Proceedings of Machine Learning Research vol 99:1–29, 2019 32nd Annual Conference on Learning Theory

Consistency of Interpolation with Laplace Kernels is a High-Dimensional Phenomenon

Alexander Rakhlin
Xiyu Zhai
Massachusetts Institute of Technology

Editors: Alina Beygelzimer and Daniel Hsu

Abstract

We show that minimum-norm interpolation in the Reproducing Kernel Hilbert Space corresponding to the Laplace kernel is not consistent if input dimension is constant. The lower bound holds for any choice of kernel bandwidth, even if selected based on data. The result supports the empirical observation that minimum-norm interpolation (that is, exact fit to training data) in RKHS generalizes well for some high-dimensional datasets, but not for low-dimensional ones.

Keywords: List of keywords

1. Introduction

Can a method perfectly fitting the training data perform well out-of-sample? In the last few years, this question was raised in the context of over-parametrized neural networks (Zhang et al., 2016; Belkin et al., 2018b), kernel methods (Belkin et al., 2018b; Liang and Rakhlin, 2018), and local nonparametric rules (Belkin et al., 2018a,c). Experiments on a range of real and synthetic datasets confirm that procedures attaining zero training error do not necessarily overfit and can generalize well (Wyner et al., 2017; Zhang et al., 2016; Belkin et al., 2018b; Liang and Rakhlin, 2018). In particular, Kernel Ridge Regression

$$\widehat{f} \in \operatorname*{argmin}_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}(f(x_i)-y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2 \tag{1}$$

performs "unreasonably well" in the regime $\lambda = 0$, even though the solution (generally) interpolates the data. Here $\mathcal{H}$ is a Reproducing Kernel Hilbert Space (RKHS) corresponding to a kernel K, $\|\cdot\|_{\mathcal{H}}$ is the corresponding RKHS norm, and $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$ are the training data. Since the argmin in (1) is not unique when $\lambda = 0$, we consider the minimum-norm interpolating solution

$$\begin{aligned} &\operatorname*{argmin}_{f\in\mathcal{H}} \quad \|f\|_{\mathcal{H}} \\ &\text{s.t. } \ f(x_i) = y_i, \ \ i = 1, \ldots, n \end{aligned} \tag{2}$$

The conditions under which interpolation, such as Kernel "Ridgeless" Regression, performs well are poorly understood. (Liang and Rakhlin, 2018) studied the high-dimensional regime $n \asymp d$, explicating (under additional assumptions) a phenomenon of implicit regularization, due to the curvature of the kernel function, high dimensionality, and favorable geometric properties of the training data, as quantified by the spectral decay of the kernel and covariance matrices.

The mechanism of implicit regularization in (Liang and Rakhlin, 2018) relies on high dimensionality d of the input space, and it is unclear whether such a "blessing of high dimensionality" is necessary for good out-of-sample performance of interpolation. Perhaps there is a different mechanism that leads to generalization of minimum-norm interpolants (2) for any dimensionality of the input space? Our experiments suggest that this is not the case: *minimum-norm interpolant does not appear to perform well in low dimensions*. The present paper provides a theoretical justification for this observation. We show that the estimation error of (2) with the Laplace kernel does not converge to zero as the sample size n increases, unless d scales with n.

We chose to study the Laplace kernel

$$K_c(x, x') = c^d e^{-c\|x-x'\|} \tag{3}$$

for several reasons. First, Belkin et al. (2018b) argue that Laplace kernel regression, in comparison to Gaussian kernel regression, is more similar to ReLU neural networks. More precisely, the nonlinearities introduced by the Laplace kernel allow SGD to have a large "computational reach". For instance, as argued in (Belkin et al., 2018b), the number of epochs required to fit natural vs random labels for Laplace kernel is well-aligned with the corresponding behavior in ReLU networks. Second, for small c, the minimum-norm interpolant in $d = 1$ corresponds to simplicial interpolation of Belkin et al. (2018a), and it may be possible to borrow some of the intuition from the latter paper for higher dimensions. Finally, the RKHS norm corresponding to Laplace kernel can be related to a Sobolev norm, facilitating the development of the lower bound in this paper. We also note that non-differentiability of the kernel function at 0 puts it outside of the assumptions made by (Liang and Rakhlin, 2018); however, a closer look at (El Karoui, 2010) reveals that it is enough to assume differentiability in a neighborhood of 0. Hence, the upper bounds of (Liang and Rakhlin, 2018) can be extended to the case of Laplace kernel, under the high-dimensional scaling $d \asymp n$.

The "width" parameter c in (3) plays an important role. In particular, the upper bounds of (Liang and Rakhlin, 2018) were only shown in the specific regime of this parameter, $c \asymp \sqrt{d}$. The choice of c presents a key difficulty for proving a lower bound: perhaps a clever data-dependent choice can yield a good estimator even in low-dimensional situations? We prove a strong lower bound: *no choice of c can make the interpolation method* (2) *consistent if d is a constant*.

The main theorem can be informally summarized as follows. If Y_i are noisy observations of $f^*(X_i)$ at random points X_i, $i = 1, \dots, n$, the minimum-norm interpolant $\widehat{f}_c$ — for the case of the Laplace kernel with any data-dependent choice of width c — is inconsistent, in the sense that with probability close to 1,

$$\mathbb{E}_{X\sim\mathcal{P}}(\widehat{f}_c(X) - f^*(X))^2 \geq \Omega_d(1).$$

Here $\mathcal{P}$ is the marginal distribution of X and $X_1, \dots, X_n$, f^* is the regression function, and the order notation Ω_d stresses the fact that d is a constant. The standard decomposition

$$\mathbb{E}(\widehat{f}_c(X) - f^*(X))^2 = \mathbb{E}(\widehat{f}_c(X) - Y)^2 - \mathbb{E}(f^*(X) - Y)^2$$

implies the same lower bound for excess loss.

2. Main Results

Let f^* be an unknown smooth function over $\Omega = \overline{B_{\mathbb{R}^d}(0,1)}$ that is not identically zero, and $\mathcal{P}$ an unknown distribution over Ω with probability density function ρ bounded as

$$0 < c_\rho \leq \rho \leq C_\rho. \tag{4}$$

Suppose $X_1, \cdots, X_n$ are sampled i.i.d. according to $\mathcal{P}$, and

$$Y_i = f^*(X_i) + \xi_i \tag{5}$$

with ξ_i assumed to be i.i.d. noise with $\mathbb{P}(\xi_i = +1) = \mathbb{P}(\xi_i = -1) = \frac{1}{2}$. We shall use $\mathcal{S}$ to denote the collection $\{(X_i, Y_i)\}_{i=1}^n$.

Theorem 1 *Let $\widehat{f}_c$ be the minimum-norm solution (2) interpolating (X_i, Y_i), with respect to Laplace kernel $K_c(x, y) = c^d e^{-c\|x-y\|}$. For fixed n and odd dimension d, with probability at least $1 - O\left(\frac{1}{\sqrt{n}}\right)$ over the draw of $\mathcal{S}$,*

$$\forall c > 0, \quad \mathbb{E}_{X\sim\mathcal{P}}(\widehat{f}_c(X) - f^*(X))^2 \geq \Omega_d(1). \tag{6}$$

Remark 2 *We emphasize that the lower bound holds for any data-dependent choice c. The requirement that d be odd is for technical simplicity, and we believe that our results can be extended to even dimensions by using more complicated tools in harmonic analysis. The assumption of binary noise process is for brevity, and the noise magnitude can be changed by simple rescaling.*

For regularized least squares (1), the parameter $\lambda > 0$ leads to a control of the norm of $\widehat{f}$. In the absence of explicit regularization, such a complexity control is more difficult to establish. Intuitively, the norm of the solution can be greatly affected by distances between datapoints, since the interpolating solution fits the noisy function values (separated by a constant), implying a large derivative if datapoints are close. More precisely, given the values $X_1, \ldots, X_n$, we define

$$r_i := \min(\min_{j\neq i} \|X_i - X_j\|, \text{dist}(X_i, \partial\Omega)) \tag{7}$$

for each $i = 1, \ldots, n$. Analyzing the behavior of the random variables r_i underlies the main proofs in this paper. While it is known that $\mathbb{E}[r_i] \lesssim n^{-1/d}$ (Györfi et al., 2006), our proofs require more delicate control of the tails of powers of these variables, including control of the inverse r_i^{-1}. As we show, the estimation error can be related to these random quantities, via Gagliardo-Nirenberg interpolation inequalities and control of higher-order derivatives.

More precisely, we show the following estimates on the random variables r_i:

Proposition 3 *There are constants C_1, C_2 depending on d, such that with probability $1 - O(\frac{1}{\sqrt{n}})$, the following holds for all $-1 \leq k \leq d$:*

$$C_1 n^{-\frac{k}{d}} \leq \frac{1}{n}\sum_{i=1}^n r_i^k \leq C_2 n^{-\frac{k}{d}}. \tag{8}$$

As a consequence, we can show that, with high probability, for at least a constant proportion of the dataset, the minimal distances r_i are, up to an absolute constant, of size $n^{-1/d}$.

Proposition 4 *For any $0 < \alpha < 1$, there is constant C_1', C_2' depending on α, d, such that with probability $1 - O(\frac{1}{\sqrt{n}})$, we have*

$$|\{i : C_1'/\sqrt[d]{n} \leq r_i \leq C_2'/\sqrt[d]{n}\}| \geq \alpha n. \tag{9}$$

In particular, the Lipschitz constant of the interpolating solution is necessarily at least $n^{1/d}$. On the other hand, the tight control of r_i's, together with properties of the RKHS corresponding to Laplace kernel, implies that the RKHS norm squared of the solution is $O(n^{1+\frac{1}{d}})$, as we prove in Proposition 19. This should be contrasted with the lower bound of $\Omega(\exp\{cn^{1/d}\})$ for the norm of any interpolating solution with respect to the Gaussian kernel given in (Belkin et al., 2018b, Theorem 1).

3. Proof

We start with a high-level outline of the proof:

(i) We show that in odd dimension d, the RKHS norm has an explicit form, equal to a Sobolev norm.

(ii) As the RKHS norm becomes the Sobolev norm, we can control "smoothness" of $\widehat{f}_c$ by controlling the RKHS norm. Since $\widehat{f}_c$ and f^* differ on points X_i by the amount ξ_i, and both functions are "smooth", we can choose small regions around X_i such that the squared loss over these regions can be lower bounded. Unfortunately, the lower bound becomes vacuous as c goes to infinity. Hence, we need a different strategy for "large" c.

(iii) When c is large, the RKHS norm approximates the L^2-norm of $\mathbb{R}^d$. We then show that after c passes a certain threshold, the L^2-norm of $\widehat{f}_c$ becomes smaller than a constant fraction of the norm of f^*, implying a lower bound on the total squared loss.

(iv) Remarkably, the two distinct lower bounds in (ii) and (iii) cover all the choices of c, a result that is not immediately evident.

More specifically, we shall show that

Proposition 5 (First Method) *Fix a positive constant $A > 0$. Then with probability at least $1 - O_{d,\rho,A}\left(\frac{1}{\sqrt{n}}\right)$, for any $c \leq A\sqrt[d]{n}$ we have*

$$L(\widehat{f}_c) \triangleq \mathbb{E}\left(\widehat{f}_c(X) - f^*(X)\right)^2 \geq \Omega_{d,\rho,f^*,A}\left(1\right). \tag{10}$$

Proposition 6 (Second Method) *There exists a constant $B = B(d, \rho, f^*) > 0$ independent of n such that with probability at least $1 - O_{d,\rho}\left(\frac{1}{\sqrt{n}}\right)$, for any $c > B\sqrt[d]{n}$ we have*

$$\mathbb{E}\left(\widehat{f}_c(X) - f^*(X)\right)^2 \geq \Omega_{d,\rho,f^*}\left(1\right). \tag{11}$$

Now we take the constant A in the first method to be equal to B and combine the two propositions, concluding that with high probability

$$\forall c \in \mathbb{R},\ \ L(\widehat{f}_c) \geq \Omega(1), \tag{12}$$

concluding the proof of Theorem 1.

