Proceedings of Machine Learning Research vol 178:1–31, 2022 35th Annual Conference on Learning Theory

Kernel interpolation in Sobolev spaces is not consistent in low dimensions

Simon Buchholz SBUCHHOLZ@TUE.MPG.DE
Max Planck Institute for Intelligent Systems, Tübingen

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

We consider kernel ridgeless ridge regression with kernels whose associated RKHS is a Sobolev space H^s. We show for $d/2 < s < 3d/4$ that interpolation is not consistent in fixed dimension, extending earlier results for the Laplace kernel in odd dimensions and underlining again that benign overfitting is rare in low dimensions. The proof proceeds by deriving sharp bounds on the spectrum of random kernel matrices using results from the theory of radial basis functions which might be of independent interest.

Keywords: Kernel ridge regression, generalization, benign overfitting, radial basis functions

1. Introduction

In the past years, numerous works starting with Zhang et al. (2017) showed empirically that overparamatrized models that interpolate the training data can nevertheless generalize. This posed a challenge to the traditional bias-variance tradeoff picture Belkin et al. (2019a) and to the learning theory bounds based on uniform convergence Nagarajan and Kolter (2019). Many works then established conditions for benign overfitting for different model classes, e.g., kernel smoothing Belkin et al. (2019b), linear regression Bartlett et al. (2020); Hastie et al. (2019); Koehler et al. (2021), kernel ridge regression Liang and Rakhlin (2020); Liang et al. (2020), and random feature models Mei and Montanari; Mei et al. (2021), see also Bartlett et al. (2021) for an overview. A complementary line focuses on the characterization of implicit regularization Soudry et al. (2018); Azulay et al. (2021). For (deep) neural nets less rigorous results are known and most of them rely on the relation to kernel methods for wide networks through the neural tangent kernel Jacot et al. (2018); Du et al. (2019). Thus, a precise understanding of generalization in kernel ridge regression is desirable and potentially necessary to study more complex neural nets Belkin et al. (2018). Here the Laplace kernel $k(x-y) = \exp(-|x-y|)$ is of particular interest because it has been shown that the Laplace kernel is more similar to the neural tangent kernel than, e.g., the Gaussian kernel, in particular they have a similar spectral behaviour Geifman et al. (2020).

Almost all results establishing benign overfitting study the high dimensional regime where dimension d and sample size n diverge jointly $n \propto d^\alpha$. In contrast, Rakhlin and Zhai (2019) showed that for fixed (odd) dimension minimum norm interpolants for the Laplace kernel cannot have vanishing error for noisy data as $n \to \infty$ even if the kernel-bandwidth is chosen depending on the data. This result provides evidence that high dimension is a requirement for benign overfitting. However, their proof is specific to the Laplace kernel and relies on a rather explicit constructions of small norm interpolants. To derive more general results it is desirable to directly control the kernel matrix without using the specific structure of the Reproducing Kernel Hilbert Space (RKHS). This is also the approach used in the asymptotic statistical theory of regularized ridge regression. There the convergence of the kernel matrix to the Mercer operator is established and optimal convergence rates can be derived Caponnetto and Vito (2007); Rosasco et al. (2010). Similar arguments can be applied in high dimensions even tough one obtains a very different limiting operator Karoui (2010). Here we consider an intermediate setting where the dimension is fixed but the bandwidth varies and we add no regularization.

It was argued in Belkin (2018) that using results from approximation theory and radial basis functions (see, e.g., Buhmann (2003); Wendland (2004)) gives much tighter control of the kernel spectrum for the Gaussian kernel than using standard concentration results for random matrices which become void for small eigenvalues. We find that this is also true in our setting and we derive lower bounds on the spectrum of kernel matrices with Sobolev spaces as an RKHS using a well-known result from Schaback (1995). Together with results for fractional Sobolev spaces this allows us to generalize the results of Rakhlin and Zhai (2019) to more kernels and remove the restriction on the dimension. We summarize the main contributions of this work as follows:

- We show how results from approximation theory can be used to get a precise control on the spectrum of kernel matrices that improves upon concentration bounds.
- The previous point allows us to considerably simplify the proof of Rakhlin and Zhai (2019), in particular our geometric arguments on the pairwise distances of data points are considerably simpler.
- We generalize their result to (fractional) Sobolev spaces in all dimensions.
- Finally, we note that the proof strategy is limited to Sobolev spaces with a moderate degree of smoothness and we clarify that new approaches are required to address interpolation in low dimensions for RKHS consisting of very smooth functions like the Gaussian kernel.

2. Setting and main result

In this section we introduce the setting, our main result and give an overview of the main ingredients of the proof.

2.1. Setting

We consider kernel ridge regression for some dataset $\{(x_1, y_1), \ldots, (x_n, y_n)\} \subset \mathbb{R}^d \times \mathbb{R}$

$$f_{\text{ridge}} = \arg\min \frac{1}{n} \sum_i (f(x_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2 \tag{1}$$

where $\mathcal{H}$ denotes the RKHS for some kernel k and $\lambda > 0$ is a regularization parameter. As $\lambda \to 0$ one obtains minimum norm interpolation given by

$$f_{\text{interpolation}} = \arg\min_{f \in \mathcal{H}} \|f\|_{\mathcal{H}} \text{ such that } f(x_i) = y_i \text{ for all } i. \tag{2}$$

We will restrict our attention to translation invariant kernels $k(x, y) = k(x - y)$ whose Fourier transform is given by

$$\hat{k}(\xi) = (1 + |\xi|^2)^{-s} \tag{3}$$

for some $s > d/2$. In this case the RKHS norm can be written as

$$\|f\|_{\mathcal{H}}^2 = \int \frac{|\hat{f}(\xi)|^2}{\hat{k}(\xi)} \, \mathrm{d}\xi = \int |\hat{f}(\xi)|^2 (1 + |\xi|^2)^s \, \mathrm{d}\xi = \|f\|_{H^s}^2 \tag{4}$$

where $H^s = W^{2,s}$ denotes the Sobolev space consisting of L^2 functions whose derivatives up to order s are square integrable. An important special case is $s = (d+1)/2$ where $k(x-y) = e^{-|x-y|}$ is the popular Laplace kernel (here we neglect scaling constants depending on the choice of the Fourier transform normalization, we refer to Rakhlin and Zhai (2019) for a proof). A standard procedure in kernel methods is to use a data dependent choice of the bandwidth selected by cross-validation. We define the kernel with bandwidth $\gamma > 0$ by

$$k_\gamma(x-y) = \gamma^{-d} k\left(\frac{x-y}{\gamma}\right) \tag{5}$$

The scaling relation $\mathcal{F}(f(x/\gamma)) = \gamma^d \hat{f}(\gamma x)$ implies that $\hat{k}_\gamma(\xi) = \hat{k}(\gamma\xi)$. In particular for kernels as in (3) we obtain $\hat{k}(\xi) = (1 + \gamma^2|\xi|^2)^{-s}$ such that changing the bandwidth just changes the relative weight of the derivative norms compared to the L^2 norm. For the first part of the paper we only need the following slightly less restrictive bound on the kernel.

Assumption 1 *The kernel k satisfies the bound*

$$c_k(1 + |\xi|)^{-\alpha} \leq \hat{k}(\xi) \tag{6}$$

for some $\alpha > d$ and constants $c_k > 0$.

Note that the kernels in (3) satisfy the assumption with $\alpha = 2s$. All the results until Section 3.4 require only this assumption and not the specific form of the kernels in (3). Next, we state a standard assumption on the regularity of the data distribution.

Assumption 2 *Let $\Omega \subset \mathbb{R}^d$ be a bounded open Lipschitz domain. The measure ρ has support $\bar{\Omega}$ and a bounded density $c_\rho \leq \rho(x) \leq C_\rho$.*

The last assumption is about the data generation.

Assumption 3 *The training data $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ consists of i.i.d. points x_i distributed according to ρ and*

$$y_i = f^*(x_i) + \varepsilon_i \tag{7}$$

where $f^* \in C_c^\infty$ is a smooth function that does not vanish identically on Ω and ε_i is i.i.d. Gaussian noise with variance σ^2.

2.2. Main results

We now state our main result.

Theorem 1 *Assume training data $\mathcal{D}$ with $|\mathcal{D}| = n$ satisfies Assumption 3 where the measure ρ satisfies Assumption 2. Let k be a kernel as in (3) with $d/2 < s < 3d/4$. Then, with probability $1 - O(1/n)$, the kernel interpolants f_γ of $\mathcal{D}$ for the kernel k_γ with bandwidth $\gamma < 1$ satisfy*

$$\mathbb{E}_\rho(f_\gamma(x) - f^*(x))^2 \geq c > 0 \tag{8}$$

uniformly in γ where c is a constant depending on everything except the sample size n and the bandwidth γ (i.e., d, s, Ω, c_ρ, C_ρ, f^*, and σ).

Let us informally state the special case for the Laplace kernel as a separate corollary.

Corollary 2 *Kernel interpolation with the Laplace kernel and adaptive bandwidth is not consistent as $n \to \infty$ in $d > 2$.*

Remark 3 *The condition $\gamma < 1$ could be removed but would require minor additional technicalities. The restriction on s could probably be loosened to $s < d$. but this condition is of more fundamental nature. While we conjecture that learning with smoother kernels in fixed dimension is not consistent, the proof given in this paper and also the proof from Rakhlin and Zhai (2019) do not extend to $s \geq d$. The reason is, loosely speaking, that the RKHS norm of the interpolant is determined by the quantity $\min_{i \neq j} |x_i - x_j|$ for $s > d$ while it is governed by the average value $n^{-1} \sum_i \min_{j \neq i} |x_i - x_j|$ for $s < d$. See also Remark 13 below.*

2.3. Proof overview

Now we give an overview of the approach used to prove the theorem. As already explained in Rakhlin and Zhai (2019) there are two different failure modes of kernel interpolation as also shown in Figure 1 in their paper. For bandwidths $\gamma > n^{1/d}$, i.e., the typical distance between neighbouring points, kernel interpolation will be smooth on scale $n^{-1/d}$ and we will make an error of order $(n^{-1/d})^d = n^{-1}$ in the ball $B(x_i, n^{1/d})$ around each data-point. So the total error is bounded from below by a constant. On the other hand, for $\gamma \ll n^{-1/d}$, the minimum norm interpolant will be a sum of little hat functions around each data-point and the interpolant will be very small away from the data points. As in Rakhlin and Zhai (2019) we state the result for the different regimes as separate propositions.

Proposition 4 *Under the same assumption as Theorem 1 there is for every $A > 0$ a constant $c > 0$ such that with probability $1 - O(1/n)$ the kernel interpolants f_γ with bandwidth $1 \geq \gamma > An^{-1/d}$ satisfy*

$$\mathbb{E}_\rho(f_\gamma(x) - f^*(x))^2 \geq c > 0. \tag{9}$$

The constant c depends on everything except the sample size n and bandwidth γ.

Proposition 5 *Under the same assumption as Theorem 1 there is a constant $B > 0$ and a constant $c > 0$ such that with probability $1 - O(1/n)$ the kernel interpolants f_γ with bandwidth $Bn^{-1/d} > \gamma$ satisfy*

$$\mathbb{E}_\rho(f_\gamma(x) - f^*(x))^2 \geq c > 0. \tag{10}$$

The constant c depends on everything except the sample size n and bandwidth γ.

Clearly, the two propositions together imply the main result. Next, we give an informal overview of the proof strategy. The general strategy in both regimes is to bound the norm of the interpolant f_γ. For small γ this norm bound will imply that f_γ has small $L2$-norm which allows us to prove Proposition 5. On the other hand, for $\gamma > An^{-1/d}$ the control of the RKHS norm allows us to conclude that the interpolant is Hoelder continuous around most data points which will imply Proposition 4.

We remark that the RKHS norm of the kernel interpolant is given by

$$\|f_\gamma\|^2_{\mathcal{H}_\gamma} = \langle y, K_\gamma(X,X)^{-1}y\rangle \tag{11}$$

where $K_\gamma(X,X)_{i,j} = k_\gamma(x_i,x_j) \in \mathbb{R}^{n\times n}$ denotes the kernel matrix and $y = (y_1,\ldots,y_n)^\top \in \mathbb{R}^n$ the vector of outputs. Thus we need to find lower bounds on the spectrum of the kernel matrix to prove the result. A substantial part of the proof only works with the inputs x_i and does not use the outputs y_i. To clarify this, it is convenient to introduce the notation $\mathcal{P} = \{x_1,\ldots,x_n\}$ for the data points without their labels. There are four main ingredients.

