}\rangle_{L^2(\mathcal{M})}\phi_{\lambda,\ell}$, we know that*

$$\|\nabla_g f\|^2_{L^2(\mathcal{M})} = \sum_{i=0}^{\infty}\sum_{\ell=1}^{m_{\lambda_i}} \lambda_i\langle f, \phi_{\lambda_i,\ell}\rangle^2_{L^2(\mathcal{M})},$$

for convergent summations.

Proof. By Equation (2),

$$\begin{aligned}
\|\nabla_g f\|^2_{L^2(\mathcal{M})} &= \int_{\mathcal{M}} \langle \nabla_g f(x), \nabla_g f(x)\rangle_g \, d\,\mathrm{vol}_g(x) \\
&= -\int_{\mathcal{M}} f(x)\Delta_g f(x) \, d\,\mathrm{vol}_g(x) \\
&= \sum_{i=0}^{\infty}\sum_{\ell=1}^{m_{\lambda_i}} \lambda_i\langle f, \phi_{\lambda_i,\ell}\rangle^2_{L^2(\mathcal{M})}.
\end{aligned}$$

□

A.5 Commutativity of Laplacian and Isometric Group Actions

Let G be a group acting isometrically on a compact, smooth, boundaryless manifold $\mathcal{M}$. As we stated in the main body of the paper, we have $\Delta_{\mathcal{M}}(T_g\phi) = T_g(\Delta_{\mathcal{M}}(\phi))$ for each smooth function ϕ on manifold $\mathcal{M}$, where $T_g\phi = \phi(gx)$. To see why, note that by Equation (2), this is equivalent to showing that

$$\int_{\mathcal{M}} h\Delta_{\mathcal{M}}(T_g\phi) d\,\mathrm{vol}_g(x) = \int_{\mathcal{M}} hT_g(\Delta_{\mathcal{M}}(\phi)) d\,\mathrm{vol}_g(x), \tag{3}$$

for each smooth function h on manifold $\mathcal{M}$. By changing the variables in the integrable and noting that $dx = d(gx)$ from isometry, we have

$$\begin{aligned}
\int_{\mathcal{M}} h\Delta_{\mathcal{M}}(T_g\phi) d\,\mathrm{vol}_g(x) &= -\int_{\mathcal{M}} \langle \nabla h, \nabla T_g\phi\rangle_g d\,\mathrm{vol}_g(x) &(4)\\
&= -\int_{\mathcal{M}} \langle \nabla T_{g^{-1}} h, \nabla\phi\rangle_g d\,\mathrm{vol}_g(x) &(5)\\
&= \int_{\mathcal{M}} T_{g^{-1}} h\Delta_{\mathcal{M}}(\phi) d\,\mathrm{vol}_g(x) &(6)\\
&= \int_{\mathcal{M}} hT_g(\Delta_{\mathcal{M}}(\phi)) d\,\mathrm{vol}_g(x). &(7)
\end{aligned}$$

A.6 Weyl's Law under Invariances

Weyl's law characterizes the asymptotic distribution of the eigenvalues in a closed-form formula [23, 47, 9]. Let us denote dimension of the space spanned by the eigenvectors corresponding to eigenvalue of the Laplace-Beltrami operator up to λ as

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

Theorem 3 (Weyl's law [23, 47, 9])**.** *Let $(\mathcal{M}, g)$ be a compact, boundaryless d-dimensional Riemannian manifold. The asymptotic behavior of dimension count D_λ follows*

$$D_\lambda = \frac{\omega_d \ \mathrm{vol}(\mathcal{M})}{(2\pi)^d} \lambda^{d/2} + \mathcal{O}(\lambda^{(d-1)/2}),$$

where $\omega_d = \frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)}$ is the volume of the unit d-dimensional ball in $\mathbb{R}^d$, $vol(\mathcal{M})$ is the Riemannian volume of $\mathcal{M}$, and $\mathcal{O}(\lambda^{(d-1)/2})$ represents the error term.

