The Exact Sample Complexity Gain from Invariances for Kernel Regression

Behrooz Tahmasebi [1] **Stefanie Jegelka** [1]

Abstract

In practice, encoding invariances into models improves sample complexity. In this work, we study this phenomenon from a theoretical perspective. In particular, we provide minimax optimal rates for kernel ridge regression on compact manifolds, with a target function that is invariant to a group action on the manifold. Our results hold for any smooth compact Lie group action, even groups of positive dimension. For a finite group, the gain effectively multiplies the number of samples by the group size. For groups of positive dimension, the gain is observed by a reduction in the manifold's dimension, in addition to a factor proportional to the volume of the quotient space. Our proof takes the viewpoint of differential geometry, in contrast to the more common strategy of using invariant polynomials. This new geometric viewpoint on learning with invariances may be of independent interest.

1. Introduction

In a broad range of applications, including machine learning for physics, molecular biology, point clouds, and social networks, the underlying learning problems are invariant with respect to a group action. The invariances are observed widely in practice, for instance, in the study of high energy particle physics (Fenton et al., 2022; Lee et al., 2020), galaxies (González et al., 2018; Domínguez Sánchez et al., 2018), and also molecular datasets (Anderson et al., 2019; Wang et al., 2021; Li et al., 2021). In learning with invariances, one aims to develop powerful architectures that exploit the problem's invariance structure as much as possible. An essential question is thus: what are the fundamental benefits of model invariance, e.g., in terms of sample complexity?

Several architectures for learning with invariances have been proposed for various types of data and invariances, including DeepSet (Zaheer et al., 2017) for sets, Convolutional Neural Networks (CNNs) (Krizhevsky et al., 2017), tensor field neural networks (Thomas et al., 2018) for point clouds with rotations, translations, and permutations symmetries. These architectures are to exploit the invariance of data as much as possible, and are invariant/equivariant by design.

In fixed dimensions, one common feature of many invariant models, including those discussed above, is that the data lie on a compact manifold (not necessarily a sphere, e.g., the Stiefel manifold for spectral data), and are invariant with respect to a group action on that manifold. Thus, characterizing the theoretical gain of invariances corresponds to studying the gain of learning under group actions on manifolds. Adopting this view, in this paper, we answer the question: *how much gain in sample complexity is achievable by encoding invariances?* As this problem is algorithm and model dependent, it is hard to address in general. A focused version of the problem, but still interesting, is to study this sample complexity gain in kernel-based algorithms, which is what we address here.

Formally, we consider Kernel Ridge Regression (KRR) with i.i.d. data on a compact manifold $\mathcal{M}$. The target function lies in a Reproducing Kernel Hilbert space (RKHS) of Sobolev functions $\mathcal{H}^s(\mathcal{M})$, $s \geq 0$. In addition, the target function is invariant to the action of an arbitrary Lie group G on the manifold. We aim to quantify: *by varying the group G, how does the sample complexity change, and what is the precise gain as G grows?*

Main results. Our main results characterize minimax optimal rates for the convergence of the (excess) population risk (generalization error) of KRR with invariances. More precisely, for the Sobolev kernel, the most commonly studied case of kernel regression, we prove that a (excess) population risk (generalization error) $\propto \left(\frac{\sigma^2 \operatorname{vol}(\mathcal{M}/G)}{n}\right)^{s/(s+d/2)}$ is both achievable and minimax optimal, where σ^2 is the variance of the observation noise, $\operatorname{vol}(\mathcal{M}/G)$ is the volume[1] of the corresponding quotient space, and d is the effective dimension of the *quotient space* (see Section 3 for a precise definition). This result shows a reduction in sample

[1]MIT CSAIL. Correspondence to: Behrooz Tahmasebi <bzt@mit.edu>.

Presented at the 2^{nd} Annual Workshop on Topology, Algebra, and Geometry in Machine Learning (TAG-ML) at the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. 2023.

[1]In general, the quotient space is not a manifold; but we can still define a notion of volume for it.

complexity in *two* intuitive ways: (1) scaling the effective number of samples, and (2) reducing dimension and hence exponent. First, for finite groups, the factor $\mathrm{vol}(\mathcal{M}/G)$ reduces to $\mathrm{vol}(\mathcal{M})/|G|$, and can hence be interpreted as scaling the "effective" number of samples by the size of the group. That is, each data point conveys the information of $|G|$ data points due to the invariance. Second, and importantly, the parameter d in the exponent can in general be much smaller than $\dim(\mathcal{M})$, which would be the correspondent of d in the non-invariant case. In the best case, $d = \dim(\mathcal{M}) - \dim(G)$, where $\dim(G)$ is the dimension of the Lie group G. Hence, the second gain shows a gain in the dimensionality of the space, and hence in the exponent.