Lower bound on kernel matrices for radial basis functions. The radial basis function community has developed several lower bounds for the spectrum of kernel matrices which are used to show approximation results. Those are of the form

$$K_\gamma(X,X) \geq h(\delta) \tag{12}$$

where h is a function and $\delta = \min|x_i - x_j|$ denotes the minimal distance between two points. This bound is not directly useful when applied to the full kernel matrix but we show that we can find good subsets where this bound is useful.

Geometric arguments. We show that we can find large subsets $\mathcal{P}' \subset \mathcal{P}$ so that the bound $\min_{x,y\in\mathcal{P}'}|x-y| > \delta$ holds for varying δ. The construction of these sets will be based on a martingale argument. Applying (12) to such subsets we obtain lower bounds for subblocks of $K_\gamma(X,X)$.

Spectral arguments. Combining the previous two arguments provides us with lower bounds on (almost) the entire spectrum of $K_\gamma(X,X)$ using varying values of δ. Now the relation (11) and the lower bound on the eigenvalues of $K_\gamma(X,X)$ allow us to control the RKHS norm of the minimum norm interpolant of the noise with high probability using concentration bounds.

Functional analytic arguments. We bound the RKHS norm of the interpolant of f^* from above for small values of γ using a truncation argument. Together with the norm-bounds of the noise interpolant we conclude that $\|f_\gamma\|_{\mathcal{H}} \to 0$ as $\gamma \to 0$ which allows us to finish the proof of Proposition 5.

Since we restrict our attention to kernels whose RKHS is a Sobolev space (potentially with non-standard weights for the derivative terms due to the bandwidth) and we can bound the Sobolev norm of the interpolant we can apply standard results from functional analysis and the theory of Sobolev spaces to also prove Proposition 4. In particular, Sobolev embedding implies that $H^s \to C^{k-d/2}$, i.e., the interpolant will be Hoelder continuous. We apply this locally around the data points (where we make an error of order 1) and patch these estimates using truncation.

Let us finally clarify the relation to the proof for the Laplace kernel in odd dimensions d.

Relation to Rakhlin and Zhai (2019). While we use a similar structure in our proof, e.g., we separate the cases in Proposition 4 and 5, there are differences in several key aspects. Their proof relies on a strong control of $\sum r_i^{-1}$ where $r_i = \min_{j\neq i} |x_i - x_j|$. This is based on a careful analysis of the expectation and covariance of r_i^{-1}. Then they construct an explicit interpolating function of the dataset using cut-off functions, i.e.,

$$f = \sum y_i \eta \left(\frac{x - x_i}{r_i} \right) \tag{13}$$

and control its Sobolev norm. In contrast, our use of the radial basis function lower bound requires us only to find well separated subsets and not to control the minimal distances r_i. Therefore our geometric arguments are much shorter and simpler than the arguments in Appendix A in Rakhlin and Zhai (2019) which involve substantial amount of computations. Their argument bounding the RKHS norm of interpolations as in (13) does not extend to $s \notin \mathbb{N}$ because then the RKHS norm is not local and a control of the distance to the closest data point is not sufficient. Our bound based on the spectrum of the kernel matrix is robust to non-locality. Our functional analytic arguments are simplified and more general and rely essentially only on Sobolev embeddings.

3. Proof

In this section we collect the key steps of the proof of the main result.

3.1. The radial basis function part

We now state a lower bound on the spectrum of kernel matrices from the theory of radial basis function. There those estimates are used to control approximation properties of kernel interpolation and the lower bound is typically in terms of the minimal distance between any two points. We use the seminal lower bound from Schaback (1995).

Theorem 6 (Theorem 3.1 in Schaback (1995)) *Let $\mathcal{P} = \{x_1, \ldots, x_m\} \subset \Omega$ be a set of points such that $\min_{i\neq j} |x_i - x_j| \geq \delta$. Suppose the kernel k satisfies the lower bound from Assumption 1. Then the kernel matrix $K_\gamma(X, X) = k_\gamma(x_i - x_j)_{1\leq i,j\leq m}$ satisfies*

$$\lambda_{\min}(K_\gamma(X, X)) \geq c_1 \gamma^{-d} \min\left(\gamma^{-(\alpha-d)} \delta^{\alpha-d}, 1\right) \tag{14}$$

where $\lambda_{\min}$ denotes the smallest eigenvalue of a matrix and c_1 is a constant depending on c_k and α from Assumption 1 and d.

Sketch of the proof. As the proof is simple and instructive and is seemingly not too well known in kernel theory we sketch it here. For translation invariant kernels we can introduce the Fourier transform to express

$$v^\top K(X, X) v = (2\pi)^{-1} \sum_{i,j} v_i v_j \int \hat{k}(\xi) e^{i\xi(x_i - x_j)} \,\mathrm{d}\xi = (2\pi)^{-1} \int \hat{k}(\xi) \left| \sum_i v_i v_j e^{i\xi x_i} \right|^2 \mathrm{d}\xi. \tag{15}$$

Thus, when replacing k by a kernel k' such that $\hat{k}(\xi) \geq \hat{k'}(\xi)$ for all ξ, then $K(X, X) \geq K'(X, X)$ where $K(X, X) = k(x_i, x_j)_{1\leq i,j\leq n}$ denotes the kernel matrix for any dataset $\{x_1, \ldots, x_n\}$ and similarly for K'. Now we choose a kernel k' with compact support in $(0, \delta)$ and scale it such that $\hat{k} \geq \hat{k}'$. It is essentially sufficient to use a smooth bump function for k', its Fourier transform will then decay super-polynomially. As k' has support in $(0, \delta)$ the kernel matrix K' is diagonal because $|x_i - x_j| > \delta$. Now it is trivial to lower bound its spectrum and this ends the proof. Surprisingly, this already gives sharp bounds. A full proof can be found in Section C.1.

3.2. The geometric part

In this section we show some geometric results on the distribution of the data points, in particular on the distribution of their distances. The main result of this section is Lemma 7 below. There we show that with high probability it is possible to remove a small fraction of the points from the dataset such that all pairwise distances between the remaining points are of order $O(n^{-1/d})$. Later, we will apply the bound from Theorem 6 to such subsets. We denote by $\omega_d = |B(1, 0)|$ the volume of the unit ball in dimension d.

Lemma 7 *Let ρ satisfy Assumption 2. Let $\mathcal{P} = \{x_1, \ldots, x_n\}$ be a set of i.i.d. points distributed according to ρ. Then for $\kappa \in (0, 1)$ there is with probability $1 - e^{-\frac{3\kappa n}{16}}$ a subset $\mathcal{P}' \subset \mathcal{P}$ such that $|\mathcal{P}'| \geq (1 - \kappa)n$ and $\min_{x \neq y} |x - y| \geq \delta$ for $x, y \in \mathcal{P}'$ where*

$$\delta = n^{-\frac{1}{d}} \left(\frac{\kappa}{2C_\rho \omega_d} \right)^{\frac{1}{d}}. \tag{16}$$

A difficulty in deriving this type of geometric arguments is that the distance to the nearest point is not an independent variable for different points in the dataset. This was one difficulty faced in Rakhlin and Zhai (2019). Since we only need weaker geometric results about the existence of suitable subsets we can resort to an argument where we select points depending only on the previous data points. This gives us a martingale and we can use concentration results for martingales which are almost as strong as concentration results for i.i.d. variables. The full proof can be found in Section C.2.1.

We also need a bound on the minimal distance between any two points with high probability.

Lemma 8 *Let ρ satisfy Assumption 2. Let $\mathcal{P} = \{x_1, \ldots, x_n\}$ be a set of i.i.d. points distributed according to ρ. Then with probability $1 - n^{-1}$ the bound*

$$\min_{i \neq j} |x_i - x_j| \geq (C_\rho n^3 \omega_d)^{-1/d} \tag{17}$$

holds.

The simple proof can be found in Section C.2.2. Lemmas 7 and 8 are used to upper bound the RKHS norm of the minimum norm interpolant and this requires good control for the entire dataset. To then lower bound the error of the interpolant we need to show that there is a reasonably large subset of the data with good properties. We assume ρ and Ω satisfy Assumption 2. Let $c_\Omega > 0$ be a constant such that

$$\mathbb{P}_\rho(\mathbf{1}(\operatorname{dist}(x, \partial\Omega) < c_\Omega) < \frac{1}{10}. \tag{18}$$

Such a constant exists since Ω has a Lipschitz boundary. We define

$$\delta_{\min} = n^{-1/d} \left(\frac{1}{10 C_\rho \omega_d} \right)^{1/d}. \tag{19}$$

Lemma 9 *Let $1 > \nu > 0$. Then there is a constant $\Theta >$ depending on d, ν and C_ρ such that the following holds. Let $\mathcal{P} = \{x_1, \ldots, x_n\}$ be a set of i.i.d. points whose distribution satisfies Assumption 2. With probability at least $1 - e^{-n/200}$ there is a good subset $\mathcal{P}' \subset \mathcal{P}$ with the following properties. For $x \in \mathcal{P}'$ we have $\operatorname{dist}(x, \partial\Omega) \geq C_\Omega$, $|x - y| > \delta_{\min}$ for $x \neq y \in \mathcal{P}'$, $|\mathcal{P}'| \geq |\mathcal{P}|/2$, and for all $x \in \mathcal{P}'$*

$$\sum_{y \in \mathcal{P}' \setminus \{x\}} |x - y|^{-d-2\nu} \leq \Theta \delta_{\min}^{-2\nu} n. \tag{20}$$

Remark 10 *The purpose of this lemma is to construct a good subset of the dataset such that all points are well separated from each other and the boundary so that the errors made around those points accumulate gracefully. The condition (20) might appear a bit surprising and enforces a more global control of the point distances. It is not needed for integer values of s where norms are local but for fractional s we need this condition to control the effect of the non-locality.*

The proof of this result is similar to the proof of Lemma 7 and can be found in Section C.2.3.

3.3. The spectral part

Next, we show how the results in Theorem 6 on the lower bound on a block of the kernel matrix can be used to show lower bounds on parts of the spectrum of the complete kernel matrix using Lemma 7. We denote the eigenvalues of $K_\gamma(X, X)$ by μ_m and assume they are ordered in decreasing order. The key argument of this section appears in Proposition 11 below. The remaining results are mostly algebra.

Proposition 11 *Let ρ satisfy Assumption 2 and k satisfy Assumption 1. Let $\mathcal{P} = \{x_1, \ldots, x_n\}$ be a set of i.i.d. points distributed according to ρ. Then for $\kappa > 0$ with probability $1 - e^{-\frac{3\kappa n}{16}}$ the eigenvalue μ_m of $K_\gamma(X, X)$ with $m = (1 - \kappa)n$ satisfies for all γ*

$$\mu_m \geq c_2 \gamma^{-d} \min\left(\frac{\kappa^{(\alpha-d)/d}}{(n^{1/d}\gamma)^{\alpha-d}}, 1 \right) \tag{21}$$

where c_2 is a constant depending on d, C_ρ, c_k and α.

Proof Using Lemma 7 we find with probability $1 - e^{-\frac{3\kappa n}{16}}$ a set $\mathcal{P}' \subset \mathcal{P}$ with $|\mathcal{P}'| \geq (1 - \kappa)n$ such that the pairwise distances of points in $\mathcal{P}'$ are at least δ where

$$\delta = n^{-\frac{1}{d}} \left(\frac{\kappa}{2 C_\rho \omega_d} \right)^{\frac{1}{d}}. \tag{22}$$

Denote with X' the data matrix of the data points in $\mathcal{P}'$. Using Theorem 6 we conclude that

$$\lambda_{\min}(K_\gamma(X',X')) \geq c_1\gamma^{-d}\min\left(\gamma^{-(\alpha-d)}\delta^{\alpha-d},1\right) \geq c_2\gamma^{-d}\min\left(\frac{\kappa^{(\alpha-d)/d}}{(n^{1/d}\gamma)^{\alpha-d}},1\right) \tag{23}$$

where $c_2 = c_2(d, C_\rho, \alpha, c_k)$ is a constant. In particular, all eigenvalues of $K_\gamma(X',X')$ can be lower bounded by the right-hand side of the last display. Note that $K_\gamma(X',X')$ is a submatrix of $K_\gamma(X,X)$ and using the Courant-Fischer-Weyl min-max principle and $|\mathcal{P}'| \geq (1-\kappa)n$ we conclude. ■

Applying the last result for varying values of κ we can control the entire spectrum of $K_\gamma(X,X)$.