Define $D_{\lambda,G}$ as the dimension of the space induced by projection of the corresponding eigenspaces of D_λ into the space of G-invariant functions. Tahmasebi and Jegelka [48] proved the following characterization over this dimension as $\lambda \to \infty$.

Theorem 4 (Dimension counting [48])**.** *Let $(\mathcal{M}, g)$ be a compact, boundaryless d-dimensional Riemannian manifold, and G be a compact finite Lie group acting isometrically on $(\mathcal{M}, g)$. Then.*

$$D_{\lambda,G} = \frac{\omega_d \ \mathrm{vol}(\mathcal{M}/G)}{(2\pi)^d} \lambda^{d/2} + \mathcal{O}(\lambda^{(d-1)/2}),$$

as $\lambda \to \infty$, where again ω_d is the volume of the unit d-dimensional ball in $\mathbb{R}^d$.

A.7 Sobolev Spaces on Manifolds

The ordinary definition of Sobolev spaces on manifolds deals with having square-integrable derivatives up to an order s. Here, since our focus is on spectral approach, we present the spectral definition of Sobolev spaces.

Definition 16 (Sobolev spaces)**.** The Sobolev space of functions $H^s(\mathcal{M})$ on a compact, smooth, boundaryless Riemannian manifold $\mathcal{M}$ is 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 $\alpha \coloneqq 2s/d$.

Note that the above definition is equivalent to the other definition of the Sobolev spaces that involves considering λ^s instead of D_λ^α above. Using Weyl's law (see Appendix A.6), one can show that both definitions are equivalent.

A.8 Sobolev Kernels

Sobolev spaces are RKHS when $s > d/2$. Indeed, the Sobolev kernel can be defined as:

$$K_{H^s(\mathcal{M})}(x, y) \coloneqq \sum_{\lambda} \sum_{\ell=1}^{m_\lambda} D_\lambda^{-\alpha} \phi_{\lambda,\ell}(x) \phi_{\lambda,\ell}(y).$$

Note that any group G that acts isometrically on the manifold, also acts on the eigenspace of Laplacian via orthogonal matrices. Since orthogonal matrices preserve the inner product we conclude that

$$K_{H^s(\mathcal{M})}(gx, gy) = K_{H^s(\mathcal{M})}(x, y),$$

for any $g \in G$, which means that the Sobolev kernel is shift-invariant. However, this is clearly not G-invariant since it produces small bump functions, which need not be invariant.

A.9 Kernel Ridge Regresion (KRR)

Consider a Positive-Definite Symmetric (PDS) kernel $K(.,.)$ on a smooth, compact, boundaryless manifold with H denoting its RKHS. The objective of Kernel Ridge Regression (KRR) estimator is to introduce the RKHS norm to the ERM objective to make sure of finding smooth interpolators:

$$\min_{f \in H} \Big\{ \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \eta \|f\|_H^2 \Big\}, \tag{8}$$

where η denotes the regularization parameter that balances the bias and variance terms. Here, the objective function takes a closed-form solution to the represented theorem for kernels. This gives an efficient estimator, which is termed KRR in the literature.

However, this estimator need not be G-invariant even when trained on invariant data. To see why, note that as long as the space H includes non-invariant functions, there is a chances that we find a non-invariant function optimizing the above objective due to the observation noise. Thus, the only way to make sure that the KRR estimator is G-invariant is to impose the assumption of having G-invariant kernels, which translated to group averaging over the Sobolev kernel:

$$K^G_{H^s(\mathcal{M})}(x, y) \coloneqq \frac{1}{|G|} \sum_{g \in G} K_{H^s(\mathcal{M})}(gx, y). \tag{9}$$

This method is unfortunately not computationally feasible, even though it achieves minimax optimal generalization bounds for learning under invariances with kernels [48].

B Proofs

B.1 Minimal generating set of a Group

Here, we restate and prove the following lemma on the size of the minimal generating set in group theory for completeness.

Proposition 5. *The minimal generating set S of a finite group G, has a size $\rho(G) \coloneqq |S| \leq \log_2(|G|)$.*

Proof. Consider the minimal generating set $S = \{g_1, g_2, \ldots, g_{|S|}\}$ of the finite group G. For each $k \in \{1, 2, \ldots, |S|\}$, define $G_k = \langle g_1, g_2, \ldots, g_k \rangle$.

