Interpolation and Learning with Scale Dependent Kernels

Nicolò Pagliana [*1], Alessandro Rudi [†3], Ernesto De Vito[‡1], and Lorenzo Rosasco [§2,4,5]

[1]MaLGa, DIMA, Universita' degli Studi di Genova, Genova, Italy
[2]MaLGa, DIBRIS, Universita' degli Studi di Genova, Genova, Italy
[3]INRIA - Sierra project team, Paris, France
[4]Center for Brains, Minds and Machines, McGovern Institute, MIT, Cambridge, USA
[5]Istituto Italiano di Tecnologia - IIT, Genova, Italy

Abstract

We study the learning properties of nonparametric ridge-less least squares. In particular, we consider the common case of estimators defined by scale dependent kernels, and focus on the role of the scale. These estimators interpolate the data and the scale can be shown to control their stability through the condition number. Our analysis shows that there are different regimes depending on the interplay between the sample size, the dimension, and the smoothness of the problem. Indeed, when the sample size is less than exponential in the ambient dimension, then the scale can be chosen so that the learning error decreases. As the sample size becomes larger, the overall error stops decreasing but interestingly the scale can be chosen in such a way that the variance due to noise remains bounded. Our analysis combines, probabilistic results with a number of analytic techniques from interpolation theory.

1 Introduction

A classical idea in statistical learning theory is that there should be a tradeoff between fitting the data and the complexity of the estimators [33, 20, 14]. Indeed, much work is devoted to characterizing this intuition through different measures of complexity [30, 11, 29]. This point of view is contrasted by the recent empirical observation that it is often possible to fit, *interpolate*, the data

[*]pagliana@dima.unige.it
[†]alessandro.rudi@inria.fr
[‡]devito@dima.unige.it
[§]lrosasco@mit.edu

arbitrarily well, without degrading learning accuracy. This is indeed, true for deep neural networks [35], but also for other models [5] including kernel methods [7], and begs the question of whether interpolation can be reconciled with classical learning theory [5] .

Kernel methods provide a generalization of many classical linear models [28] and are grounded in the theory of reproducing kernel Hilbert spaces [1]. From a theoretical point of view, they provide a natural starting point to understand the learning properties of interpolating estimators. Recent works have started considering the properties of linear models, where interpolation can be achieved as soon as the number of parameters exceed the number of available points, the so called overparameterized regime. For example, linear regression is considered in [22], while linear models based on random features are studied in [6, 19]. In this context, bounds are typically derived via random matrix theory [25], letting the number of points and parameters go to infinity, a common setting in high dimensional statistics. Kernel methods have been considered in [23, 24]. In this case, the number of parameters is infinite, and bounds are derived assuming the dimension of the points to scale with their number. Further, the kernel is assumed to be fixed. A family of estimators defined by kernels depending on a scale (bandwith) parameter is studied in [8]. These estimators are akin to classical local kernel methods [20], but based on a family of singular kernels leading to interpolation.

In this paper, we study the properties of global kernel estimators, defined by minimum norm interpolating estimators defined by scale dependent Matern kernels [34]. The focus is on deriving non asymptotic bounds and understanding the role of the scale parameter in the kernel. Finite sample bounds for kernel methods are typically studied [29, 13, 12, 30], adding penalties or constraints that prevent interpolation. Only a handful of works consider the role of kernel parameters [18, 31, 21], but also in this case penalties are added. Here, we focus on the case where no penalty is added, called kernel ridge-less regression in [23], and focus on the influence of the kernel parameters. The basic observation is that there is a wide range of problems, where the corresponding estimators are stable, even without adding any penalty or constraint. This a byproduct of sampling data in high dimension, or, more precisely, of having a number of samples which is not exponential in the dimension. In this, case, the minimum distance among the points is large and the corresponding kernel matrix is shown to have small condition number. Error bounds can be derived as consequence of this basic observation, combining learning theory results [13], with tools from interpolation theory [34, 4]. Further, the scale of the kernel can be tuned to improve stability and hence the bounds. When the number of points grows, distances shrink and stability gets worse. Interestingly, even in this case, the variance can be controlled by tuning the scale parameter of the kernel, however the error plateaus and there is no consistency. Indeed, the lack of consistency is expected in view of the lower bounds in [26]. Interestingly, necessary and sufficient conditions for consistency as well as non asymptotic bounds are given in [2] for a general infinite dimensional regression problem, but these conditions are not realized by the class of kernels we consider.

The plan of the paper follows. We introduce the setting of the problem in Section 2, our main results in Section 3, and some simple experiments in Section 4. We defer technical details to the appendix.

2 Learning and interpolating with scale dependent kernel

We introduce the problem of learning with least-squares [20, 14], and the minimum norm interpolating estimator that we study. **Setting.** Let $(\mathcal{X} \times \mathbb{R}, \rho)$ be a probability space where $\mathcal{X}$ is a closed subset of $\mathbb{R}^d$. Denote by $\rho_{\mathcal{X}}$ the marginal measure on $\mathcal{X}$, and by $\rho(\cdot|x)$ the conditional measure on $\mathbb{R}$ given $x \in \mathcal{X}$. Statistical learning with least squares corresponds to the problem of minimizing the *expected risk*, defined as

$$\mathcal{E}(f) = \int_{\mathcal{X}\times\mathbb{R}} (f(x) - y)^2 \, \mathrm{d}\rho(x, y).$$

The expected risk is minimized by the *regression function*

$$f_\rho(x) = \int_{\mathbb{R}} y \, \mathrm{d}\rho(y|x),$$

but neither the expected risk nor the regression function can be computed, because ρ is known only through a *training set* $\mathbf{z} = (\mathbf{x}, \mathbf{y}) = (x_i, y_i)_{i=1}^n \sim \rho^n$ of n i.i.d random samples. Then, the learning problem is to use the data to derive an empirical estimate $f_{\mathbf{z}}$ of f_ρ. A natural way to measure the quality of an estimate is the *excess risk*

$$\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho),$$

In the following, we study the excess risk for a class of interpolating kernel least squares solutions, that we introduce next.

Interpolation with scale dependent kernels. We consider estimators defined by a family of symmetric positive definite kernels $k_\gamma : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$, depending on a scale parameter $\gamma > 0$, see [1, 9]. Each such kernel defines a unique Hilbert space of functions $\mathcal{H}_\gamma$ called *reproducing kernel Hilbert space* (RKHS) with inner product denoted by $\langle \cdot, \cdot \rangle_\gamma$, such that for every $x \in \mathcal{X}$, $k_{\gamma,x} = k_\gamma(x, \cdot) \in \mathcal{H}_\gamma$ and for every $f \in \mathcal{H}_\gamma$ the following reproducing property holds $f(x) = \langle k_{\gamma,x}, f \rangle_\gamma$. The estimator we consider is the minimum norm interpolating solution, i.e.

$$f^\dagger_{\gamma,\mathbf{z}} = \operatorname*{arg\,min}_{f\in\mathcal{H}_\gamma} \|f\|_\gamma \quad \text{such that} \quad f(x_i) = y_i \quad \forall i = 1, \dots, n \, . \tag{1}$$

The above estimator can be explicitly computed using the *kernel matrix* $K_{\gamma,\mathbf{x}} \in \mathbb{R}^{n\times n}$ such that $(K_{\gamma,\mathbf{x}})_{i,j} = k_\gamma(x_i, x_j)$, see Appendix D for details. Indeed, if we assume the input data to be distinct, then, for the class of kernel we consider, $K_{\gamma,\mathbf{x}}$ is invertible (see next section and [34]), and the estimator $f^\dagger_{\gamma,\mathbf{z}}$ can be computed as

$$f^\dagger_{\gamma,\mathbf{z}}(x) = k_\gamma(x, \mathbf{x})^\top K^{-1}_{\gamma,\mathbf{x}} \, \mathbf{y} \, .$$

where $k_\gamma(x, \mathbf{x})$ denotes the n-dimensional vector with entries $(k_\gamma(x, \mathbf{x}))_i = k_\gamma(x, x_i)$. In the following, we study the properties of the above estimator with a focus on the role played by the scale parameter γ. We next provide insight, comparing the minimum norm solution with the classical kernel ridge regression (KRR) estimator.

Stability and regularization with scale dependent kernels. With the above notation, the KRR estimator is defined for $\lambda > 0$ as

$$f^\dagger_{\gamma,\lambda,\mathbf{z}} = \operatorname*{arg\,min}_{f\in\mathcal{H}_\gamma} \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \|f\|^2_\gamma \,.$$

The above estimator can also be explicitly computed

$$f^\dagger_{\gamma,\lambda,\mathbf{z}}(x) = k_\gamma(x, \mathbf{x})^\top (K_{\gamma,\mathbf{x}} + n\lambda \,\mathrm{I})^{-1}\, \mathbf{y} \,.$$

Compared to 1 the inverse of the kernel matrix $K^{-1}_{\gamma,\mathbf{z}}$ is replaced by $(K_{\gamma,\mathbf{x}}+n\lambda \,\mathrm{I})^{-1}$. The idea is that adding the diagonal term can improve stability of the matrix in terms of the corresponding condition number, and also in terms of learning. Indeed, most theoretical studies of KRR focus on the role of the regularization parameter λ for learning, while assuming the kernel to be fixed [13, 32, 10]. This contrasts with common practice where both λ and the kernel parameters are tuned. This latter situation is studied only in a handful papers [18, 31, 21], but always assuming $\lambda > 0$. The case $\lambda = 0$, sometimes called ridge-less regression, has been recently considered in [26, 23], see also [2]. The key observation is that, not only the kernel matrix in (1) is invertible, but its condition number can be small depending on the distribution of the distances of the points in the training sets. This means that the corresponding estimator may converge, even without explicit regularization. Indeed, as we discuss in the following, for the Matern kernels the smallest eigenvalue depends on the minimal distance among the points, which can be large as long as the number of points is not exponential in the dimension. In this work, we also leverage the additional observation that, for the kernels we consider, the scale parameter γ further improve the smallest eigenvalue and hence the stability of the estimator. Before developing these ideas, we first introduce the class of kernels under investigation.

Matern kernels We focus on scale dependent Matern kernels [34], that are radial basis kernels with Fourier decay characterized by a smoothness parameter $s > d/2$ (see Appendix B for details),

$$k_\gamma(x, x') = Q_\gamma(x - x') \quad \text{with} \quad Q_\gamma(z) = \left(\frac{\|z\|}{\gamma}\right)^{s-d/2} \mathbf{K}_{s-d/2}\left(\frac{\|z\|}{\gamma}\right). \tag{2}$$

Here $\mathbf{K}_\alpha$ denotes the modified Bessel function of the second kind with parameter α. Matern kernels are uniformly bounded, since for every $\gamma > 0$ (see Lemma 8 in the Appendix)

$$\sup_{x,x'\in\mathcal{X}} k_\gamma(x, x') \le \kappa_s, \qquad \kappa_s = 2^{\frac{2s-d-2}{2}}\,\Gamma(s - d/2). \tag{3}$$

A peculiar and important property of these kernels is that for $s > d/2$, the RKHS corresponding to different scales are all equivalent to the Sobolev space $W_2^s(\mathcal{X})$ for every γ. This means that, if a function belongs to $\mathcal{H}_\gamma$ for some γ, then it also belong to $\mathcal{H}_{\gamma'}$ for any other γ'. However, the same function will have different norms in spaces corresponding to different scale parameters. Also, the scale parameter effectively rescales the distance of the points, and, as we see next, it changes the spectral properties of the matrices and operators defined by the kernels. The main example of kernels we consider are Laplace kernels ($s = d/2 + 1/2$), $Q_\gamma(z) = e^{-\|z\|/\gamma}$. Note that Gaussian kernels $e^{-\|z\|^2/(2\gamma^2)}$ do not satisfy the above assumptions [34] and in particular the associated RKHS are nested, i.e. spaces with larger scales are included in those with smaller scales. Further, Gaussian kernels have exponential Fourier decay, that makes it hard to control the eigenvalues of the corresponding kernel matrix. With the above comments in mind, we next present our main results.