Theorem 12 *Let ρ satisfy assumption 2 and k satisfy Assumption 1. Let $\mathcal{P} = \{x_1,\ldots,x_n\}$ be a set of i.i.d. points distributed according to ρ. With probability at least $1-2/n$ the following bound holds for the spectrum of $K_\gamma(X,X)$ for all $\gamma > 0$*

$$\mu_m^{-1} \leq \begin{cases} c_3\gamma^d\left(\frac{(n^{2/d}\gamma)^{\alpha-d}}{(n-m)^{(\alpha-d)/d}}+1\right) & \text{for } m < n-32\ln(n) \\ c_3\gamma^d\left(\gamma^{(\alpha-d)}n^{3(\alpha-d)/d}+1\right) & \text{for } m > n-32\ln(n) \end{cases} \tag{24}$$

where c_3 is a constant depending on α, c_k, C_ρ and d.

Remark 13 *From this result it becomes apparent that the nature of the spectrum changes for $\alpha > 2d$ because then $x^{-(\alpha-d)/d}$ is not integrable at 0 and thus the smallest eigenvalues dominate the trace while for $\alpha < 2d$ the bulk dominates. This becomes clearer in the proof of Corollary 14.*

The proof of this Theorem can be found in Section C.3.1. For future reference we introduce the event

$$E_{\text{spec}} = \{\text{The conclusion of Theorem 12. holds for all } \gamma > 0\}. \tag{25}$$

We have shown that $\mathbb{P}(E_{\text{spec}}) \geq 1-2/n$. As a consequence of Theorem 6 we can deduce a bound on the trace of the inverse kernel matrix. This is (up to σ^2) the expectation of the squared RKHS-norm of the minimum norm interpolant of the noise.

Corollary 14 *Let k satisfy Assumption 1 and ρ satisfy Assumption 2. Let $\mathcal{P} = \{x_1,\ldots,x_n\}$ be a set of i.i.d. points distributed according to ρ. Assume that $2\alpha < 3d$. On E_{spec} the following bounds hold*

$$\operatorname{Tr} K_\gamma(X,X)^{-1} \leq c_4(n\gamma^d + (n\gamma^d)^{\alpha/d}) \tag{26}$$

$$\left(\operatorname{Tr} K_\gamma(X,X)^{-2}\right)^{1/2} \leq c_5\left(\frac{(n\gamma^d)}{\sqrt{n}} + \frac{(\gamma^d n)^{\alpha/d}}{\sqrt{n}} + \sqrt{\ln(n)}(n^{1/d}\gamma)^\alpha n^{(2\alpha-3d)/d}\right) \tag{27}$$

for some constants $c_4, c_5 > 0$.

The proof of this corollary is a bit technical so we provide a short heuristic. Note that a typical eigenvalue satisfies the bound

$$\mu_m^{-1} \leq \gamma^d(1+(n^{1/d}\gamma)^{\alpha-d}). \tag{28}$$

So we indeed expect a trace of order $n\gamma^d(1+(n\gamma^d)^{(\alpha-d)/d})$. The actual proof shows that this is indeed true by controlling the tail eigenvalues. It can be found in Section C.3.2.

We now state the consequence for the norm of minimum norm interpolants of random noise.

Proposition 15 *Let $\varepsilon \in \mathbb{R}^n$ be a random vector with Gaussian distribution $N(0, \sigma^2 \mathrm{Id})$. Suppose that $\mathcal{P} = \{x_1, \ldots, x_n\}$ is a set of i.i.d. points distributed according to ρ where ρ satisfies Assumption 2. Denote the minimum norm interpolant defined in equation (2) of $\{(x_i, \varepsilon_i) : 1 \leq i \leq n\}$ for the kernel k_γ by $f_{\gamma,\varepsilon}$ for some γ. Assume that E_{spec} holds. Let $0 < \tau = (3d - 2\alpha)/d < 1$. Then there are constants $c_6, c_7 > 0$ such that with probability at least $1 - e^{-c_6 n^\tau}$ the bound*

$$\|f_{\gamma,\varepsilon}\|_{\mathcal{H}_\gamma}^2 \leq c_7 \sigma^2 (n\gamma^d + (n\gamma^d)^{\alpha/d}) \tag{29}$$

holds conditioned on the event E_{spec}.

The proof relies on (11), Corollary 14, and concentration results for χ^2 variables. It can be found in Section C.3.3.

Since we want to show that even tuning the bandwidth to the realization of the noise cannot be consistent we need the following improvement of the previous result that shows that the same result holds uniformly over γ. This is not completely obvious because even tough the bounds on the spectrum hold for all γ simultaneously, the eigenvalues could a-priori align with the few large eigenvalues when adapting γ to the data.

Corollary 16 *Let $(x_i, \varepsilon_i)_{1 \leq i \leq n}$ be as in Proposition 15. Assume that E_{spec} occurs. Let $\tau = (3d - 2\alpha)/d > 0$. There is a constant $c > 0$ such that with probability $1 - e^{-cn^{\tau/2}}$ over the realization of the noise the bound*

$$\|f_{\gamma,\varepsilon}\|_{\mathcal{H}_\gamma}^2 \leq c_8 \sigma^2 (n\gamma^d + (n\gamma^d)^{\alpha/d}) \tag{30}$$

holds for all $\gamma < 1$.

The proof is based on monotonicity of the RKHS norm and a union bound over a grid combined with a simple argument for $\gamma \to 0$ where the kernel approaches the δ-kernel. The full proof is in Section C.3.4.

3.4. The functional analytic part

In this section we collect the functional analytic results required to prove the main theorem. It has two parts, in the first part we show that the minimum norm interpolant of the ground truth function has very small RKHS norm as $\gamma \to 0$. This allows to conclude the proof of Proposition 5. In the second part we show how upper bounds on the RKHS norm can be converted to lower bounds on the error. Then we can finish the proof of Proposition 4. While all results until here relied on Assumption 1 for the kernel k we make from now on the stronger assumption that k is given by (3).

Lemma 17 *Let $f^* \in C_c^\infty$ and not identically zero on Ω. Let $\{x_1, \ldots, x_n\}$ be any n points in Ω. Denote the minimum norm interpolant of $(x_i, f^*(x_i))$ by f_{γ,f^*}. Then there is a constant $c_9 > 0$ depending on C_ρ, f^*, d, and s such that*

$$\|f_{\gamma,f^*}\|_{\mathcal{H}_\gamma}^2 \leq \frac{1}{6C_\rho^2} \|f^*\|_{L^2(\rho)}^2 + c_9 (\gamma^2 n^{2/d} + \gamma^{2s} n^{2s/d}). \tag{31}$$

The proof of this Lemma is based on multiplying f^* by a cut-off function equal to 1 at the data points and vanishing close by. This gives an interpolating function which has a small RKHS norm when choosing the scaling properly. The proof can be found in Section C.4.1.

Now we can conclude the proof of Proposition 5.

Proof [Proof of Proposition 5] Suppose $\gamma < Bn^{-1/d}$. We assume that the conclusion of Corollary 16 holds which occurs with probability $1 - O(n^{-1})$. Then we can bound

$$\begin{aligned} \|f_\gamma\|_{\mathcal{H}}^2 &= \|f_{\gamma,\varepsilon} + f_{\gamma,f^*}\|_{\mathcal{H}}^2 \le 2\|f_{\gamma,\varepsilon}\|_{\mathcal{H}}^2 + 2\|f_{\gamma,f^*}\|_{\mathcal{H}}^2 \\ &\le c_8\sigma^2(n\gamma^d + (n\gamma^d)^{\alpha/d}) + \frac{1}{3C_\rho}\|f^*\|_{L^2(\rho)}^2 + c_9(\gamma^2 n^{2/d} + \gamma^{2s}n^{2s/d}) \\ &\le \frac{1}{3C_\rho}\|f^*\|_{L^2(\rho)}^2 + c_8\sigma^2(B^d + B^\alpha) + c_9(B^2 + B^{2s}) \le \frac{2}{3C_\rho}\|f^*\|_{L^2(\rho)}^2 \end{aligned} \tag{32}$$

for B sufficiently small. Then we obtain

$$\begin{aligned} \|f^* - f_\gamma\|_{L^2(\rho)} &\ge \|f^*\|_{L^2(\rho)} - \|f_\gamma\|_{L^2(\rho)} \ge \|f^*\|_{L^2(\rho)} - \sqrt{C_\rho}\|f_\gamma\|_2 \\ &\ge \|f^*\|_{L^2(\rho)} - \sqrt{C_\rho}\|f_\gamma\|_{\mathcal{H}_\gamma} \ge (1 - \sqrt{2/3})\|f^*\|_{L^2(\rho)}. \end{aligned} \tag{33}$$

Here we used $\|f_\gamma\|_2 \le \|f_\gamma\|_{H_\gamma}$ (which follows from $\hat{k}_\gamma(\xi) \le 1$ and (4)) and (32) in the last step. ■

Next we will use results from the theory of Sobolev spaces to address the second failure case. A very brief overview of Sobolev spaces can be found in Section B

The main idea is that interpolation of Sobolev spaces shows that any function u with $u(x_0) = 1$ cannot have the two norms $\|u\|_2$ and $\|u\|_{H^s}$ both arbitrarily small because, loosely speaking the smoother the function becomes the larger its total integral will be. After formalizing this result we will apply this argument to each data point and patch the result together to get a global estimate.

Lemma 18 *Let $u \in H^s(\mathbb{R}^d)$ with $s > d/2$ with $u(x_0) = 1$ for some $x_0 \in \mathbb{R}^d$. Then there is a constant $C = C(d,s) > 0$ such that the bound*

$$\|u\|_2^{-\frac{4s-2d}{d}} \le C\|u\|_{H^s}^2 \tag{34}$$

holds.

As a heuristic argument, we note that if the endpoint Sobolev embedding $H^{d/2} \to L^\infty$ would hold then the claim would follow from the interpolation inequality (63) with $0 \le d/2 < s$. Indeed, this would imply

$$\|u\|_\infty \le C\|u\|_{H^{d/2}} \le C\|u\|_2^{1-d/(2s)}\|u\|_{H^s}^{d/(2s)}. \tag{35}$$

We can fix this by using that for $s > d/2$ we embed into a Hoelder space and we can lower bound the L^2 norm of a function with $u(0) = 1$ in terms of the Hoelder constant. The full proof is in Section C.4.2

Next, we apply this lemma locally around each point and patch the local inequalities together. We fix a cut-off function $\eta \in C_c^\infty(\mathbb{R}^d)$ such that $\operatorname{supp}\eta \subset B(0,1)$ and $\eta|_{B(0,1/3)} = 1$. We use the shorthand $\eta_\delta(x) = \eta(x/\delta)$ in the following. To increase readability, we provide two versions of the following Proposition. Here we give a simpler one for $s \in \mathbb{N}$. The general case for fractional spaces that involves additional technicalities is moved to Section C.4.4.

Proposition 19 *Let $\mathcal{P}' = \{x_1, \ldots, x_m\} \subset \Omega$ be points such that $|x_i - x_j| \geq 3\delta$ for $i \neq j$ and $\mathrm{dist}(x_i, \partial\Omega) \geq \delta$ for some $\delta > 0$. Let $u \in H^s$ with $s > d/2$, $s \in \mathbb{N}$ be a function such that $|u(x_i)| \geq c_\ell$ for $1 \leq i \leq m$. Then there is a constant $c_{10} > 0$ depending on d, s, and c_ℓ such that*

$$1 \leq c_{10} m^{-\frac{2s}{d}} \delta^{-2s} \|u\|_{L^2(\Omega)}^{\frac{4s}{d}} + c_{10} m^{-\frac{2s}{d}} \|u\|_{L^2(\Omega)}^{\frac{4s-2d}{d}} \|D^s u\|_2^2. \tag{36}$$

The main idea of the proof is to consider the functions $u_i = u\eta_\delta(\cdot - x_i)$ where we truncate u around each data point. We can apply Lemma 18 to each of the u_i. Then we can patch those estimates using generalized mean inequalities and additivity of norms for disjoint sets (for $s \in \mathbb{N}$). The full proof is in Section C.4.3.

We can finally finish the proof of Theorem 1 by showing Proposition 4.