The identity e is not equal to any of g_k, and hence cannot be a member of any G_n, since it can always be produced by combining an element with its inverse. Moreover, for all $k \in \{1, 2, \ldots, |S|\}$, we know that $g_{k+1} \notin G_k$, since otherwise $\langle g_1, g_2, \ldots, g_k, g_{k+2}, \ldots g_{|S|} \rangle = G$ which contradicts the minimality of the generating set S for the group G. Therefore, $g_{n+1}G_n$ the left coset of G_n is disjoint from G_n. Additionally, by definition, we know that $g_{n+1}G_n \bigcup G_n \subseteq G_{n+1}$. Hence, $|G_{n+1}| \geq |g_{n+1}G_n| + |G_n| = 2|G_n|$. By induction, $2^{|S|} = 2^{|S|}|G_1| \leq |G_{|S|}| = |G|$ which establishes the claim. □

B.2 Constrained Optimization

In this section, we preset a detailed analysis of the constrained quadratic optimization problem that is used in Algorithm 1.

Proposition 6 (Projection into invariant subspace of eigenspaces)**.** *The optimization problem,*

$$\widehat{f}_\lambda \coloneqq \arg\min_{f_\lambda} \sum_{\ell=1}^{m_\lambda} (f_{\lambda,\ell} - \widetilde{f}_{\lambda,\ell})^2, \quad \text{s.t.} \quad \forall g \in S : D^\lambda(g) f_\lambda = f_\lambda, \tag{10}$$

with $|S| = m$, has a closed form solution,

$$\widehat{f}_\lambda = \widetilde{f}_\lambda - {B^\lambda}^\top (B^\lambda {B^\lambda}^\top)^\dagger (B^\lambda \widetilde{f}_\lambda),$$

where

$$B^\lambda = \begin{bmatrix} D^\lambda(g_1) - I \\ D^\lambda(g_2) - I \\ \vdots \\ D^\lambda(g_m) - I \end{bmatrix},$$

and † denotes Moore–Penrose inverse.

Proof. For better readability, we define $B(g_i) \coloneqq D^\lambda(g_i) - I$, where $I \in \mathbb{R}^{m_\lambda \times m_\lambda}$ is the identity matrix of size m_λ, then,

$$B^\lambda = \begin{bmatrix} B(g_1) \\ B(g_2) \\ \vdots \\ B(g_m) \end{bmatrix}.$$

For ease of notation, let $a \coloneqq \widetilde{f}_\lambda \in \mathbb{R}^{m_\lambda}$ and $a^* \coloneqq \widehat{f}_\lambda \in \mathbb{R}^{m_\lambda}$, then the optimization problem (10) can be rewritten as,

$$a^* = \min_{a'} \frac{1}{2} \|a - a'\|^2 \text{ subject to } B^\lambda a' = 0. \tag{11}$$

Now, we need to show that the projection of a onto the subspace defined by $\{a' \mid Ba' = 0\}$ has the following analytical form,

$$a^* = a - {B^\lambda}^\top (B^\lambda {B^\lambda}^\top)^\dagger (B^\lambda a).$$

We form the Lagrangian,

$$\mathcal{L}(a', \lambda) = \frac{1}{2}\|a - a'\|^2 + \xi^\top B^\lambda a,$$

where $\xi \in \mathbb{R}^{m_\lambda}$ is the vector of Lagrange multipliers. By taking gradients,

$$\frac{\partial \mathcal{L}}{\partial a'} = (a' - a) + {B^\lambda}^\top \xi = 0,$$

thus,

$$a^* = a - {B^\lambda}^\top \xi. \tag{12}$$

Substituting back into the constraint $B^\lambda a^* = 0$,

$$B^\lambda(a - {B^\lambda}^\top \xi) = B^\lambda a - B^\lambda {B^\lambda}^\top \xi = 0. \tag{13}$$