3 From interpolation to learning: non asymptotic bounds

We introduce the main assumptions we consider and then state and discuss our main results.

3.1 Main assumptions

First, we introduce a basic assumption on the marginal distribution $\rho_\mathcal{X}$.

Assumption 1. *The marginal distribution $\rho_\mathcal{X}$ is the uniform distribution over $\mathcal{X}$, where $\mathcal{X}$ is a bounded subset of $\mathbb{R}^d$, with Lipschitz boundary and satisfying an internal cone condition, i.e. there exists an angle $\alpha \in (0, \pi/2)$ and a radius $r > 0$ such that for every $x \in \mathcal{X}$ a unit vector $\xi(x)$ exists such that the cone*

$$C(x, \xi(x), \alpha, r) := \left\{x + \lambda y : y \in \mathbb{R}^d,\ \|y\|_2 = 1,\ y^T \xi(x) \geq \cos\alpha,\ \lambda \in [0, r]\right\}$$

is contained in $\mathcal{X}$. We let $\mathrm{vol}(\mathcal{X}) = \int_\mathcal{X} \mathrm{d}x < +\infty$, where $\mathrm{d}x$ is the Lebesgue measure on $\mathbb{R}^d$.

The geometric assumptions on $\mathcal{X}$ are standard in approximation theory and satisfied for many subsets of $\mathbb{R}^d$ (such as balls or cubes) [34]. They are crucial to use tools from interpolation theory and Fourier analysis to study the spectral properties of matrices/operators defined by the kernel, and in particular to derive upper and lower bounds on their eigenvalues. Assuming a uniform distribution is needed only when studying the minimum distance among points and is done for simplicity. Similar conclusions should hold more generally. Next, we make a standard assumption on the output distribution.

Assumption 2. *We assume there exist $\sigma > 0$, such that for almost all $x \in \mathcal{X}$*

$$\mathbb{E}\left[y^2\right] < +\infty, \qquad \textit{and} \qquad \mathbb{E}\left[\exp\left(\lambda(y - f_\rho(x))\right) \mid x\right] \leq \exp\left(\sigma^2\lambda^2/2\right) .$$

The first condition implies that the regression function f_ρ is well defined, while the second is a condition on the variance of the outputs. This assumption is equivalent to a random design regression model $y = f_\rho(x) + \varepsilon$, where ε is a σ^2-subgaussian random variable.

Finally, we assume a well-specified model. As seen in Section 2, the spaces $\mathcal{H}_\gamma$ are equivalent for all scales γ, and also they are equivalent to Sobolev space $W_2^s(\mathcal{X})$ (see appendix B for details).

Assumption 3. *It holds that $f_\rho \in W_2^s(\mathcal{X})$.*

This assumption simplifies the analysis and the presentation. It can be relaxed to a miss-specified case where $f_\rho \notin \mathcal{H}_\gamma$, potentially leading to different learning behavior of the estimator. We leave this to future work. We comment instead on the role of the scale parameter. We will see in the following that considering a well specified model implies that the the scale parameter mainly influences the variance, rather than the bias/interpolation error, of the considered estimator.

3.2 Main results

Provided with the above discussion, we begin presenting our results. We first state a simplified version of our main theorem, and then provide details in the next section. In the following theorem we denote by $a(n) \lesssim b(n)$ the fact that there exists a constant C not depending on n such that $a(n) \leq Cb(n)$ for $n \in \mathbb{N}$ (analogously for $\approx$).

Theorem 1 (Main result: simplified)**.** *Under Assumptions 1, 2, 3, let $\delta > 0$, $p > 0$ and choose $s = \frac{d}{2} + p$ then, for every $\gamma \gtrsim C_2 n^{-2/d}$, the following holds with probability greater than $1 - 4\delta$*

$$\mathcal{E}(f^\dagger_{\gamma,\mathbf{z}}) - \mathcal{E}(f_\rho) \lesssim \begin{cases} \gamma^{\frac{4p^2}{d+2p}}\ n^{-\frac{2p}{d+2p}} & \text{if} \quad n \lesssim \left(\exp\left(\frac{1-2p}{2p}\right)\right)^{\frac{d}{2}} \\ \gamma^{\frac{4p^2}{d+2p}}\ n^{\frac{(d+4p)(2p)}{d(d+2p)}} & \text{otherwise .} \end{cases} \tag{4}$$

With the choice $\gamma \approx n^{-2/d}$, that is of the order of the minimum distance between the points in $\mathbf{x}$, then with probability greater than $1 - 4\delta$

$$\mathcal{E}(f^\dagger_{\gamma,\mathbf{z}}) - \mathcal{E}(f_\rho) \lesssim \begin{cases} n^{-4p/d} & \text{if} \quad n \lesssim \left(\exp\left(\frac{1-2p}{2p}\right)\right)^{\frac{d}{2}} \\ C & \text{otherwise} \end{cases} \tag{5}$$

where C and the other constants depend only on δ, $\mathcal{X}$, d and p (see appendix for their definition).

The main difference with recent results [23, 24] is that the above bounds are non asymptotic. In particular, the dimension d is kept fixed and the number of points n varied. Our results highlight two different phases in the learning curve crucially depending on n and d. If the kernel is not too smooth, i.e.,

$d/2 < s < d/2 + 1/2$, then, when n is less than exponential in the dimension d, the error improves with n. However, as n increases further, the improvements stop, in particular preventing the consistency of the estimator. This latter result aligns with the findings in [26], showing that ridge-less regression cannot be consistent for the class of kernels we considered. Our non asymptotic bounds further indicate the potential benefits of tuning the scale parameter. The bound in (4) holds for $\gamma \gtrsim n^{-2/d}$, and suggests γ can be tuned to improve the results. Indeed, from the bounds, the best choice is taking γ as small as possible that is scaling as $n^{-2/d}$. The bounds thus obtained depend on the dimensionality as well as the smoothness of the problem. In particular, they suggest a saturation effect [3], where the rates do not improve if p (hence s) is large. These results can be compared to those in [8] where local singular kernel dependent estimators are studied. Later we also compare with known bounds in nonparametric statistics. Here, we discuss the relation with the results in [2], deriving finite sample bounds for random design regression in an infinite dimensional setting. Interestingly, these results provide necessary and sufficient conditions for consistency of unregularized estimators. The setting in this paper provides a specific instance of the abstract setting of [2] in the context of kernel methods. In our notation, the condition for consistency is related to a condition on the smoothness s, which, roughly speaking, needs to be very close to $d/2$. Indeed, this turns out to be a critical regime also in our analysis, where the conditions for consistency are not met, even if tuning γ. In this view, It would be interesting to see if there are classes of kernels where consistency can be achieved. In the rest of the section, we discuss the main result leading to Theorem 1 as well as provide its complete statement.

3.3 Variance and interpolation error

We discuss an error decomposition for the excess risk of $f^\dagger_{\gamma,\mathbf{z}}$ and study the corresponding error terms. It is useful to first introduce the space L^2_ρ of square integrable functions on $\mathcal{X}$ with respect to the marginal $\rho_\mathcal{X}$ where $\|\cdot\|_\rho$ denotes the corresponding norm. Indeed, it is well known [15, 20] that

$$\mathcal{E}(f) - \mathcal{E}(f_\rho) = \|f - f_\rho\|^2_\rho .$$

Using the above identity and denoting by $\widehat{f}_\rho, \widehat{\varepsilon} \in \mathbb{R}^n$ the vectors which entries $f_\rho(x_i)$ and $y_i - f_\rho(x_i)$, we can decompose the excess risk of $f^\dagger_{\gamma,\mathbf{z}}$ as

$$\begin{aligned}
\mathcal{E}(f^\dagger_{\gamma,\mathbf{z}}) - \mathcal{E}(f_\rho) &= \left\| f^\dagger_{\gamma,\mathbf{z}} - f_\rho \right\|^2_\rho = \left\| k_\gamma(\cdot,\mathbf{x})^\top K^{-1}_{\gamma,\mathbf{x}}\, \mathbf{y} - f_\rho \right\|^2_\rho \qquad (6)\\
&= \left\| k_\gamma(\cdot,\mathbf{x})^\top K^{-1}_{\gamma,\mathbf{x}}\, \widehat{f}_\rho - f_\rho + k_\gamma(\cdot,\mathbf{x})^\top K^{-1}_{\gamma,\mathbf{x}}\, \widehat{\varepsilon} \right\|^2_\rho \\
&\leq 2\Big[\underbrace{\left\| k_\gamma(\cdot,\mathbf{x})^\top K^{-1}_{\gamma,\mathbf{x}}\, \widehat{f}_\rho - f_\rho \right\|^2_\rho}_{\text{interpolation error}} + \underbrace{\mathbb{E}_\mathbf{y} \left\| k_\gamma(\cdot,\mathbf{x})^\top K^{-1}_{\gamma,\mathbf{x}}\, \widehat{\varepsilon} \right\|^2_\rho}_{\text{variance}} \Big] \qquad (7)
\end{aligned}$$

where the last equality follows since the expectation of $\widehat{\varepsilon}$ conditionally to $\mathbf{x}$ is zero by definition. The first term in the decomposition is the error of approximating a function in $\mathcal{H}_\gamma$ with its interpolant over the discrete set of points $\mathbf{x}$ [34]. The second term is the variance of the estimator and depends on the noise. Both terms depend on the random sampling of the inputs.

Our main technical result is a bound on the variance. A key quantity is the minimum distance among the input points $\mathbf{x} = \{x_1, \ldots, x_n\}$, also known as *separation distance* [34], and defined as

$$q_\infty = \frac{1}{2} \min_{i \neq j} \|x_i - x_j\|_\infty \ ,$$

where $\|x\|_\infty$ is the sup norm of the vector x. The separation distance is crucial to control the minimum eigenvalue of the kernel matrix in terms of the scale γ. In turn, the minimum eigenvalue governs the variance of the estimator. The next lemma generalizes a result in [17] to the case of Matern kernels.

Lemma 1 (Lower bound on $\sigma_{\min}(K_{\gamma,\mathbf{x}})$)**.** *Let $K_{\gamma,\mathbf{x}}$ be the kernel matrix and q_∞ be the separation distance of the set $\mathbf{x}$. Then for every $0 < q \leq q_\infty$ there exist two constants $c_2(d,s)$ and $c_1(d,s)$ (defined in the proof E) depending on d and s such that*

$$\sigma_{\min}(K_{\gamma,\mathbf{x}}) \geq \begin{cases} c_1(d,s) \left(\frac{\gamma}{q}\right)^d & \text{if} \quad \gamma < \gamma_q \\ c_2(d,s) \left(\frac{q}{\gamma}\right)^{2s-d} & \text{if} \quad \gamma \geq \gamma_q \end{cases} \tag{8}$$

where $\gamma_q = \frac{2q}{\sqrt{d}\pi} \sqrt{\frac{s}{s+1}}$.

Note that, if γ is larger than a quantity of the order of the separation distance q_∞, then the lower bound increases with q. This means that the more distant are the points, the better is the condition number of the kernel matrix. Moreover, the lower bound is increasing and then decreasing in γ with a maximum in $\gamma = \gamma_q$. This shows that the scale γ allows a control on the condition number of the empirical kernel matrix. In particular with the choice $\gamma = \gamma_q$ we obtain a constant lower bound,

$$\sigma_{\min}(K_{\gamma,\mathbf{x}}) \geq c_2(d,s) \left(\frac{\sqrt{d}\pi(s+1)}{2s}\right)^{2s-d} =: c_3(d,s) \ . \tag{9}$$

Also note that the above bound suggests a diverging behavior as γ becomes smaller than γ_q, which in practice seems pessimistic, see Section 4. Improving the lower bound in this regime is left for future work. We note that the smallest eigenvalue of the kernel matrix is also the key quantity studied in [23, 24] using different techniques. In our approach, lower bound in Lemma 1 is completely deterministic, and probability enters only in finding a lower bound q_n on the separation distance. For the sake of simplicity and since considering $\gamma < \gamma_q$ does

not lead to any improvement, we next consider $\gamma \geq \gamma_q$. In particular, for a given $0 < \delta < 1$, we have the following lower bound on the separation distance, that holds with probability greater than $1 - \delta$

$$q_n := \frac{\sqrt[d]{\delta \ \mathrm{vol}\,(\mathcal{X})}}{4} \cdot \frac{1}{n^{2/d}} \leq q_\infty$$

and define

$$\gamma_n := \gamma_{q_n} = \frac{\sqrt[d]{\delta \ \mathrm{vol}\,(\mathcal{X})}}{2\pi\sqrt{d}} \sqrt{\frac{s}{s+1}} \cdot \frac{1}{n^{2/d}} \,. \tag{10}$$

The next proposition gives the bound on the variance.

