Report on "On the Sample Complexity of Learning under Invariance and Geometric Stability"

Son Luu

April 12, 2024

1 Summary

1.1 Technical summary and analysis

1.1.1 Set up

• **Statistical learning and Kernel ridge regression (KRR):** Consider the supervised learning problem where the data are pairs of input-label $(x_i, y_i)_{i=1}^n$ satisfying

$$\forall\ i \in \{1, \dots, n\} : \begin{cases} x_i \in \mathcal{X} := \mathbb{S}^{d-1} \subset \mathbb{R}, y_i \in \mathbb{R} \\ (x_i, y_i) \sim \rho \\ E(y_i|x_i) = f^*(x_i),\ f^* : \mathbb{S}^{d-1} \to \mathbb{R},\ f^* \in L^2(\mathbb{S}^{d-1}) \end{cases} \tag{1}$$

where $\mathbb{S}^{d-1}$ is the unit sphere in d dimension, f^* is the target function, $L^2(\mathbb{S}^{d-1})$ is the space of square-integrable real-valued functions on $\mathbb{S}^{d-1}$ and ρ is a distribution with the marginal for the input component being the uniform distribution on $\mathbb{S}^{d-1}$. Denote $d\tau$ the uniform measure on $\mathbb{S}^{d-1}$ and write $L^2(\mathbb{S}^{d-1})$ as $L^2(d\tau)$ or L^2 for simplicity. Our goal is to obtain generalization bounds as a function of the sample size n on the excess risk

$$E[R(\hat{f}_n)] - R(f^*) = E[\|\hat{f}_n - f^*\|_{L^2}^2] \tag{2}$$

where $\hat{f}_n$ is the estimator based on the data, $\|\cdot\|_{L^2}$ is the L^2 norm, the expectation is over the n samples and the L^2 risk R is defined as

$$R(f) := E_{(x,y)\sim\rho}[(f(x) - y)^2]. \tag{3}$$

In [BVB21], the authors consider the KRR estimator

$$\hat{f}_\lambda := \arg\min_{f \in \mathcal{H}_K} \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}_K}^2 \tag{4}$$

where $\lambda > 0$, $K(x, x') := \kappa(\langle x, x' \rangle)$ is a dot-product kernel and $\mathcal{H}_K$ is the associated reproducible kernel Hilbert space (RKHS).

• **Harmonic analysis on the sphere:** For any $f \in L^2$, we can decompose it in the spherical harmonic polynomial basis $\{Y_{k,j} : k \geq 0, j = 1, \dots, N(d, k)\}$ of L^2 as follows

$$f(x) = \sum_{k \geq 0} \sum_{j=1}^{N(d,k)} a_{k,j} Y_{k,j}(x) \tag{5}$$

where $a_{k,j} \in \mathbb{R}, \sum_k \sum_{j=1}^{N(d,k)} a_{k,j}^2 < \infty$ and for each k, the collection $\{Y_{k,j}\}_{j=1}^{N(d,k)}$ form an orthonormal basis of the space $V_{d,k}$ of degree k spherical harmonics. Using the same basis, any positive definite dot-product kernel K can be written as

$$K(x, x') = \sum_{k \geq 0} \mu_k \sum_{j=1}^{N(d,k)} Y_{k,j}(x) Y_{k,j}(x') \tag{6}$$

where $\{\mu_k\}_{k \geq 0}$ are the eigenvalues of the following operator

$$T_K f(x) := \int K(x, x') f(x') d\tau(x'). \tag{7}$$

Investigating the decays of μ_k is crucial to obtaining bounds on the excess risk.

• **Group invariant and geometrically stable functions:** Consider a finite set G of transformations $\sigma : \mathcal{X} \to \mathcal{X}$. Define the smoothing operator S_G as follows

- If G is a group

$$S_G f(x) := \frac{1}{|G|} \sum_{\sigma \in G} f(\sigma \cdot x) \tag{8}$$

where $\sigma \cdot x := \sigma(x)$ and $|G|$ is the size of G.

- If G is not a group

$$S_G f(x) := \sum_{\sigma \in G} h(\sigma) f(\sigma \cdot x) \tag{9}$$

where $h(\sigma) \geq 0$ for all $\sigma \in G$ and $\sum_{\sigma \in G} h(\sigma) = 1$. In what follows, assume that G is symmetric, i.e. $\sigma^{-1} \in G$ if $\sigma \in G$ and $h(\sigma) = h(\sigma^{-1})$ so that S_G is self-adjoint.