Hence, ξ satisfies the above linear system which may have infinite number of solutions. We claim that the choice of $\xi^* = (B^\lambda {B^\lambda}^\top)^\dagger B^\lambda a$ leads to the optimal solution of optimization problem (11). The objective of optimization (11) is $\|a - a^*\|^2 = \|{B^\lambda}^\top \xi\|_2^2$. Any solution ξ to the linear system (13), can be decomposed as $\xi = \xi^* + \xi_0$ where ξ_0 is in the nullspace of $B^\lambda {B^\lambda}^\top$. Hence,

$$\|{B^\lambda}^\top \xi\|_2^2 = \|{B^\lambda}^\top (\xi^* + \xi_0)\|_2^2 \overset{\text{(I)}}{=} \|{B^\lambda}^\top \xi^*\|_2^2 + \|{B^\lambda}^\top \xi_0\|_2^2 \geq \|{B^\lambda}^\top \xi^*\|_2^2.$$

(I) follows since ${B^\lambda}^\top \xi^*$ and ${B^\lambda}^\top \xi_0$ are orthogonal w.r.t. each other. Placing $\xi^* = (B^\lambda {B^\lambda}^\top)^\dagger B^\lambda a$ in Equation (12) concludes the proof. □

Remark 7 (Time complexity of optimization in each eigenspace). We arbitrarily chose to use the closed form solution of optimization problem (10) instead of iterative approaches. In the closed form solution, we need to calculate the pseuedoinverse of matrix $B^\lambda {B^\lambda}^\top \in \mathbb{R}^{|S|m_\lambda \times |S|m_\lambda}$ which can be done via singular value decomposition (SVD) in $O(|S|^3 m_\lambda^3)$. The other operations are matrix multiplications that are dominated by this part in terms of computational complexity.

B.3 Main Theorem

Proof. To prove Theorem 1, we use Algorithm 1. Let us start by calculating the time and oracle complexity of the algorithm. Given a dataset $\mathcal{S}$ of size n, we first compute

$$\widetilde{f}_{\lambda,\ell} = \frac{1}{n}\sum_{i=1}^{n} y_i \phi_{\lambda,\ell}(x_i), \tag{14}$$

for each λ such that $D_\lambda \leq D = n^{1/(1+\alpha)}$, and each $\ell \in [m_\lambda]$. This requires $\mathcal{O}(n^{1+1/(1+\alpha)})$ oracle calls and can be accomplished in time $\mathcal{O}(n^{1+1/(1+\alpha)})$.

Next, we solve the following constrained quadratic program:

$$\widehat{f}_{\lambda,\ell} \leftarrow \arg\min_{f_{\lambda,\ell}} \sum_{\ell=1}^{m_\lambda} (f_{\lambda,\ell} - \widetilde{f}_{\lambda,\ell})^2, \tag{15}$$

$$\text{s.t.} \quad \forall g \in S : D^\lambda(g) f_\lambda = f_\lambda. \tag{16}$$

This is done for each λ such that $D_\lambda \leq n^{1/(1+\alpha)}$. Note that to even set up this program, we need $\mathcal{O}(|S|m_\lambda^2)$ oracle calls to find the constraints. We have

$$(D^\lambda(g))_{\ell,\ell'} = \langle \phi_{\lambda,\ell}(x), \phi_{\lambda,\ell'}(gx) \rangle_{L^2(\mathcal{M})} \tag{17}$$

for each $\ell, \ell' \in [m_\lambda]$.

Therefore, the total oracle complexity of the proposed algorithm is

$$\mathcal{O}\left(\sum_{\lambda: D_\lambda \leq n^{1/(1+\alpha)}} |S| m_\lambda^2 + n^{(2+\alpha)/(1+\alpha)} \right). \tag{18}$$

We have already shown in Proposition 5 that one can use a generator set with $\rho(G) \leq \log(|G|)$. Moreover, note that

$$\sum_{\lambda: D_\lambda \leq n^{1/(1+\alpha)}} m_\lambda^2 = \mathcal{O}(n^{2/(1+\alpha)}). \tag{19}$$