Proposition 1 (Variance)**.** *Let $\delta > 0$, $s > d/2$, $\gamma > \gamma_n$ and*

$$N_{s,d} = \sqrt{\delta \ \mathrm{vol}(\mathcal{X})} \left(\frac{8^{\frac{1}{2s-d}}}{4} \right)^{\frac{d}{2}} \,. \tag{11}$$

Under Assumptions 1, 2 there exist a constant $c > 0$ depending on $d, s, \mathcal{X}$ such that with probability greater than $1 - 3\delta$

$$\left\| k_\gamma(\cdot, \mathbf{x})^\top K_{\gamma,\mathbf{x}}^{-1} \, \widehat{\varepsilon} \, \right\|_\rho^2 \leq \begin{cases} \frac{32 c_\delta \sigma^2 c}{c_2(d,s)\delta} \gamma^{\frac{(2s-d)^2}{2s}} n^{-\frac{2s-d}{2s}} & \text{if} \quad n \leq N_{s,d} \\ \frac{4 c_\delta \sigma^2 c}{c_2(d,s) \delta^{\frac{2s}{d}}} \gamma^{\frac{(2s-d)^2}{2s}} n^{\frac{(4s-d)(2s-d)}{2sd}} & \text{if} \quad n > N_{s,d} \end{cases}$$

where $c_2(d, s)$ is defined in Lemma 1. If $\gamma = \gamma_n$, then with probability greater than $1 - 3\delta$

$$\left\| k_\gamma(\cdot, \mathbf{x})^\top K_{\gamma,\mathbf{x}}^{-1} \, \widehat{\varepsilon} \, \right\|_\rho^2 \leq \begin{cases} \frac{32 c_\delta \sigma^2 \kappa_s}{c_4(d,s)\delta} n^{-\frac{2(2s-d)}{d}} & \text{if} \quad n \leq N_{s,d} \\ \frac{4 c_\delta \sigma^2 \kappa_s}{c_4(d,s) \delta^{\frac{2s}{d}}} & \text{if} \quad n > N_{s,d} \end{cases}$$

where $c_4(d, s) = c_3(d, s) \left(\frac{4}{\sqrt[d]{\delta \, \mathrm{vol}(\mathcal{X})}} \right)^{2s-d}$, $c_3(d, s)$ is defined in (9) *and κ_s is defined in* (13)*.*

The proof of Proposition 1 is in Appendix F. It combines Lemma 1 with a probabilistic bound on the separation distance and a new estimate of the effective dimension of the RKHS defined by Matern kernels. The condition $n \leq N_{s,d}$ is meaningful if the kernel is not too smooth, meaning that at most $s = d/2 + 1/2$ (that is $p = 1/2$). In this case, if the number of points is no larger than exponential in d, then the variance is decreasing in n, for any fixed γ and the decrease can be made faster tuning γ. Note that, $N_{s,d}$ goes to $+\infty$ as s approach $d/2$ (that is $p \sim 0$). On the other hand, if the number of points is very large, then for fixed γ the bound start to increase with n. Interestingly, tuning γ the variance stays bounded by a constant, even for very smooth kernels (meaning large s).
Next proposition gives a bound on the interpolation error.

Proposition 2 (Interpolation error)**.** *Let $\delta > 0$, $s > d/2$ and $\gamma > 0$. Under Assumption 3, with probability at least $1-\delta$ it holds that*

$$\left\| k_\gamma(\cdot, \mathbf{x})^\top K_{\gamma,\mathbf{x}}^{-1} \, \widehat{f}_\rho - f_\rho \right\|_\rho^2 \leq \frac{3\kappa_s}{\sqrt{n}} \log \frac{2}{\delta} \, \|f_\rho\|_{W_2^s}^2 \ .$$

where κ_s is defined in equation (13)*.*

This results is based on the observation that the interpolation error is the norm of the difference between f_ρ and its projection on the n-dimensional subspace $\mathcal{H}_{\gamma,\mathbf{x}} = \operatorname{span}\{k_{\gamma,x_i} \, : \, i = 1, \ldots, n\}$, and on the use of standard concentration inequalities. The main new contribution is showing that the interpolation error does not depend on the scale parameter γ. This result can be improved to get a higher convergence rate. However, note that since the variance is dominant even a higher rate would not improve the total bound on the performance of the expected excess risk. We leave this improvement for a longer version of the paper. We also note that the above quantity is standard in interpolation [34], however known bounds typically require n to be very large.
From Proposition 1 and Proposition 2 we derive the complete statement of our main result.

Theorem 2 (Bound on the excess risk)**.** *Let $\delta > 0$, $s > d/2$, $\gamma > \gamma_n$, then with probability greater than $1 - 4\delta$*

$$\mathcal{E}(f^\dagger_{\gamma,\mathbf{z}}) - \mathcal{E}(f_\rho) \leq \begin{cases} \frac{64 c_\delta \sigma^2 c}{\delta c_2(d,s)} \gamma^{\frac{(2s-d)^2}{2s}} n^{-\frac{2s-d}{2s}} + \frac{6\kappa_s}{\sqrt{n}} \log \frac{2}{\delta} \, \|f_\rho\|_{W_2^s}^2 & \text{if} \quad n \leq N_{s,d} \\ \frac{8 c_\delta \sigma^2 c}{c_2(d,s) \delta^{\frac{2s}{d}}} \gamma^{\frac{(2s-d)^2}{2s}} n^{\frac{(4s-d)(2s-d)}{2sd}} + \frac{6\kappa_s}{\sqrt{n}} \log \frac{2}{\delta} \, \|f_\rho\|_{W_2^s}^2 & \text{if} \quad n > N_{s,d} \ . \end{cases}$$

With the choice γ_n in (10) *it holds with probability greater than $1 - 4\delta$ by*

$$\mathcal{E}(f^\dagger_{\gamma,\mathbf{z}}) - \mathcal{E}(f_\rho) \leq \begin{cases} \frac{64 c_\delta \sigma^2 \kappa_s}{c_4(d,s)\delta} n^{-\frac{2(2s-d)}{d}} + \frac{6\kappa_s}{\sqrt{n}} \log \frac{2}{\delta} \, \|f_\rho\|_{W_2^s}^2 & \text{if} \quad n \leq N_{s,d} \\ \frac{8 c_\delta \sigma^2 \kappa_s}{c_4(d,s) \delta^{\frac{2s}{d}}} + \frac{6\kappa_s}{\sqrt{n}} \log \frac{2}{\delta} \, \|f_\rho\|_{W_2^s}^2 & \text{if} \quad n > N_{s,d} \ . \end{cases}$$

The variance strongly depends on the scale parameter γ, whereas the interpolation error does not. Moreover, the latter decreases with n, so that the variance dominates the error, see Thereom1. As already noted, if the kernel is not too smooth, i.e. $s < d/2 + 1/2$, then the obtained bound shows two different regimes of n, d. For fixed γ, if the number of points is not exponential in d, then the bound is decreasing in n, but when n is larger then the bound starts to increase. The intuition is that if n is not very high, then the input points are distant and the condition number of $K_{\gamma,\mathbf{x}}$ does not affect the behavior of the estimator. When n increases, the inputs start to be too close, and the condition number increases, degrading the stability of the estimator. The scale parameter γ can be tuned to improve the stability and hence the bound for "n small". Interestingly, it also to allows the variance to stay bounded for "n large". This last fact is true independently on the smoothness s of the kernel. The a main drawback of the analysis is a pessimistic dependence of the constants in the dimension d.

3.4 Related works

We further comment on the relation with related works. As already mentioned, the results in [23] [24] are close to our analysis. In [23] the authors consider $d \asymp n$, and show that the curvature of the kernel has an implicit regularization effect, proving a data-dependent bound. In [24] the authors consider $d \asymp n^\alpha$ (with $0 < \alpha < 1$) and show that for certain values of α, the error goes to zero as n, d go to $+\infty$. The difference with these works is that they use results from random matrix theory and concentration of measure to exploit the high-dimensionality of the data and control the minimum eigenvalue. In contrast, we fix the dimension d and use analytic results from interpolation theory, coupled with probabilistic estimates of the separation distance. Further, we investigate the role of the scale parameter γ, which in practice can be tuned. Our estimates are explicit but have pessimistic dependence on the d. If the number of points is not too large (meaning exponential in d), then our bound takes the form

$$\mathbb{E}_{\mathbf{z}}\left[\mathcal{E}(f^\dagger_{\gamma,\mathbf{z}}) - \mathcal{E}(f_\rho)\right] \leq \begin{cases} C\, n^{-\frac{2p}{d+2p}} & \gamma \text{ constant} \\ C\, n^{-4p/d} & \gamma = \gamma(n). \end{cases}$$

This bound is similar to bound for local methods to estimate a Lipschitz continuous function (like partitioning estimators or kernel smoothing [20]). However the dependence of our constant C on d is worse, since it increases with d. This effect is due to technicalities in the estimation of the minimum eigenvalue. Numerical simulations show that in practice the minimum eigenvalue is increasing in d. A recent work [8] studies a local estimator with singular kernels, under the well-specified model and assuming f_ρ to be in the Hölder class (with parameter β). Here, the scale γ is also tuned to of order $n^{-1/(2\beta+d)}$ and gives the rate of optimal (for that class of problems) rate $n^{-2\beta/(2\beta+d)}$. In contrast, here we study the role of the scale for the global estimators, obtained as minimum norm interpolants. For this estimator the bound start increasing (or stay bounded with an appropriate tuning of γ) with the number of points n, hence is not consistent. This is directly connected with recent results in [2] studying consistency when the marginal distribution is assumed to Gaussian and the eigenvalues $\sigma_{\ell,\gamma}$ of the integral operator $L_\gamma f(x) = \int k_\gamma(x,x')f(y)\, d\rho_{\mathcal{X}}(x')$ decay as $(\ell \log^q \ell)^{-1}$ for some $q \geq 2$. In our case the eigenvalues decay only as $\sigma_\ell \approx \ell^{-2s/d}$, $s > d/2$. The slow decay can only be approximatively achieved when $s \to (d/2)^+$. Note that in this case, the condition $n \leq N_{s,d}$ is always satisfied, since $N_{s,d} \to +\infty$ and the bound becomes decreasing (even if slowly) in n.

4 Numerical results

We perform basic experiments considering $\mathcal{X} = [0,1]^d$, ρ_X uniform and $k_\gamma(x,x') = e^{-\|x-x'\|/\gamma}$.

Smallest eigenvalue vs γ (Fig. 1). We consider the behavior of smallest singular value $\sigma_{\min}(K_{\gamma,\mathbf{x}})$ as function of γ, varying d. As expected, $\sigma_{\min}(K_{\gamma,\mathbf{x}})$

is larger for smaller γ, and as mentioned before it does not diverge for $\gamma \to 0$, but rather it stabilizes, suggesting our bound is loose for γ small. Also, it further improves as d increases, since the distance between points also increases.

Excess risk vs γ (Fig. 2). We study the behavior for the excess risk as a function of the scale γ, varying n. We assume a regression model $y_i = f_\rho(x_i) + \varepsilon_i$ where $f_\rho(x) = \arctan(\|x\|^2)$ and ε_i are Gaussian random variables with zero-mean and variance 2. The dimension d is chosen equal to 10. The plot shows that the excess risk, approximated as the average of 20 simulations on a validation set of 10^3 data, can be improved by tuning γ, but the improvements level out as n increases. Also, the results suggest that the choice $\gamma = n^{-2/d}$ is appropriate.

