New Convergence Rates for Function Approximation Using Kernel Methods

David Young
Trinity College Dublin

Douglas Leith
Trinity College Dublin

***Abstract*—We derive new bounds for function approximation using reproducing kernel Hilbert spaces by analysing kernels which are k-times continuously differentiable, instead of kernels which are norm equivalent to Sobolev spaces. This means our method allows for a much wider choice of kernels, including the popular exponential quadratic and polynomial kernels, which are widely used in practical application of kernel methods. Our analysis also reveals other possible avenues for new and improved convergence bounds, depending on how we describe our target domain $\mathcal{X}$ and set of sample points X.**

I. Introduction

Reproducing kernel Hilbert spaces (RKHS) are widely used for modeling non-linear functions given pointwise evaluations of the target functions. In function approximation the goal is to design algorithms to minimise the approximation error with as few evaluations as possible, while controlling variables such as how to choose the sample points. This problem arises in such diverse applications as reinforcement learning [1], integral approximation [2] and Gaussian Process Bandits [3], and in all cases RKHS are used to find solutions. We can use RKHS to solve these problems by defining a loss function and then choosing a function from the RKHS that minimises this loss over our target domain. One of the main advantages of RKHS is that for certain choices of loss function this optimisation problem has a simple closed form solution, making it extremely popular for many practical use cases [4]–[6]. An important question is then how the choice of evaluation points affect the quality of this solution. One method for dealing with this is looking at the theoretical worst case convergence bounds for these solutions, and analysing how the choice of parameters effects these bounds. Worst case upper bounds are also of specific importance in the case of Gaussian Process Bandits [3] mentioned above, as they are needed to bound the regret of many UCB based algorithms. We note that a bound in probability is not needed for this application in Gaussian Process Bandits, the absolute bound we provide is sufficient.

Much of the existing theoretical work on learning rates in this area focuses on Sobolev spaces [7]–[9], which results in a limited choice of kernels. This is because the results strictly require that the RKHS be norm equivalent to a Sobolev space of the form W_2^l. The specific requirement of a value of 2 in the subscript means that any kernel which satisfies these conditions can only be finitely continuously differentiable. This immediately excludes many of the most popular kernels used in practical uses cases for kernel methods, such as the exponential quadratic kernel and the polynomial kernel. The kernels covered by the Sobolev work must also be shift invariant, a condition that excludes simple linear kernels of any kind. Our work instead allows any kernel of continuously differentiable degree, including infinite, and does not require shift invariance, meaning we cover the practical kernels mentioned above that are not covered by the Sobolev work. We note also that there is much more work on RKHS interpolation than on regression, while our result generalises both cases. This is significant as many applications require approximate interpolation instead of exact [10].

The attention given to Sobolev spaces seems to be related to their usefulness for solving certain partial differential equation problems [11], [12], but in machine learning we are most often dealing with functions arising from natural processes, and so the space of k times continuously differentiable functions $C^k(\mathcal{X})$ is a more fitting domain. Even though it is possible to approximate these functions using Sobolev kernels, the literature misses that this is not what is being done in practice. Real world applications of these methods use the simpler, more practical kernels which we have mentioned already, and so our work gives convergence rate bounds on these methods that the Sobolev work cannot. Another reason for the focus of previous work on Sobolev spaces is how the Sobolev norm emerges in their analysis as a measure of the 'smoothness' of the target function. This fixed smoothness measure then requires the kernel they choose to be norm equivalent to this measure, and so results in a very limited pool of options. We are able to extend this analysis to $C^k(\mathcal{X})$, and so remove the restrictions on the kernel choice, by removing the reliance on this Sobolev norm. We do this by utilising a result from [13] which bounds the derivatives of a function in a RKHS by its kernel norm, allowing use to use any continuously differentiable kernel to measure smoothness.

II. Related Work

Our work is most related to the seminal work of [7], but we build on it to analyse k times continuously differentiable functions instead of functions belonging to Sobolev spaces. This allows us to keep explicit the values of constants in our convergence bounds, as well as extending our results to a much wider choice of kernel. The work of [9] also builds on [7], but still focused on Sobolev spaces, as does [8], with the addition of taking account of misspecified smoothness and likelihood. Some of these previous results are in the Gaussian process framework, but works such as [14] and [15] show the deep connection and equivalence from that setup and kernel methods.

There is some work analysing the problem with a general RKHS instead of a Sobolev space, but it is mostly in the domain of statistical learning theory [16] which assumes the location evaluation points are sampled from some unknown probability distribution [17]–[19]. This statistical assumption makes it distinct from our framework, the bounds are determined in term of the choice of points. This also means that it does not allow for much of the practical work mentioned below, and its analysis differs in form from ours, and the analysis mentioned above in [7]–[9].

There is much literature on the practical applications of kernel method, looking at their general use in machine learning [20], their applications in computational biology [21], analysing their use in global optimization [22], and applying them to reinforcement learning [1], [23], [24]. Significantly, many of these works utilise setups and kernels that are covered by our analysis but not by contemporary theoretical work such as [8], due to their focus on Sobolev spaces, or by [17], as the evaluation points need to be carefully chosen.

III. Outline of Result

As mentioned above the purpose of this work, and our main result, is a bound on the approximation error of RKHS interpolation or regression for continuously differentiable kernels. This result is stated in Theorem 5 in Section 8, and we will give a brief outline of its most basic version here, to make clear the goal of the intervening analysis.

We have a target function $g \in C^k(\mathcal{X})$, the set of k-times continuously differentiable functions over our target domain $\mathcal{X}$. We have a kernel $\mathcal{K}$ such that $g \in H_{\mathcal{K}}$, its reproducing kernel Hilbert space, from which we get the kernel approximation $\hat{g}_\lambda$ to be defined in the next section, where $\lambda \geq 0$ is a smoothing parameter. We have a a target domain $\mathcal{X}$ we are trying to approximate over and a set X of N sample points that we choose the location of. The fill distance of a set of points $X = \{x_1, \dots, x_N\} \subseteq \mathcal{X}$ for a bounded domain $\mathcal{X}$ is define to be

$$h_{X,\mathcal{X}} := \sup_{x \in \mathcal{X}} \min_{1 \leq j \leq N} ||x - x_j||_2$$

We show that if the fill distance $h_{X,\mathcal{X}}$ of our sample points X is less than some constant Q depending on the smoothness k of our target function and the shape of our target domain $\mathcal{X} \subset \mathbb{R}^n$, then we can bound the error of our approximation by;

$$||g - \hat{g}_\lambda||_{L_\infty(\mathcal{X})} \leq C \left(h^k_{X,\mathcal{X}} ||g||_{\mathcal{H}_{\mathcal{K}}} + \sqrt{\lambda} ||g||_{\mathcal{H}_{\mathcal{K}}} \right)$$

where C is a constant that depends on the smoothness k, the shape and dimension of $\mathcal{X}$, and the kernel $\mathcal{K}$. We can see that by careful choice of λ the convergence will go as $O(h^k_{X,\mathcal{X}})$, and so we have that the error goes to zero as $O(N^{-k/n})$. This gives us a general worst case bound for function approximation using continuously differentiable kernels, that scales well with smoothness k. This gives a bound that did not exist before in this form for practical kernels such as exponential quadratic and polynomial, and so allows for new methods of parameter design, such as choosing lambda, in their practical application. Proving this bound motivates the work in this paper.