Definition 1. *Let G be defined as above*

- *If G is a group, a function f is G-invariant if*

$$f(\sigma \cdot x) = f(x) \quad \forall\ \sigma \in G, x \in \mathcal{X} \tag{10}$$

- *A function f is geometrically stable if*

$$\exists\ g \in L^2 : f(x) = S_G g(x) \quad \forall\ x \in \mathcal{X} \tag{11}$$

Note that invariance implies geometric stability since $f = S_G f$ for any G-invariant function.

1.1.2 Bounds for invariant target

For a G-invariant target f^*, the authors derive bounds for KRR estimators using a generic dot-product kernel K and its smoothed version

$$K_G(x, x') := S_G K(x, x') = \frac{1}{|G|} \sum_{\sigma \in G} K(\sigma \cdot x, x') \tag{12}$$

In addition, denote $\bar{V}_{d,k} := S_G V_{d,k}$ and its basis of invariant spherical harmonics $\{\bar{Y}_{k,j}\}_{j=1}^{\bar{N}(d,k)}$. Then we have the following lemma

Lemma 1. *For any $k \geq 0$, we have*

$$\gamma_d(k) := \frac{\bar{N}(d,k)}{N(d,k)} = \frac{1}{|G|} \sum_{\sigma \in G} E_x[P_{d,k}(\langle \sigma \cdot x, x' \rangle)] \tag{13}$$

where $P_{d,k}$ are Legendre polynomials of degree k in d dimensions (normalized with $P_{d,k}(1) = 1$).

The quantity $\gamma_d(k)$ will determine the gain in sample complexity due to invariance. In this setting, the following assumptions are made

(A1) (capacity condition): $\mathcal{N}_K(\lambda) \leq C_K \lambda^{-1/\alpha}$ with $\alpha > 1, C_K > 0$.
Here the degrees of freedom $\mathcal{N}_K(\lambda)$ is defined as

$$\mathcal{N}_K(\lambda) := Tr(\Sigma_K(\Sigma_K + \lambda I)^{-1}) = \sum_{m \geq 0} \frac{\lambda_m}{\lambda_m + \lambda} \tag{14}$$

where $\Sigma_K := E_x[K(x, \cdot) \bigotimes_{\mathcal{H}_K} K(x, \cdot)]$ is the covariance operator $(\lambda_m)_{m \geq 0}$ its eigenvalues, taking multiplicity into account, which are the same as those of T_K when data is distributed according to $d\tau$.

(A2) (source condition): $\exists\ r > \frac{\alpha-1}{2\alpha}, g \in L^2, \|g\|_{L^2} \leq C_{f^*} : f^* = T_K^r g$ where

$$T_K^r g := \sum_{k\geq 0} \mu_k^r \sum_{j=1}^{N(d,k)} a_{k,j} Y_{k,j} \tag{15}$$

with $g = \sum_{k\geq 0} \sum_{j=1}^{N(d,k)} a_{k,j} Y_{k,j}$.

(A3) (invariance): f^* is G-invariant

(A4) (problem noise): ρ is such that $E_\rho[(y - f^*(x))^2|x] \leq \sigma_\rho^2$

The generalization bounds for invariant targets are as follow

Theorem 1. *Assume (A1-4). Let $\nu_d(l) := \sup_{k\geq l} \gamma_d(k)$ and assume $\nu_0 > 0$.*
Let $n > \max\{\|f^*\|_\infty^2/\sigma_p^2, (C_1/\nu_0)^{\frac{\alpha}{2\alpha r+1-\alpha}}\}$ and define

$$l_n := \sup\{l : D(l) < C_2 \nu_d(l)^{\frac{2\alpha r}{2\alpha r+1}} n^{\frac{1}{2\alpha r+1}}\} \tag{16}$$

where $D(l) = \sum_{k<l} \bar{N}(d,k)$.
Using KRR with kernel K_G, we have, for $\lambda = C_3(\nu_d(l_n)/n)^{\frac{\alpha}{2\alpha r+1}}$,

$$E[R(\hat{f}_\lambda)] - R(f^*) \leq C_4 \left(\frac{\nu_d(l_n)}{n}\right)^{\frac{2\alpha r}{2\alpha r+1}} \tag{17}$$

Using KRR with kernel K, we have, for $\lambda = C_3 n^{\frac{-\alpha}{2\alpha r+1}}$,

$$E[R(\hat{f}_\lambda)] - R(f^*) \leq C_4 \left(\frac{1}{n}\right)^{\frac{2\alpha r}{2\alpha r+1}} \tag{18}$$

Here the constants $C_{1:4}$ only depend on the parameters of assumptions (A1-4).