Excess risk vs n (Fig. 3). With the same setting as before, we study the behavior for the excess risk as a function of n, varying γ. The plot shows that, for any fixed γ, the excess error decreases with n and then reaches a constant value as predicted by theory. The result improve as we choose a smaller scale and in particular the result improve as the scale approaches the value $\gamma = 1.082\ n^{-2/d}$.

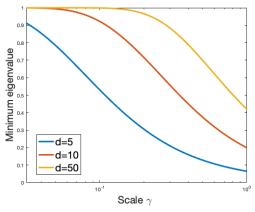

Figure 1: Plot of $\sigma_{\min}(K_{\gamma,\mathbf{x}})$ as function of γ with fixed $n = 500$ and different choices of d.

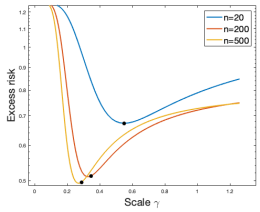

Figure 2: Approximation of the excess risk of $f^\dagger_{\gamma,\mathbf{z}}$. Black dot show the choice $\gamma = n^{-2/d}$.

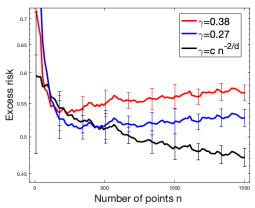

Figure 3: Approximation of the excess risk of $f^\dagger_{\gamma,\mathbf{z}}$.

5 Conclusions

We study excess risk bounds for kernel ridge-less regression keeping the data dimension fixed and focusing on understanding the role the kernel parameters. Our results suggest that error decreases as long as the number of point is smaller than exponential d and then flattens out preventing consistency. The result is a direct consequence of the interplay between the minimum distance among points and the smallest eigenvalue of the kernel matrix. As points are further away, the condition number improves, highlighting the key role played by the input dimension, rather than the number of parameters in the model. The scale of the kernel effectively changes the distance used in the kernel, providing further control on the stability of the estimator. A number of developments are left for future work, like sharpening the estimates at small scales and improving the dependence of the constants on d. Considering different class of kernels and comparing ridge and ridge-less regression would be interesting.

References

[1] N. Aronszajn. Theory of reproducing kernels. Trans. Amer. Math. Soc, 68(3):337–404, 1950.

[2] Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. Proceedings of the National Academy of Sciences, 2020.

[3] Frank Bauer, Sergei V. Pereverzev, and Lorenzo Rosasco. On regularization algorithms in learning theory. J. Complex., 23(1):52–72, 2007.

[4] 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.

[5] 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, 2019.

[6] Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. arXiv preprint arXiv:1903.07571, 2019.

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

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

[9] Alain Berlinet and Christine Thomas-Agnan. Reproducing kernel Hilbert spaces in probability and statistics. Springer Science & Business Media, 2011.

[10] Gilles Blanchard and Nicole Mücke. Optimal rates for regularization of statistical inverse learning problems. Foundations of Computational Mathematics, 18(4):971–1013, 2018.

[11] O. Bousquet, S. Boucheron, and G. Lugosi. Introduction to statistical learning theory. pages 169–207, 2003.

[12] O. Bousquet and A. Elisseeff. Stability and generalization. Journal Machine Learning Research, 2001.

[13] Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. Foundations of Computational Mathematics, 7(3):331–368, 2007.

[14] F. Cucker and S. Smale. On the mathematical foundations of learning. Bulletin of the AMS, 39:1–49, 2002.

[15] Felipe Cucker and Ding Xuan Zhou. Learning theory: an approximation theory viewpoint, volume 24. Cambridge University Press, 2007.

[16] Christine De Mol, Ernesto De Vito, and Lorenzo Rosasco. Elastic-net regularization in learning theory. Journal of Complexity, 25(2):201–230, 2009.

[17] Benedikt Diederichs and Armin Iske. Improved estimates for condition numbers of radial basis function interpolation matrices. Journal of Approximation Theory, 2017.

[18] Mona Eberts, Ingo Steinwart, et al. Optimal regression rates for svms using gaussian kernels. Electronic Journal of Statistics, 7:1–42, 2013.

[19] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. arXiv preprint arXiv:1904.12191, 2019.

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

[21] Thomas Hamm and Ingo Steinwart. Adaptive learning rates for support vector machines working on data with low intrinsic dimension. arXiv, pages arXiv–2003, 2020.

[22] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. arXiv preprint arXiv:1903.08560, 2019.

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

[24] Tengyuan Liang, Alexander Rakhlin, and Xiyu Zhai. On the risk of minimum-norm interpolants and restricted lower isometry of kernels. arXiv preprint arXiv:1908.10292, 2019.

[25] V. A. Marchenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. Mat. Sb. (N.S.), 72(114):4:457–483, 1967.

[26] Alexander Rakhlin and Xiyu Zhai. Consistency of interpolation with laplace kernels is a high-dimensional phenomenon. arXiv preprint arXiv:1812.11167, 2018.

[27] Alessandro Rudi, Guillermo D Canas, and Lorenzo Rosasco. On the sample complexity of subspace learning. In Advances in Neural Information Processing Systems, pages 2067–2075, 2013.

[28] Bernhard Scholkopf and Alexander J Smola. Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press, 2001.

[29] Shai Shalev-Shwartz and S. Ben-David. Understanding Machine Learning: From Theory to Algorithms. Cambridge eBooks, 2014.

[30] Ingo Steinwart and Andreas Christmann. Support vector machines. Springer Science & Business Media, 2008.

[31] Ingo Steinwart and Simon Fischer. A closer look at covering number bounds for gaussian kernels. arXiv preprint arXiv:1912.11741, 2019.

[32] Ingo Steinwart, Don R Hush, Clint Scovel, et al. Optimal rates for regularized least squares regression. In COLT, pages 79–93, 2009.

[33] V. N. Vapnik. Statistical learning theory. Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons Inc., New York, 1998. A Wiley-Interscience Publication.

[34] Holger Wendland. Scattered data approximation, volume 17. Cambridge university press, 2004.

[35] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In International Conference on Learning Representations (ICLR), 2017.

Appendix A Notation

The Euclidean norm and scalar product of $\mathbb{R}^n$ are denoted by $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$. The Lebesgue measure of $\mathbb{R}^d$ is denoted by $\mathrm{d}x$, for any Borel subset E its volume is

$$\mathrm{vol}(E) = \int_E \mathrm{d}x.$$

and the corresponding Lebesgue spaces are $L^p(\mathbb{R}^d,\ \mathrm{d}x)$ with $p \in [0, +\infty]$. If the Lebesgue measure is replaced by the marginal distribution $\rho_{\mathcal{X}}$ we use the short notation L^p_ρ.

The Fourier transform $\mathcal{F}$ is defined as

$$\mathcal{F}f(\xi) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} f(x)\ e^{-i\langle \xi, x\rangle}\ dx \qquad \forall f \in L^1\left(\mathbb{R}^d, dx\right)$$

and if $\mathcal{F}f \in L^1\left(\mathbb{R}^d, d\xi\right)$ the inversion formula

$$h(x) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \mathcal{F}f(\xi)\ e^{i\langle \xi, x\rangle}\ d\xi\ .$$

If A is an bounded operator between to Hilbert spaces, we denote by $A^\dagger$ the Moore-Penrose inverse. If A is a semi-positive definite square matrix we denote by $\sigma_{\min}(A)$ the smallest non-zero eigenvalue of M.

Appendix B Sobolev spaces and Matern kernels

We recall some properties, see [34] for more details.

1. We remember the definition of the Bessel function of the second kind $\mathbf{K}_\alpha$ with parameter α, for $x > 0$ and $\alpha > 0$ it holds

$$\mathbf{K}_\alpha(x) = \left(\frac{\pi}{2x}\right)^{1/2} \frac{e^{-x}}{\Gamma(\alpha + 1/2)} \int_0^\infty e^{-u} u^{\alpha - 1/2} \left(1 + \frac{u}{2x}\right)^{\alpha - 1/2}\ \mathrm{d}u\ .$$

2. The Sobolev space of function on $\mathbb{R}^d$ of smoothness s can be defined as

$$W_2^s\left(\mathbb{R}^d\right) = \left\{ f \in L^2\left(\mathbb{R}^d\right) : \mathcal{F}f(\cdot)\left(1 + \|\cdot\|_2^2\right)^{s/2} \in L^2\left(\mathbb{R}^d\right) \right\}$$

   where $L^2\left(\mathbb{R}^d\right)$ denotes the square integrable functions on $\mathbb{R}^d$ with respect to the Lebesgue measure $\mathrm{d}x$. This space can be equipped with the inner product

$$\langle f, g\rangle_{W_2^s(\mathbb{R}^d)} := (2\pi)^{-d/2} \int_{\mathbb{R}^d} (\mathcal{F}f)(\omega)\overline{(\mathcal{F}g)(\omega)}\left(1 + \|\omega\|_2^2\right)^s d\omega$$

   and the respective norm

$$\|f\|^2_{W_2^s(\mathbb{R}^d)} := (2\pi)^{-d} \int_{\mathbb{R}^d} |(\mathcal{F}f)(\omega)|^2 \left(1 + \|\omega\|_2^2\right)^s d\omega$$

   Now we can define the Sobolev spaces $W_2^s\left(\mathcal{X}\right)$ over a sufficiently regular domain $\mathcal{X} \subseteq \mathbb{R}^d$ as the restriction of the functions in $W_2^s\left(\mathbb{R}^d\right)$ to $\mathcal{X}$.

3. the Fourier trasform of Q_γ is

$$\mathcal{F}Q_\gamma(\xi) = 2^{s-1}\Gamma(s)\frac{\gamma^d}{(1+\gamma^2\left\|\xi\right\|^2)^s}\ ; \tag{12}$$

4. since $\mathcal{F}Q_\gamma(\xi)$ is not zero, K_γ is positive definite [34] and so the kernel matrix $K_{\gamma,\mathbf{x}}$ is invertible, provided the input data $\mathbf{x}$ are disjoint;

5. the kernel K_γ is bounded

$$\sup_{x,x'\in\mathcal{X}} k_\gamma(x,x') \leq 2^{\frac{2s-d-2}{2}}\,\Gamma(s-d/2) = \kappa_s\ ; \tag{13}$$

6. all RKHS $\mathcal{H}_\gamma$ are equal to the Sobolev space $W_2^s(\mathcal{X})$ with equivalent norms

$$c_\gamma\left\|f\right\|_{\mathcal{H}_\gamma} \leq \left\|f\right\|_{W_2^s(\mathcal{X})} \leq C_\gamma\left\|f\right\|_{\mathcal{H}_\gamma}$$

   provided suitable regularity condition on $\mathcal{X}$ as in Assumption 1 [34]. However, the constants c_γ and C_γ depend on γ and the dependence on this parameter since it allows to rescale the space $\mathcal{X}$ and so controlling the distance between the points x_i ;

7. with the choice $s = \frac{d}{2}+\frac{1}{2}$, we recover Laplace kernel

$$Q_\gamma(z) = \sqrt{\frac{\pi}{2}}\,e^{-\frac{\left\|z\right\|}{\gamma}}\ .$$

8. The norm in the RKHS $\mathcal{H}_\gamma$ is given by

$$\left\|f\right\|^2_{\mathcal{H}_\gamma} = \int_{\mathbb{R}^d}\frac{\left|\mathcal{F}f(\omega)\right|^2}{\left|\mathcal{F}\Phi_\gamma(\omega)\right|}\,\mathrm{d}\omega\ .$$

Appendix C Mathematical setting: kernel operators

In this section we define all the key operator that will be useful in the analysis of the error.
We let $S_\gamma : \mathcal{H}_\gamma \to L^2_\rho$ such that almost surely $S_\gamma f(x) = \langle f, k_{\gamma,x}\rangle_\gamma$. Let S^*_γ be the adjoint operator of S_γ and denote $T_\gamma = S^*_\gamma S_\gamma$ the covariance operator and $L_\gamma = S_\gamma S^*_\gamma$ the integral operator, defined as

$$\begin{aligned} &T_\gamma\colon \mathcal{H}_\gamma \to \mathcal{H}_\gamma && T_\gamma f = \int \langle f, k_{\gamma,x}\rangle_\gamma\, k_{\gamma,x}\; d\rho_{\mathcal{X}}(x) \\ &L_\gamma\colon L^2_\rho \to L^2_\rho && L_\gamma f(x) = \int k_\gamma(x,x')f(x')\; d\rho_{\mathcal{X}}(x')\ . \end{aligned}$$