IV. Reproducing Kernel Hilbert Spaces

We have an unknown function $g \in \mathcal{H}_{\mathcal{K}}$ we are trying to approximate over $\mathcal{X} \subset \mathbb{R}^n$, where $\mathcal{H}_{\mathcal{K}}$ is a reproducing kernel space with kernel $\mathcal{K} \in C^{2k}(\mathcal{X} \times \mathcal{X})$, $k \geq 0$. We note that $\mathcal{K} \in C^{2k}(\mathcal{X} \times \mathcal{X})$ means that $\mathcal{H}_{\mathcal{K}} \subset C^k(\mathcal{X})$. We choose a set of points $X = \{x_1, \dots, x_N\} \subset \mathcal{X}$ to evaluate our function on, which we call our centers. The interpolation problem is then defined as

$$\hat{g} \in \arg\min_{h \in \mathcal{H}_{\mathcal{K}}} ||h||^2_{\mathcal{H}_{\mathcal{K}}} \text{ such that } h(x_i) = g(x_i) \;\; \forall x_i \in X$$

and the regression problem is defined as

$$\hat{g}_\lambda \in \arg\min_{h \in \mathcal{H}_{\mathcal{K}}} \sum_{i=1}^N (h(x_i) - g(x_i))^2 + \lambda ||h||^2_{\mathcal{H}_{\mathcal{K}}}, \quad \lambda > 0$$

where λ is a regularisation parameter.

Both of these optimisation problems have the closed form solution [15] below

$$\hat{g}_\lambda(x) = \sum_{i=1}^N w_\lambda(x, x_i) g(x_i), \quad \lambda \geq 0 \tag{1}$$

with $(w_\lambda(x, x_1), \dots, w_\lambda(x, x_N))^T = K_{xX}(K_{XX} + \lambda I_N)^{-1}$ where $K_{xX} := (\mathcal{K}(x, x_1), \dots, \mathcal{K}(x, x_N))$ and $K_{XX} \in R^{N \times N}$ is the matrix whose $(i,j)'th$ element is $\mathcal{K}(x_i, x_j)$.

The interpolation solution is given by $\hat{g}_0$, and so to guarantee a solution in this case we must have that K_{XX} is invertible.

We define $f_\lambda(x) := g(x) - \hat{g}_\lambda(x)$ and so we have that $f_\lambda \in \mathcal{H}_{\mathcal{K}}$. We now want to bound $f_\lambda|X$, the function f_λ restricted just to the points X. In the interpolation setting we have by definition that $\hat{g}_0(x_i) = g(x_i) \;\; \forall x_i \in X$, and so $||f_0|X||_\infty = 0$.

For the regression setting we do not have the same condition, but we note that if we take the solution to the interpolation problem for the same target function and set X, and substitute it into the regression objective function we get

$$\sum_{i=1}^N (\hat{g}_0(x_i) - g(x_i))^2 + \lambda ||\hat{g}_0||^2_{\mathcal{H}_{\mathcal{K}}} = \lambda ||\hat{g}_0||^2_{\mathcal{H}_{\mathcal{K}}}$$

Since $\hat{g}_\lambda$ is the unique minimizer of this objective function we immediately have

$$\sum_{i=1}^N (\hat{g}_\lambda(x_i) - g(x_i))^2 + \lambda ||\hat{g}_\lambda||^2_{\mathcal{H}_{\mathcal{K}}} \leq \lambda ||\hat{g}_0||^2_{\mathcal{H}_{\mathcal{K}}}$$

$$\implies \sum_{i=1}^N (\hat{g}_\lambda(x_i) - g(x_i))^2 \leq \lambda ||\hat{g}_0||^2_{\mathcal{H}_{\mathcal{K}}}$$

$$\implies (\hat{g}_\lambda(x_i) - g(x_i))^2 \leq \lambda ||\hat{g}_0||^2_{\mathcal{H}_{\mathcal{K}}} \;\; \forall x_i \in X$$

$$\implies |\hat{g}_\lambda(x_i) - g(x_i)| \leq \sqrt{\lambda} ||\hat{g}_0||_{\mathcal{H}_{\mathcal{K}}} \;\; \forall x_i \in X$$

and so in the regression setting we have that $||f_\lambda|X||_\infty \leq \sqrt{\lambda} ||\hat{g}_0||_{\mathcal{H}_{\mathcal{K}}}$. If we note from above that $||f_0|X||_\infty = 0$, we have that $||f_\lambda|X||_\infty \leq \sqrt{\lambda} ||\hat{g}_0||_{\mathcal{H}_{\mathcal{K}}}$ for all $\lambda \geq 0$.

We now investigate bounding $||f_\lambda|\mathcal{X}||_\infty$, f_λ over our whole target domain $\mathcal{X}$.

V. Bounds on Convergence Rate

Our main result in this section will be to bound our error function f_λ in terms of it value at the points in X. We will drop the λ for this section as it does not effect any of the results, and we note that from its definition we have that its is continuously differentiable up to degree k. This bound will give us the learning rate, that is the convergence rate of our error function to zero. This result will require two intermediate steps. First we will bound the difference between f and its Taylor polynomial using a standard Taylor's theorem argument, and then we will derive conditions on the points X such that we can bound the Taylor polynomial in terms its value on X.

The result of this section will be a bound that depends on the diameter of the set it is defined over, so we must be able to control said diameter to get a useful bound. Therefore our analysis in this section will focus on a subset $\Omega \subset \mathcal{X}$, and then after we will show that we can cover $\mathcal{X}$ by many such subsets. We also define $X_\Omega = X \cap \Omega$.

To bound the difference between our function f and its Taylor Polynomial, and also the derivatives of this difference, we must first determine type of space this is defined for.

Definition 1. A set $A \subset \mathbb{R}^n$ is said to be star shaped with respect to a point $a \in A$ if $\forall x \in A$ the line segment from a to x is entirely contained in A.

A set Ω being star shaped with respect to some point a guarantees we can bound f over all of Ω relative to its Taylor polynomial at a. With this definition in hand we can now prove Theorem 1 below.

Theorem 1. *Let open bounded $\Omega \subset \mathbb{R}^n$ be star shaped with respect to some point $a \in \Omega$. Let $R = \max_{x,y\in\Omega} ||x-y||_\infty$. If $f \in C^k(\Omega)$, with $k \geq 1$ and $k > |\alpha|$ for α a multi-index then we have that*

$$||D^\alpha f - D^\alpha T_a^{k-1} f||_{L_\infty(\Omega)} \leq C_{k,n,|\alpha|} R^{k-|\alpha|} \max_{|\beta|=k} ||D^\beta f||_{L_\infty(\Omega)}$$

for $C_{k,n,|\alpha|} = \frac{n^{k-|\alpha|}}{(k-|\alpha|)!}$. $T_a^{k-1} f(x)$ here is the $k-1$-th order Taylor polynomial of f at $a \in \mathbb{R}^n$.