Since $\nu_d(l_n) \leq 1$, there will always be some improvement in sample complexity when using K_G as opposed to K. In addition, the paper shows that the rate $\frac{\nu_d(l_n)}{n}$ is asymptotically $\frac{1}{|G|n}$, which is asymptotically minimax optimal.

1.2 Bounds for geometrically stable target

For a geometrically stable target f^*, the smoothed kernel K_G is redefined as

$$K_G(x,x') := S_G K(x,x') = \sum_{\sigma\in G} h(\sigma) K(\sigma\cdot x, x') \tag{19}$$

Next, we have the following lemma

Lemma 2. *There exists a basis of spherical harmonics $\bar{Y}_{k,j}$, for $k \geq 0$, and $j = 1,\ldots,N(d,k)$, in which the operator S_G is diagonal, with eigenvalues $\lambda_{k,j} \geq 0$. In addition, we have*

$$\gamma_d(k) := N(d,k)^{-1} \sum_{j=1}^{N(d,k)} \lambda_{k,j} = \sum_{\sigma\in G} h(\sigma) E_x[P_{d,k}(\langle \sigma\cdot x, x'\rangle)] \tag{20}$$

In this setting, assumptions (A1-3) is replaced by

(A5) (capacity condition): the eigenvalues $(\xi_m)_{m\geq 0}$ of T_K satisfy $\xi_m \leq C(m+1)^{-\alpha}$ where $\alpha > 1$.

(A6) (source condition): $\exists\ r > \frac{\alpha-1}{2\alpha}, g \in L^2, \|g\|_{L^2} \leq C_{f^*} : f^* = S_G^r T_K^r g$.

Note that assumption (A6) is the same as a standard source condition for K_G since $T_{K_G} = S_G T_K$. With this, the generalization bounds for geometrically stable targets are as follow

Theorem 2. *Assume (A4-6). Let $\nu_d(l) := \sup_{k \geq l} \gamma_d(k)$ and assume $\nu_0 > 0$.*
Let $n > \max\{\|f^*\|_\infty^2/\sigma_p^2, (C_1/\nu_0)^{\frac{1}{2\alpha r+1-\alpha}}\}$ and define

$$l_n := \sup\{l : D(l) < C_2 \nu_d(l)^{\frac{2r}{2\alpha r+1}} n^{\frac{1}{2\alpha r+1}}\} \tag{21}$$

where $D(l) = \sum_{k<l} \bar{N}(d,k)$.
Using KRR with kernel K_G, we have, for $\lambda = C_3(\nu_d(l_n)^{1/\alpha}/n)^{\frac{\alpha}{2\alpha r+1}}$,

$$E[R(\hat{f}_\lambda)] - R(f^*) \leq C_4 \left(\frac{\nu_d(l_n)^{1/\alpha}}{n}\right)^{\frac{2\alpha r}{2\alpha r+1}}. \tag{22}$$

Using KRR with kernel K, we have the same bound with $\nu_d(l_n)$ replaced by 1, but with a possibly smaller constant C_4. Here the constants $C_{1:4}$ only depend on the parameters of assumptions (A4-6).

The bounds here are essentially the same as in Theorem 1 with $\nu_d(l_n)$ replaced by its new definition and an added $1/\alpha$ exponential reducing the improvement to sample complexity.

1.2.1 Proof techniques

The proof techniques for both the inivariant and geometrically stable cases follow a similar path. First, by Proposition 7.2 from [Bac21], we have, for $\lambda \leq 1$, assuming $K_G(x,x) \leq 1$ almost surely and $n \geq \frac{5}{\lambda}(1 + \log(1/\lambda))$,

$$E[R(\hat{f}_\lambda)] - R(f^*) \leq 16 \mathcal{A}_{\mathcal{H}_{K_G}}(\lambda, f^*) + 16\frac{\sigma_\rho^2}{n}\mathcal{N}_{K_G}(\lambda) + \frac{24}{n^2}\|f^*\|_\infty^2 \tag{23}$$

where the degrees of freedom $\mathcal{N}_{K_G}(\lambda)$ is defined as in assumption (A1) and the approximation error $\mathcal{A}_{\mathcal{H}_{K_G}}(\lambda, f^*)$ is defined as

$$\mathcal{A}_{\mathcal{H}_{K_G}}(\lambda, f^*) := \inf_{f \in \mathcal{H}_{K_G}} \|f - f^*\|_{L^2}^2 + \lambda \|f\|_{\mathcal{H}_{K_G}} \tag{24}$$