Intuition The two modes of failure described in Propositions 5 and 6 are illustrated in Figure 1. For "small" value of c, the solution creates an overly smooth (essentially piece-wise linear) interpolation, while for "large" values, the function behaves more similarly to a collection of thin spikes. In the first case, the non-vanishing (with n) MSE is due to the inability of the interpolated solution to smoothly track the true regression function, while in the second case the solution has an $L_2(\Omega)$ norm that is only a fraction of the corresponding norm of the true regression function. The key message of the paper is that in low dimensions there is no "middle ground" (that is, a choice of c) that would make the interpolation rule consistent as n increases. It is worth emphasizing again that the low-dimensional intuition does not carry over to high dimensions, and the MSE of the interpolated solution can be small, under various conditions on the eigenvalue decay of the sample covariance matrix (Liang and Rakhlin, 2018).

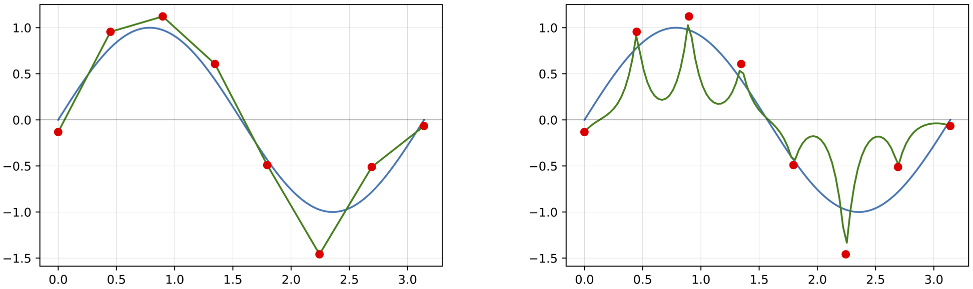

Figure 1: The two modes of failure of minimum-norm interpolation in low dimension. Regression function f^* depicted in blue, noisy observations are depicted in red, and the minimum-norm interpolation with respect to Laplace kernel – in green. Left: $c = 1/10$. Right: $c = 10$.

3.1. Notation

We work with the RKHS $\mathcal{H}_c$ corresponding to the Laplace kernel (3). The subscript emphasizes our focus on the width c. The inner product in $\mathcal{H}_c$ is denoted by $\langle f, g\rangle_{\mathcal{H}_c}$, and $\|f\|^2_{\mathcal{H}_c} = \langle f\rangle_{\mathcal{H}_c}$ denotes the squared norm. We will be using the scaling as described in Proposition 7 in Section 4 so that

$$\langle f\rangle_{\mathcal{H}_c} = \sum_{i=0}^{\frac{d+1}{2}} \binom{\frac{d+1}{2}}{i} c^{-2i} \langle f\rangle_i = \|f\|_{L^2(\mathbb{R}^d)} + \sum_{i=1}^{\frac{d+1}{2}} \binom{\frac{d+1}{2}}{i} c^{-2i} \langle f\rangle_i \tag{13}$$

where

$$\langle f\rangle_i \triangleq \int_{\mathbb{R}^d} |\mathcal{F}f|^2 \|p\|^{2i} dp = C_{d,i} \|D^i f\|^2_{L^2(\mathbb{R}^d)}, \tag{14}$$

with $\mathcal{F}f$ denoting the Fourier transform of f.

3.2. First Method: Control of Hölder Continuity

Proof [of Proposition 5]

Denoting $f \triangleq \widehat{f}_c - f^*$,

$$\mathbb{E}\left(\widehat{f}_c(X) - f^*(X)\right)^2 \geq \Omega_{d,\rho}\left(\|f\|^2_{L^2(\Omega)}\right). \tag{15}$$

Hence, we need only to give a lower bound for $\|f\|^2_{L^2(\Omega)}$. From Proposition 18, for any $I \subset [n]$,

$$\|f\|^2_{L^2(\Omega)} \geq \min\left\{1, \Omega_d\left(\left(\frac{\min_{i\in I} r_i^{-d-1} \sum_{i\in I} r_i^d f(X_i)^2}{\max_{i\in I} r_i^{-d-1} + c^{d+1}\langle f\rangle_{\mathcal{H}_c}}\right)^d \sum_{i\in I} r_i^d f(X_i)^2\right)\right\}. \tag{16}$$

We will now prove Proposition 5 by giving upper bounds for $\max_{i\in I} r_i^{-d-1}$ and $c^{d+1}\langle f\rangle_{\mathcal{H}_c}$ and lower bounds for $\min_{i\in I} r_i^{-d-1}$ and $\sum_{i\in I} r_i^d f(X_i)^2$.

Estimate A. From Proposition 4, with probability $1 - O_{d,\rho}(\frac{1}{\sqrt{n}})$ there is a subset $I \subset [n]$ of size at least $\frac{9}{10}n$ such that

$$\Omega_{d,\rho}\left(n^{-\frac{1}{d}}\right) \leq \min_{i\in I} r_i \leq \max_{i\in I} r_i \leq O_{d,\rho}\left(n^{-\frac{1}{d}}\right). \tag{17}$$

Hence,

$$\Omega_{d,\rho}\left(n^{\frac{d+1}{d}}\right) \leq \min_{i\in I} r_i^{-d-1} \leq \max_{i\in I} r_i^{-d-1} \leq O_{d,\rho}\left(n^{\frac{d+1}{d}}\right). \tag{18}$$

Estimate B. Note that for any i,

$$f(X_i)^2 = (\widehat{f}_c(X_i) - f^*(X_i)^2 = (Y_i - f^*(X_i))^2 = \xi_i^2 = 1, \tag{19}$$

Then applying equation (17) we get

$$\sum_{i\in I} r_i^d f(X_i)^2 \geq \Omega_{d,\rho}\left(\sum_{i\in I}\left(n^{-\frac{1}{d}}\right)^d \cdot 1\right) \geq \Omega_{d,\rho,f^*,A}(1). \tag{20}$$

Estimate C. From Proposition 19 in the Appendix, with probability $1 - O_{d,\rho}\left(\frac{1}{\sqrt{n}}\right)$

$$\begin{aligned} c^{d+1}\langle\widehat{f}_c\rangle_{\mathcal{H}_c} &\leq c^{d+1}\left(\frac{1}{3}\|f^*\|^2_{L^2(\Omega)} + O_{d,\rho,f^*}\left(\frac{\sqrt[d]{n}}{c}\left(1 + \frac{\sqrt[d]{n}}{c}\right)^d\right)\right) \\ &\leq O_{d,\rho,f^*}\left(c^{d+1} + \sqrt[d]{n}\left(c + \sqrt[d]{n}\right)^d\right) \\ &\leq O_{d,\rho,f^*}\left(A^{d+1}n^{\frac{d+1}{d}} + \sqrt[d]{n}\left(A\sqrt[d]{n} + \sqrt[d]{n}\right)^d\right) \\ &= O_{d,\rho,f^*,A}\left(n^{\frac{d+1}{d}}\right). \end{aligned} \tag{21}$$

It then follows that

$$c^{d+1}\langle f\rangle_{\mathcal{H}_c} \leq 2c^{d+1}\langle\widehat{f}_c\rangle_{\mathcal{H}_c} + 2c^{d+1}\langle f^*\rangle_{\mathcal{H}_c} \leq O_{d,\rho,f^*,A}(n^{\frac{d+1}{d}}). \tag{22}$$

We are now ready to put all these estimates together. With probability $1 - O_{d,\rho}\left(\frac{1}{\sqrt{n}}\right)$

$$\left(\frac{\min\limits_{i\in I} r_i^{-d-1} \sum_{i\in I} r_i^d f(X_i)^2}{\max\limits_{i\in I} r_i^{-d-1} + c^{d+1}\langle f\rangle_{\mathcal{H}_c}}\right)^d \sum_{i\in I} r_i^d f(X_i)^2 \geq \Omega_{d,\rho,f^*,A}(1). \tag{23}$$

As a result, with probability at least $1 - O_{d,\rho,f^*,A}\left(\frac{1}{\sqrt{n}}\right)$,

$$L(\widehat{f}_c) = \mathbb{E}\left(\widehat{f}_c(X) - f^*(X)\right)^2 \geq \Omega_{d,\rho,f^*,A}\left(1\right). \tag{24}$$

■

3.3. Second Method: Control of L^2 norm

The lower bound in this regime boils down to proving an upper bound on the $L_2(\Omega)$ norm of the interpolated solution as compared to the $L_2(\Omega)$ norm of f^*. Proposition 19 proves this fact by constructing another interpolating solution whose RKHS norm can be explicitly controlled. Since $\widehat{f}_c$ is the minimal norm solution, the result follows by triangle inequality.

Proof [**of Proposition 6**]

We need only to show the existence of B such that

$$\forall c > B\sqrt[d]{n},\ \ \|\widehat{f}_c - f^*\|^2_{L^2(\Omega)} \geq \Omega_{d,\rho,f^*}\left(1\right). \tag{25}$$

From equation (131) in Proposition 19 in the Appendix,

$$\begin{aligned}\langle\widehat{f}_c\rangle_{\mathcal{H}_c} &\leq \frac{1}{3}\|f^*\|^2_{L^2(\Omega)} + O_{d,\rho,f^*}\left(\frac{\sqrt[d]{n}}{c}\left(1 + \frac{\sqrt[d]{n}}{c}\right)^d\right)\\ &\leq \frac{1}{3}\|f^*\|^2_{L^2(\Omega)} + O_{d,\rho,f^*}\left(\frac{1}{B}\left(1 + \frac{1}{B}\right)^d\right).\end{aligned} \tag{26}$$

Then for $B = B(d,\rho,f^*)$ large enough,

$$\langle\widehat{f}_c\rangle_{\mathcal{H}_c} \leq \frac{1}{3}\|f^*\|^2_{L^2(\Omega)} + \frac{1}{3}\|f^*\|^2_{L^2(\Omega)} \leq \frac{2}{3}\|f^*\|^2_{L^2(\Omega)}. \tag{27}$$

Now, by triangle inequality,

$$\|\widehat{f}_c - f^*\|_{L^2(\Omega)} \geq \|f^*\|_{L^2(\Omega)} - \|\widehat{f}_c\|_{L^2(\Omega)} \geq \left(1 - \sqrt{\frac{2}{3}}\right)\|f^*\|_{L^2(\Omega)} = \Omega_{d,\rho,f^*}\left(1\right) \tag{28}$$

as desired. This completes the proof.

■

We presented brief proofs of the lower bounds, postponing much of the technical details to the Appendix. Next section is devoted to analyzing the RKHS corresponding to the Laplace kernel, and, in particular, to proving a succinct expression for the RKHS norm.

4. Explicit form of the RKHS norm

In this section, we provide an expression, up to constant factors, for the RKHS norm corresponding to the Laplace kernel, along with the associated eigenfunctions and eigenvalues. We believe these estimates will be useful for future study of interpolation and other methods with Laplace kernels. Notably, the expansions provided in the next proposition are finite (with only $(d+2)/2$ terms), given the choice of the basis, as opposed to infinite-dimensional expansions for the Gaussian kernel.