Proof [Proof of Proposition 4 for $s \in \mathbb{N}$] We consider the minimum norm interpolant $f_\gamma = f_{\gamma,\varepsilon} + f_{\gamma,f^*}$. We assume that Corollary 16 holds which occurs with probability $1 - O(n^{-1})$. We can bound

$$\|f_{\gamma,f^*}\|_{\mathcal{H}_\gamma} \leq \|f^*\|_{\mathcal{H}_\gamma} \leq C_{f^*}(1+\gamma^s) \leq 2C_{f^*}. \tag{37}$$

Here the first step follows from minimum norm interpolation and the second step follows from $\|f^*\|^2 \leq C(\|f^*\|_2^2 + \gamma^{2s}\|D^s f^*\|_2^2)$. We set $u = f_\gamma - f^*$ and infer using the last display and Corollary 16 that

$$\|u\|_{\mathcal{H}_\gamma}^2 = \|f_\gamma - f^*\|_{\mathcal{H}_\gamma}^2 \leq 3\left(\|f_{\gamma,\varepsilon}\|_{\mathcal{H}_\gamma}^2 + \|f_{\gamma,f^*}\|_{\mathcal{H}_\gamma}^2 + \|f^*\|_{\mathcal{H}_\gamma}^2\right) \leq C(n\gamma^d + (n\gamma^d)^{2s/d} + 1). \tag{38}$$

This implies

$$\|D^s u\|_2^2 \leq \gamma^{-2s}\|u\|_{\mathcal{H}_\gamma}^2 \leq C\gamma^{-2s}(n\gamma^d + (n\gamma^d)^{2s/d} + 1). \tag{39}$$

We assume that $\mathcal{P}'$ is a set as in Lemma 9. By choosing $c_\sigma > 0$ sufficiently small we find a further subset still denoted by $\mathcal{P}'$ such that $|\varepsilon_i| > c_\sigma$ on $\mathcal{P}'$ and $\mathcal{P}' \geq n/4$. We can now apply Proposition 19 with $\mathcal{P}'$ and u (note that $|u(x_i)| = |f^*(x_i) + \varepsilon_i - f^*(x_i)| \geq c_\sigma$ for $x_i \in \mathcal{P}'$) and we get

$$1 \leq c_{10} 4^{-2s/d} n^{-2s/d} 3^{-2s} \delta_{\min}^{-2s} \|u\|_{L^2(\Omega)}^{\frac{4s}{d}} + c_{10} 4^{-2s/d} n^{-2s/d} \|u\|_{L^2(\Omega)}^{\frac{4s-2d}{d}} \|D^s u\|_2^2 \tag{40}$$

where $\delta_{\min} > cn^{-1/d}$ was defined in (19). From the last display we conclude that either

$$\frac{1}{2} \leq c_{10} 4^{-2s/d} 3^{-2s} n^{-2s/d} \delta_{\min}^{-2s} \|u\|_{L^2(\Omega)}^{\frac{4s}{d}} \leq C\|u\|_{L^2(\Omega)}^{\frac{4s}{d}} \tag{41}$$

or

$$\begin{aligned}
\frac{1}{2} &\leq c_{10} 2^{-2s/d} n^{-2s/d} \|u\|_{L^2(\Omega)}^{\frac{4s-2d}{d}} \|D^s u\|_2^2 \\
&\leq Cn^{-2s/d}\gamma^{-2s}(n\gamma^d + (n\gamma^d)^{2s/d} + 1)\|u\|_{L^2(\Omega)}^{\frac{4s-2d}{d}} \\
&\leq C((n^{1/d}\gamma)^{d-2s} + 1 + (n^{1/d}\gamma)^{-2s})\|u\|_{L^2(\Omega)}^{\frac{4s-2d}{d}} \\
&\leq C(A^{d-2s} + A^{-2s} + 1)\|u\|_{L^2(\Omega)}^{\frac{4s-2d}{d}}
\end{aligned} \tag{42}$$

if $\gamma > An^{-1/d}$. In either case we conclude (since $4s - 2d > 0$) that

$$\|u\|_{L^2(\rho)}^2 \geq c_\rho \|u\|_2^2 \geq C_A. \tag{43}$$

This ends the proof for $s \in \mathbb{N}$. ■

Having proved Proposition 5 and 4 the main result Theorem 1 follows directly.

4. Conclusion

We investigated minimum norm interpolation in fixed dimension for kernels whose RKHS is a Sobolev space. We showed that the minimum norm interpolant is not consistent as $n \to \infty$ extending results from Rakhlin and Zhai (2019). Our results provide a small step towards a better understanding of the relation between the problem parameters (dimensions, sample size, data distribution, method class) that govern success and failures of interpolating estimators which is overall still a poorly understood question. We gave evidence that tools from approximation theory can be useful to tackle these questions for kernel minimum-norm interpolation, in particular for regimes where more classical approaches from the kernel literature fail. Natural questions that arise from our result is whether they can be extended to slowly growing dimension d and how they can be generalized to smoother kernels. It is known that the RKHS norm grows exponentially with sample size in this case Belkin et al. (2018), thus likely entirely different techniques are required in this case.

References

R.A. Adams and J.J.F. Fournier. *Sobolev Spaces*. ISSN. Elsevier Science, 2003. ISBN 9780080541297. URL https://books.google.de/books?id=R5A65Koh-EoC.

Shahar Azulay, Edward Moroshko, Mor Shpigel Nacson, Blake E. Woodworth, Nathan Srebro, Amir Globerson, and Daniel Soudry. On the implicit bias of initialization shape: Beyond infinitesimal mirror descent. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 468–477. PMLR, 2021. URL http://proceedings.mlr.press/v139/azulay21a.html.

Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020. ISSN 0027-8424. doi: 10.1073/pnas.1907378117. URL https://www.pnas.org/content/117/48/30063.

Peter L. Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: a statistical viewpoint. *Acta Numerica*, 30:87–201, 2021. doi: 10.1017/S0962492921000027.

Mikhail Belkin. Approximation beats concentration? an approximation view on inference with smooth radial kernels. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018*, volume 75 of *Proceedings of Machine Learning Research*, pages 1348–1361. PMLR, 2018. URL http://proceedings.mlr.press/v75/belkin18a.html.

Mikhail Belkin, Siyuan Ma, and Soumik Mandal. To understand deep learning we need to understand kernel learning. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 540–548. PMLR, 2018. URL http://proceedings.mlr.press/v80/belkin18a.html.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019a. ISSN 0027-8424. doi: 10.1073/pnas.1903070116. URL https://www.pnas.org/content/116/32/15849.

Mikhail Belkin, Alexander Rakhlin, and Alexandre B. Tsybakov. Does data interpolation contradict statistical optimality? In Kamalika Chaudhuri and Masashi Sugiyama, editors, *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, volume 89 of *Proceedings of Machine Learning Research*, pages 1611–1619. PMLR, 2019b. URL http://proceedings.mlr.press/v89/belkin19a.html.

Martin D. Buhmann. *Radial Basis Functions: Theory and Implementations*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2003. doi: 10.1017/CBO9780511543241.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Found. Comput. Math.*, 7(3):331–368, 2007. doi: 10.1007/s10208-006-0196-8. URL https://doi.org/10.1007/s10208-006-0196-8.

Fan Chung and Linyuan Lu. Concentration inequalities and martingale inequalities: a survey. *Internet Mathematics*, 3(1):79 – 127, 2006. doi: im/1175266369. URL https://doi.org/.

Eleonora Di Nezza, Giampiero Palatucci, and Enrico Valdinoci. Hitchhiker's guide to the fractional sobolev spaces. *Bulletin des Sciences Mathématiques*, 136(5):521–573, 2012. ISSN 0007-4497. doi: https://doi.org/10.1016/j.bulsci.2011.12.004. URL https://www.sciencedirect.com/science/article/pii/S0007449711001254.

Simon S. Du, Xiyu Zhai, Barnabás Póczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=S1eK3i09YQ.

Amnon Geifman, Abhay Kumar Yadav, Yoni Kasten, Meirav Galun, David W. Jacobs, and Ronen Basri. On the similarity between the laplace and neural tangent kernels. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1006ff12c465532f8c574aeaa4461b16-Abstract.html.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *CoRR*, abs/1903.08560, 2019. URL http://arxiv.org/abs/1903.08560.

Arthur Jacot, Clément Hongler, and Franck Gabriel. Neural tangent kernel: Convergence and generalization in neural networks. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 8580–8589, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/5a4be1fa34e62bb8a6ec6b91d2462f5a-Abstract.html.

Noureddine El Karoui. The spectrum of kernel random matrices. *The Annals of Statistics*, 38(1):1 – 50, 2010. doi: 10.1214/08-AOS648. URL https://doi.org/10.1214/08-AOS648.

Frederic Koehler, Lijia Zhou, Danica J. Sutherland, and Nathan Srebro. Uniform convergence of interpolators: Gaussian width, norm bounds and benign overfitting. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=FyOhThdDBM.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000. ISSN 00905364. URL http://www.jstor.org/stable/2674095.

G. Leoni. *A First Course in Sobolev Spaces*. Graduate studies in mathematics. American Mathematical Society, 2009. ISBN 9780821847688. URL https://books.google.de/books?id=eE2ZAwAAQBAJ.

Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel "Ridgeless" regression can generalize. *The Annals of Statistics*, 48(3):1329 – 1347, 2020. doi: 10.1214/19-AOS1849. URL https://doi.org/10.1214/19-AOS1849.

Tengyuan Liang, Alexander Rakhlin, and Xiyu Zhai. On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. In Jacob D. Abernethy and Shivani Agarwal, editors, *Conference on Learning Theory, COLT 2020, 9-12 July 2020, Virtual Event [Graz, Austria]*, volume 125 of *Proceedings of Machine Learning Research*, pages 2683–2711. PMLR, 2020. URL http://proceedings.mlr.press/v125/liang20a.html.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, n/a(n/a). doi: https://doi.org/10.1002/cpa.22008. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.22008.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Learning with invariances in random features and kernel models. In Mikhail Belkin and Samory Kpotufe, editors, *Conference on Learning Theory, COLT 2021, 15-19 August 2021, Boulder, Colorado, USA*, volume 134 of *Proceedings of Machine Learning Research*, pages 3351–3418. PMLR, 2021. URL http://proceedings.mlr.press/v134/mei21a.html.

Vaishnavh Nagarajan and J. Zico Kolter. Uniform convergence may be unable to explain generalization in deep learning. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 11611–11622, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/05e97c207235d63ceb1db43c60db7bbb-Abstract.html.

Alexander Rakhlin and Xiyu Zhai. Consistency of interpolation with laplace kernels is a high-dimensional phenomenon. In Alina Beygelzimer and Daniel Hsu, editors, *Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA*, volume 99 of *Proceedings of Machine Learning Research*, pages 2595–2623. PMLR, 2019. URL http://proceedings.mlr.press/v99/rakhlin19a.html.

Lorenzo Rosasco, Mikhail Belkin, and Ernesto De Vito. On learning with integral operators. *J. Mach. Learn. Res.*, 11:905–934, 2010. URL https://dl.acm.org/citation.cfm?id=1756036.

Robert Schaback. Error estimates and condition numbers for radial basis function interpolation. *Advances in Computational Mathematics*, 3(3):251–264, 1995. doi: 10.1007/BF02432002. URL https://doi.org/10.1007/BF02432002.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *J. Mach. Learn. Res.*, 19:70:1–70:57, 2018. URL http://jmlr.org/papers/v19/18-188.html.

Holger Wendland. *Scattered Data Approximation*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2004. doi: 10.1017/CBO9780511617539.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=Sy8gdB9xx.

Appendix A. Concentration inequalities

We collect some concentration inequalities.

A.1. Martingale concentration bounds

We state two well known martingale concentration inequalities. Azuma's inequality is the extension of Hoeffding's inequality from independent random variables to martingales.

Theorem 20 (Azuma's inequality, Theorem 5.2 in Chung and Lu (2006)) *Let M_i be a discrete martingale with $M_0 = 0$ that satisfies for all $i \geq 0$*

$$|M_i - M_{i-1}| \leq c_i. \tag{44}$$

Then

$$\mathbb{P}(|M_n - \mathbb{E}(M_n)| \geq \lambda) \leq 2e^{-\frac{\lambda^2}{2\sum_{i=1}^n c_i^2}}. \tag{45}$$

The second bound is the martingale version of Bernstein type inequalities that is stronger when the conditional variance of the martingale is small.

Theorem 21 (Freedman's inequality, Theorem 6.1 in Chung and Lu (2006)) *Let M_i be a discrete martingale adapted to the filtration $\mathcal{F}_n$ with $M_0 = 0$ that satisfies for all $i \geq 0$*

$$|M_{i+1} - M_i| \leq K, \tag{46}$$

$$\mathrm{Var}(M_i|\mathcal{F}_{i-1}) \leq \sigma_i^2. \tag{47}$$

Then

$$\mathbb{P}(M_n - \mathbb{E}(M_n) \geq \lambda) \leq e^{-\frac{\lambda^2}{2\sum_{i=1}^n \sigma_i^2 + K\lambda/3}}. \tag{48}$$

A.2. Chi-squared concentration lemma

The following Lemma can be used to bound the tail of a sum of weighted χ_1^2-variables.