Proof. From Taylors theorem [26] we have;

$$|f(x) - T_a^{k-1} f(x)| \leq \sum_{|\omega|=k} (x-a)^\omega \frac{1}{\omega!} \max_{|\beta|=k} \max_{y\in\Omega} |D^\beta f(y)|$$

Now we have that for $x, a \in \Omega$

$$\begin{aligned}(x-a)^\omega &= (x_1-a_1)^{\omega_1}(x_2-a_2)^{\omega_2}\dots(x_n-a_n)^{\omega_n} \\ &\leq R^{\omega_1}R^{\omega_2}\dots R^{\omega_n} = R^{|\omega|} = R^k\end{aligned}$$

$$\implies ||f - T_a^{k-1} f||_{L_\infty(\Omega)} \leq R^k \max_{|\beta|=k} ||D^\beta f||_{L_\infty(\Omega)} \sum_{|\omega|=k} \frac{1}{\omega!}$$

The value of the remaining sum can be found by starting from the sum of multinomial coefficients

$$\sum_{\omega_1+\ldots+\omega_n=k} \binom{k}{\omega_1,\ldots,\omega_n} = \sum_{\omega_1+\ldots+\omega_n=k} \frac{k!}{\omega_1!\ldots!\omega_n} = n^k$$

Since by definition $|\omega| = \omega_1 + \ldots + \omega_n$ and $\omega! = \omega_1!\ldots!\omega_n$ we immediately get

$$\sum_{|\omega|=k} \frac{1}{\omega!} = \frac{1}{k!} \sum_{\omega_1+\ldots+\omega_n=k} \frac{k!}{\omega_1!\ldots!\omega_n} = \frac{n^k}{k!}$$

Therefore we have

$$||f - T_a^{k-1} f||_{L_\infty(\Omega)} \leq \frac{n^k}{k!} R^k \max_{|\beta|=k} ||D^\beta f||_{L_\infty(\Omega)}$$

Using the identity $D^\alpha T_a^{k-1} f = T^{k-|\alpha|-1} D^\alpha f$ we have

$$\begin{aligned}||D^\alpha f - D^\alpha T^{k-1} f||_{L_\infty(\Omega)} &= ||D^\alpha f - T^{k-|\alpha|-1} D^\alpha f||_{L_\infty(\Omega)} \\ &\leq \frac{n^{k-|\alpha|}}{(k-|\alpha|)!} R^{k-|\alpha|} \max_{|\beta|=k-|\alpha|} ||D^\beta D^\alpha f||_{L_\infty(\Omega)}\end{aligned}$$

And since $\max_{|\beta|=k-|\alpha|} ||D^\beta D^\alpha f||_{L_\infty(\Omega)} \leq \max_{|\beta|=k} ||D^\beta f||_{L_\infty(\Omega)}$ we have the desired result:

$$||D^\alpha f - D^\alpha T_a^{k-1} f||_{L_\infty(\Omega)} \leq C_{k,n,|\alpha|} R^{k-|\alpha|} \max_{|\beta|=k} ||D^\beta f||_{L_\infty(\Omega)}$$

where $C_{k,n,|\alpha|} = \frac{n^{k-|\alpha|}}{(k-|\alpha|)!}$ □

With this in hand we now turn to the second intermediate step, bounding the Taylor Polynomial $T_a^k f$ and its derivatives, in terms of its value on our set of centers X_Ω. Since $T_a^k f$ is by definition a polynomial of degree at most k, we need to determine conditions on X_Ω so that the value of any such possible polynomial at any point $x \in \Omega$ can be expressed in terms of its value on X_Ω. We do this by deriving conditions on X_Ω such that $\forall x \in \Omega$ there exists a bounded and well defined mapping from any polynomial up to degree k's value on X_Ω to its value on x. We introduce the notion of norming sets to establish sufficient conditions for our sets Ω and X_Ω, using the definition from [25].

Definition 2. Let V be a finite dimensional linear space with norm $||\cdot||_V$ and dual V^*. Given two subspaces $W \subseteq V$ and $Z \subseteq V^*$, the set Z is called a norming set of W under the norm $||\cdot||_V$ if there exists some $c > 0$ such that

$$\sup_{z\in Z,\ ||z||_V=1} |z(w)| \geq c||w||_V \quad \forall w \in W$$

The values c is the norming constant for Z on W.

We can see the utility of norming sets in Lemma 1 below from [7].

Lemma 1. *Suppose $W \subset V$ a finite dimensional linear space and $Z = \{z_1, \ldots, z_M\}$ is a norming set for W under the L_∞ norm with norming constant c. For every $\varphi \in W^*$, the dual of W, there exists a vector $u \in \mathbb{R}^M$ depending only on φ, such that for every $w \in W$*

$$\varphi(w) = \sum_{j=1}^M u_j z_j(w)$$

and

$$\sum_{j=1}^M |u_j| \leq \frac{1}{c} ||\varphi||_{L_\infty(W^*)}$$

Proof. See Theorem 3.4 in [7] □

This result shows us that given a finite norming set Z for a finite dimensional linear space W, any function from the dual space W^* can be expressed as a linear combination of functionals in Z, with bounded coefficient vectors u_j.

We can apply this to our problem by letting $W = \pi_k(\mathbb{R}^n)|\Omega$, the space of n-variate polynomials up to degree k restricted to the set Ω. Our objective is to be able to express the value of any $p \in W$ at any $x \in \Omega$ in terms its values at our centers X_Ω. For any function f its value at a point x can be obtained by applying the point evaluation function δ_x to f, where $\delta_x(f) = f(x)$ by definition. Since clearly $\delta_x \in W^*$ for every $x \in \Omega$, our goal now becomes showing that $\forall x \in \Omega$, δ_x can be expressed as a linear combination of the functions $\delta_{x_i}, x_i \in X_\Omega$. Lemma 1 tells us that if the functions δ_{x_i} form a norming set for W, such a mapping exists and its coefficient vectors are bounded.

We now prove conditions on X_Ω to guarantee that X forms a norming set for $\pi_k(\mathbb{R}^n)|\Omega$.

Lemma 2. *Let $p \in \pi_k(\mathbb{R}^n)$ the space of n-variate polynomials of absolute degree at most k and let $\Omega \subset \mathbb{R}^n$ be bounded. Let the set $X_\Omega = \{x_1, ..., x_N\} \subseteq \Omega$ satisfy:*

- *$\forall x \in \Omega\ \exists x_j \in X_\Omega$ such that there exists a line segment of length ℓ passing through x and x_j and entirely contained in Ω, with $\ell \geq Ak^2||x_j - x||_2$ for some constant $A > 2$*

then the set $Z = \{\delta_{x_1}, \dots, \delta_{x_N}\}$ is a norming set for $\pi_k(\mathbb{R}^n)|\Omega$ under the L_∞ norm with norming constant $\left(1 - \frac{2}{A}\right)$.

Proof. Let $p \in \pi_k(\mathbb{R}^n)$ and let $L = ||p||_{L_\infty(\mathcal{X})}$. Since Ω is bounded $\exists x \in \overline{\Omega}$ such that $p(x) = L$. Since X_Ω satisfies our condition above we have that $\exists a, b \geq 0,\ a + b = \ell$ such that the whole line segment

$$x + t\frac{x_j - x}{||x_j - z||_2}, \quad t \in [-a, b]$$

lies in $\overline{\Omega}$. From this line segment we define

$$\tilde{p}(t) := p(x + t\frac{x_j - x}{||x_j - z||_2}), \quad t \in [-a, b]$$

From [27, p. 91] we have Markov's inequality for a polynomial $q \in \pi_k(\mathbb{R}^1)$ is

$$|q'(t)| \leq k^2||q||_{L_\infty([-1,1])}, \quad t \in [-1, 1]$$

We can easily rescale this to get that $\forall a, b \geq 0,\ a + b = \ell$ and $\forall q \in \pi_k(\mathbb{R}^1)$

$$|q'(t)| \leq \frac{2}{\ell}k^2||q||_{L_\infty([-a,b])}, \quad t \in [-a, b]$$

Combining this with $\tilde{p}(t)$ we have

$$\begin{aligned}
|p(x) - p(x_j)| &\leq \int_0^{||x-x_j||_2} |\tilde{p}'(t)|dt \\
&\leq \frac{2}{\ell}k^2||\tilde{p}||_{L_\infty([-a,b])}\int_0^{||x-x_j||_2} dt \\
&\leq \frac{2}{\ell}k^2||p||_{L_\infty(\mathcal{X})}||x - x_j||_2 \\
&\leq \frac{2}{\ell}k^2||p||_{L_\infty(\mathcal{X})}\frac{\ell}{Ck^2} \\
&\leq \frac{2}{C}||p||_{L_\infty(\mathcal{X})}
\end{aligned}$$