Our results generalize and greatly expand previous results by Bietti et al. (2021), which only apply to *finite* groups and *isometric* actions and are valid only on spheres. In contrast, we derive optimal rates for all compact manifolds and smooth compact Lie group actions (not necessarily isometric), including groups of positive dimension. In particular, the reduction in dimension applies to infinite groups, since for finite groups $\dim(G) = 0$. Hence, our results reveal a new perspective on the reduction in sample complexity that was not possible with previous assumptions. Our rates are consistent with the classical results for learning in Sobolev spaces on manifolds without invariances. (To illustrate our general results, in the full version, we make them explicit for kernel counterparts of popular invariant models, such as DeepSets, GNNs, PointNet, and SignNet).

Even though our theoretical results look intuitively reasonable, the proof is challenging. We study the space of invariant functions as a function space on the quotient space $\mathcal{M}/G$. To bound its complexity, we develop a dimension counting theorem for functions on the quotient space, which is at the heart of our analysis and of independent interest. The difficulty is that $\mathcal{M}/G$ is not always a manifold. Moreover, it may exhibit non-trivial boundaries that require boundary conditions to study function spaces. Different boundary conditions can lead to very different function spaces, and a priori the appropriate choice for the invariant functions is unclear. We prove that smooth invariant functions on $\mathcal{M}$ satisfy the Neumann boundary condition on the (potential) boundaries of the quotient space, thus characterizing exactly the space of invariant functions.

The ideas behind the proof are of potential independent interest: we provide a differential geometric viewpoint of the class of functions defined on manifolds and study group actions on manifolds from this perspective. This stands in contrast to the classical strategy of using polynomials generating the class of functions (Mei et al., 2021; Bietti et al., 2021), which is restricted to spheres. To the best of our knowledge, the tools used in this paper are new to the literature on learning with invariances. In short, we make the following contributions:

- We characterize the exact sample complexity gain from invariances for kernel regression on compact manifolds for an arbitrary Lie group action. Our results reveal two ways to reduce sample complexity, including a new reduction in dimensionality that was not obtainable with assumptions in prior work.
- Our proof analyzes invariant functions as a function space on the quotient space; this differential geometric perspective and our new dimension counting theorem, which is at the heart of our analysis, may be of independent interest.

2. Preliminaries and Problem Statement

Consider a smooth connected compact boundaryless[2] $\dim(\mathcal{M})-$dimensional (Riemannian) manifold $(\mathcal{M}, g)$, where g is the Riemannian metric. Let G denote an arbitrary compact Lie group of dimension $\dim(G)$ (i.e., a group with a smooth manifold structure), and assume that G acts smoothly on the manifold $(\mathcal{M}, g)$; this means that each $\tau \in G$ corresponds to a diffeomorphism $\tau : \mathcal{M} \to \mathcal{M}$, i.e., an invertible smooth map. Without loss of generality, we can assume that G acts *isometrically* on $(\mathcal{M}, g)$, i.e., G is a Lie subgroup of the isometry group $\mathrm{ISO}(\mathcal{M}, g)$. To see why this is not restrictive, given a base metric g, consider a new metric $\tilde{g} = \mu_G(\tau^* g)$, where μ_G is the left-invariant Haar (uniform) measure of G, and $\tau^* g$ is the pullback of the metric g by τ. Under the new metric, G acts isometrically on $(\mathcal{M}, \tilde{g})$.

We are given a dataset $\mathcal{S} = \{(x_i, y_i) : i = 1, 2, \ldots, n\} \subseteq (\mathcal{M} \times \mathbb{R})^n$ of n labeled samples, where $x_i \sim_{\text{i.i.d.}} \mu$, for the uniform (Borel) probability measure $d\mu(x) := \frac{1}{\mathrm{vol}(\mathcal{M})} d\,\mathrm{vol}_g(x)$. Here, $d\,\mathrm{vol}_g(x)$ denotes the volume element of the manifold defined using the Riemannian metric g. We assume the uniform sampling for simplicity; our results hold for non-uniform cases, too. The hypothesis class is a set $\mathcal{F} \subseteq L^2_{\text{inv}}(\mathcal{M}, G) \subseteq L^2(\mathcal{M})$ including only G-invariant square-integrable functions on the manifold, i.e., those $f \in L^2(\mathcal{M})$ satisfying $f(\tau(x)) = f(x)$ for all $\tau \in G$. We assume that there exists a function $f^\star \in \mathcal{F}$ such that $y_i = f^\star(x_i) + \epsilon_i$ for each $(x_i, y_i) \in \mathcal{S}$, where ϵ_i's are conditionally zero-mean random variables with variance σ^2, i.e., $\mathbb{E}[\epsilon_i | x_i] = 0$ and $\mathbb{E}[\epsilon_i^2 | x_i] \le \sigma^2$.