Lemma 22 (Lemma 1 in Laurent and Massart (2000)) *Let X_i be i.i.d. standard normal variables and $a_i > 0$ for $1 \leq i \leq n$. Denote the ℓ^2 and ℓ^2 norm as usual by $|a|_2$ and $|a|_\infty$. Consider*

$$Z = \sum_{i=1}^n a_i(X_i^2 - 1). \tag{49}$$

Then, for any $x > 0$

$$\mathbb{P}(Z \geq 2|a|_2\sqrt{x} + 2|a|_\infty x) \leq e^{-x}. \tag{50}$$

Appendix B. Functional analytic tools

In this Appendix we collect results from functional analysis and Sobolev spaces that are used in the proof. A completely self-contained introduction is beyond the scope of this paper and we refer to Leoni (2009); Adams and Fournier (2003) for an introduction and to Di Nezza et al. (2012) for the extension to fractional Sobolev spaces.

B.1. Sobolev spaces

We only consider the Hilbert space case $p = 2$ on $\mathbb{R}^d$. In this case $H^s(\mathbb{R}^d)$ can be defined as the set of all functions in $L^2(\mathbb{R}^d)$ such that the following norm is finite

$$\|u\|_{H^s}^2 = \int (1 + |\xi|^2)^s |\hat{u}|^2 \, \mathrm{d}\xi. \tag{51}$$

In Fourier space it is easy to see that using $(1 + |\xi|)^{2s}$ or $1 + |\xi|^{2s}$ as Fourier multipliers leads to equivalent norms and we switch between without explicitly noting that. In particular, we use the equivalent norm

$$\|u\|_{H^s}^2 = \|u\|_2^2 + \|D^s u\|_2^2 \tag{52}$$

where $D^s u = \mathcal{F}^{-1}(|\xi|^s \hat{u})$. Equivalently one can define the fractional derivative for $\nu \in (0, 1)$ by

$$\|D^\nu u\|_2^2 = \int \int \frac{(u(x) - u(y))^2}{|x - y|^{d+2\nu}} \, \mathrm{d}x \mathrm{d}y. \tag{53}$$

And for $s = k + \nu$ with $\nu \in (0, 1)$ we can define

$$\|D^s u\|_2 = \|D^\nu (D^k u)\|_2^2. \tag{54}$$

We only need (53) in Lemma 23 below. Otherwise we can resort to the Fourier picture and use that the same Sobolev embedding results as in the integer case hold (see Theorem 6.5 in Di Nezza et al. (2012)).

B.2. Some basic results for fractional derivatives

We state two well-known results on fractional derivatives.

Lemma 23 *Let* $\nu \in (0, 1)$ *and* $d > 1$. *Let* $u, v \in H^\nu(\mathbb{R}^d)$ *such that* $\mathrm{supp}(u) = U$ *and* $\mathrm{supp}(v) = V$ *with* $r = \mathrm{dist}(U, V) > r > 0$. *Then*

$$\langle D^\nu u, D^\nu v \rangle \leq C \frac{|U|^{\frac{d+2\nu}{2d}} |V|^{\frac{d+2\nu}{2d}}}{r^{d+2\nu}} \|u\|_{H^\nu} \|v\|_{H^\nu}. \tag{55}$$

Proof By definition

$$\langle D^\nu u, D^\nu v \rangle = \int_U \int_V \frac{(u(x) - u(y))(v(x) - v(y))}{|x - y|^{d+2\nu}} \, \mathrm{d}x \, \mathrm{d}y. \tag{56}$$

For the numerator not to vanish we need to have $x \in U$, $y \in V$ or vice versa. We now set $p = 2d/(d - 2\nu)$ and $q = 2d/(d + 2\nu)$ such that $p^{-1} + q^{-1} = 1$ and by Sobolev embedding $\|u\|_p \leq \|u\|_{H^\nu}$. Then we can bound using Hoelder's inequality

$$\begin{aligned} \langle D^\nu u, D^\nu v \rangle &\leq 2 \int \int \frac{|u(x)| \cdot |v(y)|}{r^{d+2\nu}} \, \mathrm{d}x \, \mathrm{d}y. \\ &\leq 2 \|u\|_p \|\mathbf{1}_U\|_q \|\mathbf{1}_V\|_q \|v\|_p \\ &\leq C \|u\|_{H^s} \|v\|_{H^s} |U|^{\frac{d+2\nu}{2d}} |V|^{\frac{d+2\nu}{2d}}. \end{aligned} \tag{57}$$

■

The following standard result gives a simplified 'product rule' for Sobolev functions.

Lemma 24 *Let $u, v \in H^s(\mathbb{R}^d)$ with $s > d/2$. Then the following bound holds*

$$\|D^s(uv)\|_2^2 \le 2^{2s}(\|uD^s v\|_2^2 + \|vD^s u\|_2^2). \tag{58}$$

Proof We observe

$$\begin{aligned}\|D^s(uv)\|_2^2 &= \int |\xi|^{2s} |\widehat{uv}|^2 \,\mathrm{d}\xi = \int |\xi|^{2s} |\hat{u} * \hat{v}|^2 \,\mathrm{d}\xi \\ &= \int |\xi|^{2s} \left| \int \hat{v}(\zeta)\hat{u}(\xi - \zeta)\,\mathrm{d}\zeta \right|^2 \mathrm{d}\xi.\end{aligned} \tag{59}$$

Noting that $|\xi|^{2s} \le (|\xi - \zeta| + |\zeta|)^{2s} \le 2^{2s}(|\xi - \zeta|^{2s} + |\zeta|^{2s})$ for all $\xi, \zeta \in \mathbb{R}^d$ we can continue to estimate

$$\begin{aligned}\|D^s(uv)\|_2^2 \le +2^{2s} &\int \left| \int \hat{u}(\zeta)|\zeta|^s \hat{v}(\xi - \zeta)\,\mathrm{d}\zeta \right|^2 \mathrm{d}\xi \\ &+ 2^{2s} \int \left| \int \hat{u}(\zeta)\hat{v}(\xi - \zeta)|\xi - \zeta|^s \,\mathrm{d}\zeta \right|^2 \mathrm{d}\xi \\ &= 2^{2s} \left\| (|\cdot|^s \hat{u}) * \hat{v} \right\|_2^2 + 2^{2s} \left\| \hat{u} * (|\cdot|^s \hat{v}) \right\|_2^2 \\ &= 2^{2s} \left\| (D^s u)v \right\|_2^2 + 2^{2s} \left\| uD^s v \right\|_2^2 .\end{aligned} \tag{60}$$

■

B.3. An interpolation result for Sobolev spaces

We need an interpolation inequality to interpolate between Sobolev spaces of different smoothness parameters. It is a special case of general interpolation results which we state here for completeness.

Lemma 25 *For $0 < \beta < 1$ and $u \in H^s(\mathbb{R}^d)$ with $s \ge d/2 + \beta$ the bound*

$$\|u\|_{C^\beta} \le C \|u\|_2^{\frac{2s-d-2\beta}{2s}} \|u\|_{H^s}^{\frac{d+2\beta}{2s}} \tag{61}$$

holds.

Proof This is a special case of the general Gagliardo-Nirenberg interpolation theorem for fractional Sobolev spaces (see Leoni (2009) for the non-fractional setting). The proof of this special case is much simpler than the general case. Sobolev embedding theory for fractional Sobolev spaces implies that there is a constant such that for $t = \beta + d/2$

$$\|u\|_{C^\beta} \le \|u\|_{H^t}. \tag{62}$$

The spaces H^s satisfy for $0 \le s_1 \le s_m \le s_2$ the interpolation inequality

$$\|u\|_{H^{s_m}} \le \|u\|_{H^{s_1}}^{\frac{s_2 - s_m}{s_2 - s_1}} \|u\|_{H_{s_2}}^{\frac{s_m - s_1}{s_2 - s_1}}. \tag{63}$$

While this is also contained in the Gagliardo-Nirenberg inequality it can also be shown using the Fourier transform and a Hoelder estimate (just like the interpolation of L^p norms). The claim follows by setting $s_m = t$, $s_1 = 0$, and $s_2 = s$ ■

Appendix C. Omitted proofs

Here we collect the full proofs of all results in the paper. We use the same structure as in Section 3 to facilitate orientation.

C.1. Proofs for the radial basis function arguments

For completeness, we give a streamlined version of the proof of Theorem 6 (Theorem 3.1 in Schaback (1995)) which also takes care of the dependence on γ which is not considered there.

C.1.1. Proof of Theorem 6

Proof The main idea is to lower bound the Fourier transform of the kernel function by the Fourier transform of a compactly supported distribution.

Recall that by assumption $\hat{k}(\xi) \geq c(1+|\xi|)^{-\alpha}$ for all ξ and note that $\hat{k}_\gamma(\xi) = \hat{k}(\gamma\xi)$ which implies

$$\hat{k}_\gamma(\xi) = \hat{k}(\gamma\xi) \geq (1+\gamma|\xi|)^{-\alpha} \tag{64}$$

Let $\varphi \in C_c^\infty(\mathbb{R})$ with $\operatorname{supp}\varphi \subset (-\frac{1}{2}, \frac{1}{2})$ be a symmetric bump function such that $\|\varphi\|_2 = 1$. Since smooth functions with compact support are in the Schwartz class there are constants C_k such that

$$|\hat{\varphi}(\xi)|^2 \leq \frac{C_k}{(1+|x|)^k}. \tag{65}$$

Consider the function

$$\chi_\delta(x) = M(\varphi * \varphi)(x/\delta) \tag{66}$$

where $M > 0$ is chosen later. Note that $\operatorname{supp}\chi_\delta \subset (-\delta, \delta)$. We can now bound

$$0 \leq \hat{\chi}_\delta(\xi) = M\delta^d|\hat{\varphi}(\xi\delta)|^2 \leq \frac{MC_\alpha\delta^d}{(1+\delta|\xi|)^\alpha} \leq \frac{MC_\alpha\delta^d}{c_k}\frac{(1+\gamma|\xi|)^\alpha}{(1+\delta|\xi|)^\alpha}\hat{k}_\gamma(\xi). \tag{67}$$

For $\gamma|\xi| \leq 1$ we can bound

$$\hat{\chi}_\delta(\xi) \leq \frac{MC_\alpha}{c_k}2^\alpha\delta^d\hat{k}_\gamma(\xi) \tag{68}$$

and similarly for $\gamma|\xi| \geq 1$

$$\hat{\chi}_\delta(\xi) \leq \frac{MC_\alpha\delta^d}{c_k}\frac{2^\alpha(\gamma|\xi|)^\alpha}{(1+\delta|\xi|)^\alpha}\hat{k}_\gamma(\xi) \tag{69}$$

so we conclude

$$\hat{\chi}_\delta(\xi) \leq \frac{MC_\alpha 2^\alpha\gamma^d}{c_k}\max\left((\gamma\delta^{-1})^{\alpha-d}, (\gamma\delta^{-1})^{-d}\right)\hat{k}_\gamma(\xi). \tag{70}$$

Now, for $M = c_k \left(C_\alpha 2^\alpha \gamma^d \max\left((\gamma\delta^{-1})^{\alpha-d}, (\gamma\delta^{-1})^{-d}\right)\right)^{-1}$ we infer that

$$\hat{\chi}_\delta(|\xi|) \leq \hat{k}_\gamma(\xi). \tag{71}$$

After this preparation the proof is straightforward. We write

$$\begin{aligned} vK_\gamma(X,X)v &= \sum_{i,j} v_i v_j k_\gamma(x_i - x_j) \\ &= (2\pi)^{-1} \int \sum_{i,j} v_i v_j \hat{k}_\gamma(\xi) e^{i\xi(x_i - x_j)}\, \mathrm{d}\xi \\ &= (2\pi)^{-1} \int \left| \sum_i v_i e^{i\xi x_i} \right|^2 \hat{k}_\gamma(\xi)\, \mathrm{d}\xi. \end{aligned} \tag{72}$$

Using the lower bound (71) we can further estimate

$$vK_\gamma(X,X)v \geq (2\pi)^{-1} \int \left| \sum_i v_i e^{i\xi x_i} \right|^2 \hat{\chi}_\delta(\xi)\, \mathrm{d}\xi = \sum_{i,j} v_i v_j \chi_\delta(x_i - x_j) = \chi_\delta(0)|v|^2 \tag{73}$$

where we used $|x_i - x_j| \geq \delta$ and $\operatorname{supp} \chi_\delta \subset (-\delta, \delta)$ in the last step. Then

$$\chi_\delta(0) = M(\phi * \phi)(0) = M$$

implies $K_\gamma(X,X) \geq M$. Note that for $\gamma < \delta$ we can apply the same argument with $\delta' = \gamma$ to conclude that

$$K_\gamma(X,X) \geq c_k \left(C_\alpha 2^\alpha \gamma^d \max\left(\gamma^{\alpha-d}\delta^{-(\alpha-d)}, 1\right)\right)^{-1} \tag{74}$$

Setting $c_1 = c_k/(C_\alpha 2^\alpha)$ finishes the proof. ■

C.2. Proofs for the geometric arguments

Here we collect the omitted proofs from Section 3.2.