Now combining this result with $|p(x)| = ||p||_{L_\infty(\mathcal{X})}$ we can use the reverse triangle inequality to get

$$\begin{aligned}
&|p(x) - p(x_j)| \geq |\ |p(x)| - |p(x_j)|\ | \\
\implies &\frac{2}{C}||p||_{L_\infty(\mathcal{X})} \geq ||p(x)||_{L_\infty(\mathcal{X})} - |p(x_j)| \\
\implies &|p(x_j)| \geq \left(1 - \frac{2}{C}\right)||p||_{L_\infty(\mathcal{X})}
\end{aligned}$$

Recall that $\delta_{x_j}(p) = p(x_j)$ by definition, giving us

$$|\delta_{x_j}(p)| \geq \left(1 - \frac{2}{C}\right)||p||_{L_\infty(\mathcal{X})}$$

We note that $C > 2$ and so $\left(1 - \frac{2}{C}\right) > 0$, and that $\forall x\ \ ||\delta_x||_{L_\infty} = 1$ by definition. Therefore we have that

$$\sup_{\delta_{x_j} \in Z, ||\delta_{x_j}||_{L_\infty}=1} |\delta_{x_j}(p)| \geq \left(1 - \frac{2}{C}\right)||p||_{L_\infty} \quad \forall p \in \pi_k(\mathbb{R}^n)|\Omega$$

and so the set Z is a norming set for $\pi_k(\mathbb{R}^n)|\Omega$ under the L_∞ norm, with norming constant $(1 - \frac{2}{C})$ ☐

Combining Lemmas 1 and 2 we immediately get conditions on Ω and X_Ω such that any polynomial $p \in \pi_k(\mathbb{R}^n)|\Omega$ can be expressed in terms of its value on X_Ω, with bounded coefficient vectors. However, since we also want to do the same for the derivatives of any polynomial, we will additionally require that our set $\mathcal{X}$ satisfies an interior cone condition, the definition of which below is taken from [7].

Definition 3. A set $\Omega \subseteq \mathbb{R}^n$ is said to satisfy an interior cone condition if there exists an angle $\theta \in (0, \frac{\pi}{2})$ and radius $r > 0$ such that for every $x \in \Omega$ a unit vector $\xi(x)$ exists such that the cone

$$C(x\xi(x), \theta, r) := \{x + \lambda y : y \in R^n, ||y||_2 = 1, y^T\xi(x) \geq \cos\theta, \lambda \in [0, r]\} \tag{2}$$

is contained in Ω.

We can see that any convex set will automatically satisfy an interior cone condition, as will many non-convex sets. The intuition behind a set satisfying some interior cone condition is that for any point in the set there must be some non-zero length vector emanating from the point contained entirely in the intersection of the point and the interior of the set.

We can then use Lemma 3 below from [7] to bound the norm of the derivative of our polynomial by its norm

Lemma 3. *Suppose that $\Omega \subset \mathbb{R}^n$ is bounded and satisfies an interior cone condition with radius $r > 0$ and angle θ. If $p \in \pi_k(\mathbb{R}^n)$ and α a multi-index for which $|\alpha| \leq k$ then*

$$||D^\alpha p||_{L_\infty(\Omega)} \leq \left(\frac{2k^2}{r\sin\theta}\right)^{|\alpha|}||p||_{L_\infty(\Omega)}$$

Proof. See Proposition 11.6 in [7] ☐

We can immediately combine this with Lemmas 1 and 2 to get the following result.

Theorem 2. *Let $p \in \pi_k(\mathbb{R}^n)$ the space of n-variate polynomials of absolute degree at most k, with $k > |\alpha|$ for α a multi-index and let $\Omega \subset \mathbb{R}^n$ be bounded. Let the sets Ω and $X_\Omega = \{x_1, ..., x_N\} \subseteq \Omega$ satisfy:*

1) *$\forall x \in \Omega\ \exists x_j \in X_\Omega$ such that there exists a line segment of length ℓ passing through x and x_j and entirely contained in Ω, with $\ell \geq Ak^2||x_j - x||_2$ for some constant $A > 2$*
2) *if $|\alpha| > 0$ then Ω satisfies an interior cone condition for some $r > 0, \theta \in (0, \frac{\pi}{2})$*

then there exists real numbers $a_j^\alpha(x)$ such that

$$D^\alpha p(x) = \sum_{j=1}^N a_j^\alpha(x)p(x_j)$$

and

$$\sum_{j=1}^N |a_j^\alpha(x)| \leq \frac{A}{A-2}\left(\frac{2k^2}{r\sin\theta}\right)^{|\alpha|}$$

Proof. This proof is a combination of what we sketched out after Lemma 1 and the result of Lemma 3. We have that Ω and X_Ω satisfy the conditions of Lemma 2 and so the set $Z = \{\delta_{x_1}, \dots, \delta_{x_N}\}$ is a norming set for $\pi_k(\mathbb{R}^n)|\Omega$ under the L_∞ norm with norming constant $(1 - \frac{2}{A})$. We note that $|\alpha| < k$ means that $D^\alpha p \in \pi_k(\mathbb{R}^n)|\Omega$. Then applying Lemma 1 to this we immediately have that $\forall \varphi \in (\pi_k(\mathbb{R}^n)|\mathcal{X})^*$ there exists a vector $a \in \mathbb{R}^N$ such that

$$\varphi(D^\alpha p) = \sum_{j=1}^{N} a_j \delta_{x_j}(D^\alpha p), \;\; \forall p \in \pi_k(\mathbb{R}^n)|\Omega$$

Since $\forall x \in \mathcal{X} \;\; \delta_x \in (\pi_k(\mathbb{R}^n)|\Omega)^*$ we have that $\forall x \in \Omega \exists a_j \in \mathbb{R}^N$ such that

$$\delta_x(D^\alpha p) = \sum_{j=1}^{N} a_j D^\alpha p(x_j), \;\; \forall p \in \pi_k(\mathbb{R}^n)|\Omega$$

and so we have that there exists real numbers $a_j(x)$ such that

$$D^\alpha p(x) = \sum_{j=1}^{N} a_j D^\alpha p(x_j), \;\; \forall p \in \pi_k(\mathbb{R}^n)|\Omega$$

if $\alpha = 0$ then the bound follows immediately from $||\delta_x||_{L_\infty} = 1$ and Lemma 1

$$\sum_{j=1}^{N} |a_j(x)| \leq (1 - \frac{2}{A})^{-1} = \frac{A}{A-2}$$

and we are done. If $|\alpha| > 0$ then

$$D^\alpha p(x) = \sum_{j=1}^{N} a_j^\alpha p(x_j), \;\; \forall p \in \pi_k(\mathbb{R}^n)|\Omega$$

with $a_j^\alpha(x) = a_j(x) D^\alpha p(x_j)/p(x_j)$. The bound then follows same as above but with the addition of assuming Ω satisfies some interior cone condition with radius r and angle θ and applying Lemma 3

$$\sum_{j=1}^{N} |a_j^\alpha(x)| \leq \sum_{j=1}^{N} (|a_j(x)|) \, || \frac{D^\alpha p}{p} ||_{L_\infty(\Omega)} \leq \frac{A}{A-2} \left(\frac{2k^2}{r \sin\theta} \right)^{|\alpha|}$$

and we're done. □

We can now combine this result with Theorem 1 to get the main result of this section and our most general result, which bounds f on Ω in terms of the diameter R of the set.