For any set of point $\mathbf{x} = (x_1, \dots, x_n)$ we also introduce finite rank operators by replacing ρ by

$$\widehat{\rho} = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$$

and consider the space $\mathbb{R}^n$ with the normalized norm $\|\cdot\|_n = \frac{1}{\sqrt{n}} \|\cdot\|$ and scalar product $\langle \cdot, \cdot \rangle_n = \frac{1}{n} \langle \cdot, \cdot \rangle$

$$\begin{aligned}
&S_{\gamma,\mathbf{x}} : \mathcal{H}_\gamma \to \mathbb{R}^n \qquad (S_{\gamma,\mathbf{x}} f)_i = \langle k_{\gamma,x_i}, f \rangle_\gamma \quad \forall f \in \mathcal{H}_\gamma,\ \forall i \in \{1, \dots, n\} \\
&S^*_{\gamma,\mathbf{x}} : \mathbb{R}^n \to \mathcal{H}_\gamma \qquad S^*_{\gamma,\mathbf{x}} w = \frac{1}{n} \sum_{i=1}^{n} w_i k_{\gamma,x_i} \quad \forall w \in \mathbb{R}^n \\
&T_{\gamma,\mathbf{x}} \colon \mathcal{H}_\gamma \to \mathcal{H}_\gamma \qquad T_{\gamma,\mathbf{x}} = S^*_{\gamma,\mathbf{x}} S_{\gamma,\mathbf{x}} = \frac{1}{n} \sum_{i=1}^{n} k_{\gamma,x_i} \otimes k_{\gamma,x_i} \\
&L_{\gamma,\mathbf{x}} \colon \mathbb{R}^n \to \mathbb{R}^n \qquad L_{\gamma,\mathbf{x}} = S_{\gamma,\mathbf{x}} S^*_{\gamma,\mathbf{x}} = \frac{1}{n} K_{\gamma,\mathbf{x}}
\end{aligned} \tag{14}$$

where $K_{\gamma,\mathbf{x}}$ denotes the kernel matrix which entries are $(K_{\gamma,\mathbf{x}})_{i,j} = k_\gamma(x_i, x_j)$ for every $i, j \in \{1, \dots, n\}$.

Since k_γ is bounded and measurable, all above operators are bounded, $T_\gamma, T_{\gamma,\mathbf{x}}, L_\gamma, L_{\gamma,\mathbf{x}}$ are positive trace class operators and the pairs T_γ, L_γ and $T_{\gamma,\mathbf{x}}, L_{\gamma,\mathbf{x}}$ have the same positive eigenvalues. Furthermore, $\mathcal{H}_\gamma = \overline{\text{Range}}(S^*_\gamma) = \overline{\text{Range}}\left(S^*_\gamma S_\gamma\right)$ and $\mathcal{H}_{\gamma,\mathbf{x}} = \text{Range}(S^*_{\gamma,\mathbf{x}}) = \text{Range}(S^*_{\gamma,\mathbf{x}} S_{\gamma,\mathbf{x}})$.
Observe that with these operator the condition on the regression function (Assumption 3) is equivalent to assume there exists $g \in \mathcal{H}_1$

$$f_\rho = S_1 g$$

and that the estimated solution $f^\dagger_{\gamma,\mathbf{z}}$ can be written as (see section D)

$$f^\dagger_{\gamma,\mathbf{z}} = T^\dagger_{\gamma,\mathbf{x}} S^*_{\gamma,\mathbf{x}} \mathbf{y} \, .$$

and that the vector $\widehat{f_\rho} = S_{\gamma,\mathbf{x}} f_\rho$. Finally the error decomposition (6) can be rewritten as

$$\begin{aligned}
\left\| S_\gamma f^\dagger_{\gamma,\mathbf{z}} - f_\rho \right\|^2_\rho &= \left\| S_\gamma \left(T^\dagger_{\gamma,\mathbf{x}} S^*_{\gamma,\mathbf{x}} S_{\gamma,\mathbf{x}} f_\rho + T^\dagger_{\gamma,\mathbf{x}} S^*_{\gamma,\mathbf{x}} \widehat{\varepsilon} \right) - f_\rho \right\|^2_\rho \\
&\leq 2 \left\| S_\gamma T^\dagger_{\gamma,\mathbf{x}} S^*_{\gamma,\mathbf{x}} S_{\gamma,\mathbf{x}} f_\rho - f_\rho \right\|^2_\rho + 2 \left\| S_\gamma T^\dagger_{\gamma,\mathbf{x}} S^*_{\gamma,\mathbf{x}} \widehat{\varepsilon} \right\|^2_\rho \\
&= 2 \left(\underbrace{\left\| S_\gamma \left(P_{\gamma,\mathbf{x}} - I \right) f_\rho \right\|^2_\rho}_{\text{interpolation error}} + \underbrace{\left\| S_\gamma T^\dagger_{\gamma,\mathbf{x}} S^*_{\gamma,\mathbf{x}} \widehat{\varepsilon} \right\|^2_\rho}_{\text{variance}} \right) ,
\end{aligned} \tag{15}$$

where $P_{\gamma,\mathbf{x}} = T^\dagger_{\gamma,\mathbf{x}} S^*_{\gamma,\mathbf{x}} S_{\gamma,\mathbf{x}} = T^\dagger_{\gamma,\mathbf{x}} T_{\gamma,\mathbf{x}}$ is the projection onto the interpolation space $\mathcal{H}_{\gamma,\mathbf{x}}$.

Appendix D Nonparametric ordinary least squares

The family of estimators we consider are given by

$$f^{\dagger}_{\gamma,\mathbf{z}} = {S_{\gamma,\mathbf{x}}}^{\dagger}\mathbf{y}\ . \tag{16}$$

We describe how such estimators can be derived and implemented numerically. Consider the set of linear equations

$$f(x_i) = y_i, \qquad i = 1, \ldots, n$$

for $f \in \mathcal{H}_\gamma$. Then, the above equations can be written as

$$\langle f, k_{\gamma,x_i}\rangle_\gamma = y_i, \qquad i = 1, \ldots, n.$$

With the aid of the empirical kernel operators they correspond to the finite dimensional inverse problem

$$S_{\gamma,\mathbf{x}} f = \mathbf{y}.$$

From property 2 of section B we know that if the kernel k_γ is positive-definite (which is our case) then for every $n \in \mathbb{N}$ the dimension of Range$(S^*_{\gamma,\mathbf{x}})$ is n.

Then the above problem has multiple solutions. However, it is a standard fact that there is a minimal norm solution solving

$$\min_{S_{\gamma,\mathbf{x}} f = \mathbf{y}} \|f\|_\gamma\ .$$

The form of the solution is easily derived using Lagrange multipliers, considering

$$\min_{\alpha \in L^2_\rho, f \in \mathcal{H}_\gamma} \mathcal{L}(\alpha, f), \qquad \mathcal{L}(\alpha, f) = \frac{1}{2}\|f\|^2_\gamma - \langle \alpha, S_{\gamma,\mathbf{x}} f - \mathbf{y}\rangle_n\ .$$

Setting the partial derivative w.r.t. f to zero gives

$$\partial_f \mathcal{L}(\alpha, f) = f - S^*_{\gamma,\mathbf{x}}\alpha = 0 \qquad \Rightarrow \qquad f = S^*_{\gamma,\mathbf{x}}\alpha.$$

Setting the partial derivative w.r.t. α t Setting the partial derivative w.r.t. f to zero gives

$$\partial_\alpha \mathcal{L}(\alpha, f) = -(S_{\gamma,\mathbf{x}} f - \mathbf{y}) = 0 \qquad \Rightarrow \qquad S_{\gamma,\mathbf{x}} f = \mathbf{y}$$

And combining the two conditions

$$S_{\gamma,\mathbf{x}} S^*_{\gamma,\mathbf{x}}\alpha = \mathbf{y} \qquad \Rightarrow \qquad \alpha = (S_{\gamma,\mathbf{x}} S^*_{\gamma,\mathbf{x}})^{-1}\mathbf{y}$$

where, since the dimension of Range$(S^*_{\gamma,\mathbf{x}})$ is n then $(\hat{S}_\gamma {\hat{S}_\gamma}^*)^{-1}$ is invertible and

$$f = S^*_{\gamma,\mathbf{x}}(S_{\gamma,\mathbf{x}} S^*_{\gamma,\mathbf{x}})^{-1}\mathbf{y}$$

that is (16). For the numerical realization of the above method, by using the reproducing property and the definition of the empirical operators, it follows that

$$f^{\dagger}_{\gamma,\mathbf{z}} = \frac{1}{n}\sum_{i=1}^{n} k_\gamma(x, x_i)\alpha_i \qquad \alpha = (S_{\gamma,\mathbf{x}} S^*_{\gamma,\mathbf{x}})^{-1}\mathbf{y} = n{K_{\gamma,\mathbf{x}}}^{-1}\mathbf{y},$$

or equivalently

$$f^{\dagger}_{\gamma,\mathbf{z}} = \sum_{i=1}^{n} k_\gamma(x, x_i)c_i \qquad c = (K_{\gamma,\mathbf{x}})^{-1}\mathbf{y}.$$

On the other hand without assuming a positive definite kernel than the linear systems does not have a solution and least squares need be considered,

$$\min_{f\in\mathcal{H}_\gamma} \sum_{i=1}^{n} (y_i - f(x_i))^2.$$

that is

$$\min_{f\in\mathcal{H}_\gamma} \|\mathbf{y} - S_{\gamma,\mathbf{x}} f\|_n^2 ,$$

where $\|\cdot\|_n$ denotes the norm in $L^2_{\hat{\rho}}$. The optimality condition gives the following equivalent linear system,

$$S^*_{\gamma,\mathbf{x}} S_{\gamma,\mathbf{x}} f = S^*_{\gamma,\mathbf{x}}\mathbf{y}.$$

and the estimator can be written as

$$f^{\dagger}_{\gamma,\mathbf{z}} = {T_{\gamma,\mathbf{x}}}^{\dagger} S^*_{\gamma,\mathbf{x}}\mathbf{y} .$$

Appendix E Proof of Lemma 1: Control of the condition number

Proof. As in [17] the following identity holds true for all $c \in \mathbb{R}^n$

$$c^T K_{\gamma,\mathbf{x}} c = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \mathcal{F}Q_\gamma(\xi) \left| \sum_{i=1}^{n} c_i e^{i\langle x_i,\xi\rangle} \right| d\xi .$$

Fix $R > 0$ and set

$$\varphi_\gamma(R) := \inf_{\|\xi\|\le R} \mathcal{F}Q_\gamma(\xi),$$

since $\mathcal{F}Q_\gamma \ge 0$

$$\begin{aligned} c^T K_{\gamma,\mathbf{x}} c \ge & \frac{1}{(2\pi)^{d/2}} \int_{B(0,R)} \mathcal{F}Q_\gamma(\xi) \left| \sum_{i=1}^{n} c_i e^{i\langle x_i,\xi\rangle} \right| d\xi \\ \ge & \frac{1}{(2\pi)^{d/2}} \inf_{\|\xi\|\le R} \mathcal{F}Q_\gamma(\xi) \int_{B(0,R)} \left| \sum_{i=1}^{n} c_i e^{i\langle x_i,\xi\rangle} \right| d\xi \\ \ge & \frac{1}{(2\pi)^{d/2}} \varphi_\gamma(R) \left(1 - \frac{d\pi^2}{4q^2R^2}\right) q^{-d}(2\pi)^d \left(\frac{\pi}{4}\right)^{2d} \|c\|_2^2 \end{aligned}$$

where the last bound follows by Corollary 2.3 of [17] by replacing q_∞ with a lower bound $0 < q \leq q_\infty$ (see Theorem 8.1 in [**?**] for further details) and provided that

$$R \geq \frac{\sqrt{d}\pi}{2q}. \tag{17}$$

Hence we have the lower bound for the smallest eigenvalues of the kernel matrix

$$\sigma_{\min}(K_{\gamma,\mathbf{x}}) \geq \left(\frac{\pi^5}{2^7}\right)^{d/2} \varphi_\gamma(R) \left(1 - \frac{d\pi^2}{4q^2R^2}\right) q^{-d} .$$