Let $K : \mathcal{M} \times \mathcal{M}$ denote a continuous positive-definite symmetric (PDS) kernel on the manifold $\mathcal{M}$, and $\mathcal{H} \subseteq L^2(\mathcal{M})$ its Reproducing Kernel Hilbert Space (RKHS). The kernel K is called $G-$*invariant* if and only if for all $x, y \in$

[2] Although the results in this paper can be easily extended to manifolds with boundaries, for simplicity, we focus on the boundaryless case.

$\mathcal{M}$, one has $K(x, y) = K(\tau(x), \tau'(y))$, for any $\tau, \tau' \in G$.

Kernel Ridge Regression (KRR) on the data $\mathcal{S}$ with a G-invariant kernel K asks for the function $\hat{f}$ that minimizes

$$\hat{f} := \arg\min_{f \in \mathcal{H}} \Big\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \eta \|f\|_{\mathcal{H}}^2 \Big\}. \quad (1)$$

By the representer theorem (Mohri et al., 2018), the optimal solution $\hat{f} \in \mathcal{H}$ is of the form $\hat{f} = \sum_{i=1}^{n} a_i K(x_i, .)$ for a weight vector $\mathbf{a} \in \mathbb{R}^n$. The objective function $\hat{\mathcal{R}}(\hat{f})$ can thus be written as

$$\hat{\mathcal{R}}(\hat{f}) = \tfrac{1}{n} \|\mathbf{y} - \mathbf{K}\mathbf{a}\|_2^2 + \eta \mathbf{a}^T \mathbf{K} \mathbf{a}, \quad (2)$$

where $\mathbf{y} = (y_1, y_2, \ldots, y_n)^n$ and $\mathbf{K} = \{K(x_i, x_j)\}_{i,j=1}^n$ is the Gram matrix. This gives the closed-form solution $\mathbf{a} = (\mathbf{K} + n\eta I)^{-1}\mathbf{y}$. Also, define the population risk $\mathcal{R}(f) := \mathbb{E}_{x \sim \mu}[(y - f(x))^2]$.

This paper focuses on the RKHS of Sobolev functions, $\mathcal{H} = \mathcal{H}^s_{\text{inv}}(\mathcal{M}) = \mathcal{H}^s(\mathcal{M}) \cap L^2_{\text{inv}}(\mathcal{M}, G)$, $s > 0$. This includes all functions having square-integrable derivatives up to order s. Note that $\mathcal{H}^s(\mathcal{M})$ includes only continuous functions when $s > \dim(\mathcal{M})/2$. Moreover, it contains only continuously differentiable functions up to order k when $s > \dim(\mathcal{M})/2 + k$. Note that $\mathcal{H}^s(\mathcal{M})$ is an RKHS if and only if $s > \dim(\mathcal{M})/2$.

3. Main Results

Our first theorem provides an upper bound on the excess population risk, or the generalization error, of KRR with invariances.

Theorem 3.1 (Convergence rate of KRR with invariances)**.** *Consider KRR with invariance with respect to a group G on the Sobolev space $\mathcal{H}^s_{\text{inv}}(\mathcal{M})$, $s > d/2$, with $d = \dim(\mathcal{M}/G)$. Assume that $f^\star \in \mathcal{H}^{s\theta}_{\text{inv}}(\mathcal{M})$ for some $\theta \in (0, 1]$, and let $s = \frac{d}{2}(\kappa + 1)$ for a positive κ. Then,*

$$\mathbb{E}\Big[\mathcal{R}(\hat{f}) - \mathcal{R}(f^\star)\Big] \leq 32 \Big(\frac{1}{\kappa\theta} \frac{\omega_d}{(2\pi)^d} \frac{\sigma^2 \operatorname{vol}(\mathcal{M}/G)}{n} \Big)^{\theta s/(\theta s + d/2)} \|f^\star\|^{d/(\theta s + d/2)}_{\mathcal{H}^{s\theta}_{\text{inv}}(\mathcal{M})},$$

with the optimal regularization parameter

$$\eta = \Big(\frac{1}{2\kappa\theta \|f^\star\|^2_{\mathcal{H}^{s\theta}_{\text{inv}}(\mathcal{M})}} \frac{\omega_d}{(2\pi)^d} \frac{\sigma^2 \operatorname{vol}(\mathcal{M}/G)}{n} \Big)^{\theta s/(\theta s + d/2)},$$

where $\omega_d = \frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)}$ is the volume of the unit $d-$ball in the Euclidean space $\mathbb{R}^d$.