Therefore, the oracle complexity is

$$\mathcal{O}\left(\log(|G|) n^{2/(1+\alpha)} + n^{(2+\alpha)/(1+\alpha)} \right). \tag{20}$$

Let us now calculate the time complexity of finding the estimator. We have already established that we can compute the empirical estimation in time $\mathcal{O}(n^{1+1/(1+\alpha)})$. Next, we need to solve the constrained quadratic program with $\log(|G|)$ constraints and m_λ variables for each λ such that $D_\lambda \leq n^{1/(1+\alpha)}$. Using the proposed algorithm in Appendix B.2 and also Remark 7, we can solve each of these constrained quadratic programs in time $\mathcal{O}(\log^3(|G|) m_\lambda^3)$. Therefore, the total time complexity of this step is bounded by

$$\mathcal{O}\left(\sum_{\lambda: D_\lambda \leq n^{1/(1+\alpha)}} \log^3(|G|) m_\lambda^3 \right) = \mathcal{O}\left(\log^3(|G|) n^{3/(1+\alpha)} \right). \tag{21}$$

This proves that the total time complexity of Algorithm 1 is

$$\mathcal{O}\left(\log^3(|G|) n^{3/(1+\alpha)} + n^{(2+\alpha)/(1+\alpha)} \right). \tag{22}$$

Finally, note that given $\widehat{f}$, one can evaluate it on new unlabeled data $x \in \mathcal{M}$ using the formula:

$$\widehat{f}(x) = \sum_{\lambda: D_\lambda \leq D} \sum_{\ell=1}^{m_\lambda} \widehat{f}_{\lambda,\ell} \phi_{\lambda,\ell}(x), \tag{23}$$

with $D = n^{1/(1+\alpha)}$, which requires both time and oracle complexity of $\mathcal{O}(n^{1/(1+\alpha)})$.

To complete the proof, we need to study the convergence of the population risk of the proposed estimator. We first note that

$$\mathcal{R}(\widehat{f}) = \mathbb{E}[\|\widehat{f} - f^\star\|_{L^2(\mathcal{M})}^2] \leq 2\mathbb{E}[\|\widehat{f} - f^\star_{\leq D}\|_{L^2(\mathcal{M})}^2] + 2\mathbb{E}[\|f^\star_{>D}\|_{L^2(\mathcal{M})}^2], \tag{24}$$

where $f^\star_{\leq D}$ denotes the orthogonal projection of the function $f^\star$ onto the space of eigenfunctions with eigenvalues satisfying $D_\lambda \leq D$. Moreover, $f^\star_{>D} = f^\star - f^\star_{\leq D}$.

First, let us upper bound the second term. Note that, according to the assumption, $f^\star \in H^s(\mathcal{M})$. Thus,

$$\mathbb{E}[\|f^\star_{>D}\|^2_{L^2(\mathcal{M})}] = \sum_{\lambda: D_\lambda > D} \sum_{\ell=1}^{m_\lambda} (f^\star_{\lambda,\ell})^2 \tag{25}$$

$$= \sum_{\lambda: D_\lambda > D} \sum_{\ell=1}^{m_\lambda} D_\lambda^{-\alpha} D_\lambda^{\alpha} (f^\star_{\lambda,\ell})^2 \tag{26}$$

$$\leq D^{-\alpha} \sum_{\lambda: D_\lambda > D} \sum_{\ell=1}^{m_\lambda} D_\lambda^{\alpha} (f^\star_{\lambda,\ell})^2 \tag{27}$$

$$\leq D^{-\alpha} \sum_{\lambda} \sum_{\ell=1}^{m_\lambda} D_\lambda^{\alpha} (f^\star_{\lambda,\ell})^2 \tag{28}$$

$$= D^{-\alpha} \|f^\star\|^2_{H^s(\mathcal{M})}. \tag{29}$$

Now we focus on the first term. Note that

$$\mathbb{E}[\|\widehat{f} - f^\star_{\leq D}\|^2_{L^2(\mathcal{M})}] = \sum_{\lambda: D_\lambda \leq D} \sum_{\ell=1}^{m_\lambda} \mathbb{E}[|\widehat{f}_{\lambda,\ell} - f^\star_{\lambda,\ell}|^2]. \tag{30}$$