Recalling that Q_γ is Matern kernel (2), by 12

$$\varphi_\gamma(R) = 2^{s-1}\Gamma(s)\frac{\gamma^d}{(1+\gamma^2R^2)^s} \geq 2^{s-1}\Gamma(s)\frac{\gamma^d}{(2\gamma^2)^s R^{2s}} \qquad \text{assuming} \quad \gamma^2 \geq \frac{1}{R^2}$$

son that

$$\sigma_{\min}(K_{\gamma,\mathbf{x}}) \geq \left(\frac{\pi^5}{2^7}\right)^{d/2} 2^{s-1}\Gamma(s)\frac{\gamma^d}{(2\gamma^2)^s} \underbrace{\frac{1}{R^{2s}}\left(1 - \frac{d\pi^2}{4q^2R^2}\right)}_{h(R)} q^{-d} .$$

With the choice

$$R^2 = \frac{d\pi^2}{4q^2}\left(1 + \frac{1}{s}\right) = \frac{d\pi^2}{4q^2}\left(\frac{s+1}{s}\right) =: \frac{1}{\gamma_q^2}$$

we obtain

$$\sigma_{\min}(K_{\gamma,\mathbf{x}}) \geq \frac{c_2(d,s)}{\gamma^{2s-d}}\, q^{2s-d} \qquad \gamma \geq \gamma_q$$

where

$$c_2(d,s) = \left(\frac{\pi^5}{2^7}\right)^{d/2} 2^{s-1}\Gamma(s)\left(\frac{2s}{d\pi^2(s+1)}\right)^s \left(\frac{1}{s+1}\right)$$

which conclude the proof of the second case.
For the first case, observe that

$$\varphi_\gamma(R) = 2^{s-1}\Gamma(s)\frac{\gamma^d}{(1+\gamma^2R^2)^s} \geq 2^{s-1}\Gamma(s)\frac{\gamma^d}{(2)^s} = \frac{1}{2}\Gamma(s)\,\gamma^d . \qquad 0 < \gamma < \gamma_q$$

so that, as above,

$$\sigma_{\min}(K_{\gamma,\mathbf{x}}) \geq c_1(d,s)\gamma^d\, q^{-d} \qquad 0 < \gamma < \gamma_q$$

where

$$c_1(d,s) = \left(\frac{\pi^5}{2^7}\right)^{d/2} \frac{1}{2}\,\Gamma(s)\left(\frac{1}{s+1}\right) .$$

□

Remark 1. *The constant $c_2(d,s)$ can be approximated as*

$$\begin{aligned}
c_2(d,s) &= \left(\frac{\pi^5}{2^7}\right)^{d/2} 2^{s-1}\Gamma(s)\left(\frac{2s}{d\pi^2(s+1)}\right)^s \frac{1}{s+1} \\
&\approx \frac{1}{2}\left(\frac{\pi^5}{2^7}\right)^{d/2} 2^s\ \Gamma(s)\left(\frac{2}{d\pi^2}\right)^s \frac{1}{s+1} \\
&\approx \frac{1}{2}\left(\frac{\pi^5}{2^7}\right)^{d/2} 2^s\ \frac{(s-1)^{s-1}}{e^{s-1}}\sqrt{2\pi(s-1)}\left(\frac{2}{d\pi^2}\right)^s \frac{1}{s+1} \\
&\approx \frac{1}{2}\left(\frac{\pi^5}{2^7}\right)^{d/2} 2^s\ \frac{(s-1)^{s}}{e^{s-1}}\sqrt{2\pi(s-1)}\left(\frac{2}{d\pi^2}\right)^s \frac{1}{(s+1)(s-1)} \\
&\approx \frac{1}{2}\left(\frac{\pi^5}{2^7}\right)^{d/2} 2^s\ \frac{1}{e^{s-1}}\left(\frac{2(s-1)}{d\pi^2}\right)^s \frac{\sqrt{2\pi(s-1)}}{(s+1)(s-1)}
\end{aligned}$$

where in the third line we approximate the Gamma function with Stirling's formula. If s is sufficiently close to $d/2$ this constant can be approximated as $10^{-d/2}$.

In the next corollary we specialize Lemma 1 by exploiting Assumption 1 on the data points.

Corollary 1.
Under the same assumptions of Lemma 1 and Assumption 1, let $0 < \delta < 1$ and assuming

$$\gamma \geq \frac{\sqrt[d]{\delta\ \mathrm{vol}\,(\mathcal{X})}}{2\pi\sqrt{d}}\sqrt{\frac{s}{s+1}} \cdot \frac{1}{n^{2/d}}$$

then with probability greater than $1-\delta$ it holds

$$\sigma_{\min}(K_{\gamma,\mathbf{x}}) \geq \frac{c_2(d,s)\delta^{\frac{2s-d}{d}}}{\gamma^{2s-d}} \frac{1}{n^{\frac{2}{d}(2s-d)}}\ . \tag{18}$$

Moreover if

$$n \leq \sqrt{\delta\ \mathrm{vol}(\mathcal{X})}\left(\frac{8^{\frac{1}{2s-d}}}{4}\right)^{\frac{d}{2}} \tag{19}$$

then

$$\sigma_{\min}(K_{\gamma,\mathbf{x}}) \geq \frac{c_2(d,s)}{8\gamma^{2s-d}} \tag{20}$$

where $c_2(d,s)$ defined in Lemma 1 .

Proof.
The first part follows directly from Lemma 1 and choosing the lower bound q on q_∞ given in (31).
For the second part, with the choice of $t = 8^{-\frac{1}{2s-d}}$ (30) reads

$$\mathbb{P}\left[q_\infty \leq 8^{-\frac{1}{2s-d}}\right] \leq n^2\ \frac{1}{\mathrm{vol}(\mathcal{X})}\Big(\frac{4}{8^{\frac{1}{2s-d}}}\Big)^d \leq \delta$$

where the last inequality follows from the assumption on n. Then, from Lemma 1 with probability greater than $1-\delta$

$$\sigma_{\min}(K_{\gamma,\mathbf{x}}) \geq \frac{c_2(d,s)}{\gamma^{2s-d}}\, q_\infty^{2s-d} \geq \frac{c_2(d,s)}{8\gamma^{2s-d}}$$

□

Appendix F Proof of Proposition 1: Variance

Proof. Given $0 < \delta < 1$, under Assumption 2 it holds that with probability grater than $1-\delta$

$$\begin{aligned}
\left\|S_\gamma T_{\gamma,\mathbf{x}}^\dagger S_{\gamma,\mathbf{x}}^* \widehat{\varepsilon}\right\|_\rho^2 &= \left\langle S_\gamma T_{\gamma,\mathbf{x}}^\dagger S_{\gamma,\mathbf{x}}^* \widehat{\varepsilon}, S_\gamma T_{\gamma,\mathbf{x}}^\dagger S_{\gamma,\mathbf{x}}^* \widehat{\varepsilon}\right\rangle_\rho \\
&= \left\langle \widehat{\varepsilon}, S_{\gamma,\mathbf{x}} T_{\gamma,\mathbf{x}}^\dagger T_\gamma T_{\gamma,\mathbf{x}}^\dagger S_{\gamma,\mathbf{x}}^* \widehat{\varepsilon}\right\rangle_n \\
&= \frac{1}{n}\left\langle \widehat{\varepsilon}, S_{\gamma,\mathbf{x}} T_{\gamma,\mathbf{x}}^\dagger T_\gamma T_{\gamma,\mathbf{x}}^\dagger S_{\gamma,\mathbf{x}}^* \widehat{\varepsilon}\right\rangle \\
&\leq \frac{c_\delta \sigma^2}{n}\left[\mathrm{Tr}\left(S_{\gamma,\mathbf{x}} T_{\gamma,\mathbf{x}}^\dagger T_\gamma T_{\gamma,\mathbf{x}}^\dagger S_{\gamma,\mathbf{x}}^*\right)\right] \\
&= \frac{c_\delta \sigma^2}{n}\left[\mathrm{Tr}\left(T_\gamma T_{\gamma,\mathbf{x}}^\dagger\right)\right]
\end{aligned}$$

where the inequality follows from Lemma 10 with $c_\delta = 4\log\left(\frac{1}{\delta}\right) + 2$.
Observe that the key quantity $\mathrm{Tr}\left(T_\gamma T_{\gamma,\mathbf{x}}^\dagger\right)$ depends on the ratio between the eigenvalues of the covariance operator T_γ and its empirical approximation $T_{\gamma,\mathbf{x}}$. Observe that

$$T_\gamma {T_{\gamma,\mathbf{x}}}^\dagger = (T_\gamma + \lambda \mathrm{I})^{1/2} T_\gamma (T_\gamma + \lambda \mathrm{I})^{-1} (T_\gamma + \lambda \mathrm{I})^{1/2} (T_{\gamma,\mathbf{x}} + \lambda \mathrm{I})^{-1/2} (T_{\gamma,\mathbf{x}} + \lambda \mathrm{I}) {T_{\gamma,\mathbf{x}}}^\dagger (T_{\gamma,\mathbf{x}} + \lambda \mathrm{I})^{-1/2},$$

then the cyclic property of the trace and Hölder inequality for Schatten norms imply

$$\begin{aligned}
\text{Variance} &\leq \frac{c_\delta \sigma^2}{n\delta}\,\mathrm{Tr}\left(T_\gamma T_{\gamma,\mathbf{x}}^\dagger\right) \\
&\leq \frac{c_\delta \sigma^2}{n\delta}\,\mathrm{Tr}\left(T_\gamma (T_\gamma + \lambda \mathrm{I})^{-1}\right) \left\|(T_\gamma + \lambda \mathrm{I})^{1/2} (T_{\gamma,\mathbf{x}} + \lambda \mathrm{I})^{-1/2}\right\|_\infty^2 \left\|(T_{\gamma,\mathbf{x}} + \lambda \mathrm{I}) T_{\gamma,\mathbf{x}}^\dagger\right\|_\infty \\
&= \frac{c_\delta \sigma^2}{n\delta} \mathcal{N}_\gamma(\lambda) \left\|(T_\gamma + \lambda \mathrm{I})^{1/2} (T_{\gamma,\mathbf{x}} + \lambda \mathrm{I})^{-1/2}\right\|_\infty^2 \left(1 + \frac{\lambda}{\sigma_{\min}\left(T_{\gamma,\mathbf{x}}\right)}\right) \\
&= \frac{c_\delta \sigma^2}{\delta} \left\|(T_\gamma + \lambda \mathrm{I})^{1/2} (T_{\gamma,\mathbf{x}} + \lambda \mathrm{I})^{-1/2}\right\|_\infty^2 \left(\frac{\mathcal{N}_\gamma(\lambda)}{n} + \frac{\lambda \mathcal{N}_\gamma(\lambda)}{\sigma_{\min}\left(K_{\gamma,\mathbf{x}}\right)}\right)
\end{aligned} \tag{21}$$

where $\mathcal{N}_\gamma(\lambda) = \mathrm{Tr}\left(T_\gamma (T_\gamma + \lambda \mathrm{I})^{-1}\right)$ are the degrees of freedom of T_γ, $\sigma_{\min}\left(T_{\gamma,\mathbf{x}}\right)$ is the smallest non-zero eigenvalue of $T_{\gamma,\mathbf{x}}$ and in the last equality we use the following fact

$$\sigma_{\min}\left(T_{\gamma,\mathbf{x}}\right) = \frac{1}{n}\sigma_{\min}\left(K_{\gamma,\mathbf{x}}\right)$$

where $\sigma_{\min}(K_{\gamma,\mathbf{x}})$ denotes the smallest eigenvalue of the kernel matrix $K_{\gamma,\mathbf{x}}$. Lemma 7 implies that, given $0 < \delta < 1$, then for any

$$\lambda \geq \frac{9}{n} \log \frac{n}{\delta}$$

with probability greater than $1 - \delta$

$$\frac{2}{3} \leq \left\| (T_\gamma + \lambda \, \mathrm{I})^{\frac{1}{2}} (T_{\gamma,\mathbf{x}} + \lambda \, \mathrm{I})^{-\frac{1}{2}} \right\|_\infty^2 \leq 2 \, .$$