Proposition 7 *Consider the kernel $K_c(x,y) = c^d e^{-c\|x-y\|}$ in $\mathbb{R}^d$ with d odd. The corresponding RKHS norm is given by*

$$\langle f\rangle_{\mathcal{H}_c} \sim \int_{\mathbb{R}^d} |\mathcal{F}f|^2(1+\|p\|^2/c^2)^{\frac{d+1}{2}}dp \sim \sum_{i=0}^{\frac{d+1}{2}} \binom{\frac{d+1}{2}}{i} c^{-2i}\langle f\rangle_i. \tag{29}$$

where

$$\langle f\rangle_i = \int_{\mathbb{R}^d} |\mathcal{F}f|^2\|p\|^{2i}dp = C_{d,i}\|D^i f\|^2_{L^2(\mathbb{R}^d)}. \tag{30}$$

and the Fourier transformation $\mathcal{F}$ is chosen such that

$$\langle f\rangle_0 = \|f\|^2_{L^2(\Omega)}. \tag{31}$$

As scaling does not change the output of the algorithm, we take the convention that

$$\langle f\rangle_{\mathcal{H}_c} = \sum_{i=0}^{\frac{d+1}{2}} \binom{\frac{d+1}{2}}{i} c^{-2i}\langle f\rangle_i = \|f\|_{L^2(\mathbb{R}^d)} + \sum_{i=1}^{\frac{d+1}{2}} \binom{\frac{d+1}{2}}{i} c^{-2i}\langle f\rangle_i \tag{32}$$

Proof Consider the integral operator

$$T_K f(x) = \int_y K(x,y)f(y)dy. \tag{33}$$

We have

$$\langle f,g\rangle_{\mathcal{H}_c} = \langle f, T_K^{-1}g\rangle_{L^2(\mathbb{R}^d)}. \tag{34}$$

An eigenspace-decomposition of T_K immediately gives the form of the inner product in the RKHS. Since $K_c(x,y) = k(x-y)$ with $k(x) = c^d e^{-c\|x\|}$, it is easy to verify that the family $\{h_p(x) = e^{ip\cdot x}\}_{p\in\mathbb{R}^d}$ are eigenfunctions of T_K:

$$T_K h_p(x) = \int_y k(x-y)e^{ip\cdot y}dy = \lambda(p)h_p(x) \tag{35}$$

where

$$\lambda(p) = \int_y k(x-y)e^{ip\cdot(y-x)}dy = \int_x k(x)e^{-ip\cdot x}dx. \tag{36}$$

Therefore, the inner product of RKHS can be written as

$$\langle f,g\rangle_{\mathcal{H}_c} = \int_{x,p,y} \frac{1}{\lambda(p)^{-1}} f(x)^* h_p(x)h_p(y)^* g(y)dxdpdy \tag{37}$$

which can be further rewritten as:

$$\langle f, g\rangle_{\mathcal{H}_c} = \int_p \frac{1}{\lambda(p)^{-1}} \mathcal{F}f(p)^* \mathcal{F}g(p) dp. \tag{38}$$

Now for $\lambda(p)$, we have

$$\lambda = \mathcal{F}k. \tag{39}$$

In fact, $\lambda(p)$ can be explicitly computed (see e.g. (Stein and Weiss, 1971, Thm 1.4)):

$$\begin{aligned}\lambda(p) &= c^d \int_{\mathbb{R}^d} e^{-c\|x\|} e^{-ipx} dx \\ &= \int_{\mathbb{R}^d} e^{-\|x\|} e^{-ipx/c} dx\end{aligned}$$

The last expression is equal to

$$\begin{aligned}&\int_{\mathbb{R}^d} \left(\frac{1}{\sqrt{\pi}} \int_0^\infty \frac{e^{-\eta}}{\sqrt{\eta}} e^{-\|x\|^2/4\eta} d\eta\right) e^{-ipx/c} dx \\ &= \frac{1}{\sqrt{\pi}} \int_0^\infty \frac{e^{-\eta}}{\sqrt{\eta}} \left(\int_{\mathbb{R}^d} e^{-\|x\|^2/4\eta} e^{-ipx/c} dx\right) d\eta \\ &= \frac{1}{\sqrt{\pi}} \int_0^\infty \frac{e^{-\eta}}{\sqrt{\eta}} (4\pi\eta)^{d/2} e^{-\eta\|p\|^2/c^2} d\eta \\ &= \frac{2^d \pi^{(d-1)/2} \Gamma(\frac{d+1}{2})}{(1 + \|p\|^2/c^2)^{(d+1)/2}}.\end{aligned}$$

Then

$$\lambda(p)^{-1} = \frac{(1 + \|p\|^2/c^2)^{(d+1)/2}}{2^d \pi^{(d-1)/2} \Gamma(\frac{d+1}{2})} = \frac{\sum_{i=0}^{(d+1)/2} \binom{\frac{d+1}{2}}{i} \|p\|^{2i}/c^{2i}}{2^d \pi^{(d-1)/2} \Gamma(\frac{d+1}{2})} \tag{40}$$

and

$$\begin{aligned}\int_p \frac{1}{\lambda(p)^{-1}} \mathcal{F}f(p)^* \mathcal{F}g(p) dp &= \int_p \sum_{i=0}^{(d+1)/2} \frac{\binom{\frac{d+1}{2}}{i} \|p\|^{2i}/c^{2i}}{2^d \pi^{(d-1)/2} \Gamma(\frac{d+1}{2})} \mathcal{F}f(p)^* \mathcal{F}g(p) dp \qquad (41) \\ &= \sum_{i=0}^{(d+1)/2} \frac{\binom{\frac{d+1}{2}}{i}/c^{2i}}{2^d \pi^{(d-1)/2} \Gamma(\frac{d+1}{2})} \int_p \|p\|^{2i} \mathcal{F}f(p)^* \mathcal{F}g(p) dp, \qquad (42)\end{aligned}$$

implying the result.

■

5. Discussion

We have presented theoretical evidence that minimum-norm interpolation with Laplacian kernel is not consistent if d does not scale with n. On the other hand, in the high-dimensional scaling regime $n \asymp d$, (Liang and Rakhlin, 2018) exhibited a phenomenon of implicit regularization that allows, under a number of additional assumptions, the estimation error to be small. The interaction of dimensionality, sample size, and eigenvalue decays for the population and sample covariance matrices is complex, and identifying all the regimes when interpolation succeeds is still a largely unexplored area. In particular, our lower bound becomes vacuous as soon as d starts to scale with n. It would be interesting to understand the minimal scaling of d along with assumptions on the underlying distribution that allow minimum-norm interpolation to succeed.

Partial motivation for the study of interpolation methods comes from the recent successes of neural networks. These overparametrized models are typically trained to achieve zero error on the training data (Zhang et al., 2016; Belkin et al., 2018b), yet perform well out-of-sample. Recent work connecting sufficiently wide neural networks and the effective kernel (Mei et al., 2018; Chizat and Bach, 2018; Daniely, 2017; Jacot et al., 2018; Du et al., 2018) suggests that interpolating neural networks can be studied through the lens of kernel methods. In particular, it can be shown that the limiting solutions in such cases are, in fact, *minimum-norm interpolants* with respect to the corresponding kernel. Hence, further study of strengths and limitations of minimum-norm interpolation can shed light on performance of neural networks.

Acknowledgments

This work was partly supported by the DARPA Lagrange program and the MIT-Sensetime Alliance on Artificial Intelligence.

References

Mikhail Belkin, Daniel Hsu, and Partha Mitra. Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. *arXiv preprint arXiv:1806.05161*, 2018a.

Mikhail Belkin, Siyuan Ma, and Soumik Mandal. To understand deep learning we need to understand kernel learning. *arXiv preprint arXiv:1802.01396*, 2018b.

Mikhail Belkin, Alexander Rakhlin, and Alexandre B Tsybakov. Does data interpolation contradict statistical optimality? *arXiv preprint arXiv:1806.09471*, 2018c.

Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *arXiv preprint arXiv:1805.09545*, 2018.

Amit Daniely. Sgd learns the conjugate kernel class of the network. In *Advances in Neural Information Processing Systems*, pages 2422–2430, 2017.

Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018.

Noureddine El Karoui. The spectrum of kernel random matrices. *The Annals of Statistics*, 38(1): 1–50, 2010.

László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *arXiv preprint arXiv:1806.07572*, 2018.

Giovanni Leoni. *A First Course in Sobolev Spaces, Second Edition*. American Mathematical Society, 2017.

Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel" ridgeless" regression can generalize. *arXiv preprint arXiv:1808.00387*, 2018.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layers neural networks. *arXiv preprint arXiv:1804.06561*, 2018.

Elias M Stein and Guido Weiss. *Introduction to Fourier analysis on Euclidean spaces (PMS-32)*. Princeton University Press, 1971.

Abraham J Wyner, Matthew Olson, Justin Bleich, and David Mease. Explaining the success of adaboost and random forests as interpolating classifiers. *The Journal of Machine Learning Research*, 18(1):1558–1590, 2017.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

Appendix A. Bounds of Average Separation

A.1. Main Claims

Proof [**of Proposition 4**] With probability at least $1 - O(\frac{1}{\sqrt{n}})$ for all $-1 \leq k \leq d$, for constants C_1, C_2,

$$C_1 n^{-\frac{k}{d}} \leq \frac{1}{n}\sum_{i=1}^{n} r_i^k \leq C_2 n^{-\frac{k}{d}}. \tag{43}$$

Let $\beta = 1 - \frac{1}{2}(1-\alpha) = \frac{1+\alpha}{2}$. Let I_1 be a subset of $[n]$ of size ceil(βn) such that $\forall i \in I_1, j \in [n] \setminus I_1, r_i \geq r_j$. Let $r = \min_{i \in I_1} r_i$. Then

$$C_2 \sqrt[d]{n} \geq \frac{1}{n}\sum_i r_i^{-1} \geq \frac{1}{n}\sum_{i \in I_1} r_i^{-1} \geq \frac{1}{n}\frac{\beta n}{r} = \frac{\beta}{r}. \tag{44}$$

It then follows that $r \geq C_2/(\beta\sqrt[d]{n})$. Take $C_1' = C_2/\beta$. Hence, for any $i \in I_1, r_i \geq C_1'/\sqrt[d]{n}$. Similarly, there is a subset I_2 of $[n]$ of size ceil(βn) such that $\forall i \in I_2, r_i \geq C_2'/\sqrt[d]{n}$. Note that $|I_1 \cap I_2| \geq \alpha n$, concluding the proof.

■

In the rest of this section, we prove Proposition 3. Since we have the inequality

$$\left(\frac{1}{n}\sum_{i=1}^{n} r_i^{-1}\right)^{-k} \leq \frac{1}{n}\sum_{i=1}^{n} r_i^k \leq \left(\frac{1}{n}\sum_{i=1}^{n} r_i^d\right)^{\frac{k}{d}} \tag{45}$$

for all $-1 \leq k \leq d$, we need only to prove that with high probability

$$\frac{1}{n}\sum_{i=1}^{n} r_i^d \lesssim n^{-1} \tag{46}$$

and

$$\frac{1}{n}\sum_{i=1}^{n} r_i^{-1} \lesssim n^{\frac{1}{d}}. \tag{47}$$

A.2. Average of r_i^d

The following is always true:

$$\sum_{i=1}^{n} r_i^d \lesssim \sum_{i=1}^{n} m(B(X_i, \frac{1}{2}r_i)) \leq m(\Omega) \lesssim 1 \tag{48}$$

Then the result follows.

A.3. Average of r_i^{-1}

A.3.1. STRATEGY

We shall use Chebyshev's inequality to bound average of r_i^{-1}, and thus we need to estimate $\text{Cov}(r_i^{-1}, r_j^{-1})$. This step is not direct because r_i, r_j are not independent: both depend on X_i and X_j.

We define $\tilde{r}_i, \tilde{r}_j$ for any fixed pair of (i, j) such that

- $\tilde{r}_i = r_i, \tilde{r}_j = r_j$ with high probability
- $\tilde{r}_i$ is independent w.r.t X_j, $\tilde{r}_j$ is independent w.r.t X_i

We will then show that $Cov(\tilde{r}_i, \tilde{r}_j)$ is small and that the difference between $Cov(r_i, r_j)$ and $Cov(\tilde{r}_i, \tilde{r}_j)$ is small. Applying Chebyshev's inequality then yields the result.