C.2.1. PROOF OF LEMMA 7

Proof We define the filtration $\mathcal{F}_i = \sigma(x_1, \ldots, x_i)$. We consider the events

$$E_i = \{x_i \in \bigcup_{j<i} B(x_j, \delta)\} \tag{75}$$

and define $X_i = \mathbf{1}_{E_i}$. Note that X_i is measurable with respect to $\mathcal{F}_i$ and therefore $S_i = \sum_{j\leq i} X_j$ defines an adapted stochastic process. We denote the Doob decomposition of S_i by $S_i = M_i + A_i$ where $S_0 = M_0 = A_0 = 0$ and M_i is a martingale and A_i predictable (w.r.t. $\mathcal{F}_i$). Our goal is to use concentration inequalities to bound S_n with high probability, which will quickly imply the claim. We observe that

$$P(E_i|\mathcal{F}_{i-1}) = \mathbb{P}_\rho\left(x_i \in \bigcup_{j<i} B(x_j,\delta) |\, x_1, \ldots, x_{i-1}\right) \leq C_\rho \left| \bigcup_{j\neq i} B(x_j, \delta) \right| \leq C_\rho n \omega_d \delta^d \tag{76}$$

and this implies

$$A_i - A_{i-1} = \mathbb{E}(X_i|\mathcal{F}_{i-1}) = \mathbb{P}(E_i|\,\mathcal{F}_{i-1}) \leq C_\rho n\omega_d\delta^d. \tag{77}$$

Now fix the value of δ to be as in (16) which implies

$$A_n = \sum_{i=1}^n A_i - A_{i-1} \leq C_\rho n\omega_d\delta^d n = \frac{\kappa n}{2}. \tag{78}$$

Note that $|M_i - M_{i-1}| \leq 1$ and thus using Azuma's inequality (for reference the statement is Theorem 20 in the appendix) implies

$$\mathbb{P}\left(M_n \geq \frac{\kappa n}{2}\right) \leq e^{-\frac{\kappa^2 n^2}{8n}}. \tag{79}$$

As we want to apply this to small values κ of order almost n^{-1} we need a stronger concentration result. We note that

$$\mathrm{Var}(M_i|\mathcal{F}_{i-1}) = \mathrm{Var}(X_i|\mathcal{F}_{i-1}) \leq \mathbb{E}(X_i^2|\mathcal{F}_i) = \mathbb{E}(X_i|\mathcal{F}_i) \leq C_\rho n\omega_d\delta^d = \frac{\kappa}{2}. \tag{80}$$

Then the concentration inequality stated in Theorem 21 implies

$$\mathbb{P}\left(M_n \geq \frac{\kappa n}{2}\right) \leq e^{-\frac{\kappa^2 n^2}{8(n\kappa/2 + n\kappa/6)}} = e^{-\frac{3\kappa n}{16}}. \tag{81}$$

We conclude that

$$\mathbb{P}\left(S_n \geq \kappa n\right) \leq \mathbb{P}\left(A_n \geq \frac{\kappa n}{2}\right) + \mathbb{P}\left(M_n \geq \frac{\kappa n}{2}\right) \leq e^{-\frac{3\kappa n}{16}}. \tag{82}$$

We consider the set $\mathcal{P}' = \{x_i \in \mathcal{P}|\, X_i = 0\}$ and claim that $|x_i - x_j| \geq \delta$ for $x_i, x_j \in \mathcal{P}'$. Indeed, assume $j < i$. Then $X_i = 0$ implies that $x_i \notin \bigcup_{j<i} B(x_j, \delta)$ and thus $|x_i - x_j| \geq \delta$. If $S_n = \sum X_i \leq \kappa n$ we conclude that $|\mathcal{P}'| \geq n(1-\kappa)$ and together with (82) this ends the proof. ■

C.2.2. Proof of Lemma 8

Proof As before, we can bound the probability that a fixed point x_i is close to any other point by

$$\mathbb{P}(\min_{j:\, j\neq i} |x_i - x_j| \leq \delta) \leq C_\rho \left|\bigcup_{j\neq i} B(x_j, \delta)\right| \leq C_\rho n\omega_d\delta^d. \tag{83}$$

The union bound implies

$$\mathbb{P}(\min_{j,i:\, j\neq i} |x_i - x_j| \leq \delta) \leq C_\rho n^2\omega_d\delta^d. \tag{84}$$

Setting $\delta = (C_\rho n^3\omega_d)^{-1/d}$ finishes the proof. ■

C.2.3. PROOF OF LEMMA 9

Proof We argue similar to the proof of Lemma 7. Since the structure of the proof is similar, we do not spell out all the details of the martingale construction. We assume $x_i \in \Omega$ for $i < j$ are given and x_j follows the distribution ρ and control the probability of certain events uniform in x_i. Note that by assumption on ρ ($\mathbf{1}$ denotes the indicator function)

$$\mathbb{P}_\rho(\mathbf{1}(x_j \in \bigcup_{i<j} B(x_i, \delta_{\min}))) \leq C_\rho \omega_d \delta_{\min}^d n \leq \frac{1}{10}. \tag{85}$$

Also for all $y \in \Omega$

$$\begin{aligned} \mathbb{E}_\rho\left((x-y)^{-d-2s}\mathbf{1}(|x-y| \geq \delta_{\min}\right) &= \int_{B(y,\delta_{\min})^c} |x-y|^{-d-2\nu}\rho(x)\,\mathrm{d}x \\ &\leq C_\rho \int_{B(x,\delta_{\min})^c} |x-y|^{-d-2\nu}\,\mathrm{d}y \leq \frac{\Theta}{200}\delta_{\min}^{-2\nu} \end{aligned} \tag{86}$$

for some $\Theta > 0$ depending only on C_ρ, d, and ν. We conclude that for each j

$$\mathbb{P}_\rho\left(\sum_{i<j} |x_i - x_j|^{-d-2\nu}\mathbf{1}(|x_i - x_j| > \delta_{\min}) > \frac{\Theta\delta_{\min}^{-2\nu} n}{20}\right) \leq \frac{1}{10}. \tag{87}$$

Also $\mathbb{P}_\rho(\operatorname{dist}(x_j, \partial\Omega) < c_\Omega) < \frac{1}{10}$. The union bound implies that

$$\begin{aligned} &\mathbb{P}_\rho\left(x_j \notin \bigcup_{i<j} B(x_i, \delta_{\min}), \sum_{i<j} |x_i - x_j|^{-d-2\nu}\mathbf{1}_{|x_i - x_j| > \delta_{\min}} < \frac{\Theta\delta_{\min}^{-2\nu} n}{20}, \operatorname{dist}(x_j, \partial\Omega) > c_\Omega\right) \\ &= \mathbb{P}_\rho\left(x_j \notin \bigcup_{i<j} B(x_i, \delta_{\min}), \sum_{i<j} |x_i - x_j|^{-d-2\nu} < \frac{\Theta\delta_{\min}^{-2\nu} n}{20}, \operatorname{dist}(x_j, \partial\Omega) > C_\Omega\right) > \frac{7}{10}. \end{aligned} \tag{88}$$

Using Azuma's bound from Theorem 20 just like in the proof of Lemma 7 we conclude that with probability at least $1 - 2e^{-|\mathcal{P}|/200}$ we can find a subset $\mathcal{P}_s = \{z_1, \ldots, z_m\}$ with $|\mathcal{P}_s| \geq \frac{6n}{10}$ and $\min_{i \neq j} |z_i - z_j| \geq \delta$, $\operatorname{dist}(x_j, \partial\Omega) > c_\Omega$ and

$$\sum_{i \neq j} |z_i - z_j|^{-d-2\nu} < \frac{\Theta\delta_{\min}^{-2\nu} n^2}{10}. \tag{89}$$

Using Markov's inequality we see that there are at most $n/10$ points in $\mathcal{P}_s$ such that

$$\sum_{z' \in \mathcal{P}_s, z \neq z'} |z - z'|^{-d-2\nu} > \Theta\delta_{\min}^{-2\nu} n \tag{90}$$

Removing those points we find a subset $\mathcal{P}' \subset \mathcal{P}_s$ such that $|\mathcal{P}'| \geq |\mathcal{P}|/2$ and for each $z \in \mathcal{P}'$

$$\sum_{z' \in \mathcal{P}_s, z \neq z'} |z - z'|^{-d-2\nu} \leq \Theta\delta_{\min}^{-2\nu} n. \tag{91}$$

■

C.3. Proofs for the spectral arguments

Here we collect the omitted proofs from Section 3.3.

C.3.1. Proof of Theorem 6

Proof Let $m_i = n - 2^{-i}n$ with $1 \le i \le i_0$ where i_0 is chosen such that $16\ln(n)/n \le 2^{-i_0} \le 32\ln(n)/n$. We set $\kappa_i = (n-m_i)/n$ such that $m_i = (1-\kappa_i)n$. Using Proposition 11 we conclude that with probability $1 - e^{-3(n-m_i)/16} = 1 - e^{-3\cdot 2^{-i}n}$ the bound

$$\mu_{m_i} \ge c_2\gamma^{-d}\min\left(\frac{(n-m_i)^{(\alpha-d)/d}}{(n^{2/d}\gamma)^{\alpha-d}}, 1\right) \tag{92}$$

holds for all γ. Using a union bound we conclude that with probability $1 - 2e^{-3(n-m_{i_0})/16} \ge 1 - e^{-3\ln(n)} = 1 - n^{-3}$ the bound

$$\mu_m \ge c_2 2^{-(\alpha-d)/d}\gamma^{-d}\min\left(\frac{(n-m)^{(\alpha-d)/d}}{(n^{2/d}\gamma)^{\alpha-d}}, 1\right) \tag{93}$$

holds for all $m \le n - 2^{-i_0}n$ where we used that if $m_i \ge m$ and i is chosen minimal we have $n - m_i \ge (n-m)/2$. For $m \ge n - 32\ln(n)$ we use Lemma 8 (together with Theorem 6) to conclude that with probability $1 - n^{-1}$ all eigenvalues satisfy

$$\mu_m \ge c_1(C_\rho\omega_d)^{-1/d}\gamma^{-d}\min\left(\gamma^{-(\alpha-d)}n^{-3(\alpha-d)/d}, 1\right). \tag{94}$$

Passing to the inverse ends the proof. Note that all results are uniform in γ as they only rely on geometric conditions. ■

C.3.2. Proof of Corollary 11

Proof We have

$$\operatorname{Tr} K_\gamma(X,X)^{-1} = \sum_m \mu_m^{-1} \tag{95}$$

so we can use Theorem 12 to control the sum.