Denoting

$$\lambda_n = \frac{9}{n} \log \frac{n}{\delta} \propto \frac{\log n}{n}$$

and we obtain

$$\text{Variance} \leq \frac{2c_\delta \sigma^2}{\delta} \left(\frac{\mathcal{N}_\gamma(\lambda_n)}{n} + \frac{\lambda_n \mathcal{N}_\gamma(\lambda_n)}{\sigma_{\min}(K_{\gamma,\mathbf{x}})} \right) \leq 4c_\delta \sigma^2 \frac{\lambda_n \mathcal{N}_\gamma(\lambda_n)}{\sigma_{\min}(K_{\gamma,\mathbf{x}})} \, . \tag{22}$$

By Lemma 4

$$\mathcal{N}_\gamma(\lambda) \leq c \, \gamma^{-\frac{d}{2s}(2s-d)} \, \lambda^{-\frac{d}{2s}}$$

with a suitable constant c that depends only on $d, s, \mathcal{X}$, but not on γ and λ. Moreover since

$$\lambda \mathcal{N}_\gamma(\lambda) \leq \operatorname{Tr} T_\gamma \leq \kappa_s$$

then

$$\text{Variance} \leq \frac{4c_\delta \sigma^2 \kappa_s}{\delta \sigma_{\min}(K_{\gamma,\mathbf{x}})} \, . \tag{23}$$

By combining these two bounds we have that

$$\text{Variance} \leq \frac{4c_\delta \sigma^2}{\delta} \frac{\min\left\{ c \, \gamma^{-\frac{d}{2s}(2s-d)} \, \lambda^{1-\frac{d}{2s}}, \kappa_s \right\}}{\sigma_{\min}(K_{\gamma,\mathbf{x}})} \, .$$

We split the proof according to the condition on n.
If $n \leq N_{s,d}$ from 20 with probability at least $1 - \delta$ it holds

$$\sigma_{\min}(K_{\gamma,\mathbf{x}}) \geq \frac{c_2(d,s)}{8\gamma^{2s-d}} \qquad \text{if} \quad \gamma \geq \frac{\sqrt[d]{\delta \ \operatorname{vol}(\mathcal{X})}}{2\pi\sqrt{d}} \sqrt{\frac{s}{s+1}} \cdot \frac{1}{n^{2/d}} \, . \tag{24}$$

Hence, we get with probability greater than $1 - 3\delta$

$$\begin{aligned}
\text{Variance} &\leq \frac{4c_\delta \sigma^2}{\delta} \frac{\lambda_n \mathcal{N}_\gamma(\lambda_n)}{\sigma_{\min}(K_{\gamma,\mathbf{x}})} \\
&\leq \frac{32 c_\delta \sigma^2}{\delta c_2(d,s)} \gamma^{(2s-d)} \min\left\{ c\gamma^{-\frac{d}{2s}(2s-d)} \lambda_n^{1-\frac{d}{2s}}, \kappa_s \right\} \\
&\leq \frac{32 c_\delta \sigma^2}{\delta c_2(d,s)} \gamma^{(2s-d)} \min\left\{ c\gamma^{-\frac{d}{2s}(2s-d)} \left(\frac{\log n}{n} \right)^{\frac{1}{2s}(2s-d)}, \kappa_s \right\} \, .
\end{aligned}$$

If $n \geq N_{s,d}$, bound (18) implies that with probability greater than $1 - 3\delta$

$$\text{Variance} \leq \frac{4c_\delta \sigma^2}{c_2(d,s)\delta^{\frac{2s}{d}}} \gamma^{(2s-d)} n^{\frac{2}{d}(2s-d)} \min\left\{ c\gamma^{-\frac{d}{2s}(2s-d)} \left(\frac{\log n}{n}\right)^{\frac{1}{2s}(2s-d)}, \kappa_s \right\} .$$

□

Appendix G Proof of Proposition 2: Interpolation error

Proof. In order to study the interpolation error, we first show that the choice of $\mathcal{H}_\gamma$ as hypothesis space is equivalent to fix $\mathcal{H}_1$ and to rescale the input space $\mathcal{X}$. Given $\gamma > 0$ define the dilation diffeomorphism

$$R_\gamma \colon \mathbb{R}^d \to \mathbb{R}^d \qquad x \mapsto \frac{x}{\gamma}$$

so that $Q_\gamma = Q_1 \circ R_\gamma$. Define the feature map

$$F_\gamma \colon \mathbb{R}^d \to \mathcal{H}_1 \qquad F_\gamma(x) = k_{1,R_\gamma(x)}.$$

Since

$$\langle F_\gamma(x), F_\gamma(x') \rangle_{\mathcal{H}_1} = k_\gamma(x,x') \qquad \text{and} \qquad \langle f, F_\gamma(x) \rangle_{\mathcal{H}_1} = 0 \ \forall x \in \mathcal{X} \implies f = 0$$

then the map

$$\begin{aligned} U_\gamma \colon \mathcal{H}_1 &\to \mathcal{H}_\gamma \\ f(x) &\mapsto \langle f, F_\gamma(x) \rangle_{\mathcal{H}_1} = f\left(R_\gamma(x)\right) = f \circ R_\gamma \end{aligned}$$

is a unitary operator. Denoted by $(e_i)_i$ the canonical base of $\mathbb{R}^d$, then

$$S_{\gamma,\mathbf{z}}(\cdot) = \sum_{i=1}^n \langle k_{\gamma,x_i}, \cdot \rangle_{\mathcal{H}_\gamma} e_i = \sum_{i=1}^n \left\langle U_\gamma k_{1,R_\gamma(x_i)}, \cdot \right\rangle_{\mathcal{H}_\gamma} e_i = \sum_{i=1}^n \left\langle k_{1,R_\gamma(x_i)}, U_\gamma^*(\cdot) \right\rangle_{\mathcal{H}_1} e_i$$

where $R_\gamma(\mathbf{z}) = (R_\gamma(x_1), y_1, \ldots, R_\gamma(x_n), y_n)$, so that

$$\begin{aligned} S_{\gamma,\mathbf{z}} &= S_{1,R_\gamma(\mathbf{z})} U_\gamma^* \\ S_{\gamma,\mathbf{z}}^* &= U_\gamma S_{1,R_\gamma(\mathbf{z})}^* \\ T_{\gamma,\mathbf{z}} &= S_{\gamma,\mathbf{z}}^* S_{\gamma,\mathbf{z}} = U_\gamma S_{1,R_\gamma(\mathbf{z})}^* S_{1,R_\gamma(\mathbf{z})} U_\gamma^* = U_\gamma T_{1,R_\gamma(\mathbf{z})} U_\gamma^* . \end{aligned}$$

Hence

$$f_{\gamma,\mathbf{z}}^\dagger = T_{\gamma,\mathbf{z}}^\dagger S_{\gamma,\mathbf{z}}^* \mathbf{y} = U_\gamma T_{1,R_\gamma(\mathbf{z})}^\dagger U_\gamma^* U_\gamma S_{1,R_\gamma(\mathbf{z})}^* \mathbf{y} = U_\gamma T_{1,R_\gamma(\mathbf{z})}^\dagger S_{1,R_\gamma(\mathbf{z})}^* \mathbf{y} = U_\gamma f_{1,R_\gamma(\mathbf{z})}^\dagger . \tag{25}$$

Define the rescaled input space $\mathcal{X}_\gamma = R_\gamma(\mathcal{X})$, so that R_γ is a diffeomorphism from $\mathcal{X}$ to $\mathcal{X}_\gamma$, and define the probability distribution ρ_γ on $\mathcal{X}_\gamma \times \mathbb{R}$ as the pushforward measure of ρ

$$\rho_\gamma = (R_\gamma \times \mathrm{I})_* (\rho)$$

so that

$$\int_{\mathbb{R}^d \times \mathbb{R}} f(x,y) \ \mathrm{d}\rho_\gamma(x,y) = \int_{\mathbb{R}^d \times \mathbb{R}} f\left(R_\gamma(x), y\right) \ \mathrm{d}\rho(x,y).$$

A simple computation shows that the marginal distribution $\rho_{\mathcal{X}_\gamma}$ and the conditional distribution $\rho_\gamma(\cdot \mid x)$ of $\rho_\mathcal{X}$ are

$$\rho_{\mathcal{X}_\rho,\gamma} = (R_\gamma)_* \rho_\mathcal{X} \qquad \rho_\gamma(\cdot \mid x) = \rho(\cdot \mid R_\gamma^{-1}(x)), \tag{26}$$

so that it holds true

$$f_\rho(x) = f_{\rho_\gamma}(R_\gamma(x))$$

where f_ρ and f_{ρ_γ} are the regression functions with respect to ρ and ρ_γ, respectively.
Finally, by (26) the operator

$$\begin{aligned} W_\gamma \colon L^2_{\rho_\gamma} &\to L^2_\rho \\ f &\mapsto f \circ R_\gamma \end{aligned}$$

is unitary and

$$f_\rho = W_\gamma f_{\rho_\gamma} \tag{27}$$

Now regarding the population operators we have

$$\begin{aligned} S_\gamma \colon \mathcal{H}_\gamma &\to L^2_\rho \\ S_{1,\gamma} \colon \mathcal{H}_1 &\to L^2_{\rho_\gamma} \end{aligned}$$

and it holds

$$S_\gamma U_\gamma = W_\gamma S_{1,\gamma} \ . \tag{28}$$

By (25), (27) and (28), the bias term becomes

$$\begin{aligned} \left\| S_\gamma f^\dagger_{\gamma,\mathbf{z}} - f_\rho \right\|^2_{L^2_\rho} &= \left\| S_\gamma U_\gamma f^\dagger_{1,R_\gamma(\mathbf{z})} - W_\gamma f_{\rho_\gamma} \right\|^2_{L^2_\rho} \\ &= \left\| W_\gamma S_{1,\gamma} f^\dagger_{1,R_\gamma(\mathbf{z})} - W_\gamma f_{\rho_\gamma} \right\|^2_{L^2_\rho} \\ &= \left\| W_\gamma \left(S_{1,\gamma} f^\dagger_{1,R_\gamma(\mathbf{z})} - f_{\rho_\gamma} \right) \right\|^2_{L^2_\rho} \\ &= \left\| S_{1,\gamma} f^\dagger_{1,R_\gamma(\mathbf{z})} - f_{\rho_\gamma} \right\|^2_{L^2_{\rho_\gamma}} \ . \end{aligned}$$

Since the Sobolev norm $\|\cdot\|_{W^s_2}$ is equivalent to the RKHS norm $\|\mathcal{H}_1\|$ then Assumption 3 implies exists $g \in \mathcal{H}_1$ such that $\|f_\rho\|_{W^s_2} = \|g\|_{\mathcal{H}_1}$ and

$$f_\rho = S_\gamma g_\gamma \quad \text{where} \quad g_\gamma = U_\gamma g \ ,$$

due to (27) and (28) can be rewritten as

$$W_\gamma f_{\rho_\gamma} = f_\rho = S_\gamma U_\gamma g = W_\gamma S_{1,\gamma}\, g \implies f_{\rho_\gamma} = S_{1,\gamma}\, g$$

moreover, since

$$\|g_\gamma\|_{\mathcal{H}_\gamma} = \|U_\gamma g\|_{\mathcal{H}_\gamma} = \|g\|_{\mathcal{H}_1}$$

we obtain that

$$\left\|S_\gamma f^\dagger_{\gamma,\mathbf{z}} - f_\rho\right\|^2_{L^2_\rho} = \left\|S_{1,\gamma}\left(f^\dagger_{1,R_\gamma(\mathbf{z})} - g\right)\right\|^2_{L^2_{\rho_\gamma}} .$$