Setting $G = \{\text{id}_G\}$ (i.e., the trivial group) recovers the standard generalization bound without group invariances. In particular, without invariances, the dimension d becomes $\dim(\mathcal{M})$, and the volume $\operatorname{vol}(\mathcal{M}/G)$ becomes $\operatorname{vol}(\mathcal{M})$. Hence, group invariance can lead to a two-fold gain:

- **Exponent**: the dimension d in the exponent can be much smaller than the corresponding $\dim(\mathcal{M})$.
- **Effective number of samples**: the number of samples is multiplied by

$$\frac{\omega_{\dim(\mathcal{M})}/(2\pi)^{\dim(\mathcal{M})}}{\omega_d/(2\pi)^d} . \frac{\operatorname{vol}(\mathcal{M})}{\operatorname{vol}(\mathcal{M}/G)}. \quad (3)$$

The quantity (3) reduces to $|G|$ if G is a finite group that efficiently acts on $\mathcal{M}$ (i.e., if any group element acts non-trivially on the manifold). Intuitively, any sample conveys the information of $|G|$ data points, which can be interpreted as having effectively $n \times |G|$ samples (compared to non-invariant KRR with n samples). For groups of positive dimension, it measures how the group is contracting the volume of the manifold. Note that for finite groups, one always has $\frac{\operatorname{vol}(\mathcal{M})}{\operatorname{vol}(\mathcal{M}/G)} \geq 1$.

Dimension and volume for quotient spaces. In Theorem 3.1, the quotient space $\mathcal{M}/G$ is defined as the set of all orbits $[x] := \{\tau(x) : \tau \in G\}$, $x \in \mathcal{M}$, but $\mathcal{M}/G$ is not always a (boundaryless) manifold. Thus, it is not immediately possible to define its dimension and volume. The quotient space is a finite disjoint union of manifolds, each with its specific dimension/volume. In the full version, we review the theory of quotients of manifolds, and observe that there exists an open dense subset $\mathcal{M}_0 \subseteq \mathcal{M}$ such that $\mathcal{M}_0/G$ is open and dense in $\mathcal{M}/G$, and more importantly, it is a connected precompact manifold. $\mathcal{M}_0/G$ is called the *principal* part of the quotient space. It has the largest dimension among all the manifolds that make up the quotient space.

The projection map $\pi : \mathcal{M}_0 \to \mathcal{M}_0/G$ induces a metric on $\mathcal{M}_0/G$ and this allows us to define $\operatorname{vol}(\mathcal{M}/G) := \operatorname{vol}(\mathcal{M}_0/G)$. Note that $\operatorname{vol}(\mathcal{M}/G)$ depends on the Riemannian metric, which itself might depend on the group G if we start from a base metric and then deform it to make the action isometric. The volume $\operatorname{vol}(\mathcal{M}_0/G)$ is computed with respect to the dimension of $\mathcal{M}_0/G$, thus being nonzero even if $\dim(\mathcal{M}_0/G) < \dim(\mathcal{M})$.

The effective dimension of the quotient space is defined as $d := \dim(\mathcal{M}_0/G)$. Alternatively, one can define the effective dimension as

$$d := \dim(\mathcal{M}) - \dim(G) + \min_{x \in \mathcal{M}} \dim(G_x), \quad (4)$$

where $G_x := \{\tau \in G : \tau(x) = x\}$ is called the isotropic group of the action at point $x \in \mathcal{M}$. For example, if there exists a point $x \in \mathcal{M}$ with the trivial isotropy group $G_x = \{\text{id}_G\}$, then $d = \dim(\mathcal{M}) - \dim(G)$.