We bound

$$\begin{aligned}\operatorname{Tr} K_\gamma(X,X)^{-1} &\le c_3\gamma^d\left(\sum_{m=1}^n 1 + \sum_{m=1}^n \frac{(n^{2/d}\gamma)^{\alpha-d}}{(n-m)^{(\alpha-d)/d}} + 32\ln(n)\gamma^{\alpha-d}n^{3\frac{\alpha-d}{d}}\right)\\ &\le c_3\gamma^d n c_3 C n^{-2}(n^{2/d}\gamma)^\alpha \sum_{i=1}^{n-m_0}\frac{1}{i^{(\alpha-d)/d}} + c_3\ln(n)(\gamma n^{1/d})^\alpha n^{\frac{2\alpha-3d}{d}}.\end{aligned} \tag{96}$$

The last term can be bounded by

$$c_3\ln(n)(\gamma n^{1/d})^\alpha n^{\frac{2\alpha-3d}{d}} \le c_3(\gamma n^{1/d})^\alpha \tag{97}$$

under the assumption $2\alpha < 3d$. We control the sum in (96) using the following approximation for $\alpha < 2d$

$$\sum_{i=1}^{n} \frac{1}{i^{(\alpha-d)/d}} \leq \int_0^n x^{-(\alpha-d)/d}\,\mathrm{d}x \leq Cn^{-(\alpha-2d)/d} = Cn^{(2d-\alpha)/d}. \tag{98}$$

This implies

$$\begin{aligned} \operatorname{Tr} K_\gamma(X,X)^{-1} &\leq C\left(n\gamma^d + n^{-2}(n^{2/d}\gamma)^\alpha n^{(2d-\alpha)/d} + (\gamma n^{1/d})^\alpha\right) \\ &\leq C\left(n\gamma^d + (n^{1/d}\gamma)^\alpha + (\gamma n^{1/d})^\alpha\right) \\ &\leq C(n\gamma^d + (n\gamma^d)^{\alpha/d}). \end{aligned} \tag{99}$$

Exactly the same argument can be used to bound

$$\begin{aligned} \operatorname{Tr} K_\gamma(X,X)^{-2} &= \sum \mu_m^{-2} \\ &\leq Cn\gamma^{2d} + Cn^{-2}(n^{2/d}\gamma)^\alpha n^{(3d-2\alpha)/d} + C\ln(n)(n^{1/d}\gamma)^{2\alpha} n^{(4\alpha-6d)/d} \\ &\leq C\frac{(n\gamma^d)^2}{n} + C\frac{(\gamma^d n)^{2\alpha/d}}{n} + C\ln(n)(n^{1/d}\gamma)^{2\alpha} n^{(4\alpha-6d)/d} \end{aligned} \tag{100}$$

where we require $(2\alpha - 2d)/d \leq 1$, i.e. $2\alpha < 3d$ to control the integral $\int x^{-2(\alpha-d)/d}\,\mathrm{d}x$ that appears in the bound. Now the bound (27) follows using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b > 0$. ■

C.3.3. PROOF OF PROPOSITION 15

Proof Denote the eigendecomposition of $K_\gamma(X,X)$ by $K_\gamma(X,X) = \sum \mu_m v_m \otimes v_m$ where v_m are normalized. Then we have

$$\|f_{\gamma,\varepsilon}\|_{\mathcal{H}}^2 = \langle \varepsilon, K_\gamma(X,X)^{-1}\varepsilon\rangle = \sum \mu_m^{-1}(\varepsilon v_m)^2. \tag{101}$$

As v_m constitutes an orthonormal basis we conclude that the variables εv_m are independent Gaussian variables with variance σ^2. Thus,

$$\|f_{\gamma,\varepsilon}\|_{\mathcal{H}}^2 \overset{\mathcal{D}}{=} \sum_m \frac{\sigma^2}{\mu_m}(Z_i^2 - 1) + \sigma^2 \operatorname{Tr} K_\gamma(X,X)^{-1} \tag{102}$$

where Z_i are standard normal variables. Now we use a result for the concentration of sums of rescaled χ^2 variables stated in Lemma 22 where $a_m = \sigma^2\mu_m^{-1}$. Note that using (27) in Corollary 14 we get

$$|a|_2 \leq \sigma^2\sqrt{\operatorname{Tr} K_\gamma(X,X)^{-2}} \leq c_5\sigma^2\left(\frac{(n\gamma^d)}{\sqrt{n}} + \frac{(\gamma^d n)^{\alpha/d}}{\sqrt{n}} + \sqrt{\ln(n)}(n^{1/d}\gamma)^\alpha n^{(2\alpha-3d)/d}\right). \tag{103}$$

From Theorem 12 we moreover find that

$$|a|_\infty \leq \sigma^2 c_3 (n\gamma^d)^{\alpha/d} n^{(2\alpha-3d)/d}. \tag{104}$$

We now set $\tau = (3d - 2\alpha)/d > 0$ (note that $\alpha > d$ implies $\tau < 1$) and

$$c_6 = \min\left(\sup_{z>1} \frac{\ln(z)}{z^\tau}, 1\right). \tag{105}$$

Let $x = c_6 n^\tau$. Then we get from (103)

$$\begin{aligned} |a|_2\sqrt{x} &\le c_5\sigma^2\left(n\gamma^d + (\gamma^d n)^{\alpha/d}\right) n^{\tau/2-1/2} + c_5\sigma^2\sqrt{\ln(n)}(n^{1/d}\gamma)^\alpha n^{-\tau}\sup_{z>1}\frac{\sqrt{\ln(z)}}{z^{\tau/2}} n^{\tau/2} \\ &\le 2c_5\sigma^2\left(n\gamma^d + (\gamma^d n)^{\alpha/d}\right). \end{aligned} \tag{106}$$

Similarly we obtain

$$|a|_\infty x \le \sigma^2 c_3^{-1}(n\gamma^d)^{\alpha/d} n^{-\tau} n^\tau \le \sigma^2 c_3 (n\gamma^d)^{\alpha/d}. \tag{107}$$

Applying the concentration bound in Lemma 22 and (102) we get

$$\begin{aligned} &\mathbb{P}\left(\|f_{\gamma,\varepsilon}\|_{\mathcal{H}}^2 \ge 2\sqrt{x}|a|_2 + 2x|a|_\infty + \operatorname{Tr} K_\gamma(X,X)^{-1}\right) \\ &\qquad \le \mathbb{P}\left(\|f_{\gamma,\varepsilon}\|_{\mathcal{H}}^2 \ge c_7\sigma^2(n\gamma^d + (\gamma^d n)^{\alpha/d}\right) \le e^{-c_6 n^\tau} \end{aligned} \tag{108}$$

where $c_7 = c_4 + 2c_3 + 4c_5$. ■

C.3.4. PROOF OF COROLLARY 16

Proof We first show the result for very small γ. In this case the kernel matrix becomes diagonally dominant and it is easy to lower bound it. By assumption we have $|x - y| \ge cn^{-3/d}$ for $x \ne y \in \mathcal{P}$. Moreover, Assumption 1 implies that $k(x) \le e^{-Cx}$. This implies that for $\gamma < \gamma_{\min} < (Ccn^{-1-3/d})^{1/d}$ we have

$$\sum_{i\ne j} k_\gamma(x_i - x_j) \le n\gamma^{-d}e^{-n} < \gamma^{-d}/2. \tag{109}$$

This implies $K_\gamma(X,X) \ge \gamma^{-d}/2$ and thus

$$\langle \varepsilon, K_\gamma(X,X)^{-1}\varepsilon\rangle \le 2|\varepsilon|^2\gamma^d. \tag{110}$$

Using the concentration bound in Lemma 22 we conclude that

$$\mathbb{P}(|\varepsilon|^2 \ge 4\sigma^2 n) \le e^{-n}. \tag{111}$$

This shows the result uniformly for $\gamma < \gamma_{\min}$. To conclude we note that $\|f\|_{\mathcal{H}_\gamma} \le \|f\|_{\mathcal{H}_{\gamma'}}$ for $\gamma < \gamma'$ and the fact that $f_{\gamma,\varepsilon}$ has minimal RKHS norm of all interpolating functions implies that it is sufficient to establish the result for a grid $\gamma_i = 2^{-i}$ and $i = 0, \ldots, i_0$ where i_0 satisfies $\gamma_{\min}/2 < 2^{-i_0} < \gamma_{\min}$. Applying Proposition 15 to each γ_i we can conclude with the union bound that with probability at least $1 - e^{-n} - \log_2(\gamma_{\min})e^{-c_6 n^\tau} \ge 1 - e^{-cn^{\tau/2}}$ the bound

$$\|f_{\gamma,\varepsilon}\|_{\mathcal{H}_\gamma}^2 \ge (c_7 2^\alpha + 4)\sigma^2(n\gamma^d + (n\gamma^d)^{\alpha/d}) \tag{112}$$

holds for all $\gamma < 1$. With $c_8 = c_7 2^\alpha + 4$ the proof is finished. ■

C.4. Proofs of the functional analytic arguments

Here we collect the omitted proofs from Section 3.4 and the extension to $s \notin \mathbb{N}$.

C.4.1. PROOF OF LEMMA 17

Proof Let $a > 0$ be such that

$$\int_U |f^*(x)|^2 \, \mathrm{d}x \leq \frac{1}{6C_\rho} \int |f^*(x)|^2 \rho(x) \, \mathrm{d}x \tag{113}$$

holds for all measurable sets $U \subset \mathbb{R}^n$ such that $|U| \leq a$. Note that a only depends on f^* and C_ρ. Pick

$$\delta = \frac{1}{4} \left(\frac{a}{\omega^d n} \right)^{1/d} \tag{114}$$

We now consider the function η defined by

$$\eta(x) = \begin{cases} 1 & \text{if } \min_i |x - x_i| < 2\delta \\ 3\delta - \min_i |x - x_i| & \text{if } 2\delta \leq \min_i |x - x_i| < 3\delta \\ 0 & \text{if } 3\delta \leq \min_i |x - x_i| \end{cases} \tag{115}$$

which is a continuous cut-off function around the data points. Let φ_δ be a smooth mollifier with support in $B(0, \delta)$ with $\int \varphi_\delta = 1$. Then $\tilde{\eta} = \varphi_\delta * \eta$ satisfies $\tilde{\eta} \equiv 1$ on $\bigcup_i B(x_i, \delta)$, $\tilde{\eta} \equiv 0$ on $(\bigcup_i B(x_i, 5\delta))^{\mathrm{c}}$ and $\|D^s \tilde{\eta}\|_\infty \leq C\delta^{-s}$. We conclude that $\tilde{\eta} f^*$ interpolates the points $(x_i, f^*(x_i))$ and we control its RKHS norm as follows

$$\begin{aligned} \|\tilde{\eta} f^*\|_{\mathcal{H}_\gamma}^2 &\leq \|\tilde{\eta} f^*\|_2^2 + \sum_{i=1}^{\lfloor s \rfloor} c_i \gamma^i \|D^i(\tilde{\eta} f^*)\|_2^2 + c_s \gamma^s \|D^s(\tilde{\eta} f^*)\|_2^2 \\ &= A_1 + A_2 + A_3. \end{aligned} \tag{116}$$

We estimate A_2 as follows

$$\begin{aligned} A_2 &\leq \sum_{i=1}^{\lfloor s \rfloor} c_i 2^i \gamma^{2i} (\|f^* D^i \tilde{\eta}\|_2^2 + \|\tilde{\eta} D^i f^*\|_2^2) \\ &\leq \sum_{i=1}^{\lfloor s \rfloor} c_i 2^{2i} \gamma^{2i} \left(\|D^i \tilde{\eta}\|_\infty^2 \|f^*\|_2^2 + \|D^i f^*\|_2^2 \right) \\ &\leq C_{f^*} \sum_{i=1}^{\lfloor s \rfloor} c_i 2^{2i} \gamma^{2i} (\delta^{-2i} + 1) \end{aligned} \tag{117}$$

where we used that f^* is smooth in the last step to bound all its norms. Similarly, we obtain for the last term

$$A_3 \leq C_{f^*} c_s 2^{2s} \gamma^{2s} (\delta^{-2s} + 1). \tag{118}$$

To bound the first term we note that

$$\left|\bigcup_i B(x_i, 4\delta)\right| \leq (4\delta)^d n\omega_d \leq a. \tag{119}$$

Thus we can estimate using (113) to obtain

$$A_1 \leq \int |f^*(x)|^2 \mathbf{1}\left(x \in \bigcup_i B(x_i, 4\delta)\right) \mathrm{d}x \leq \frac{1}{6C_p}\|f^*\|^2_{L^2(\rho)}. \tag{120}$$

Combining (117), (118), and (120) together with $\delta \geq cn^{-1/d}$ and the definition of the minimum norm interpolant we conclude that

$$\|f^2_{\gamma,f^*}\| \leq \|\tilde{\eta}f^*\|_{\mathcal{H}_\gamma} \leq \frac{1}{6C_\rho^2}\|f^*\|^2_{L^2(\rho)} + C\sum_{i=1}^{\lfloor s\rfloor} \gamma^{2i} n^{2i/d} + C\gamma^{2s} n^{2s/d}. \tag{121}$$

This ends the proof. ■

C.4.2. PROOF OF LEMMA 18

Proof The proof has two steps. We bound the Hoelder norm using an interpolation inequality and then show that the Hoelder norm lower bounds the L^2 norm. First, we bound the Hoelder norm of u. Let $0 < \beta < 1$ such that $d/2 + \beta \leq s$. Using the Sobolev interpolation result for fractional Sobolev spaces stated in Lemma 25 gives us

$$\|u\|_{C^\beta} \leq C\|u\|_2^{\frac{2s-d-2\beta}{2s}} \|u\|_{H^s}^{\frac{d+2\beta}{2s}}. \tag{122}$$

We now lower bound the L^2 norm of u in terms of its Hoelder constant. By definition

$$\frac{|u(x) - u(x_0)|}{|x - x_0|^\beta} \leq \|u\|_{C^\beta}. \tag{123}$$

This implies $u(x) \geq 1/2$ for x such that $|x - x_0| \leq (2\|u\|_{C^\beta})^{-1/\beta}$ and therefore

$$\|u\|_2^2 \geq 2^{-d}\omega_d(2\|u\|_{C^\beta})^{-d/\beta}. \tag{124}$$

This implies that there is a constant C' such that

$$\|u\|_2^{-2\beta/d} \leq C'\|u\|_{C^\beta} \leq C\|u\|_2^{\frac{2s-d-2\beta}{2s}} \|u\|_{H^s}^{\frac{d+2\beta}{2s}}. \tag{125}$$