Next, $\mathcal{A}_{\mathcal{H}_{K_G}}(\lambda, f^*)$ is bounded by a function of λ and $\mathcal{N}_{K_G}(\lambda)$ is bounded by a function of λ and l. Specifically, the bound for $\mathcal{A}_{\mathcal{H}_{K_G}}(\lambda, f^*)$ uses the fact that $\mathcal{A}_{\mathcal{H}_{K_G}}(\lambda, f^*) = \mathcal{O}(\mathcal{A}_{\mathcal{H}_K}(\lambda, f^*))$ (Lemma 3 of [BVB21]), a bound on $\mathcal{A}_{\mathcal{H}_K}(\lambda, f^*)$ (Theorem 3, p.33 of [CS02]) and the source condition. The bound on $\mathcal{N}_{K_G}(\lambda)$ relies on the capacity condition and the expression the expression in (14). Then the sum of these bounds is bounded by choosing the appropriate $\lambda = \lambda_n$. In short, the process can be expressed as follows

$$\begin{aligned}
E[R(\hat{f}_\lambda)] - R(f^*) &\leq 16 \mathcal{A}_{\mathcal{H}_{K_G}}(\lambda, f^*) + 16\frac{\sigma_\rho^2}{n}\mathcal{N}_{K_G}(\lambda) + \frac{24}{n^2}\|f^*\|_\infty^2 \\
&\leq F_1(\lambda) + 16\frac{\sigma_\rho^2}{n} F_2(\lambda, l) + \frac{24}{n^2}\|f^*\|_\infty^2 \\
&\leq F_1(\lambda_n) + 16\frac{\sigma_\rho^2}{n} \underbrace{F_2(\lambda_n, l)}_{} + \frac{24}{n^2}\|f^*\|_\infty^2 \\
&= \underbrace{F_1(\lambda_n) + 16\frac{\sigma_\rho^2}{n}[G_1(\lambda_n, l)}_{} + G_2(l)] + \frac{24}{n^2}\|f^*\|_\infty^2 \\
&= \qquad F_3(\lambda_n, l, n) \qquad + 16\frac{\sigma_\rho^2}{n} G_2(l) + \frac{24}{n^2}\|f^*\|_\infty^2
\end{aligned}$$

where F_1, F_2 are functions corresponding to the bounds on $\mathcal{A}_{\mathcal{H}_{K_G}}(\lambda, f^*), \mathcal{N}_{K_G}(\lambda)$ and $F_3(\lambda_n, l, n)$ have the same form as the bounds in the main results. Finally, conditions are imposed on l and n to ensure $G_2(l,n)$ and $\frac{24}{n^2}\|f^*\|_\infty^2$ are smaller than $F_3(\lambda_n, l)$.

1.2.2 Discussion

Overall, the generalization bounds on the excess risk is obtained by decomposing it into the approximation error, degrees of freedom and an additional error then bound each of these terms. All assumptions are either made by previous works (capacity and source conditions) or made to achieve the desired bound without much deeper meaning (conditions for l, n and λ). The main challenge is that the exponent of the sample complexity rate is of order $1/d$ the target is Lipschitz, which is not optimal.
The paper's results can be generalized to infinite groups since the finite group argument is only used to defined the smoothing operator S_G. We can change this definition to accommodate infinite groups as follows

$$S_G f(x) := \int f(\sigma \cdot x) d\mu_G(\sigma) \tag{25}$$

where μ_G is the left-invariant Haar (uniform) measure of G. In addition, the input space can be changed from the unit sphere $\mathbb{S}^{d-1}$ to some kind of compact manifold. In this setting, the spherical harmonic basis can be changed to a more general basis and $\gamma_d(k)$ can monitor the change in dimension of the subspaces after projecting to the space of G-invariant functions.

1.3 Conceptual summary

In supervised learning problems, it is often beneficial to utilize properties such as invariance or stability to certain groups of transformations. The paper studies the sample complexity of learning problems for target functions with these properties. The authors do this by examining spherical harmonic decompositions of such functions on the sphere. They show that the non-parametric rate of convergence for kernel methods improves by a factor equal to the size of the group asymptotically when an invariant kernel is used compare to a non-invariant kernel. These results also apply to geometrically stable targets but with a less significant improvement.