Remark 3.2. The invariant Sobolev space $\mathcal{H}^s_{\text{inv}}(\mathcal{M}) \subseteq \mathcal{H}^{s\theta}_{\text{inv}}(\mathcal{M}) \subseteq L^2_{\text{inv}}(\mathcal{M})$. If the regression function $f^\star$ does not belong to the Sobolev space $\mathcal{H}^s_{\text{inv}}(\mathcal{M})$ (i.e., $\theta \in (0, 1)$),

the achieved exponent only depends on θs (i.e., the smoothness of the regression function $f^\star$ and not the smoothness of the kernel). The bound decreases monotonically as s increases; smoother functions are easier to learn.

The next theorem states our minimax optimality result. For simplicity, we assume $\theta = 1$.

Theorem 3.3 (Minimax optimality)**.** *For any estimator $\hat{f}$,*

$$\sup_{\substack{f^\star \in \mathcal{H}^s_{\mathrm{inv}}(\mathcal{M}) \\ \|f^\star\|_{\mathcal{H}^s_{\mathrm{inv}}(\mathcal{M})}=1}} \mathbb{E}\Big[\mathcal{R}(\hat{f}) - \mathcal{R}(f^\star)\Big] \geq C_\kappa \Big(\frac{\omega_d}{(2\pi)^d} \frac{\sigma^2 \operatorname{vol}(\mathcal{M}/G)}{n} \Big)^{s/(s+d/2)},$$

where C_κ is a constant only depending on κ.

An explicit formula for C_κ is given in the full version. Note that the above minimax lower bound not only proves that the achieved bound by the KRR estimator is optimal but also shows the optimality of the prefactor characterized in Theorem 3.1 with respect to the effective dimension d (up to multiplicative constants depending on κ).

4. Conclusion

In this work, we derived new generalization bounds for learning with invariances. These generalization bounds show a two-fold gain in sample complexity: (1) in the dimension term in the exponent, and (2) in the effective number of samples. Our results significantly generalize the range of settings where the bounds apply. In particular, (1) goes beyond prior work, since it applies only to groups of positive dimension, whereas prior work assumed finite dimensions. At the heart of our analysis is a new dimension counting bound for invariant functions on manifolds, which we expect to be useful more generally for analyzing learning with invariances. We prove this bound via differential geometry, and show how to overcome several technical challenges related to the properties of the quotient space.

Acknowledgements

This research was supported by the NSF award CCF-2112665, the NSF Award 2134108, and the ONR grant N00014-20-1-2023 (MURI ML-SCOPE).

References

Anderson, B., Hy, T.-S., and Kondor, R. Cormorant: Covariant molecular neural networks. *Advances in Neural Information Processing Systems*, 2019.

Bietti, A., Venturi, L., and Bruna, J. On the sample complexity of learning under geometric stability. *Advances in Neural Information Processing Systems*, 34:18673–18684, 2021.

Domínguez Sánchez, H., Huertas-Company, M., Bernardi, M., Tuccillo, D., and Fischer, J. Improving galaxy morphologies for sdss with deep learning. *Monthly Notices of the Royal Astronomical Society*, 476(3):3661–3676, 2018.

Fenton, M. J., Shmakov, A., Ho, T.-W., Hsu, S.-C., Whiteson, D., and Baldi, P. Permutationless many-jet event reconstruction with symmetry preserving attention networks. *Physical Review D*, 105(11):112008, 2022.

González, R. E., Munoz, R. P., and Hernández, C. A. Galaxy detection and identification using deep learning and data augmentation. *Astronomy and computing*, 25:103–109, 2018.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

Lee, J. S. H., Park, I., Watson, I. J., and Yang, S. Zero-permutation jet-parton assignment using a self-attention network. *arXiv preprint arXiv:2012.03542*, 2020.

Li, Z., Yang, S., Song, G., and Cai, L. Hamnet: Conformation-guided molecular representation with hamiltonian neural networks. *arXiv preprint arXiv:2105.03688*, 2021.

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

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of machine learning*. MIT press, 2018.

Thomas, N., Smidt, T., Kearnes, S., Yang, L., Li, L., Kohlhoff, K., and Riley, P. Tensor field networks: Rotation-and translation-equivariant neural networks for 3d point clouds. *arXiv preprint arXiv:1802.08219*, 2018.

Wang, Z., Wang, C., Zhao, S., Du, S., Xu, Y., Gu, B.-L., and Duan, W. Symmetry-adapted graph neural networks for constructing molecular dynamics force fields. *Science China Physics, Mechanics & Astronomy*, 64(11): 1–9, 2021.

Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R. R., and Smola, A. J. Deep sets. *Advances in neural information processing systems*, 30, 2017.