A.3.2. UPPER BOUND FOR $\mathbb{E}[r_i^{-1}]$ AND $\mathbb{E}[r_i^{-2}]$

$$\mathbb{P}(r_i < r) = 1 - m_{\mathcal{P}}(B(X_i, r)^c)^n \leq n m_{\mathcal{P}}(B(X_i, r)) \lesssim n r^d. \tag{49}$$

Then

$$\begin{aligned}
\mathbb{E}r_i^{-1} &= \mathbb{E}\int_0^\infty \mathbb{I}(r_i^{-1} > s)ds \\
&= \int_0^\infty \mathbb{E}\mathbb{I}(r_i^{-1} > s)ds \\
&= \int_0^\infty \mathbb{P}(r_i^{-1} > s)ds \\
&\leq \int_0^\infty \min(1, C_d n s^{-d})ds \\
&= s_0 + C_d n s_0^{1-d}/(d-1) \text{ where } C_d n s_0^{-d} = 1 \\
&= \frac{d}{d-1}s_0 \\
&= \frac{d}{d-1}\sqrt[d]{C_d n}
\end{aligned} \tag{50}$$

and

$$
\begin{aligned}
\frac{1}{2}\mathbb{E}r_i^{-2} &= \mathbb{E}\int_0^\infty s\mathbb{I}(r_i^{-1} > s)ds \\
&= \int_0^\infty s\mathbb{E}\mathbb{I}(r_i^{-1} > s)ds \\
&= \int_0^\infty s\mathbb{P}(r_i^{-1} > s)ds \\
&\le \int_0^\infty s\min(1, C_d n s^{-d})ds \qquad (51) \\
&= \frac{1}{2}s_0^2 + C_d n s_0^{2-d}/(d-2) \text{ where } C_d n s_0^{-d} = 1 \\
&= \left(\frac{1}{2} + \frac{1}{d-2}\right)s_0^2 \\
&= \left(\frac{1}{2} + \frac{1}{d-2}\right)(\sqrt[d]{C_d n})^2.
\end{aligned}
$$

Hence,

$$\mathbb{E}r_i^{-2} \le \frac{d}{d-2}(C_d n)^{\frac{2}{d}} \qquad (52)$$

A.3.3. ESTIMATE OF COV($\frac{1}{\tilde{r}_i}, \frac{1}{\tilde{r}_j}$)

Define

$$\tilde{r}_i := \min(\min_{k\neq i,j}|X_k - X_i|, \text{dist}(\partial\Omega, X_i)) \qquad (53)$$

and

$$\tilde{r}_j := \min(\min_{k\neq i,j}|X_k - X_j|, \text{dist}(\partial\Omega, X_j)) \qquad (54)$$

Then

$$r_i = \min(\tilde{r}_i, |X_i - X_j|), r_j = \min(\tilde{r}_j, |X_i - X_j|) \qquad (55)$$

and $\tilde{r}_j$ is independent of X_i and $\tilde{r}_i$ is independent of X_j.

$$
\begin{aligned}
&\mathbb{E}[\frac{1}{\tilde{r}_i\tilde{r}_j}] - \mathbb{E}[\frac{1}{\tilde{r}_i}]\mathbb{E}[\frac{1}{\tilde{r}_j}] \\
&= \mathbb{E}_{X_i,X_j}[\mathbb{E}[\frac{1}{\tilde{r}_i\tilde{r}_j}|X_i,X_j]] - \mathbb{E}_{X_i}[\mathbb{E}[\frac{1}{\tilde{r}_i}|X_i]]\mathbb{E}_{X_j}[\mathbb{E}[\frac{1}{\tilde{r}_j}|X_j]] \\
&= \mathbb{E}_{X_i,X_j}[\mathbb{E}[\frac{1}{\tilde{r}_i\tilde{r}_j}|X_i,X_j]] - \mathbb{E}_{X_i}\left[\mathbb{E}\left[\frac{1}{\tilde{r}_i}|X_i\right]\mathbb{E}_{X_j}\left[\mathbb{E}[\frac{1}{\tilde{r}_j}|X_j]\right]\right] \\
&= \mathbb{E}_{X_i,X_j}\left[\mathbb{E}[\frac{1}{\tilde{r}_i\tilde{r}_j}|X_i,X_j]\right] - \mathbb{E}_{X_i,X_j}\left[\mathbb{E}\left[\frac{1}{\tilde{r}_i}|X_i\right]\mathbb{E}\left[\frac{1}{\tilde{r}_j}|X_j\right]\right] \text{ (indep. between } X_i \text{ and } X_j\text{)} \\
&= \mathbb{E}_{X_i,X_j}\left[\mathbb{E}\left[\frac{1}{\tilde{r}_i\tilde{r}_j}\Big|X_i,X_j\right] - \mathbb{E}\left[\frac{1}{\tilde{r}_i}\Big|X_i,X_j\right]\mathbb{E}\left[\frac{1}{\tilde{r}_j}\Big|X_i,X_j\right]\right]
\end{aligned}
$$

where we used independence between $\tilde{r}_i$ and X_j and between $\tilde{r}_j$ and X_i. The last expression can be written as

$$\begin{aligned}
&= \mathbb{E}_{X_i,X_j}\Bigg[\mathbb{E}\left[\int_0^\infty ds\mathbb{I}(\tilde{r}_i^{-1} > s)\int_0^\infty dt\mathbb{I}(\tilde{r}_j^{-1} > t)\Big|X_i, X_j\right] \\
&- \left(\mathbb{E}\int_0^\infty ds\mathbb{I}(\tilde{r}_i^{-1} > s)\Big|X_i, X_j\right)\left(\mathbb{E}\int_0^\infty dt\mathbb{I}(\tilde{r}_j^{-1} > t)\Big|X_i, X_j\right)\Bigg] \\
&= \mathbb{E}_{X_i,X_j}\Bigg[\int_0^\infty ds\int_0^\infty dt\mathbb{E}\left[\mathbb{I}(\tilde{r}_i^{-1} > s, \tilde{r}_j^{-1} > t)\Big|X_i, X_j\right] \\
&- \int_0^\infty \mathbb{E}\left[\mathbb{I}(\tilde{r}_i^{-1} > s)\Big|X_i, X_j\right]ds\int_0^\infty \mathbb{E}\left[\mathbb{I}(\tilde{r}_j^{-1} > t)\Big|X_i, X_j\right]dt\Bigg] \\
&= \mathbb{E}_{X_i,X_j}\left[\int_0^\infty\int_0^\infty\left(\mathbb{P}\left[\tilde{r}_i^{-1} > s, \tilde{r}_j^{-1} > t\Big|X_i, X_j\right] - \mathbb{P}\left[\tilde{r}_i^{-1} > s\Big|X_i, X_j\right]\mathbb{P}\left[\tilde{r}_j^{-1} > t\Big|X_i, X_j\right]\right)dsdt\right].
\end{aligned}$$

Now,

$$\begin{aligned}
&\mathbb{P}[r_i^{-1} > s, r_j^{-1} > t|X_i, X_j]] \\
&\quad = 1 - \mathbb{P}[r_i^{-1} < s|X_i, X_j]] - \mathbb{P}[r_j^{-1} < t|X_i, X_j]] + \mathbb{P}[r_i^{-1} < s, r_j^{-1} < t|X_i, X_j] \\
&\mathbb{P}[r_i^{-1} > s|X_i, X_j]\mathbb{P}[r_j^{-1} > t|X_i, X_j] \\
&\quad = 1 - \mathbb{P}[r_i^{-1} < s|X_i, X_j] - \mathbb{P}[r_j^{-1} < t|X_i, X_j] + \mathbb{P}[r_i^{-1} < s|X_i, X_j]\mathbb{P}[r_j^{-1} < t|X_i, X_j]
\end{aligned} \tag{56}$$

Then

$$\begin{aligned}
&\mathbb{E}[\frac{1}{\tilde{r}_i\tilde{r}_j}] - \mathbb{E}[\frac{1}{\tilde{r}_i}]\mathbb{E}[\frac{1}{\tilde{r}_j}] \\
&= \mathbb{E}_{X_i,X_j}\left[\int_0^\infty\int_0^\infty\left(\mathbb{P}\left[\tilde{r}_i^{-1} > s, \tilde{r}_j^{-1} > t\Big|X_i, X_j\right] - \mathbb{P}\left[\tilde{r}_i^{-1} > s\Big|X_i, X_j\right]\mathbb{P}\left[\tilde{r}_j^{-1} > t\Big|X_i, X_j\right]\right)dsdt\right] \\
&= \mathbb{E}_{X_i,X_j}\left[\int_0^\infty\int_0^\infty\left(\mathbb{P}\left[\tilde{r}_i^{-1} < s\Big|X_i, X_j\right]\mathbb{P}\left[\tilde{r}_j^{-1} < t\Big|X_i, X_j\right] - \mathbb{P}\left[\tilde{r}_i^{-1} < s, \tilde{r}_j^{-1} < t\Big|X_i, X_j\right]\right)dsdt\right] \\
&= \mathbb{E}_{X_i,X_j}\Bigg[\int_0^\infty\int_0^\infty\Big(m_{\mathcal{P}}(B(X_i, s^{-1})^c)^{n-2}m_{\mathcal{P}}(B(X_j, t^{-1})^c)^{n-2} \\
&\qquad\qquad - m_{\mathcal{P}}((B(X_i, s^{-1}) \cup B(X_j, t^{-1}))^c)^{n-2}\Big)dsdt\Bigg] \\
&= \mathbb{E}_{X_i,X_j}\Bigg[\int_{R_0^{-1}}^\infty\int_{R_0^{-1}}^\infty\Big(m_{\mathcal{P}}(B(X_i, s^{-1})^c)^{n-2}m_{\mathcal{P}}(B(X_j, t^{-1})^c)^{n-2} \\
&\qquad\qquad - m_{\mathcal{P}}((B(X_i, s^{-1}) \cup B(X_j, t^{-1}))^c)^{n-2}\Big)dsdt\Bigg]
\end{aligned} \tag{57}$$

where $R_0 = \text{diam}(\Omega)$ is a constant depending only on d.

When $s^{-1} + t^{-1} < |X_i - X_j|$, we have

$$B(X_i, s^{-1}) \cup B(X_j, t^{-1}) = B(X_i, s^{-1}) \sqcup B(X_j, t^{-1}) \tag{58}$$

where $\sqcup$ means disjoint union. Then

$$\begin{aligned}
&m_{\mathcal{P}}(B(X_i,s^{-1})^c)m_{\mathcal{P}}(B(X_j,t^{-1})^c)-m_{\mathcal{P}}((B(X_i,s^{-1})\cup B(X_j,t^{-1}))^c)\\
&=m_{\mathcal{P}}(B(X_i,s^{-1})^c)m_{\mathcal{P}}(B(X_j,t^{-1})^c)-m_{\mathcal{P}}((B(X_i,s^{-1})\sqcup B(X_j,t^{-1}))^c)\\
&=(1-m_{\mathcal{P}}(B(X_i,s^{-1})))(1-m_{\mathcal{P}}(B(X_j,t^{-1})))-(1-m_{\mathcal{P}}(B(X_i,s^{-1}))-m_{\mathcal{P}}(B(X_j,t^{-1})))\\
&=m_{\mathcal{P}}(B(X_i,s^{-1}))m_{\mathcal{P}}(B(X_j,t^{-1}))\\
&\geq 0
\end{aligned} \tag{59}$$