1.3.1 Relation to other works

The results of [BVB21] are based on three strategies: bounding the excess risk by the approximation error and estimation error using a bound in [Bac21], then bounding the approximation error using a bound in [CS02] and the estimation error using a bound in [MMM21]. The bound in [MMM21] examines the fraction of invariant eigenfunctions and total eigenfunctions of each eigenspace in the spherical harmonic decomposition, is an effective way to establish the convergence rate for the generalization bound. [BVB21] also improves the convergence rate found in [MMM21] in the sense that the improvement for invariant kernels in [BVB21] can be exponential in dimension while the rate improvement in [MMM21] is only polynomial in dimension. However, in the class of groups consider by [MMM21] (subsets of the orthogonal group $\mathcal{O}(d)$), the improvement is minimax for the excess risk while the improvement in [BVB21] is only minimax for the rate of convergence on the generalization bound, which is a weaker optimality property. Note also that [MMM21] considers the high-dimensional regime where $d \to \infty$ as $n \to \infty$ while [BVB21] considers d to be fixed. For the input space, the paper considers the uniform distribution on the unit sphere and uses spherical harmonic decompositions, which are relatively restrictive. [TJ23] generalizes this setting by considering a compact manifold input space and considers smooth compact Lie groups instead of finite groups. Their bounds improve upon the exponent of rate of convergence, which suffers from the curse of dimensionality in [BVB21], and are minimax for the excess risk. [EZ21] also studies benefits of group invariance, but focuses on linear models, and only considers interpolating estimators.

1.3.2 Contributions and impact

The task of finding the sample complexity for the generalization bound on the excess risk is a relatively well studied problem but studying the sample complexity benefit of incorporating prior information on invariance and stability is a relatively new problem. In addition, the paper focuses on kernel ridge regression in a fixed dimensional regime, which is a setting not yet considered for the addressed problem. For proving techniques,

the paper mostly utilizes results and assumptions from previous works for the bounds on invariant targets. For geometrically stable target, there are some innovations such as formulating of the smoothing operator as a weighted sum instead of an average, and redefining key quantities and assumptions based on this new operator.
The paper is technically sound overall. The results for invariant target is adequately supported by experiment. There are also optimality discussions for invariant target showing that it is the best result one can hope to achieve with current assumptions. The results for geometrically stable target, on the other hand, have no supporting experiments and little discussion on the tightness of the bounds. For clarity, the paper is well organized. However, some notations are not clearly defined such as taking power of an operator.
The paper provides several contributions. The first is the minimax rate of convergence for the generalization bounds. It sets up the intuition that the improvement in rate can be as large as the "size" of the group and has inspired results for more general settings and groups [TJ23]. This solidifies the benefits seen from leveraging invariance in kernel methods and hopefully can be extended to other learning methods as well. In addition, there are many properties that can be explored further for geometrically stable targets such as approximation-estimation trade-off or optimality. In practice, however, it is not clear how the paper's results can help in finding new, more effective learning methods.

2 Mini-proposals

2.1 Proposal 1: Generalization bounds for geometrically stable targets in more general settings

The results from [BVB21] apply to kernel ridge regression (KRR) with input space being the unit sphere and target function being invariant to a finite group. This setting is fairly limited and the finite group assumption is not crucial to the proof. Therefore, it is natural to extend these results to more general groups such as Lie groups and more general input space such as a manifold. Unfortunately, this has already been done in [TJ23].
The results for geometrically stable targets, on the other hand, have not been generalized. Therefore, the subject of this proposal is to generalize the results for geometrically stable targets to a similar setting as in [TJ23].

2.2 Proposal 2: Leveraging invariance in different learning regimes

The results from [BVB21] apply to KRR so another extension is to establish the generalization bounds for different learning regimes. Some examples include kernel density estimation for invariant or geometrically stable probability densities or locally adaptive estimation method such as wavelet method.

3 Project report

For this mini-project, we shall focus on proposal 1.

3.1 Set up

Consider a smooth connected compact boundaryless $dim(\mathcal{M})$-dimensional (Riemannian) manifold $\mathcal{M}$ equipped with the Riemannian metric. Let G denote a symmetric Borel set of diffeomorphisms $\sigma : \mathcal{M} \to \mathcal{M}$ and assume there exists a compact Lie group $\langle G \rangle$ (i.e., a group with a smooth manifold structure) containing G. Further assume that $\langle G \rangle$ is a subgroup of the isometry group $ISO(\mathcal{M})$.
Let the notations and assumptions on the data be the same as Section 1, i.e. the data are pairs of input-label $(x_i, y_i)_{i=1}^n$ satisfying

$$\forall\, i \in \{1, \dots, n\} : \begin{cases} x_i \in \mathcal{M}, y_i \in \mathbb{R} \\ (x_i, y_i) \sim \rho \\ E(y_i|x_i) = f^*(x_i),\ f^* : \mathcal{M} \to \mathbb{R},\ f^* \in L^2(\mathcal{M}) \\ E((f^*(x_i) - y_i)^2 | x_i) \leq \sigma_\rho \end{cases} \tag{26}$$