We now take the last display to the power $d/(d + 2\beta)$. Note that

$$\frac{d}{d+2\beta}\left(\frac{2\beta}{d} + \frac{2s-d-2\beta}{2s}\right) = \frac{4s\beta}{2s(d+2\beta)} + \frac{2sd}{2s(d+2\beta)} - \frac{d}{2s} = 1 - \frac{d}{2s} \tag{126}$$

$$\frac{d}{d+2\beta}\frac{d+2\beta}{2s} = \frac{d}{2s} \tag{127}$$

implying that

$$1 \leq C\|u\|_2^{1-d/(2s)}\|u\|_{H^s}^{d/(2s)}. \tag{128}$$

This finishes the proof. ■

C.4.3. PROOF OF PROPOSITION 19

Proof We apply Lemma 18 to each of the functions $u_i = u\eta_\delta(\cdot - x_i)$ where we cut off u around each data point. This implies

$$\sum_{i:x_i\in\mathcal{P}'} \|u_i\|_2^{-\frac{4s-2d}{d}} \leq C \sum_{i:x_i\in\mathcal{P}'} \|u_i\|_{H^s}^2. \tag{129}$$

Since $s \in \mathbb{N}$ the H^s norm has a local expression and we conclude

$$\sum_{i:x_i\in\mathcal{P}'} \|u_i\|_{H^s}^2 = \left\| \sum_{i:x_i\in\mathcal{P}'} u_i \right\|_{H^s}^2 \tag{130}$$

because the functions u_i have disjoint support. For $s \notin \mathbb{N}$ we will need a more involved argument because the H^s norm is not local. We use the shorthand $\tilde{u} = \sum_{i:x_i\in\mathcal{P}'} u_i$ and $\tilde{\eta} = \sum_{i:x_i\in\mathcal{P}'} \eta_\delta(\cdot - x_i)$ so that $\tilde{u} = \tilde{\eta}u$. We now bound using Lemma 24

$$\begin{aligned} \|\tilde{u}\|_{H^s}^2 &= \|\tilde{u}\|_2^2 + \|D^s\tilde{u}\|_2^2 \\ &\leq \|\tilde{u}\|_2^2 + 2^{2s}\|uD^s\tilde{\eta}\|_2^2 + 2^{2s}\|\tilde{\eta}D^s u\|_2^2 \\ &\leq \|\tilde{u}\|_2^2 + 2^{2s}\|D^s\tilde{\eta}\|_\infty^2\|u\|_{L^2(\Omega)}^2 + 2^{2s}\|\tilde{\eta}\|_\infty^2\|D^s u\|_2^2 \end{aligned} \tag{131}$$

In the last step we used that by assumption $\mathrm{supp}(\tilde{\eta}) \subset \Omega$. This again relies on $s \in \mathbb{N}$ because otherwise D^s can increase the support. Now we can bound $\|D^s\tilde{\eta}\|_\infty \leq C\delta^{-s}$ and obtain

$$\|\tilde{u}\|_{H^s}^2 \leq C\delta^{-2s}\|u\|_{L^2(\Omega)}^2 + C\|D^s u\|_2^2. \tag{132}$$

We continue to lower bound the left-hand side of equation (129) using the generalized mean inequality which implies

$$\frac{\sum_{i:x_i\in\mathcal{P}'} \|u_i\|_2^2}{|\mathcal{P}'|} \geq \left(\frac{\sum_{i:x_i\in\mathcal{P}'} \left(\|u_i\|_2^2\right)^{-\frac{2s-d}{d}}}{|\mathcal{P}'|} \right)^{-\frac{d}{2s-d}}. \tag{133}$$

We infer that

$$\sum_{i:x_i\in\mathcal{P}'} \|u_i\|_2^{-\frac{4s-2d}{d}} \geq |\mathcal{P}'|^{\frac{2s}{d}} \left(\sum_{i:x_i\in\mathcal{P}'} \|u_i\|_2^2 \right)^{-\frac{2s-d}{d}} \geq m^{\frac{2s}{d}} \|u\|_{L^2(\Omega)}^{-\frac{4s-2d}{d}}. \tag{134}$$

Here we again used that u_i have disjoint support. Combining the estimates (129), (130), (132), and (134) we obtain the bound

$$1 \leq Cm^{-\frac{2s}{d}}\delta^{-2s}\|u\|_{L^2(\Omega)}^{\frac{4s}{d}} + Cm^{-\frac{2s}{d}}\|u\|_{L^2(\Omega)}^{\frac{4s-2d}{d}}\|D^s u\|_2^2. \tag{135}$$

■

C.4.4. EXTENSION OF THE RESULTS FROM SECTION 3.4 TO $s \notin \mathbb{N}$

In this section we discuss the extension of Proposition 19 to fractional Sobolev spaces. Instead of Proposition 19 we use the following extension to fractional spaces that takes into account the non-locality of the Sobolev norm.

Proposition 26 *Let $\mathcal{P}' = \{x_1, \ldots, x_m\} \subset \Omega$. Assume there are constants $c_\Omega > 0$, $\delta_{\min} > 0$, $\Theta > 0$ such that for $x \in \mathcal{P}'$ we have $\mathrm{dist}(x, \partial\Omega) \geq c_\Omega$, $|x - y| > \delta_{\min}$ for $x \neq y \in \mathcal{P}'$, and for all $x \in \mathcal{P}'$*

$$\sum_{y \in \mathcal{P}' \setminus \{x\}} |x - y|^{-d-2\nu} \leq \Theta' \tag{136}$$

Let $u \in H^s$ with $s > d/2$, $s \in \mathbb{N}$ be a function such that $|u(x_i)| \geq c_\ell$ for $1 \leq i \leq m$. Then there is a constant c_{11} depending on d, ν, c_Ω, and c_ℓ such that

$$1 \leq c_{11} m^{-\frac{2s}{d}} \Theta'^{2s/(d+2\nu)} \|u\|_{L^2(\Omega)}^{\frac{4s}{d}} + c_{11} m^{-\frac{2s}{d}} \|u\|_{L^2(\Omega)}^{\frac{4s-2d}{d}} (\|D^s u\|_2^2 + \|u\|_2^2). \tag{137}$$

Proof Let $0 < \delta < \delta_{\min}/4$ be a constant to be chosen later. As in the proof of Proposition 19 we consider the functions $u_i = u\eta_\delta(\cdot - x_i)$ where we cut off u around each data point. We now write $s = k + \nu$. To emulate the step (130) we bound using Lemma 23 (and $\mathrm{dist}(B(x_i, \delta), B(x_j, \delta)) > |x_i - x_j|/2$ as $|x_i - x_j| > \delta_{\min} \geq 4\delta$)

$$\begin{aligned}
\sum_{i \neq j} |\langle D^s u_i, D^s u_j \rangle| &\leq \sum_{i \neq j} C \|D^k u_i\|_{H^\nu} \|D^k u_j\|_{H^\nu} (\omega_d \delta^d)^{(d+2\nu)/d} \left(\frac{x_i - x_j}{2} \right)^{-(d+2\nu)} \\
&\leq \sum_{i \neq j} C (\|u_i\|_{H^s}^2 + \|u_j\|_{H^s}^2) \delta^{d+2\nu} |x_i - x_j|^{-(d+2\nu)} \\
&= 2C \sum_i \|u_i\|_{H^s}^2 \delta^{d+2\nu} \sum_{j \neq i} |x_i - x_j|^{-(d+2\nu)} \\
&\leq C \sum_i \|u_i\|_{H^s}^2 \delta^{d+2\nu} \Theta'.
\end{aligned} \tag{138}$$

We now fix $\delta = (2C\Theta')^{-1/(d+2\nu)}$ and conclude that

$$\sum_{i \neq j} |\langle D^s u_i, D^s u_j \rangle| \leq \frac{1}{2} \sum_i \|u_i\|_{H^s}^2. \tag{139}$$

This implies using that the u_i have disjoint support

$$\begin{aligned}
\|\sum_i u_i\|_{H^s}^2 &= \|\sum_i u_i\|_2^2 + \|D^s \sum_i u_i\|_2^2 \\
&\geq \sum_i \|u_i\|_2^2 + \sum_i \|D^s u_i\|_2^2 - \sum_{i \neq j} |\langle D^s u_i, D^s u_j \rangle| \\
&\geq \frac{1}{2} \sum_i \|u_i\|_{H^s}^2.
\end{aligned} \tag{140}$$

The same reasoning as in the proof of Proposition 19 (see equation (129) and (134)) implies then

$$m^{\frac{2s}{d}}\|u\|_{L^2(\Omega)}^{-\frac{4s-2d}{d}} \le C\|\sum u_i\|_{H^s}^2. \tag{141}$$

We use the shorthand $\tilde{u} = \sum_i u_i$ and $\tilde{\eta} = \sum_i \eta_\delta(\cdot - x_i)$ Moreover we consider an additional cut-off function $\chi \in C_c^\infty(\Omega)$ that satisfies $\chi(x) = 1$ if $\mathrm{dist}(x, \partial\Omega) > c_\Omega/2$ in particular χ is constant 1 on the support of each of the η_i and its derivatives can be bound with constants only depending on c_Ω. Then we get $\tilde{u} = \tilde{\eta}\chi u$. We now bound similarly to (131)

$$\begin{aligned}\|\tilde{u}\|_{H^s}^2 &= \|\tilde{u}\|_2^2 + \|D^s\tilde{u}\|_2^2 \\ &\le \|\tilde{u}\|_2^2 + 2^{2s}\|u\chi D^s\tilde{\eta}\|_2^2 + 2^{2s}\|\tilde{\eta}D^s(u\chi)\|_2^2 \\ &\le \|\tilde{u}\|_2^2 + 2^{2s}\|D^s\tilde{\eta}\|_\infty^2\|u\chi\|_2^2 + 2^{4s}\|\tilde{\eta}\|_\infty^2\|\chi\|_\infty^2\|D^s u\|_2^2 + 2^{4s}\|\tilde{\eta}\|_\infty^2\|D^s\chi\|_\infty^2\|u\|_2^2\end{aligned} \tag{142}$$

We can estimate $\|D^s\tilde{\eta}\|_\infty \le C\delta^{-s}$ and $\|D^s\tilde{\chi}\|_\infty \le C$ where the constant only depends on c_Ω and d. We get using that $\mathrm{supp}(\chi) \cup \mathrm{supp}(\tilde{\eta}) \subset \Omega$

$$\|\tilde{u}\|_{H^s}^2 \le C\delta^{-2s}\|u\|_{L^2(\Omega)}^2 + C\|D^s u\|_2^2 + C\|u\|_2^2. \tag{143}$$

Combining (141) and (143) we obtain

$$1 \le Cm^{-\frac{2s}{d}}\delta^{-2s}\|u\|_{L^2(\Omega)}^{\frac{4s}{d}} + Cm^{-\frac{2s}{d}}\|u\|_{L^2(\Omega)}^{\frac{4s-2d}{d}}(\|D^s u\|_2^2 + \|u\|_2^2). \tag{144}$$

Plugging in the definition of δ the proof is completed. ■

Finally, we indicate the necessary modifications in the proof of Proposition 4 for general s using Proposition 26.

Proof [Proof of Proposition 4 for arbitrary s] The proof is essentially the same as for $s \in \mathbb{N}$. The only difference is that we apply Proposition 26 and get instead of (40) the bound

$$1 \le c_{11}m^{-\frac{2s}{d}}\Theta'^{2s/(d+2\nu)}\|u\|_{L^2(\Omega)}^{\frac{4s}{d}} + c_{11}m^{-\frac{2s}{d}}\|u\|_{L^2(\Omega)}^{\frac{4s-2d}{d}}(\|D^s u\|_2^2 + \|u\|_2^2) \tag{145}$$

where $m \ge n/4$ and $\Theta' = \Theta\delta_{\min}^{-2\nu}n$ with Θ is the absolute constant from Lemma 9 that depends only on ν, d, and C_ρ. Using $\delta_{\min} \ge cn^{1/d}$ we obtain

$$\Theta'^{2s/(d+2\nu)} \le C\left(n^{(2\nu+d)/d}\right)^{2s/(d+2\nu)} \le Cn^{\frac{2s}{d}}. \tag{146}$$

The rest of the proof is the same. ■