Theorem 3. *Let $f \in C^k(\Omega)$ for open bounded $\Omega \subset \mathbb{R}^n$ which is star shaped, with $k \geq 1$ and $k \geq |\alpha|$ for α a multi-index. Let $R = \max_{x,y \in \Omega} ||x - y||_\infty$. Let the sets Ω and $X_\Omega = \{x_1, ..., x_N\} \subseteq \Omega$ satisfy:*

1) *$\forall x \in \Omega \; \exists x_j \in X_\Omega$ such that there exists a line segment of length ℓ passing through x and x_j and entirely contained in Ω, with $\ell \geq Ak^2 ||x_j - x||_2$ for some constant $A > 2$*
2) *if $|\alpha| > 0$ then Ω satisfies an interior cone condition for some $r > 0, \theta \in (0, \frac{\pi}{2})$*

Then we have that:

$$||D^\alpha f||_{L_\infty(\Omega)} \leq C_{k,n,|\alpha|,R,r,\theta} \left(R^{k-|\alpha|} \max_{|\beta|=k} ||D^\beta f||_{L_\infty(\Omega)} + R^{-|\alpha|} ||f|X_\Omega||_\infty \right)$$

where $f|X_\Omega$ represents the function f restricted to the set X_Ω and

$$C_{k,n,|\alpha|,R,r,\theta} = \max\Bigg\{ \{ \frac{n^{k-|\alpha|}}{(k-|\alpha|)!} + \frac{A}{A-2} \left(\frac{2(k-1)^2 R}{r \sin\theta} \right)^{|\alpha|} \left(\frac{n^k}{k!} \right), \frac{A}{A-2} \left(\frac{2(k-1)^2 R}{r \sin\theta} \right)^{|\alpha|} \Bigg\}$$

Proof. Recall that by definition of star shaped, Ω is star shaped with respect to some point $a \in \Omega$. Then from the triangle inequality we have that

$$||D^\alpha f||_{L_\infty(\Omega)} \leq ||D^\alpha f - D^\alpha T_a^k f||_{L_\infty(\Omega)} + ||D^\alpha T_a^k f||_{L_\infty(\Omega)} \quad (3)$$

Applying Lemma 1 to this we immediately get

$$||D^\alpha f||_{L_\infty(\Omega)} \quad (4)$$
$$\leq \frac{n^{k-|\alpha|}}{(k-|\alpha|)!} R^{k-|\alpha|} \max_{|\beta|=k} ||D^\beta f||_{L_\infty(\Omega)} + ||D^\alpha T_a^k f||_{L_\infty(\Omega)} \quad (5)$$

We can bound the second term in (3) using Theorems 1 and 2 together

$$\begin{aligned}
|D^\alpha T_a^k f(x)| &\leq \sum_{j=1}^{N} |a_j^\alpha(x)| \cdot |T_a^k f(x_j)| \\
&\leq \frac{A}{A-2} \left(\frac{2k^2}{r \sin\theta} \right)^{|\alpha|} \max_{1 \leq j \leq N} |f(x_j) - f(x_j) + T_a^k f(x_j)| \\
&\leq \frac{A}{A-2} \left(\frac{2k^2}{r \sin\theta} \right)^{|\alpha|} \left(||f - T_a^k f||_{L_\infty(\Omega)} + ||f|X_\Omega||_\infty \right) \\
&\leq \frac{A}{A-2} \left(\frac{2k^2}{r \sin\theta} \right)^{|\alpha|} \left(\frac{n^k}{k!} R^k \max_{|\beta|=k} ||D^\beta f||_{L_\infty(\Omega)} + ||f|X_\Omega||_\infty \right)
\end{aligned}$$

$$\therefore ||D^\alpha T^k f||_{L_\infty(\Omega)} \leq \frac{A}{A-2} \left(\frac{2k^2 R}{r \sin\theta} \right)^{|\alpha|} \cdot \left(\frac{n^k}{k!} R^{k-|\alpha|} \max_{|\beta|=k} ||D^\beta f||_{L_\infty(\Omega)} + R^{-|\alpha|} ||f|X_\Omega||_\infty \right)$$

Combining the above with (4) we then have

$$\begin{aligned}
&||D^\alpha f||_{L_\infty(\Omega)} \\
&\leq \frac{n^{k-|\alpha|}}{(k-|\alpha|)!} R^{k-|\alpha|} \max_{|\beta|=k} ||D^\beta f||_{L_\infty(\Omega)} \\
&+ \frac{A}{A-2} \left(\frac{2k^2 R}{r \sin\theta} \right)^{|\alpha|} \left(\frac{n^k}{k!} R^{k-|\alpha|} \max_{|\beta|=k} ||D^\beta f||_{L_\infty(\Omega)} + R^{-|\alpha|} ||f|X_\Omega||_\infty \right) \\
&= \left(\frac{n^{k-|\alpha|}}{(k-|\alpha|)!} + \frac{A}{A-2} \left(\frac{2k^2 R}{r \sin\theta} \right)^{|\alpha|} \left(\frac{n^k}{k!} \right) \right) R^{k-|\alpha|} \max_{|\beta|=k} ||D^\beta f||_{L_\infty(\Omega)} \\
&+ \frac{A}{A-2} \left(\frac{2k^2 R}{r \sin\theta} \right)^{|\alpha|} R^{-|\alpha|} ||f|X_\Omega||_\infty \\
&\leq C_{k,n,|\alpha|,R,r,\theta} \left(R^{k-|\alpha|} \max_{|\beta|=k} ||D^\beta f||_{L_\infty(\Omega)} + R^{-|\alpha|} ||f|X_\Omega||_\infty \right)
\end{aligned}$$

where

$$C_{k,n,|\alpha|,R,r\theta} = \max\left\{ \frac{n^{k-|\alpha|}}{(k-|\alpha|)!} + \frac{A}{A-2} \left(\frac{2k^2 R}{r \sin\theta} \right)^{|\alpha|} \left(\frac{n^k}{k!} \right), \frac{A}{A-2} \left(\frac{2k^2 R}{r \sin\theta} \right)^{|\alpha|} \right\}$$

□

We can see that this result shows that the convergence has two terms that go as $O(R^{k-|\alpha|})$ and $O(R^{-|\alpha|})$ respectively. Assuming small R, convergence will increase with smoothness k and for $|\alpha| = 0$ i.e. just bounding f itself, we can see the convergence rate goes just as $O(R^k)$. However, for increasing values of $|\alpha|$ the $O(R^{-|\alpha|})$ term will be growing larger, and so we must be more careful to keep $||f|X_\Omega||_\infty$ small to get good convergence.

VI. Bounds in Terms of Fill Distance

We can now see that if we can show that $\mathcal{X}$ can be covered by subsets Ω satisfying the above conditions then we can bound f on $\mathcal{X}$. How we define these subsets depends on how we want to define our set of centers X. One commonly used metric for describing how a set of centers X covers a set $\mathcal{X}$ is the fill distance, who's definition we take from [7]

Definition 4. The fill distance of a set of points $X = \{x_1, \ldots, x_N\} \subseteq \mathcal{X}$ for a bounded domain $\mathcal{X}$ is define to be

$$h_{X,\mathcal{X}} := \sup_{x \in \mathcal{X}} \min_{1 \le j \le N} ||x - x_j||_2$$

One way to interpret this fill distance is as the radius of the largest ball which is contained in $\mathcal{X}$ and does not contain a center x_j. This means that any ball contained in $\mathcal{X}$ with radius larger than $h_{X,\mathcal{X}}$ is guaranteed to contain a center x_j. We note that for any subset Ω of $\mathcal{X}$, any ball contained in Ω with radius larger than $h_{X,\mathcal{X}}$ is also guaranteed to contain a center x_j. This property can be used to define subsets that satisfy our conditions by noting one simple property of cones, that we outline below.