where f^* is the target function, $L^2(\mathcal{M})$ is the space of square-integrable real-valued functions on $\mathcal{M}$ and ρ is a distribution with the marginal for the input component being the uniform distribution on $\mathcal{M}$. Now denote K and $\mathcal{H}_K$ a continuous positive-definite symmetric dot-product kernel on $\mathcal{M} \times \mathcal{M}$ and its corresponding reproducible kernel Hilbert space (RKHS). Next define the smoothing operators S_G and $S_{\langle G \rangle}$ as follows:

$$S_G f := \int_G \sigma f d\mu_G(\sigma) \tag{27}$$

$$S_{\langle G \rangle} f := \int_{\langle G \rangle} \sigma f d\mu_G(\sigma) \tag{28}$$

where $\mu_G(\sigma)$ is the left-invariant Haar (uniform) measure of $\langle G \rangle$ and

$$\sigma f := \begin{cases} f(\sigma(\cdot)) & \text{if } f : \mathcal{M} \to \mathbb{R} \\ f(\sigma(\cdot), \cdot) & \text{if } f : \mathcal{M} \times \mathcal{M} \to \mathbb{R} \end{cases} . \tag{29}$$

We again consider the kernel ridge regression (KRR) problem with $K_G := S_G K$ where the goal is to estimate

$$\hat{f}_\lambda := \arg \min_{f \in \mathcal{H}_{K_G}} \frac{1}{n} \sum_{i=1}^n (f(x_i) - y_i)^2 + \lambda \|f\|^2_{\mathcal{H}_{K_G}} \tag{30}$$

For this mini-project, we will focus on geometrically stable target functions, i.e. functions f^* satisfying the source condition

$$\exists\, r \geq 1/2, g \in L^2, \|g\|_{L^2} \leq C_{f^*} : f^* = T^r_{K_G} g \tag{31}$$

where T_{K_G} has the same definition as in Section 1 but with the measure used in the integration is changed to the Riemannian metric.

3.2 Result

With the above set up, we have the following generalization bound on the excess risk

Theorem 3. *Assume the source condition in (31). Consider KRR the Sobolev space $\mathcal{H}_s(\mathcal{M}), s > d/2$, with $d = dim(\mathcal{M}/\langle G\rangle)$. Then*

$$E[\mathcal{R}(\hat{f}_\lambda)] - \mathcal{R}(f^*) \leq 32 \left(\frac{d\sigma_\rho^2}{4r(2s-d)n} \frac{\omega_d}{(2\pi)^d} vol(\mathcal{M}/\langle G\rangle) \right)^{2rs/(2rs+d/2)} \|g\|_{L^2}^{d/(2rs+d/2)} + o\left(n^{-2rs/(2rs+d/2)}\right). \tag{32}$$

Here the optimal regularization parameter is

$$\lambda = \left(\frac{d\sigma_\rho^2}{4r\|g\|_{L^2}^2(2s-d)n} \frac{\omega_d}{(2\pi)^d} vol(\mathcal{M}/\langle G\rangle) \right)^{s/(2rs+d/2)}, \tag{33}$$

ω_d is the volume of the unit sphere in $\mathbb{R}^d$ and $vol(\mathcal{M}/\langle G\rangle)$ is the volume of the quotient space $\mathcal{M}/\langle G\rangle$.

The first step to prove Theorem 3 is to use the following initial bound found in Proposition 7.3 of [Bac21]:

$$E[\mathcal{R}(\hat{f}_\lambda)] - \mathcal{R}(f^*) \leq 16\frac{\sigma^2}{n}\mathcal{N}_{K_G}(\lambda) + 16\mathcal{A}(\lambda, f^*) + \frac{24}{n^2}\|f^*\|_{L^\infty} \tag{34}$$

where the approximation error $\mathcal{A}(\lambda, f^*)$ and degrees of freedom $\mathcal{N}_{K_G}(\lambda)$ have the same definitions as in Section 1.