Since for $0\leq x\leq y\leq 1$, $x^{n-2}-y^{n-2}\leq(n-2)x^{n-3}(x-y)$, we have

$$\begin{aligned}
&0\leq m_{\mathcal{P}}(B(X_i,s^{-1})^c)^{n-2}m_{\mathcal{P}}(B(X_j,t^{-1})^c)^{n-2}-m_{\mathcal{P}}((B(X_i,s^{-1})\cup B(X_j,t^{-1}))^c)^{n-2}\\
&\leq(n-3)m_{\mathcal{P}}(B(X_i,s^{-1})^c)^{n-3}m_{\mathcal{P}}(B(X_j,t^{-1})^c)^{n-3}m_{\mathcal{P}}(B(X_i,s^{-1}))m_{\mathcal{P}}(B(X_j,t^{-1}))\\
&\leq(n-3)\left(\max(0,1-\frac{C_d}{s^d})\right)^{n-3}\left(\max(0,1-\frac{C_d}{t^d})\right)^{n-3}\frac{C'_d}{s^d}\frac{C'_d}{t^d}
\end{aligned} \tag{60}$$

where C_d, C'_d are constants such that for any $B(x,r)\subset\Omega$

$$C_d r^d\leq m_{\mathcal{P}}(B(x,r))\leq C'_d r^d. \tag{61}$$

When $s^{-1}>\frac{\|X_i-X_j\|}{2}$, we have

$$\begin{aligned}
&m_{\mathcal{P}}(B(X_i,s^{-1})^c)^{n-2}m_{\mathcal{P}}(B(X_j,t^{-1})^c)^{n-2}-m_{\mathcal{P}}((B(X_i,s^{-1})\cup B(X_j,t^{-1}))^c)^{n-2}\\
&\geq m_{\mathcal{P}}(B(X_i,s^{-1})^c)^{n-2}m_{\mathcal{P}}(B(X_j,t^{-1})^c)^{n-2}-m_{\mathcal{P}}((B(X_i,s^{-1}))^c)^{n-2}\\
&\geq m_{\mathcal{P}}(B(X_i,s^{-1})^c)^{n-2}\cdot\min\{1,(n-2)m_{\mathcal{P}}(B(X_j,t^{-1}))\}\\
&\geq-(\max(0,1-C_d s^{-d}))^{n-2}\min(1,(n-2)C'_d t^{-d})
\end{aligned} \tag{62}$$

and

$$\begin{aligned}
&m_{\mathcal{P}}(B(X_i,s^{-1})^c)^{n-2}m_{\mathcal{P}}(B(X_j,t^{-1})^c)^{n-2}-m_{\mathcal{P}}((B(X_i,s^{-1})\cup B(X_j,t^{-1}))^c)^{n-2}\\
&\leq m_{\mathcal{P}}(B(X_i,s^{-1})^c)^{n-2}m_{\mathcal{P}}(B(X_j,t^{-1})^c)^{n-2}-(1-m_{\mathcal{P}}(B(X_i,s^{-1}))-m_{\mathcal{P}}(B(X_j,t^{-1})))^{n-2}\\
&\leq(n-2)m_{\mathcal{P}}(B(X_i,s^{-1})^c)^{n-3}m_{\mathcal{P}}(B(X_j,t^{-1})^c)^{n-3}m_{\mathcal{P}}(B(X_i,s^{-1}))m_{\mathcal{P}}(B(X_j,t^{-1}))\\
&\leq(n-3)\left(\max(0,1-\frac{C_d}{s^d})\right)^{n-3}\left(\max(0,1-\frac{C_d}{t^d})\right)^{n-3}\frac{C'_d}{s^d}\frac{C'_d}{t^d}
\end{aligned} \tag{63}$$

Then

$$\begin{aligned}
&-(\max(0,1-C_d s^{-d}))^{n-2}\min(1,(n-2)C'_d t^{-d})\\
&\leq m_{\mathcal{P}}(B(X_i,s^{-1})^c)^{n-2}m_{\mathcal{P}}(B(X_j,t^{-1})^c)^{n-2}-m_{\mathcal{P}}((B(X_i,s^{-1})\cup B(X_j,t^{-1}))^c)^{n-2}\\
&\leq(n-3)\left(\max(0,1-\frac{C_d}{s^d})\right)^{n-3}\left(\max(0,1-\frac{C_d}{t^d})\right)^{n-3}\frac{C'_d}{s^d}\frac{C'_d}{t^d}
\end{aligned} \tag{64}$$

Similarly for $t^{-1} > \frac{\|X_i - X_j\|}{2}$, we have

$$\begin{aligned}&-(\max(0, 1 - C_d t^{-d}))^{n-2} \min(1, (n-2)C'_d s^{-d})\\ &\leq m_{\mathcal{P}}(B(X_i, s^{-1})^c)^{n-2} m_{\mathcal{P}}(B(X_j, t^{-1})^c)^{n-2} - m_{\mathcal{P}}((B(X_i, s^{-1}) \cup B(X_j, t^{-1}))^c)^{n-2}\\ &\leq (n-3)\left(\max(0, 1 - \frac{C_d}{s^d})\right)^{n-3} \left(\max(0, 1 - \frac{C_d}{t^d})\right)^{n-3} \frac{C'_d}{s^d}\frac{C'_d}{t^d}\end{aligned} \tag{65}$$

The upper bound are the same in all three cases, but the lower bounds are different.

Upper bound for Cov$(\tilde{r}_i^{-1}, \tilde{r}_j^{-1})$ We now put the above calculations together and estimate

$$\begin{aligned}&\mathrm{Cov}(\tilde{r}_i^{-1}, \tilde{r}_j^{-1})\\ &= \mathbb{E}[\frac{1}{\tilde{r}_i\tilde{r}_j}] - \mathbb{E}[\frac{1}{\tilde{r}_i}]\mathbb{E}[\frac{1}{\tilde{r}_j}]\\ &= \mathbb{E}_{X_i,X_j}\left[\int_0^\infty\int_0^\infty \Big(m_{\mathcal{P}}(B(X_i, s^{-1})^c)^{n-2} m_{\mathcal{P}}(B(X_j, t^{-1})^c)^{n-2}\right.\\ &\qquad\qquad\left. - m_{\mathcal{P}}((B(X_i, s^{-1}) \cup B(X_j, t^{-1}))^c)^{n-2}\Big) dsdt\right]\\ &= \mathbb{E}_{X_i,X_j}\left[\int_{R_0^{-1}}^\infty\int_{R_0^{-1}}^\infty \Big(m_{\mathcal{P}}(B(X_i, s^{-1})^c)^{n-2} m_{\mathcal{P}}(B(X_j, t^{-1})^c)^{n-2}\right.\\ &\qquad\qquad\left. - m_{\mathcal{P}}((B(X_i, s^{-1}) \cup B(X_j, t^{-1}))^c)^{n-2}\Big) dsdt\right]\\ &\leq \mathbb{E}_{X_i,X_j}\int_{R_0^{-1}}^\infty\int_{R_0^{-1}}^\infty (n-3)\left(\max(0, 1 - \frac{C_d}{s^d})\right)^{n-3}\left(\max(0, 1 - \frac{C_d}{t^d})\right)^{n-3}\frac{C'_d}{s^d}\frac{C'_d}{t^d}dsdt\\ &\leq \mathbb{E}_{X_i,X_j}\frac{R_0^2}{4}\int_{R_0^{-1}}^\infty\int_{R_0^{-1}}^\infty (n-3)\left(\max(0, 1 - \frac{C_d}{s^d})\right)^{n-3}\left(\max(0, 1 - \frac{C_d}{t^d})\right)^{n-3}\frac{C'_d}{s^{d-1}}\frac{C'_d}{t^{d-1}}dsdt\\ &\leq \mathbb{E}_{X_i,X_j}\frac{R_0^2}{4}\int_0^\infty\int_0^\infty (n-3)\left(\max(0, 1 - \frac{C_d}{s^d})\right)^{n-3}\left(\max(0, 1 - \frac{C_d}{t^d})\right)^{n-3}\frac{C'_d}{s^{d-1}}\frac{C'_d}{t^{d-1}}dsdt\\ &= \frac{R_0^2}{4}\frac{n-3}{d^2(n-2)^2}(C'_d/C_d)^2\\ &= O\left(\frac{1}{n}\right)\end{aligned} \tag{66}$$

Lower bound for Cov$(\tilde{r}_i^{-1}, \tilde{r}_j^{-1})$

$$
\begin{aligned}
&\mathbb{E}[\frac{1}{\tilde{r}_i\tilde{r}_j}] - \mathbb{E}[\frac{1}{\tilde{r}_i}]\mathbb{E}[\frac{1}{\tilde{r}_j}] \\
&= \mathbb{E}_{X_i,X_j}\Big[\int_0^\infty\int_0^\infty \Big(m_{\mathcal{P}}(B(X_i,s^{-1})^c)^{n-2}m_{\mathcal{P}}(B(X_j,t^{-1})^c)^{n-2} \\
&\qquad\qquad\qquad -m_{\mathcal{P}}((B(X_i,s^{-1})\cup B(X_j,t^{-1}))^c)^{n-2}\Big)dsdt\Big] \\
&= \mathbb{E}_{X_i,X_j}\left[\underbrace{\int_{\frac{2}{\|X_i-X_j\|}}^\infty\int_{\frac{2}{\|X_i-X_j\|}}^\infty \cdots dsdt}_{A} + \underbrace{\int_0^\infty\int_0^{\frac{2}{\|X_i-X_j\|}} \cdots dsdt}_{B} + \underbrace{\int_0^{\frac{2}{\|X_i-X_j\|}}\int_0^\infty \cdots dsdt}_{C}\right]
\end{aligned}
\tag{67}
$$

(a) lower bound of A.

$$
A \geq 0 \tag{68}
$$

(b) lower bound of B.

$$
\begin{aligned}
B &\geq -\int_0^{\frac{2}{\|X_i-X_j\|}}\int_0^\infty (\max(0,1-C_d s^{-d}))^{n-2}\min(1,(n-2)C_d' t^{-d})dtds \\
&\geq -\frac{2}{\|X_i-X_j\|}\left(\max\left\{0, 1-C_d\left(\frac{\|X_i-X_j\|}{2}\right)^d\right\}\right)^{n-2}\int_0^\infty \min(1,(n-2)C_d t^{-d})dt \\
&\geq -\frac{2}{\|X_i-X_j\|}\left(\max\left\{0, 1-C_d\left(\frac{\|X_i-X_j\|}{2}\right)^d\right\}\right)^{n-2}\frac{1}{d-1}((n-2)C_d)^{\frac{1}{d}}.
\end{aligned}
\tag{69}
$$

Note that

$$\begin{aligned}
&\mathbb{E}\left[\frac{2}{\|X_i - X_j\|}\left(\max\left\{0, 1 - C_d\left(\frac{\|X_i - X_j\|}{2}\right)^d\right\}\right)^{n-2}\middle| X_i\right] \\
&= \int_0^{R_0} \frac{2}{r}\left(\max\left\{0, 1 - C_d\left(\frac{r}{2}\right)^d\right\}\right)^{n-2} d\mu_{\|X_i - X_j\| | X_i}(r) \\
&\lesssim \int_0^{R_0} \frac{2}{r}\left(\max\left\{0, 1 - C_d\left(\frac{r}{2}\right)^d\right\}\right)^{n-2} r^{d-1} dr \\
&\lesssim \int_0^{R_0} \left(\max\left\{0, 1 - \frac{R_0}{2^d} r^{d-1}\right\}\right)^{n-2} r^{d-2} dr \\
&= \int_0^{R_0^{d-1}} \left(\max\left\{0, 1 - \frac{R_0}{2^d} r^{d-1}\right\}\right)^{n-2} \frac{1}{d-1} d(r^{d-1}) \\
&\lesssim \frac{1}{n}
\end{aligned} \tag{70}$$

As a result,

$$\mathbb{E}_{X_i, X_j} B \gtrsim \frac{1}{n} \tag{71}$$

(c) Similarly for C, we have

$$\mathbb{E}_{X_i, X_j} C \gtrsim \frac{1}{n} \tag{72}$$

Combining all the above inequalities, we have

$$\mathbb{E}[\frac{1}{\tilde{r}_i \tilde{r}_j}] - \mathbb{E}[\frac{1}{\tilde{r}_i}]\mathbb{E}[\frac{1}{\tilde{r}_j}] \gtrsim \frac{1}{n}. \tag{73}$$

Upper bound for $|\mathbf{Cov}(\frac{1}{\tilde{r}_i}, \frac{1}{r_j})|$

$$|\text{Cov}(\frac{1}{\tilde{r}_i}, \frac{1}{r_j})| = |\mathbb{E}[\frac{1}{\tilde{r}_i \tilde{r}_j}] - \mathbb{E}[\frac{1}{\tilde{r}_i}]\mathbb{E}[\frac{1}{\tilde{r}_j}]| \lesssim \frac{1}{n}. \tag{74}$$

A.3.4. ESTIMATE FOR THE DIFFERENCE BETWEEN $\text{COV}(\frac{1}{\tilde{r}_i}, \frac{1}{r_j})$ AND $\text{COV}(\frac{1}{\tilde{r}_i}, \frac{1}{r_j})$

Upper bound for $\mathbb{E}|\tilde{r}_i^{-1} - r_i^{-1}|^2$ We have

$$|\tilde{r}_i^{-1} - r_i^{-1}| \leq \frac{1}{\|X_i - X_j\|}\mathbb{I}\{\|X_i - X_j\| < \tilde{r}_i\}. \tag{75}$$