Lemma 4. *Suppose that $K(x,\xi,\theta,r)$ is a cone defined as in* (2). *Then for every $h \le r/(1+\sin\theta)$ the closed ball $B_{h\sin\theta}(y)$ with center $y = x + h\epsilon$ and radius $h\sin\theta$ is contained in $K(x,\xi,\theta,r)$. In particular if z is a point from this ball then the whole line segment $x + t(z-x)/||z-x||_2$, $t \in [0,r]$ is contained in the cone.*

Proof. See Lemma 3.7 in [7] □

We can now prove conditions on a subset Ω of $\mathcal{X}$, for X with a fill distance $h_{X,\mathcal{X}}$.

Lemma 5. *Let $f \in C^k(\mathcal{X})$ for bounded $\mathcal{X} \subset \mathbb{R}^n$ with $k \ge |\alpha|$ for α a multi-index, and let X have fill distance $h_{X,\mathcal{X}}$ over $\mathcal{X}$. Let $\Omega \subseteq \mathcal{X}$ be open, star shaped with respect to some point a, and satisfy a cone condition with radius $r > 0$ and angle $\theta \in (0, \pi/2)$. If the fill distance satisfies*

$$h_{X,\mathcal{X}} \le \frac{r\sin\theta}{Ak^2(1+\sin\theta)}$$

for constant $A > 2$, then Ω satisfies the conditions of Theorem 3.

Proof. Since Ω satisfies an interior cone condition we have that for any $x \in \Omega$ and any $h \le r/(1+\sin\theta)$ there is a line segment from x to any point in some ball of radius $h\sin\theta$. We have

$$h_{X,\mathcal{X}} \le \frac{r\sin\theta}{4k^2(1+\sin\theta)}$$
$$\implies \frac{h_{X,\mathcal{X}}}{\sin\theta} \le r/(1+\sin\theta)$$

and so for any point $x \in \Omega$ there is a line segment of length r from x passing through any point in the closed ball $B_{h_{X,\mathcal{X}}}(x + h_{X,\mathcal{X}})$. By definition any closed ball of radius $h_{X,\mathcal{X}}$ contains a center, and so there exists a line segment of length r from any point x passing through some center x_j. We also have

$$||x - x_j|| \le ||x - (x+h_{X,\mathcal{X}})|| + ||x + h_{X,\mathcal{X}} - x_j||$$
$$\le h_{X,\mathcal{X}} + \frac{h_{X,\mathcal{X}}}{\sin\theta} = h_{X,\mathcal{X}}\frac{1+\sin\theta}{\sin\theta}$$

Combining these we have that for any $x \in \Omega$ there exists a line segment of length r contained in Ω, passing through x and some center x_j and

$$r \ge 4k^2 h_{X,\mathcal{X}} \frac{1+\sin\theta}{\sin\theta} \ge 4k^2||x - x_j||_2$$

and so clearly all conditions of Theorem 3 are satisfied. □

Finally we see that to bound f on $\mathcal{X}$ in terms of the fill distance $h_{X,\mathcal{X}}$ of X, we simply need to show that every $x \in \mathcal{X}$ belongs to some subset Ω satisfying the above conditions. We can do so by covering $\mathcal{X}$ by specifically constructed subsets Ω_t, which has been done already in [7], and so we use that construction here.

Let $\mathcal{X}$ be bounded and satisfy an interior cone condition with radius R_0 and angle φ, and let $A > 2$ some constant. We introduce the following quantities. Let

$$\begin{aligned} \theta &:= 2\sin^{-1}(\sin\varphi/4(1+\sin\varphi)) \\ Q(k,\varphi,A) &:= \frac{\sin\varphi\sin\theta}{2Ak^2(1+\sin\varphi)(1+\sin\theta)} \\ R &:= \frac{h_{X,\mathcal{X}}}{Q(k,\varphi)} \\ r &:= R\frac{\sin\varphi}{2(1+\sin\varphi)} \end{aligned} \tag{6}$$

From these we then define the sets $T_r := \{t \in \frac{2r}{\sqrt{n}}\mathbb{Z}^n : B_r(t) \in \mathcal{X}\}$ and

$$\Omega_t := \{x \in \mathcal{X} : \text{co}(\{x\} \cup B_r(t)) \subseteq \mathcal{X} \cap B_R(t)\}, \quad t \in T_r$$

where co(D) denotes the closed convex hull of the set D. With these quantities defined we have the following lemma from [7].

Lemma 6. *With the quantities just introduced, assume $h_{X,\mathcal{X}} \le Q(k,\varphi)R_0$. Then the following hold:*

1) *Each Ω_t is star shaped with respect to the point t.*
2) *Each Ω_t satisfies an interior cone condition with radius r and angle θ.*
3) *$\mathcal{X} = \cup_{t \in T_r}\Omega_t$ and $2r \le \max_{x,y\in\Omega_t}||x-y|| \le 2R = \frac{2h_{X,\mathcal{X}}}{Q(k,\varphi,A)}$*

With this is hand we can prove our bound in terms of the fill distance $h_{X,\mathcal{X}}$.

Theorem 4. *Let $f \in C^k(\mathcal{X})$ for bounded $\mathcal{X} \subset \mathbb{R}^n$ satisfying an interior cone condition with radius R_0 and angle φ, with $k \ge |\alpha|$ for α a multi-index. If the fill distance $h_{X,\mathcal{X}}$ of X over $\mathcal{X}$ satisfies $h_{X,\mathcal{X}} \le Q(k,\varphi,A)R_0$ for some constant $A > 2$ we have that*

$$||D^\alpha f||_{L_\infty(\mathcal{X})} \le C_{k,n,|\alpha|,\varphi}\left(h_{X,\mathcal{X}}^{k-|\alpha|}\max_{|\beta|=k}||D^\beta f||_{L_\infty(\mathcal{X})} + h_{X,\mathcal{X}}^{-|\alpha|}||f|X||_\infty\right)$$

where

$$C_{k,n,|\alpha|,\varphi} = \max\left\{\frac{n^{k-|\alpha|}}{(k-|\alpha|)!} + B_{k,|\alpha|,\varphi}\left(\frac{n^k}{k!}\right), B_{k,|\alpha|,\varphi}\right\}$$

and

$$B_{k,|\alpha|,\varphi} = \frac{A}{A-2}\left(\frac{4(k-1)^2(1+\sin\varphi)}{\sin\varphi\sin\left(2\sin^{-1}(\sin\varphi/4(1+\sin\varphi))\right)}\right)^{|\alpha|}$$

Proof. We use the quantities introduced in the paragraph before Lemma 6. Since we have that $h_{X,\mathcal{X}} \leq Q(k,\varphi)R_0$ we can apply Lemma 6. We also note that

$$h_{X,\mathcal{X}} = \frac{r\sin\theta}{4k^2(1+\sin\theta)}$$

by construction. Combining these two facts we see that each set Ω_t satisfies the conditions for Lemma 5, and every $x \in \mathcal{X}$ belongs to some Ω_t. Therefore

$$||D^\alpha f||_{L_\infty(\mathcal{X})} \leq C'_{k,n,|\alpha|,R,r,\theta}\left((2R)^{k-|\alpha|}\max_{|\beta|=k}||D^\beta f||_{L_\infty(\mathcal{X})} + (2R)^{-|\alpha|}||f|X||_\infty\right)$$

where

$$C'_{k,n,|\alpha|,R,r,\theta} = \max\left\{\frac{n^{k-|\alpha|}}{(k-|\alpha|)!} + 2\left(\frac{k^2(2R)}{r\sin\theta}\right)^{|\alpha|}\left(\frac{n^k}{k!}\right), 2\left(\frac{2k^2(2R)}{r\sin\theta}\right)^{|\alpha|}\right\}$$