3.2.1 Bound on the approximation error

Using the arguments from [BVB21], we have

$$\begin{aligned}
\mathcal{A}(\lambda, f^*) &\leq \lambda^{2r}\|T_{K_G}^{-r} f^*\|_{L^2} && \text{(Theorem 3 from [CS02])} && (35)\\
&= \lambda^{2r}\|T_{K_G}^{-r} S_G^r T_{K_G}^r g\|_{L^2}^2 && \text{(source condition)} && (36)\\
&= \lambda^{2r}\|T_{K_G}^{-r} T_{K_G}^r S_G^r g\|_{L^2}^2 && (S_G \text{ and } T_{K_G} \text{ commute}) && (37)\\
&= \lambda^{2r}\|S_G^r g\|_{L^2}^2 && && (38)\\
&\leq \lambda^{2r}\|S_G^r\|^2\|g\|_{L^2}^2 && && (39)\\
&\leq \lambda^{2r}\|S_G\|^{2r}\|g\|_{L^2}^2 \leq \lambda^{2r}\|g\|_{L^2}^2 && (\|S_G\| \leq 1). && (40)
\end{aligned}$$

3.2.2 Bound on dimension of function spaces

We have the following theorem regarding the eigenvalues of K_G and $K_{\langle G\rangle} := S_{\langle G\rangle}K$

Theorem 4. *For all l, we have $\lambda_l^{K_G} \leq \lambda_l^{K_{\langle G\rangle}}$ where λ_l^K denote the l^{th} smallest eigenvalue of a positive definite kernel K.*

Proof. First, we prove that the set $\langle G\rangle\backslash G$ is also symmetric. To see this, assume that there exists $\sigma \in \langle G\rangle\backslash G$ such that $\sigma^{-1} \notin \langle G\rangle\backslash G$. This means $\sigma^{-1} \in G$ but since G is symmetric, we have $\sigma \in G$, which contradicts the assumption that $\sigma \in \langle G\rangle\backslash G$. Therefore, $\langle G\rangle\backslash G$ is symmetric. This means $K_{\langle G\rangle\backslash G} := S_{\langle G\rangle\backslash G}K$ defines a positive definite kernel and

$$K_{\langle G\rangle} = K_G + K_{\langle G\rangle\backslash G} \tag{41}$$

Now using Corollary 4.11 in [Tes14] gives us

$$\lambda_l^{K_G} \leq \lambda_l^{K_{\langle G\rangle}} \quad \forall\, l. \tag{42}$$

□

3.2.3 Bound on the degrees of freedom

Using Theorem 4 and the definition of $\mathcal{N}_{K_G}(\lambda)$, we have

$$\mathcal{N}_{K_G}(\lambda) = \sum_{l\geq 0} \frac{\lambda_l^{K_G}}{\lambda_l^{K_G} + \lambda} \leq \sum_{l\geq 0} \frac{\lambda_l^{K_{\langle G\rangle}}}{\lambda_l^{K_{\langle G\rangle}} + \lambda} = \mathcal{N}_{K_{\langle G\rangle}}(\lambda). \tag{43}$$

Next, following the derivation from [TJ23] for $K_{\langle G\rangle}$, we have

$$\begin{aligned}\mathcal{N}_{K_{\langle G\rangle}}(\lambda) \leq{}& \lambda\left(\frac{\omega_d}{(2\pi)^d} vol(\mathcal{M}/\langle G\rangle) + C_{\mathcal{M}/\langle G\rangle}\right) \\ &+ \frac{s}{2s-d}\frac{\omega_d}{(2\pi)^d} vol(\mathcal{M}/\langle G\rangle)\lambda^{1-\frac{1}{s}(s+d/2)} \\ &+ \frac{s}{2s-d+1} C_{\mathcal{M}/\langle G\rangle}\lambda^{1-\frac{1}{s}(s+(d-1)/2)}.\end{aligned} \tag{44}$$

3.2.4 Bound on the excess risk

Plugging the bounds on $\mathcal{A}(\lambda, f^*)$ and $\mathcal{N}_{K_G}(\lambda)$ into (34), we have

$$\begin{aligned}E[\mathcal{R}(\hat{f}_\lambda)] - \mathcal{R}(f^*) \leq{}& \frac{24}{n^2}\|f^*\|_{L^\infty} + 16\lambda^{2r}\|g\|_{L^2}^2 \\ &+ \frac{16\sigma_\rho^2}{n}\lambda\left(\frac{\omega_d}{(2\pi)^d} vol(\mathcal{M}/\langle G\rangle) + C_{\mathcal{M}/\langle G\rangle}\right) \\ &+ \frac{16\sigma_\rho^2}{n}\frac{s}{2s-d}\frac{\omega_d}{(2\pi)^d} vol(\mathcal{M}/\langle G\rangle)\lambda^{1-\frac{1}{s}(s+d/2)} \\ &+ \frac{16\sigma_\rho^2}{n}\frac{s}{2s-d+1} C_{\mathcal{M}/\langle G\rangle}\lambda^{1-\frac{1}{s}(s+(d-1)/2)}\end{aligned} \tag{45}$$