Conditioned on X_i, $\|X_i - X_j\|$ and $\tilde{r}_i$ are, in fact, independent. Then

$$\begin{aligned}
\mathbb{E}[|\tilde{r}_i^{-1} - r_i^{-1}|^2 | X_i, \tilde{r}_i] &\leq \mathbb{E}[\frac{1}{\|X_i - X_j\|^2}\mathbb{I}\{\|X_i - X_j\| < \tilde{r}_i\} | X_i, \tilde{r}_i] \\
&\lesssim \mathbb{E}[\tilde{r}_i^{d-2} | X_i, r_i]
\end{aligned} \tag{76}$$

Hence,

$$\mathbb{E}[|\tilde{r}_i^{-1} - r_i^{-1}|^2 \leq \mathbb{E}\tilde{r}_i^{d-2} \leq \mathbb{E}[\tilde{r}_i^d]^{\frac{d-2}{d}} \lesssim n^{-\frac{d-2}{d}} \tag{77}$$

A.3.5. UPPER BOUND FOR $\mathbb{E}|\tilde{r}_i^{-1}\tilde{r}_j^{-1} - r_i^{-1}r_j^{-1}|$

$$\begin{aligned}\mathbb{E}|\tilde{r}_i^{-1}\tilde{r}_j^{-1} - r_i^{-1}r_j^{-1}| &\leq \mathbb{E}|\tilde{r}_i^{-1}\tilde{r}_j^{-1} - \tilde{r}_i^{-1}r_j^{-1}| + \mathbb{E}|\tilde{r}_i^{-1}r_j^{-1} - r_i^{-1}r_j^{-1}| \\ &\leq \sqrt{\mathbb{E}[\tilde{r}_i^{-2}]}\sqrt{\mathbb{E}[|\tilde{r}_j^{-1} - r_j^{-1}|^2]} + \sqrt{\mathbb{E}[r_j^{-2}]}\sqrt{\mathbb{E}[|\tilde{r}_i^{-1} - r_i^{-1}|^2]} \\ &\leq \sqrt{\mathbb{E}[r_i^{-2}]}\sqrt{\mathbb{E}[|\tilde{r}_j^{-1} - r_j^{-1}|^2]} + \sqrt{\mathbb{E}[r_j^{-2}]}\sqrt{\mathbb{E}[|\tilde{r}_i^{-1} - r_i^{-1}|^2]} \\ &\lesssim \sqrt{n^{2/d}}\sqrt{n^{-\frac{d-2}{d}}} \\ &\leq n^{-\frac{d-4}{2d}}\end{aligned} \tag{78}$$

A.3.6. UPPER BOUND FOR $|\mathbb{E}[\tilde{r}_i^{-1}]\mathbb{E}[\tilde{r}_j^{-1}] - \mathbb{E}[r_i^{-1}]\mathbb{E}[r_j^{-1}]|$

First,

$$\|\mathbb{E}[\tilde{r}_i^{-1}] - \mathbb{E}[r_i^{-1}]\| \leq \sqrt{\mathbb{E}[(\tilde{r}_i^{-1} - r_i^{-1})^2]} \lesssim n^{-\frac{d-2}{2d}} \tag{79}$$

and

$$\mathbb{E}\tilde{r}_i^{-1} \leq \mathbb{E}r_i^{-1} \lesssim n^{\frac{1}{d}}. \tag{80}$$

Then

$$\begin{aligned}&|\mathbb{E}[\tilde{r}_i^{-1}]\mathbb{E}[\tilde{r}_j^{-1}] - \mathbb{E}[r_i^{-1}]\mathbb{E}[r_j^{-1}]| \\ &= |\mathbb{E}[\tilde{r}_i^{-1}]\mathbb{E}[\tilde{r}_j^{-1}] - \mathbb{E}[\tilde{r}_i^{-1}]\mathbb{E}[r_j^{-1}]| + |\mathbb{E}[\tilde{r}_i^{-1}]\mathbb{E}[r_j^{-1}] - \mathbb{E}[r_i^{-1}]\mathbb{E}[r_j^{-1}]| \\ &= \mathbb{E}[\tilde{r}_i^{-1}]|\mathbb{E}[\tilde{r}_j^{-1}] - \mathbb{E}[r_j^{-1}]| + \mathbb{E}[r_j^{-1}]|\mathbb{E}[\tilde{r}_i^{-1}] - \mathbb{E}[r_i^{-1}]| \\ &\lesssim n^{-\frac{d-4}{2d}}.\end{aligned} \tag{81}$$

A.3.7. UPPER BOUND FOR THE DIFFERENCE BETWEEN $\text{Cov}(\frac{1}{\tilde{r}_i}, \frac{1}{r_j})$ AND $\text{Cov}(\frac{1}{\tilde{r}_i}, \frac{1}{r_j})$

$$\begin{aligned}&|\text{Cov}(\frac{1}{\tilde{r}_i}, \frac{1}{r_j}) - \text{Cov}(\frac{1}{\tilde{r}_i}, \frac{1}{r_j})| \\ &= |\mathbb{E}[\tilde{r}_i^{-1}]\mathbb{E}[\tilde{r}_j^{-1}] - \mathbb{E}[r_i^{-1}]\mathbb{E}[r_j^{-1}] + \mathbb{E}[\tilde{r}_i^{-1}]\mathbb{E}[\tilde{r}_j^{-1}] - \mathbb{E}[r_i^{-1}]\mathbb{E}[r_j^{-1}]| \\ &\leq |\mathbb{E}[\tilde{r}_i^{-1}]\mathbb{E}[\tilde{r}_j^{-1}] - \mathbb{E}[r_i^{-1}]\mathbb{E}[r_j^{-1}]| + |\mathbb{E}[\tilde{r}_i^{-1}]\mathbb{E}[\tilde{r}_j^{-1}] - \mathbb{E}[r_i^{-1}]\mathbb{E}[r_j^{-1}]| \\ &\lesssim n^{-\frac{d-4}{2d}}.\end{aligned} \tag{82}$$

A.3.8. ESTIMATE OF $\text{Cov}(\frac{1}{r_i}, \frac{1}{r_j})$

$$\text{Cov}(\frac{1}{r_i}, \frac{1}{r_j}) \lesssim \text{Cov}(\frac{1}{\tilde{r}_i}, \frac{1}{\tilde{r}_j}) + n^{-\frac{d-4}{2d}} \lesssim n^{-\frac{d-4}{2d}} \tag{83}$$

A.3.9. UPPER BOUND OF VAR($\frac{1}{n}\sum_{i=1}^{n} r_i^{-1}$)

$$\begin{aligned}\mathrm{Var}(\frac{1}{n}\sum_{i=1}^{n} r_i^{-1}) &\leq \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(r_i^{-1}) + \frac{1}{n^2}\sum_{i=1,j=1,i\neq j}^{n}\mathrm{Cov}(r_i^{-1}, r_j^{-1}) \\ &\leq \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}(r_i^{-2}) + \frac{1}{n^2}\sum_{i=1,j=1,i\neq j}^{n}\mathrm{Cov}(r_i^{-1}, r_j^{-1}) \\ &\lesssim n^{-\frac{2}{d}-1} + n^{-\frac{d-4}{2d}} \\ &\lesssim n^{-\frac{d-4}{2d}}\end{aligned} \tag{84}$$

A.3.10. FINAL STEP: CHEBYSHEV'S INEQUALITY

By Chebyshev's inequality

$$\mathbb{P}[\frac{1}{n}\sum_{i=1}^{n} r_i^{-1} > An^{\frac{1}{d}} + \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} r_i^{-1}]] \leq (A^2 n^{\frac{2}{d}})^{-1}\mathrm{Var}(\frac{1}{n}\sum_{i=1}^{n} r_i^{-1}) \lesssim A^{-2}n^{-\frac{1}{2}} \tag{85}$$

since

$$\mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} r_i^{-1}] = \mathbb{E}r_1^{-1} \lesssim n^{\frac{1}{d}}. \tag{86}$$

This concludes the proof.

Appendix B. Inequalities for Functions

B.1. Gagliardo-Nirenberg interpolation inequalities

Here we quote the statements of Gagliardo-Nirenberg inequalities from (Leoni, 2017). Note that here the term "interpolation" has nothing to do with our notion of interpolation.

Theorem 8 (Gagliardo-Nirenberg interpolation for $\mathbb{R}^N$, general case, Theorem 12.87 in Leoni (2017))

Let $1 \leq p, q \leq \infty, m \in \mathbb{N}, k \in \mathbb{N}_0$, with $0 \leq k < m$, and let θ, r be such that

$$0 \leq \theta \leq 1 - k/m \tag{87}$$

and

$$(1-\theta)\left(\frac{1}{p} - \frac{m-k}{N}\right) + \theta\left(\frac{1}{q} + \frac{k}{N}\right) = \frac{1}{r} \in (-\infty, 1]. \tag{88}$$

Then there exists a constant $c = c(m, N, p, q, \theta, k) > 0$ such that

$$|\nabla^k u|_r \leq c\|u\|_{L^q(\mathbb{R}^N)}^{\theta}\|\nabla^m u\|_{L^p(\mathbb{R}^N)}^{1-\theta} \tag{89}$$

for every $u \in L^q(\mathbb{R}^N) \cap \dot{W}^{m,p}(\mathbb{R}^N)$, with the following exceptional cases:

(i) If $k = 0, mp < N$, and $q = \infty$, we assume that u vanishes at infinity.

(ii) If $1 < p < \infty$ and $m - k - N/p$ is a nonnegative integer, then (89) only holds for $0 < \theta \leq 1 - k/m$

Theorem 9 (Gagliardo-Nirenberg interpolation for domains, Theorem 13.61 in Leoni (2017)) *Let $\Omega \subset \mathbb{R}^N$ be an open set with uniformly Lipschitz continuous boundary (with parameters ϵ, L, M), let $0 < l < \epsilon/(4(1+L))$, let $m, k \in \mathbb{N}$, with $m \geq 2$ and $1 \leq k < m$, and let $1 \leq p, q, r \leq \infty$ be such that $p \leq q$ and*

$$\frac{k}{m}\frac{1}{p} + \left(1 - \frac{k}{m}\right)\frac{1}{q} = \frac{1}{r}. \tag{90}$$

If $p < q$, assume further that Ω is bounded.

Then for every $u \in L^q(\Omega) \cap \dot{W}^{m,p}(\Omega)$,

$$\|\nabla^k u\|_{L^r(\Omega)} \leq cl^{-k}|\Omega|^{1/r-1/q}\|u\|_{L^q(\Omega)} + c\|u\|_{L^q(\Omega)}^{1-k/m}\|\nabla^m u\|_{L^p(\Omega)}^{k/m} \tag{91}$$

if $p < q$, while

$$\|\nabla^k u\|_{L^p(\Omega)} \leq cl^{-k}\|u\|_{L^p(\Omega)} + c\|u\|_{L^p(\Omega)}^{1-k/m}\|\nabla^m u\|_{L^p(\Omega)}^{k/m} \tag{92}$$

if $p = q$. Here, $c > 0$ is a constant depending on m, N, p, q.

Remark 10 *Two remarks about notation:*

- *the notation $|\cdot|_r$ is defined by*

$$|u|_r := \begin{cases} \|u\|_{L^r(\mathbb{R}^N)} & \text{if } r > 0, \\ \|\nabla^n u\|_{L^\infty(\mathbb{R}^N)} & \text{if } r < 0 \text{ and } a = 0, \\ |\nabla^n u|_{C^{0,a}(\mathbb{R}^N)} & \text{if } r < 0 \text{ and } 0 < a < 1, \end{cases} \tag{93}$$

  where if $r < 0$ we set $n :=$ floor$(-N/r)$ and $a := -n - N/r \in [0, 1)$, provided the right-hand sides are well-defined.

- *$\dot{W}^{m,p}(\Omega)$ is the homogeneous Sobolev space and it coincides with the Sobolev space $W^{m,p}(\Omega)$ when Ω is a domain with finite measure.*

Remark 11

For our purposes, we need the inequality in two cases:

(i) The domain is $\mathbb{R}^d$ with d odd, $r = q = 2$, $k = 1, m = \frac{d+1}{2}, \theta = 0$, then

$$1 \times \left(\frac{1}{p} - \frac{\frac{d+1}{2} - 1}{d}\right) + 0 \times \left(\frac{1}{2} + \frac{1}{d}\right) = \frac{1}{2} \in (-\infty, 1]. \tag{94}$$

which implies

$$p = 2d \tag{95}$$

Then

$$m - k - N/p = \frac{d+1}{2} - 1 - \frac{d}{2d} = \frac{d-2}{2} \tag{96}$$

is not an integer because d is odd.