Substituting in values for R, r, and θ and rearranging gives us

$$||D^\alpha f||_{L_\infty(\mathcal{X})} \leq C_{k,n,|\alpha|,\varphi}\left(h_{X,\mathcal{X}}^{k-|\alpha|}\max_{|\beta|=k}||D^\beta f||_{L_\infty(\mathcal{X})} + h_{X,\mathcal{X}}^{-|\alpha|}||f|X||_\infty\right)$$

where

$$C_{k,n,|\alpha|,\varphi} = \max\left\{\frac{n^{k-|\alpha|}}{(k-|\alpha|)!} + B_{k,|\alpha|,\varphi}\left(\frac{n^k}{k!}\right), B_{k,|\alpha|,\varphi}\right\}$$

and

$$B_{k,|\alpha|,\varphi} = \frac{A}{A-2}\left(\frac{4k^2(1+\sin\varphi)}{\sin\varphi\sin\left(2\sin^{-1}\left(\sin\varphi/4(1+\sin\varphi)\right)\right)}\right)^{|\alpha|}$$

and so we have the desired result. □

We can see this bound has the same properties as that in Theorem 3, but now goes as fill distance $h_{X,\mathcal{X}}$ instead of R, and covers the whole set $\mathcal{X}$. This means that for bounding f ($|\alpha| = 0$) our bound will improve as we decrease the fill distance, but for $|\alpha| > 0$ we must decrease $||f|X||_\infty$ in conjunction with the fill distance.

VII. Main Result

We can now return to our RKHS setup and, reintroducing the lambda, recall that $f_\lambda = g - \hat{g}_\lambda$, with g our target function and $\hat{g}_\lambda$ our approximation both in $H_\mathcal{K}$, for kernel $\mathcal{K} \in C^{2k}(\mathcal{X}\times\mathcal{X})$. We note also that we showed that $||f_\lambda|X||_\infty \leq \sqrt{\lambda}||\hat{g}_0||_{\mathcal{H}_\mathcal{K}}$. Substituting these into our bound in Theorem 4 gives us

$$||D^\alpha(g-\hat{g}_\lambda)||_{L_\infty(\mathcal{X})} \leq C\left(h_{X,\mathcal{X}}^{k-|\alpha|}\max_{|\beta|=k}||D^\beta(g-\hat{g}_\lambda)||_{L_\infty(\mathcal{X})} + \sqrt{\lambda}h_{X,\mathcal{X}}^{-|\alpha|}||\hat{g}_0||_{\mathcal{H}_\mathcal{K}}\right)$$

where C is the constant defined there.

To get our final result we want the right hand side to depend on the smoothness of the target function g, and so we note an important result from Theorem 1 in [13] which bounds the derivatives of a function in a kernel space by its kernel norm:

$$\sum_{|\beta|\leq k}||D^\beta f||_{L_\infty(\mathcal{X})} \leq \sqrt{n^k\sum_{|\beta|\leq 2k}||D^\beta\mathcal{K}||_{L_\infty(\mathcal{X})}}||f||_{\mathcal{H}_\mathcal{K}}, \forall f \in \mathcal{H}_\mathcal{K}$$

which applied in our case immediately gives us

$$\max_{|\beta|=k}||D^\beta(g-\hat{g}_\lambda)||_{L_\infty(\mathcal{X})} \leq \sqrt{n^k\sum_{|\beta|\leq 2k}||D^\beta\mathcal{K}||_{L_\infty(\mathcal{X})}}||g-\hat{g}_\lambda||_{\mathcal{H}_\mathcal{K}}$$

We see also that because $\hat{g}_\lambda$ is the minimiser to the interpolation/regression problem, we have that $||\hat{g}_\lambda||_{\mathcal{H}_\mathcal{K}} \leq ||g||_{\mathcal{H}_\mathcal{K}}$ and so combining these we finally have

$$||D^\alpha(g-\hat{g}_\lambda)||_{L_\infty(\mathcal{X})} \leq C'\left(h_{X,\mathcal{X}}^{k-|\alpha|}||g||_{\mathcal{H}_\mathcal{K}} + \sqrt{\lambda}h_{X,\mathcal{X}}^{-|\alpha|}||g||_{\mathcal{H}_\mathcal{K}}\right)$$

where

$$C' = 2\sqrt{n^k\sum_{|\beta|\leq k}||D^\beta\mathcal{K}||_{L_\infty(\mathcal{X})}} \cdot C$$

is a constant. It is this final step that allows us to avoid the requirements of a specific smoothness measure and work with any continuously differentiable kernel. We state this result finally as the below theorem, which we outlined above in Section 4.

Theorem 5. *Let $g \in \mathcal{H}_k$, be our target function with $\mathcal{K} \in C^{2k}(\mathcal{X}\times\mathcal{X})$, and $\hat{g}_\lambda$ defined as in 1, with $\lambda \geq 0$ our smoothing parameter. We have $k \geq |\alpha|$ for α a multi-index. Let $\mathcal{X} \subset \mathbb{R}^n$ be bounded and satisfying an interior cone condition with radius R_0 and angle φ.*

If the fill distance $h_{X,\mathcal{X}}$ of X over $\mathcal{X}$ satisfies $h_{X,\mathcal{X}} \leq Q(k,\varphi,A)R_0$ for $Q(k,\varphi,A)$ a constant determined by the target degree of smoothness k, and the interior cone condition parameters φ and A, and defined as in (6)*, then*

$$|\alpha|\,||D^\alpha(g-\hat{g}_\lambda)||_{L_\infty(\mathcal{X})} \leq C_{k,n,|\alpha|,\varphi,\mathcal{K}}\left(h_{X,\mathcal{X}}^{k-|\alpha|}||g||_{\mathcal{H}_\mathcal{K}} + \sqrt{\lambda}h_{X,\mathcal{X}}^{-|\alpha|}||g||_{\mathcal{H}_\mathcal{K}}\right)$$

where we have the three constants $C_{k,n,|\alpha|,\varphi,\mathcal{K}}$, $B_{k,|\alpha|,\varphi}$ and $D_{k,n,\mathcal{K}}$ defined by

$$C_{k,n,|\alpha|,\varphi,\mathcal{K}} = \max\left\{D_{k,n,\mathcal{K}}\cdot\left(\frac{n^{k-|\alpha|}}{(k-|\alpha|)!} + B_{k,|\alpha|,\varphi}\left(\frac{n^k}{k!}\right)\right), B_{k,|\alpha|,\varphi}\right\}$$

$$B_{k,|\alpha|,\varphi} = \frac{A}{A-2}\left(\frac{4(k-1)^2(1+\sin\varphi)}{\sin\varphi\sin\left(2\sin^{-1}\left(\sin\varphi/4(1+\sin\varphi)\right)\right)}\right)^{|\alpha|}$$

$$D_{k,n,\mathcal{K}} = 2\sqrt{n^k\sum_{|\beta|\leq k}||D^\beta\mathcal{K}||_{L_\infty(\mathcal{X})}}$$

VIII. Discussion

Our bound allows us to immediately extend existing results limited to Sobolev-equivalent kernels, to any continuously differentiable kernel. We will look at two important examples below.