Next, we minimize the sum of the second and forth term in (45) in terms of λ, i.e. minimizing

$$p(\lambda) = (16\|g\|_{L^2}^2)\lambda^{2r} + \left(\frac{16\sigma_\rho^2}{n}\frac{s}{2s-d}\frac{\omega_d}{(2\pi)^d} vol(\mathcal{M}/\langle G\rangle)\right)\lambda^{1-\frac{1}{s}(s+d/2)} \tag{46}$$

To do this, we use the following result. Consider the function

$$p(t) = c_a t^{-a} + c_b t^b \tag{47}$$

where $c_a, c_b, a, b > 0$ and $p(0) = p(\infty) = \infty$. Then $p(t)$ is minimized at $t = \left(\frac{ac_a}{bc_b}\right)^{1/(a+b)}$. Apply this to the function in (46), we get that $p(\lambda)$ is minimized when

$$\lambda = \left(\frac{d\sigma_\rho^2}{4r\|g\|_{L^2}^2(2s-d)n}\frac{\omega_d}{(2\pi)^d} vol(\mathcal{M}/\langle G\rangle)\right)^{s/(2rs+d/2)} \tag{48}$$

Plugging the λ in (48) into (45) and rewriting non-dominating terms as an exponent of n times a constant give us

$$E[\mathcal{R}(\hat{f}_\lambda)] - \mathcal{R}(f^*) \leq 32\left(\frac{d\sigma_\rho^2}{4r\|g\|_{L^2}^2(2s-d)n}\frac{\omega_d}{(2\pi)^d} vol(\mathcal{M}/\langle G\rangle)\right)^{2rs/(2rs+d/2)} \|g\|_{L^2}^2 \tag{49}$$

$$+ C_1 n^{-1-(2rs+1/2)/(2rs+d/2)} + C_2 n^{-1-s/(2rs+d/2)} + C_3 n^{-2} \tag{50}$$

$$= 32\left(\frac{d\sigma_\rho^2}{4r(2s-d)n}\frac{\omega_d}{(2\pi)^d} vol(\mathcal{M}/\langle G\rangle)\right)^{2rs/(2rs+d/2)} \|g\|_{L^2}^{d/(2rs+d/2)} \tag{51}$$

$$+ o\left(n^{-2rs/(2rs+d/2)}\right) \tag{52}$$

where $C_{1:3} > 0$ are constants not depending on n. This concludes the proof for Theorem 3.

3.3 Discussion

In this project, I have provided a generalization bound on the excess risk for geometrically stable targets. Noticeably, the rate of convergence in Theorem 3 is very similar to that of [TJ23] with the quotient group being defined based on $\langle G \rangle$ instead of G itself. This difference is reasonable since we should expect a decrease in learning rate for not being exactly invariant. In addition, the multiplier of s in the exponent is $2r > 1$ instead of $\theta \in (0, 1]$, making the exponent larger overall. This is interesting since it seems to imply that the source condition gives better exponent term than the additional Sobolev condition ($f^* \in \mathcal{H}_{\theta s}(\mathcal{M})$) found in [TJ23].
The challenge for this bound, however, is how to quantify the volume $vol(\mathcal{M}/\langle G \rangle)$ and the dimension d of the quotient space. Another problem is that if the Haar measure of G were zero, it would cause the kernel S_G to be zero, making the definition in (27) useless. A future direction can be trying to replace the assumption of $\langle G \rangle$ existing to a group Q such that $K_Q - K_G$ is a positive definite kernel.

References

[Bac21] F. Bach. "Learning theory from first principles". In: *Draft of a book, version of Sept* 6 (2021), p. 2021.

[BVB21] A. Bietti, L. Venturi, and J. Bruna. "On the sample complexity of learning under geometric stability". In: *Advances in neural information processing systems* 34 (2021), pp. 18673–18684.

[CS02] F. Cucker and S. Smale. "On the mathematical foundations of learning". In: *Bulletin of the American mathematical society* 39.1 (2002), pp. 1–49.

[EZ21] B. Elesedy and S. Zaidi. "Provably strict generalisation benefit for equivariant models". In: *International conference on machine learning*. PMLR. 2021, pp. 2959–2969.

[MMM21] S. Mei, T. Misiakiewicz, and A. Montanari. "Learning with invariances in random features and kernel models". In: *Conference on Learning Theory*. PMLR. 2021, pp. 3351–3418.

[TJ23] B. Tahmasebi and S. Jegelka. "The Exact Sample Complexity Gain from Invariances for Kernel Regression". In: *Advances in Neural Information Processing Systems* 36 (2023).

[Tes14] G. Teschl. *Mathematical methods in quantum mechanics*. Vol. 157. American Mathematical Soc., 2014.