Therefore, our case is not exceptional and from equation (89), we get

$$\|Du\|_{L^{2d}(\mathbb{R}^d)} \leq C_d\|D^{\frac{d+1}{2}}u\|_{L^2(\mathbb{R}^d)} \tag{97}$$

(ii) The domain is $\Omega = \operatorname{supp} \mathcal{P} = B(\mathbf{0}, 1)$, when $N = d$ is odd, $r = q = p = 2, 0 \leq k \leq \frac{d+1}{2}, m = \frac{d+1}{2}$, then

$$\frac{k}{m}\frac{1}{p} + \left(1 - \frac{k}{m}\right)\frac{1}{q} = \frac{1}{r} \tag{98}$$

holds. Then

$$\|D^k u\|_{L^2(\Omega)} \leq C_{k,d} \|D^{\frac{d+1}{2}} u\|^{\alpha}_{L^2(\Omega)} \|u\|^{1-\alpha}_{L^2(\Omega)} + C'_{k,d}\|u\|_{L^2(\Omega)}. \tag{99}$$

Since

$$\|D^{\frac{d+1}{2}} u\|_{L^2(\Omega)} \leq \|D^{\frac{d+1}{2}} u\|_{L^2(\mathbb{R}^d)}, \tag{100}$$

from equation (92) we have

$$\|D^k u\|_{L^2(\Omega)} \leq C_{k,d} \|D^{\frac{d+1}{2}} u\|^{\alpha}_{L^2(\mathbb{R}^d)} \|u\|^{1-\alpha}_{L^2(\Omega)} + C'_{k,d}\|u\|_{L^2(\Omega)}. \tag{101}$$

Note the theorem itself doesn't cover $k = 0, \frac{d+1}{2}$ but equation (101) holds trivially in the two cases when $p = q = r$.

B.2. Morrey's inequality

Theorem 12 (Morrey's inequality)
Suppose $u : \mathbb{R}^d \to \mathbb{R}$ has weak derivative Du in $L^{2d}(\mathbb{R}^d)$

$$\sup_{x\in\mathbb{R}^d, r>0} \frac{1}{\sqrt{r}} \left| u(x) - ⨍_{B(x,r)} u(y)dy \right| \le C_d \|Du\|_{L^{2d}(\mathbb{R}^d)} \tag{102}$$

If in addition, $u \in L^q(\mathbb{R}^d)$, combining with Gagliardo-Nirenberg interpolation inequality for $\mathbb{R}^d$ (equation 89), we have

$$\sup_{x\in\mathbb{R}^d, r>0} \frac{1}{\sqrt{r}} \left| u(x) - ⨍_{B(x,r)} u(y)dy \right| \le C_d \|D^{\frac{d+1}{2}} u\|_{L^2(\mathbb{R}^d)} \tag{103}$$

Remark 13 *Here the notation $⨍_{B(x,r)}$ means the average over the ball $B(x,r)$, i.e. $\frac{1}{|B(x,r)|}\int_{B(x,r)}$.*

Remark 14 *This version of Morrey's inequality is basically a middle step of Lemma 12.47 in (Leoni, 2017) (although it is a cube instead of a ball there) and the proof is simple enough to be written down below.*

Proof For any $x \in \mathbb{R}^d, r > 0$

$$\begin{aligned}
\left| u(x) - ⨍_{B(x,r)} u(y)dy \right| &= \left| ⨍_{B(x,r)} (u(x) - u(y))dy \right| \\
&= \left| ⨍_{B(x,r)} \int_0^1 \frac{d}{dt}\Big(u(x) - u(x + t(y-x))\Big) dtdy \right| \\
&\le ⨍_{B(x,r)} \int_0^1 \|y - x\| \|Du(x + t(y-x))\| dtdy \\
&= \int_0^1 \left(⨍_{B(x,r)} \|y - x\| \|Du(x + t(y-x))\| dy \right) dt \\
&= \int_0^1 t^{-1} \left(⨍_{B(x,tr)} \|y - x\| \|Du(y)\| dy \right) dt \\
&\le \int_0^1 t^{-1} \left(⨍_{B(x,tr)} \|y - x\|^{\frac{2d}{2d-1}} dy \right)^{\frac{2d-1}{2d}} \left(⨍_{B(x,tr)} \|Du(y)\|^{2d} dy \right)^{\frac{1}{2d}} dt \\
&\le O_d \left(\int_0^1 t^{-1} \left(r^{\frac{2d}{2d-1}} t^{\frac{2d}{2d-1}} \right)^{\frac{2d-1}{2d}} \left(r^{-d} t^{-d} \int_{\mathbb{R}^d} \|Du(y)\|^{2d} dy \right)^{\frac{1}{2d}} dt \right) \\
&\le O_d \left(\sqrt{r} \|Du\|_{L^{2d}(\mathbb{R}^d)} \int_0^1 t^{-\frac{1}{2}} dt \right) \\
&= O_d \left(\sqrt{r} \|Du\|_{L^{2d}(\mathbb{R}^d)} \right)
\end{aligned}$$

■

B.3. Local Hölder Continuity around Samples

Definition 15 (Measure of Local Hölder Continuity around Samples) *For sample set $\mathcal{S}$ and index set $I \subset [n]$, we introduce the following measure of local Hölder continuity around samples*

$$[f]_{\eta,\mathcal{S},I} = \sum_{i\in I} \sup_{x\in\mathbb{R}^d, r>0} \frac{1}{r}\left(f(x)\eta\left(\frac{x-X_i}{r_i}\right) - \fint_{B(x,r)} f(y)\eta\left(\frac{y-X_i}{r_i}\right)dy\right)^2 \tag{104}$$

$$\textit{where } \eta(x) = \begin{cases} 1, & \|x\| \le \frac{1}{4} \\ e^{1-\frac{1}{2-4\|x\|}}, & \frac{1}{4} < \|x\| < \frac{1}{2} \\ 0, & \|x\| \ge \frac{1}{2} \end{cases}$$

Lemma 16 *For any subset $I \subset [n], \beta \in (0,1)$ and $f \in L^2(\Omega)$*

$$\|f\|^2_{L^2(\Omega)} \ge \frac{3}{4}\frac{\beta^d \pi^{\frac{d}{2}}}{2^d\Gamma(\frac{d}{2}+1)}\left(\sum_{i\in I} r_i^d f(X_i)^2 - 4\beta[f]_{\eta,\mathcal{S},I}\max_{i\in I} r_i^{d+1}\right). \tag{105}$$

Proof We write

$$\begin{aligned} \|f\|^2_{L^2(\Omega)} &\ge \sum_{i\in I}\int_{B(X_i,\beta r_i/2)} f(x)^2 dx & (106)\\ &\ge \sum_{i\in I}\int_{B(X_i,\beta r_i/2)} f(x)^2 \eta\left(\frac{x-X_i}{r_i}\right)^2 dx & (107)\\ &\ge \sum_{i\in I}\frac{1}{|B(X_i,\beta r_i/2)|}\left(\int_{B(X_i,\beta r_i/2)} f(x)\eta\left(\frac{x-X_i}{r_i}\right)dx\right)^2. & (108) \end{aligned}$$

Writing this expression as a normalized integral, we get

$$\begin{aligned} &\sum_{i\in I}|B(X_i,\beta r_i/2)|\left(\fint_{B(X_i,\beta r_i/2)} f(x)\eta\left(\frac{x-X_i}{r_i}\right)dx\right)^2 & (109)\\ &\ge \sum_{i\in I}|B(X_i,r_i/2)|\left(\frac{3}{4}f(X_i)^2 - 3\left(f(X_i) - \fint_{B(X_i,\beta r_i/2)} f(x)\eta\left(\frac{x-X_i}{r_i}\right)dx\right)^2\right) & (110)\\ &\ge \frac{3}{4}\sum_{i\in I}|B(X_i,\beta r_i/2)|f(X_i)^2 - 3[f]_{\eta,\mathcal{S},I}\sup_{i\in I}\beta r_i B(X_i,\beta r_i/2) & (111)\\ &= \frac{3}{4}\frac{\beta^d\pi^{\frac{d}{2}}}{2^d\Gamma(\frac{d}{2}+1)}\left(\sum_{i\in I} r_i^d f(X_i)^2 - 4\beta[f]_{\eta,\mathcal{S},I}\max_{i\in I} r_i^{d+1}\right). & (112) \end{aligned}$$

■

Lemma 17 *For any subset $I \subset [n]$, we have*

$$[f]_{\eta,\mathcal{S},I} \le O_d\Big(\big(1+\|f\|_{L^2(\Omega)}^2\big)\big(c^{d+1}\langle f\rangle_{\mathcal{H}_c} + \max_{i\in I} r_i^{-d-1}\big)\Big). \tag{113}$$

Proof

Define η_i by

$$\eta_i(x) = \eta\left(\frac{x-X_i}{r_i}\right) \tag{114}$$

and

$$A = \max\{c\langle f\rangle_{\mathcal{H}_c}^{\frac{1}{d+1}}, \max_{i\in I} r_i^{-1}\} \tag{115}$$

We prove our lemma by first proving the following inequalities:

(a) $[f]_{\eta,\mathcal{S},I} \le O_d\left(\sum_{i\in I} \|D^{\frac{d+1}{2}}(f\eta_i)\|_{L^2(\mathbb{R}^d)}^2\right)$

(b) $\sum_{i\in I} \|D^{\frac{d+1}{2}}(f\eta_i)\|_{L^2(\mathbb{R}^d)}^2 \le O_d\left(\sum_{j=0}^{\frac{d+1}{2}} A^{d+1-2j}\|D^j f\|_{L^2(\mathbb{R}^d)}^2\right)$

(c) $\|D^j f\|_{L^2(\mathbb{R}^d)} \le O_d\left(\left(1+\|f\|_{L^2(\Omega)}\right)A^j\right)$

and then it follows that

$$[f]_{\eta,\mathcal{S},I} \le O_d\Big(\big(1+\|f\|_{L^2(\Omega)}^2\big)\big(c^{d+1}\langle f\rangle_{\mathcal{H}_c} + \max_{i\in I} r_i^{-d-1}\big)\Big). \tag{116}$$

Inequality (a). This is a direct application of Morrey's inequality (equation (103)).