According to the definition, we have

$$f^\star_{\lambda,\ell} = \mathbb{E}_x[f^\star(x)\phi_{\lambda,\ell}(x)] = \mathbb{E}_{x,y}[y\phi_{\lambda,\ell}(x)], \tag{31}$$

for each λ, ℓ. Moreover, $\widehat{f}_{\lambda,\ell}$ is the empirical estimation obtained from data:

$$\widehat{f}_{\lambda,\ell} = \frac{1}{n} \sum_{i=1}^{n} y_i \phi_{\lambda,\ell}(x_i). \tag{32}$$

Thus, we obtain

$$\mathbb{E}[|\widehat{f}_{\lambda,\ell} - f^\star_{\lambda,\ell}|^2] = \frac{1}{n} \mathbb{E}\left[|y\phi_{\lambda,\ell}(x) - \mathbb{E}[y\phi_{\lambda,\ell}(x)]|^2\right] \tag{33}$$

$$= \frac{1}{n} \mathbb{E}\left[|\epsilon\phi_{\lambda,\ell}(x) + f^\star(x)\phi_{\lambda,\ell}(x) - \mathbb{E}[f^\star(x)\phi_{\lambda,\ell}(x)]|^2\right] \tag{34}$$

$$= \frac{1}{n} \left(\sigma^2 \mathbb{E}[\phi^2_{\lambda,\ell}] + \mathbb{E}\left[|f^\star(x)\phi_{\lambda,\ell}(x) - \mathbb{E}[f^\star(x)\phi_{\lambda,\ell}(x)]|^2\right]\right) \tag{35}$$

$$\leq \frac{1}{n} \left(\sigma^2 + \mathbb{E}[f^\star(x)^2 \phi^2_{\lambda,\ell}(x)]\right) \tag{36}$$

$$\leq \frac{1}{n} \left(\sigma^2 + \|f^\star\|^2_{L^\infty(\mathcal{M})}\right), \tag{37}$$

where we used the orthonormality of the eigenfunctions $\phi_{\lambda,\ell}$. Then, summing this up to dimension D gives:

$$\mathbb{E}[\|\widehat{f} - f^\star_{\leq D}\|^2_{L^2(\mathcal{M})}] \leq \frac{D}{n} \left(\sigma^2 + \|f^\star\|^2_{L^\infty(\mathcal{M})}\right). \tag{38}$$

Therefore, we can combine the two terms to derive the following population risk bound:

$$\mathcal{R}(\widehat{f}) = \mathbb{E}[\|\widehat{f} - f^\star\|^2_{L^2(\mathcal{M})}] \leq \frac{D}{n} \left(\sigma^2 + \|f^\star\|^2_{L^\infty(\mathcal{M})}\right) + D^{-\alpha} \|f^\star\|^2_{H^s(\mathcal{M})}. \tag{39}$$

We can now specify the above bound to $D = n^{1/(1+\alpha)}$, which is used in the algorithm, to get:

$$\mathcal{R}(\widehat{f}) = \mathbb{E}[\|\widehat{f} - f^\star\|^2_{L^2(\mathcal{M})}] \leq n^{-\alpha/(1+\alpha)} \left(\sigma^2 + \|f^\star\|^2_{L^\infty(\mathcal{M})}\right) + n^{-\alpha/(1+\alpha)} \|f^\star\|^2_{H^s(\mathcal{M})}, \tag{40}$$

which is equivalent to

$$\mathcal{R}(\widehat{f}) = \mathcal{O}(n^{-\alpha/(1+\alpha)}). \tag{41}$$

This completes the proof.

Remark 8. Note that other choices of D may or may not yield better bounds depending on the sparsity of the solution. For this sparsity-unaware upper bound that we use, such a choice of D is optimal. Additionally, since we focus on polynomial time algorithms, we cannot choose exponentially large D even if they deliver gains in sample complexity.

□