A. Smoothing Parameter Selection

One example is the optimal choice of smoothing parameter λ, in cases where we need it to be greater than zero. This is extremely important in cases where there may be perturbations in the data, and so the interpolation problem could give undesirable solutions. Such a result for Sobolev spaces is typified by Proposition 3.6 from [10], which gives an error bound equal to the interpolation bound, by correct choice of $\lambda > 0$. We can immediately extend this result to any continuously differentiable kernel.

From Theorem 5 we have that, under the given conditions, if we choose $\lambda \leq h_{X,\mathcal{X}}^{2k}$ then

$$||D^\alpha(g - \hat{g}_\lambda)||_{L_\infty(\mathcal{X})} \leq C h_{X,\mathcal{X}}^{k-|\alpha|} ||g||_{\mathcal{H}_\mathcal{K}}$$

with C the constant defined in Theorem 5, and so we guarantee the convergence scales as it does in the interpolation case, going as $O(N^{(|\alpha|-k)/n})$, N the number of sample points, n the dimension. This gives us new convergence guarantees for kernels used in many practical applications, that were not covered by the Sobolev work.

B. Kernel-based Quadrature

Another important application of RKHS for which we can extend results is the area of kernel-based quadrature, or Bayesian quadrature as it also known. This is a method of numerical integration that assumes the integrand belongs to some RKHS, and constructs weighted function evaluation points to keep the worst case error small.

Simply stated, given some known probability distribution P on $\mathcal{X}$ and g our integrand, the goal is to choose $\{(w_i, x_i)\}_{i=1}^N$ so that the approximation

$$\int g(x)dP(x) \approx \sum_{i=1}^N w_i g(x_i)$$

has small worst case error. Choosing the weights w_i amounts to solving the RKHS interpolation problem defined above, and so this is equivalent to minimising the worst case error for

$$\int g(x)dP(x) \approx \int \hat{g}_0(x)dP(x)$$

A standard result in this field is Theorem 9 from [8], which leaving aside their addition of misspecified smoothness, gives a bound under the assumption that g belongs to some Sobolev space W_2^k. We can again immediately extend this result to general continuously differentiable kernels. From Theorem 5 with its given assumptions, applied as in [8], we have that

$$\left|\int g(x)dP(x) - \int \hat{g}_0(x)dP(x)\right| \leq C||P||_{L^2(\mathcal{X})} h_{X,\mathcal{X}}^k ||g||_{\mathcal{H}_\mathcal{K}}$$

where C is the constant in Theorem 5. This is a worst case bound for kernel-quadrature with general continuously differentiable kernels, giving us new approximation guarantees for a wide range of popular kernel choices.

IX. Conclusions and Further Work

Our work extends previous results to more useful kernel choices. This gives us new and improved bounds in areas such as parameter selection and numerical integration, with bounds for any choice of continuously differentiable kernel. Our analysis covers popular choices of kernel such as polynomial or exponential quadratic kernels, that are missed my previous function approximation work. We have also given explicit values for the constants appear throughout our proofs, unlike both [7] and [8], making clear the dependencies on the variables.

There is more work that can be done in seeing if our most general result, Theorem 3, can be used to define different bounds depending on how we describe the set of centers X. Our main result follows from using the fill distance and interior cone condition, but the generality of Theorem 3 allows for other approaches.

References

[1] J. Long and J. Han, "Reinforcement learning with function approximation: From linear to nonlinear," 2023.

[2] F. Huszar and D. Duvenaud, "Optimally-weighted herding is bayesian quadrature," 2016.

[3] N. Srinivas, A. Krause, S. M. Kakade, and M. W. Seeger, "Information-theoretic regret bounds for gaussian process optimization in the bandit setting," *IEEE Transactions on Information Theory*, vol. 58, pp. 3250–3265, may 2012.

[4] C. Rieger and B. Zwicknagl, "Deterministic error analysis of support vector regression and related regularized kernel methods," *J. Mach. Learn. Res.*, vol. 10, p. 2115–2132, dec 2009.

[5] M. Kanagawa, B. K. Sriperumbudur, and K. Fukumizu, "Convergence analysis of deterministic kernel-based quadrature rules in misspecified settings," *Foundations of Computational Mathematics*, vol. 20, no. 1, pp. 155–194, 2020.

[6] A. Christmann and I. Steinwart, "Consistency and robustness of kernel-based regression in convex risk minimization," *Bernoulli*, vol. 13, aug 2007.

[7] H. Wendland, *Scattered Data Approximation*. Cambridge Monographs on Applied and Computational Mathematics, Cambridge University Press, 2005.

[8] G. Wynne, F.-X. Briol, and M. Girolami, "Convergence guarantees for gaussian process means with misspecified likelihoods and smoothness," 2021.

[9] F. J. Narcowich, J. D. Ward, and H. Wendland, "Sobolev error estimates and a bernstein inequality for scattered data interpolation via radial basis functions," *Constructive Approximation*, vol. 24, pp. 175–186, 2006.

[10] H. Wendland and C. Rieger, "Approximate interpolation with applications to selecting smoothing parameters," *Numer. Math.*, vol. 101, p. 729–748, oct 2005.

[11] V. Maz'ya, *Sobolev spaces*. Springer, 2013.

[12] H. Brezis and H. Brézis, *Functional analysis, Sobolev spaces and partial differential equations*, vol. 2. Springer, 2011.

[13] D.-X. Zhou, "Derivative reproducing properties for kernel methods in learning theory," *Journal of Computational and Applied Mathematics*, vol. 220, no. 1, pp. 456–463, 2008.

[14] G. S. Kimeldorf and G. Wahba, "A correspondence between bayesian estimation on stochastic processes and smoothing by splines," *The Annals of Mathematical Statistics*, vol. 41, no. 2, pp. 495–502, 1970.

[15] M. Kanagawa, P. Hennig, D. Sejdinovic, and B. K. Sriperumbudur, "Gaussian processes and kernel methods: A review on connections and equivalences," 2018.

[16] F. Cucker and D.-X. Zhou, "Learning theory: An approximation theory viewpoint," 01 2007.

[17] E. D. Vito, L. Rosasco, A. Caponnetto, U. D. Giovannini, and F. Odone, "Learning from examples as an inverse problem," *Journal of Machine Learning Research*, vol. 6, no. 30, pp. 883–904, 2005.

[18] A. Caponnetto and E. De Vito, "Optimal rates for the regularized least-squares algorithm," *Foundations of Computational Mathematics*, vol. 7, pp. 331–368, 07 2007.

[19] F. Cucker and S. Smale, "Best choices for regularization parameters in learning theory: On the bias—variance problem," *Foundations of Computational Mathematics*, vol. 2, pp. 413–428, 2002.

[20] T. Hofmann, B. Schölkopf, and A. J. Smola, "Kernel methods in machine learning," 2008.

[21] B. Schölkopf, K. Tsuda, J.-P. Vert, *et al.*, *Kernel methods in computational biology*. MIT press, 2004.

[22] A. D. Bull, "Convergence rates of efficient global optimization algorithms," 2011.

[23] A. M. S. Barreto, D. Precup, and J. Pineau, "Practical kernel-based reinforcement learning," 2014.

[24] Z. Yang, C. Jin, Z. Wang, M. Wang, and M. I. Jordan, "On function approximation in reinforcement learning: Optimism in the face of large state spaces," 2020.

[25] K. Jetter, J. Stöckler, and J. D. Ward, "Error estimates for scattered data interpolation on spheres," *Math. Comput.*, vol. 68, pp. 733–747, 1999.

[26] G. B. Folland, "Remainder estimates in taylor's theorem," *The American Mathematical Monthly*, vol. 97, no. 3, pp. 233–235, 1990.

[27] E. W. Cheney, "Introduction to approximation theory," 1966.