Inequality (b). Using Leibnitz rule we have

$$\|D^{\frac{d+1}{2}}(f\eta_i)\|_{L^2(\mathbb{R}^d)}^2 \le O_d\left(\sum_{|\alpha|=\frac{d+1}{2}}\sum_{0\le\beta\le\alpha} \|D^{\alpha-\beta}\eta_i D^\beta f\|_{L^2(\mathbb{R}^d)}^2\right). \tag{117}$$

Since the function $D^{\alpha-\beta}\eta_i D^\beta f$ is supported within the ball $B(X_i, r_i)$, we have

$$\|D^{\frac{d+1}{2}}(f\eta_i)\|_{L^2(\mathbb{R}^d)}^2 = O_d\left(\sum_{|\alpha|=\frac{d+1}{2}}\sum_{0\le\beta\le\alpha} \|D^{\alpha-\beta}\eta_i D^\beta f\|_{L^2(B(X_i,r_i))}^2\right). \tag{118}$$

By Hölder inequality,

$$\|D^{\frac{d+1}{2}}(f\eta_i)\|_{L^2(\mathbb{R}^d)}^2 \le O_d\left(\sum_{|\alpha|=\frac{d+1}{2}}\sum_{0\le\beta\le\alpha} \|D^{\alpha-\beta}\eta_i\|_{L^\infty(B(X_i,r_i))}^2 \|D^\beta f\|_{L^2(B(X_i,r_i))}^2\right). \tag{119}$$

Using the fact that

$$\|D^\beta \eta_i\|_{L^\infty(\mathbb{R}^d)} \leq C_d r_i^{-|\beta|}, \tag{120}$$

we then get

$$\begin{aligned}\|D^{\frac{d+1}{2}}(f\eta_i)\|^2_{L^2(\mathbb{R}^d)} &= O_d\left(\sum_{|\alpha|=\frac{d+1}{2}} \sum_{0\leq\beta\leq\alpha} \frac{\|D^\beta f\|^2_{L^2(B(X_i,r_i))}}{r_i^{2|\alpha-\beta|}}\right) \\ &\leq O_d\left(\sum_{j=0}^{\frac{d+1}{2}} \frac{\|D^j f\|^2_{L^2(B(X_i,r_i))}}{r_i^{d+1-2j}}\right) \\ &\leq O_d\left(\sum_{j=0}^{\frac{d+1}{2}} \frac{\|D^j f\|^2_{L^2(B(X_i,r_i))}}{\min\limits_{i\in I} r_i^{d+1-2j}}\right).\end{aligned} \tag{121}$$

Then we have

$$\begin{aligned}\sum_{i\in I}\|D^{\frac{d+1}{2}}(f\eta_i)\|^2_{L^2(\mathbb{R}^d)} &\leq O_d\left(\sum_{j=0}^{\frac{d+1}{2}} \frac{\sum\limits_{i\in I}\|D^j f\|^2_{L^2(B(X_i,r_i))}}{\min\limits_{i\in I} r_i^{d+1-2j}}\right) \\ &\leq O_d\left(\sum_{j=0}^{\frac{d+1}{2}} \frac{\|D^j f\|^2_{L^2(\Omega)}}{\min\limits_{i\in I} r_i^{d+1-2j}}\right) \\ &\leq O_d\left(\sum_{j=0}^{\frac{d+1}{2}} A^{d+1-2j}\|D^j f\|^2_{L^2(\Omega)}\right).\end{aligned} \tag{122}$$

Inequality (c). Here use Gagliardo-Nirenberg interpolation inequality for domains (equation (101)) and the fact

$$\|D^{\frac{d+1}{2}} f\|^2_{L^2(\Omega)} \leq \|D^{\frac{d+1}{2}} f\|^2_{L^2(\mathbb{R}^d)} \leq c^{d+1}\langle f\rangle_{\mathcal{H}_c}, \tag{123}$$

we have

$$\begin{aligned}\|D^j f\|_{L^2(\Omega)} &\leq O_d\left(\|D^{\frac{d+1}{2}} f\|^{\frac{2j}{d+1}}_{L^2(\Omega)}\|f\|^{1-\frac{2j}{d+1}}_{L^2(\Omega)} + \|f\|_{L^2(\Omega)}\right) \\ &\leq O_d\left(c^j\langle f\rangle^{\frac{j}{d+1}}_{\mathcal{H}_c}\|f\|^{1-\frac{2j}{d+1}}_{L^2(\Omega)} + \|f\|_{L^2(\Omega)}\right) \\ &\leq O_d\left(\left(1+\|f\|_{L^2(\Omega)}\right)A^j\right).\end{aligned} \tag{124}$$

■

Proposition 18 *For any subset $I \subset [n]$ and $f \in L^2(\Omega)$, we have*

$$\|f\|^2_{L^2(\Omega)} \geq \min\left\{1, \Omega_d\left(\left(\frac{\min\limits_{i\in I} r_i^{-d-1}\sum_{i\in I} r_i^d f(X_i)^2}{\max\limits_{i\in I} r_i^{-d-1} + c^{d+1}\|f\|_{\mathcal{H}_c}}\right)^d \sum_{i\in I} r_i^d f(X_i)^2\right)\right\}. \tag{125}$$

Proof Without loss of generality suppose that $\|f\|^2_{L^2(\Omega)} \leq 1$. Then from Lemma 17, there is a constant C_d such that

$$[f]_{\eta,\mathcal{S},I} \leq C_d \left(c^{d+1}\|f\|_{\mathcal{H}_c} + \max_{i\in I} r_i^{-d-1} \right). \tag{126}$$

From Lemma 16, we have for any $\beta \in (0,1)$:

$$\begin{aligned}\|f\|^2_{L^2(\Omega)} &\geq \frac{3}{4}\frac{\beta^d\pi^{\frac{d}{2}}}{2^d\Gamma(\frac{d}{2}+1)}\left(\sum_{i\in I} r_i^d f(X_i)^2 - 4\beta[f]_{\eta,\mathcal{S},I}\max_{i\in I} r_i^{d+1}\right)\\ &\geq \frac{3}{4}\frac{\beta^d\pi^{\frac{d}{2}}}{2^d\Gamma(\frac{d}{2}+1)}\left(\sum_{i\in I} r_i^d f(X_i)^2 - 4\beta C_d\max_{i\in I} r_i^{d+1}\left(c^{d+1}\|f\|_{\mathcal{H}_c} + \max_{i\in I} r_i^{-d-1}\right)\right).\end{aligned} \tag{127}$$

Taking

$$\beta = \frac{\max\limits_{i\in I} r_i^{-d-1}\sum_{i\in I} r_i^d f(X_i)^2}{8C_d\left(c^{d+1}\|f\|_{\mathcal{H}_c} + \max\limits_{i\in I} r_i^{-d-1}\right)}, \tag{128}$$

we get

$$\begin{aligned}\|f\|^2_{L^2(\Omega)} &\geq \frac{3}{4}\frac{\beta^d\pi^{\frac{d}{2}}}{2^d\Gamma(\frac{d}{2}+1)}\left(\sum_{i\in I} r_i^d f(X_i)^2 - \frac{1}{2}\sum_{i\in I} r_i^d f(X_i)^2\right)\\ &\geq \frac{3}{8}\frac{\pi^{\frac{d}{2}}}{2^d\Gamma(\frac{d}{2}+1)}\left(\frac{\min\limits_{i\in I} r_i^{-d-1}\sum_{i\in I} r_i^d f(X_i)^2}{8C_d\left(c^{d+1}\|f\|_{\mathcal{H}_c} + \max\limits_{i\in I} r_i^{-d-1}\right)}\right)^d \sum_{i\in I} r_i^d f(X_i)^2 \\ &\geq \Omega_d\left(\left(\frac{\min\limits_{i\in I} r_i^{-d-1}\sum_{i\in I} r_i^d f(X_i)^2}{c^{d+1}\|f\|_{\mathcal{H}_c} + \max\limits_{i\in I} r_i^{-d-1}}\right)^d \sum_{i\in I} r_i^d f(X_i)^2\right).\end{aligned} \tag{129}$$

■

B.4. Upper Bound on $\langle \widehat{f}_c\rangle_{\mathcal{H}_c}$

Proposition 19 *With probability at least $1 - O_{d,\rho}(\frac{1}{\sqrt{n}})$, for any $c > 0$ there is a function g interpolating $\mathcal{S}$ such that*

$$\langle g\rangle_{\mathcal{H}_c} \leq \frac{1}{3}\|f^*\|^2_{L^2(\Omega)} + O_{d,\rho,f^*}\left(\frac{\sqrt[d]{n}}{c}\left(1+\frac{\sqrt[d]{n}}{c}\right)^d\right). \tag{130}$$

Since $\widehat{f}_c$ has the smallest RKHS norm among all interpolating functions, we have

$$\langle \widehat{f}_c\rangle_{\mathcal{H}_c} \leq \frac{1}{3}\|f^*\|^2_{L^2(\Omega)} + O_{d,\rho,f^*}\left(\frac{\sqrt[d]{n}}{c}\left(1+\frac{\sqrt[d]{n}}{c}\right)^d\right). \tag{131}$$

Proof Define $r_i = \min_{j \neq i} \|X_i - X_j\|$ and

$$\eta(x) = \begin{cases} 1, & \|x\| \leq \frac{1}{4} \\ e^{1-\frac{1}{2-4\|x\|}}, & \frac{1}{4} < \|x\| < \frac{1}{2} \\ 0, & \|x\| \geq \frac{1}{2} \end{cases} \tag{132}$$

and for $\alpha \in (0, \frac{1}{2})$ take

$$g_\alpha(x) := \sum_{i=1}^{n} Y_i \eta \left(\frac{x - X_i}{\alpha r_i} \right). \tag{133}$$

First,

$$\begin{aligned} \|g_\alpha\|_{L^2(\mathbb{R}^d)}^2 &= \sum_i Y_i^2 \|\eta_{X_i, \alpha r_i}\|_{L^2(\mathbb{R}^d)}^2 \\ &= \alpha^d \|\eta\|_{L^2(\mathbb{R}^d)}^2 \sum_i Y_i^2 r_i^d \\ &\leq \alpha^d \|\eta\|_{L^2(\mathbb{R}^d)}^2 \sum_i (\|f^*\|_{L^\infty(\Omega)} + 1)^2 r_i^d \\ &\leq \alpha^d \|\eta\|_{L^2(\mathbb{R}^d)}^2 (\|f^*\|_{L^\infty(\Omega)} + 1)^2 \sum_i r_i^d \\ &\leq \frac{2^d |\Omega|}{|B_d(1)|} \alpha^d \|\eta\|_{L^2(\mathbb{R}^d)}^2 (\|f^*\|_{L^\infty(\Omega)} + 1)^2 \\ &\leq O_d(\alpha^d) \end{aligned} \tag{134}$$

Therefore, we can take α to be a constant dependent only on d and f^* such that

$$\|g_\alpha(x)\|_{L^2(\mathbb{R}^d)}^2 \leq \frac{1}{3} \|f^*\|_{L^2(\Omega)}^2 \tag{135}$$

Since

$$\langle \eta \left(\frac{x - X_i}{\alpha r_i} \right) \rangle_k = \alpha^{d-2k} r_i^{d-2k} \langle \eta \rangle_k \tag{136}$$

and

$$\langle u, v \rangle_{\mathcal{H}_c} = 0, \text{ if supp } u \cap \text{supp } v = \emptyset \tag{137}$$

then for $k \in \mathbb{N}$ we have

$$\langle g \rangle_{\mathcal{H}_c} = \|g\|_{L^2(\mathbb{R}^d)}^2 + \sum_{i=1}^{n} Y_i^2 (\alpha r_i)^{d-2k} \langle \eta \rangle_k. \tag{138}$$

So when d is odd,

$$\begin{aligned}
\langle g\rangle_{\mathcal{H}_c} &= \|g\|^2_{L^2(\mathbb{R}^d)} + \sum_{k=1}^{\frac{d+1}{2}} \sum_{i=1}^{n} \binom{\frac{d+1}{2}}{k} Y_i^2 c^{-2k} (\alpha r_i)^{d-2k} \langle \eta \rangle_k \\
&\leq \frac{1}{3}\|f^*\|^2_{L^2(\mathbb{R}^d)} + \sum_{k=1}^{\frac{d+1}{2}} \sum_{i=1}^{n} \binom{\frac{d+1}{2}}{k} \left(\|f^*\|_{L^\infty(\Omega)} + 1\right)^2 c^{-2k} (\alpha r_i)^{d-2k} \langle \eta \rangle_k \\
&\leq \frac{1}{3}\|f^*\|^2_{L^2(\mathbb{R}^d)} + O_{d,\rho}\left(\left(\|f^*\|_{L^\infty(\Omega)} + 1\right)^2 \sum_{k=1}^{\frac{d+1}{2}} \sum_{i=1}^{n} c^{-2k} (\alpha r_i)^{d-2k} \right) \\
&\leq \frac{1}{3}\|f^*\|^2_{L^2(\mathbb{R}^d)} + O_{d,\rho,f^*}\left(\sum_{k=1}^{\frac{d+1}{2}} \sum_{i=1}^{n} c^{-2k} r_i^{d-2k} \right).
\end{aligned} \tag{139}$$

From Proposition 3, with probability at least $1 - O_{d,\rho}\left(\frac{1}{\sqrt{n}}\right)$ we have

$$\sum_{i=1}^{n} r_i^{d-2k} \leq O_{d,\rho}\left(n^{2k/d}\right). \tag{140}$$

Then with the same probability,

$$\langle g\rangle_{\mathcal{H}_c} \leq \frac{1}{3}\|f^*\|^2_{L^2(\Omega)} + O_{d,\rho,f^*}\left(\frac{\sqrt[d]{n}}{c} \left(1 + \frac{\sqrt[d]{n}}{c}\right)^d \right). \tag{141}$$

■