This shows that we can always consider $\gamma = 1$ by rescaling the probability distribution ρ with ρ_γ and the training set from $\mathbf{z}$ to $R_\gamma(\mathbf{z})$.
By denoting the projection operator $P_{1,R_\gamma(\mathbf{z})} = T^\dagger_{1,R_\gamma(\mathbf{z})} T_{1,R_\gamma(\mathbf{z})}$ and $T_1 = S_1^* S_1$ the bias becomes

$$\begin{aligned}
\left\|S_\gamma f^\dagger_{\gamma,\mathbf{z}} - f_\rho\right\|^2_{L^2_\rho} &= \left\|S_{1,\gamma}\left(P_{1,R_\gamma(\mathbf{z})} - \mathrm{I}\right) g\right\|^2_{L^2_{\rho_\gamma}} \\
&= \left\|T_1^{1/2}\left(P_{1,R_\gamma(\mathbf{z})} - \mathrm{I}\right) g\right\|^2_{\mathcal{H}_1} \\
&= \left\|\left(T_1^{1/2} - T^{1/2}_{1,R_\gamma(\mathbf{z})}\right)\left(P_{1,R_\gamma(\mathbf{z})} - \mathrm{I}\right) g\right\|^2_{\mathcal{H}_1} \\
&\leq \left\|T_1^{1/2} - T^{1/2}_{1,R_\gamma(\mathbf{z})}\right\|^2 \left\|\left(P_{1,R_\gamma(\mathbf{z})} - \mathrm{I}\right) g\right\|^2_{\mathcal{H}_1} \\
&\leq \left\|T_1 - T_{1,R_\gamma(\mathbf{z})}\right\| \|g\|^2_{\mathcal{H}_1}
\end{aligned}$$

where in the third equality we use that $T^{1/2}_{1,R_\gamma(\mathbf{z})}\left(P_{1,R_\gamma(\mathbf{z})} - \mathrm{I}\right) = 0$ and in the last inequality we use that the operator $P_{1,R_\gamma(\mathbf{z})} - \mathrm{I}$ is a projection.
We conclude applying Lemma 6, which states that, given $0 < \delta < 1/2$, then with probability greater than $1 - \delta$ it holds

$$\left\|T_1 - T_{1,R_\gamma(\mathbf{z})}\right\| \leq \frac{3\kappa_s}{\sqrt{n}} \log \frac{2}{\delta}$$

which complete the proof.

□

Appendix H Control of the effective dimension

Lemma 2. *Let $t \in \mathbb{N}_+$ and $P_\ell : \mathcal{H}_\gamma \mapsto \mathcal{H}_\gamma$ be a projection operator with rank smaller or equal than $\ell \in \mathbb{N}$. Let $\sigma_t(L_\gamma)$ be its t-th eigenvalue of the integral operator L_γ. Then it holds:*

$$\sum_{t>\ell} \sigma_t(T_\gamma) = \sum_{t>\ell} \sigma_t(L_\gamma) \leq \int_{\mathcal{X}} \|(\mathrm{I} - P_\ell) k_{\gamma,x}\|^2_{\mathcal{H}_\gamma}\, d\rho_{\mathcal{X}}(x) \leq \sup_{x\in\mathcal{X}} \|(\mathrm{I} - P_\ell) k_{\gamma,x}\|^2_{\mathcal{H}_\gamma}$$

Proof. Let $\widetilde{L_\gamma} = S_\gamma P_\ell S_\gamma^*$. Notice that $\widetilde{L_\gamma}$ is well-defined on $L^2_\rho \mapsto L^2_\rho$ since the rank of P_ℓ is smaller or equal than ℓ, then $\sigma_t(\widetilde{L_\gamma}) = 0$ for $t > \ell$ and so

$$\sum_{t>\ell} \sigma_t(L_\gamma) = \sum_{t>\ell} \sigma_t(L_\gamma) - \sigma_t(\widetilde{L_\gamma})$$

Now note that $\sigma_t(L_\gamma) \geq \sigma_t(\widetilde{L_\gamma})$ for any $t \in \mathbb{N}$ since $\mathrm{I} \succeq P_\ell$ and so $S_\gamma S_\gamma^* \succeq S_\gamma P_\ell S_\gamma^*$. Then

$$\sum_{t>\ell} \sigma_t(L_\gamma) - \sigma_t(\widetilde{L_\gamma}) \leq \sum_{t\in\mathbb{N}} \sigma_t(L_\gamma) - \sigma_t(\widetilde{L_\gamma}) = \mathrm{Tr}(L_\gamma - \widetilde{L_\gamma})$$

since $\mathrm{I} - P_\ell = (\mathrm{I} - P_\ell)^2$ by projection property so by the cyclicity of the trace and the fact that covariance operator T_γ is given by $T_\gamma = S_\gamma^* S_\gamma = \int k_{\gamma,x} \otimes k_{\gamma,x} \, \mathrm{d}\rho_{\mathcal{X}}(x)$ we have

$$\mathrm{Tr}(L_\gamma - \widetilde{L_\gamma}) = \mathrm{Tr}\left(S_\gamma(\mathrm{I}-P_\ell)S_\gamma^*\right) = \mathrm{Tr}\left(S_\gamma(\mathrm{I}-P_\ell)^2 S_\gamma^*\right) = \mathrm{Tr}\left((\mathrm{I}-P_\ell)S_\gamma^* S_\gamma(\mathrm{I}-P_\ell)\right) = \mathrm{Tr}((\mathrm{I}-P_\ell)T_\gamma(\mathrm{I}-P_\ell))$$

Finally by linearity of the trace and integral operator we have

$$\mathrm{Tr}((\mathrm{I}-P_\ell)T_\gamma(\mathrm{I}-P_\ell)) = \int \mathrm{Tr}\left((\mathrm{I}-P_\ell)\left(k_{\gamma,x} \otimes k_{\gamma,x}\right)(\mathrm{I}-P_\ell)\right) \, \mathrm{d}\rho_{\mathcal{X}}(x) = \int \|(\mathrm{I}-P_\ell)k_{\gamma,x}\|^2_{\mathcal{H}_\gamma} \, \mathrm{d}\rho_{\mathcal{X}}(x)$$

where to prove the last equality let $v = (\mathrm{I}-P_\ell)k_{\gamma,x}$, we have

$$\begin{aligned}\mathrm{Tr}\left((\mathrm{I}-P_\ell)\left(k_{\gamma,x} \otimes k_{\gamma,x}\right)(\mathrm{I}-P_\ell)\right) &= \mathrm{Tr}\left(((\mathrm{I}-P_\ell)k_{\gamma,x}) \otimes ((\mathrm{I}-P_\ell)k_{\gamma,x})\right) \\ &= \mathrm{Tr}(v \otimes v) = \langle v, v\rangle_{\mathcal{H}_\gamma} = \|v\|^2_{\mathcal{H}_\gamma}\end{aligned} .$$

□

Lemma 3. *Let $A : \mathcal{H}_\gamma \to \mathcal{H}_\gamma$ be a bounded linear operator, then*

$$\sup_{x\in\mathcal{X}} \|Ak_{\gamma,x}\|^2_{\mathcal{H}_\gamma} \leq \sup_{\|f\|_{\mathcal{H}_\gamma}\leq 1} \|A^* f\|^2_{L_\infty(\mathcal{X})}$$

Proof. Since $\mathcal{H}_\gamma$ is a RKHS so $k_{\gamma,x} \in \mathcal{H}_\gamma$. By making use of the definition of the Hilbert norm and the fact that the evaluation functional is continuous for any $x \in \mathcal{X}$, we have:

$$\begin{aligned}\sup_{x\in\mathcal{X}} \|Ak_{\gamma,x}\|_{\mathcal{H}_\gamma} &= \sup_{\substack{x\in\mathcal{X} \\ \|f\|_{\mathcal{H}_\gamma}\leq 1}} \langle f, Ak_{\gamma,x}\rangle_{\mathcal{H}_\gamma} \\ &= \sup_{\|f\|_{\mathcal{H}_\gamma}\leq 1} \sup_{x\in\mathcal{X}} \langle f, Ak_{\gamma,x}\rangle_{\mathcal{H}_\gamma} \\ &\leq \sup_{\|f\|_{\mathcal{H}_\gamma}\leq 1} \sup_{x\in\mathcal{X}} \left|\langle A^* f, k_{\gamma,x}\rangle_{\mathcal{H}_\gamma}\right| \\ &= \sup_{\|f\|_{\mathcal{H}_\gamma}\leq 1} \sup_{x\in\mathcal{X}} |A^* f(x)| \\ &= \sup_{\|f\|_{\mathcal{H}_\gamma}\leq 1} \|A^* f\|_{C(\mathcal{X})} \\ &= \sup_{\|f\|_{\mathcal{H}_\gamma}\leq 1} \|A^* f\|_{L_\infty(\mathcal{X})}\end{aligned}$$

where the last equality holds true since $f \in C(\mathcal{X})$. □

Lemma 4 (Bound on the effective dimension)**.** *Let $s > \frac{d}{2}$. Then the effective dimension $\mathcal{N}_\gamma(\lambda)$ satisfies the following upper bound:*

$$\mathcal{N}(\lambda) \leq c\, \lambda^{-d/2s} \gamma^{-\frac{d}{2s}(2s-d)} + 1$$

where c is a constant which depends on $d, s, \mathcal{X}$.

Proof. Let $t \in \mathbb{N}$, lets consider the definition of effective dimension we have

$$\begin{aligned}
\mathcal{N}_\gamma(\lambda) &= \sum_{j\geq 1} \frac{\sigma_j\left(T_\gamma\right)}{\sigma_j\left(T_\gamma\right)+\lambda} \\
&\leq \sum_{j=1}^{t} \frac{\sigma_j\left(T_\gamma\right)}{\sigma_j\left(T_\gamma\right)+\lambda} + \lambda^{-1} \sum_{j>t} \sigma_j\left(T_\gamma\right) \\
&\leq \sum_{j=1}^{t} \frac{\sigma_j\left(T_\gamma\right)}{\sigma_j\left(T_\gamma\right)+\lambda} + \lambda^{-1} \sup_{x\in\mathcal{X}} \|(\mathrm{I}-P_\ell)k_{\gamma,x}\|^2_{\mathcal{H}_\gamma} \\
&\leq t + \lambda^{-1} \sup_{x\in\mathcal{X}} \|(\mathrm{I}-P_t)k_{\gamma,x}\|^2_{\mathcal{H}_\gamma}
\end{aligned}$$

where $P_t : \mathcal{H}_\gamma \mapsto \mathcal{H}_\gamma$ is a projection operator with rank smaller or equal than t. Lets start with the case $\gamma = 1$. We can choose P_t as the projection operator on the finite dimensional subspace

$$\mathrm{span}\{k_{1,x_1}, \ldots, k_{1,x_t}\} \qquad \text{s.t.} \quad P_t f(x_i) = f(x_i) \quad \forall i = 1\ldots, t\,, \forall f \in \mathcal{H}_1$$

where $\{x_1, \ldots, x_t\}$ are distributed over a d-dimensional grid over $\mathcal{X}$ such that the fill distance

$$h_t = \sup_{x\in\mathcal{X}} \min_{i\in\{1,\ldots,t\}} \|x - x_i\| \lesssim t^{-1/d}\,.$$

Now from Lemma 3 with $A = \mathrm{I} - P_t$ we have

$$\sup_{x\in\mathcal{X}} \|(\mathrm{I}-P_t)k_{1,x}\|^2_{\mathcal{H}_1} \leq \sup_{\|f\|_{\mathcal{H}_1}\leq 1} \|(\mathrm{I}-P_t)f\|^2_{L_\infty(\mathcal{X})}$$

and from Corollary 11.33 of [34] it holds that there exist a constant C (depending on $d, s, \mathcal{X}$) such that

$$\sup_{\|f\|_{\mathcal{H}_1}\leq 1} \|(\mathrm{I}-P_t)f\|^2_{L_\infty(\mathcal{X})} \leq C h_t^{2s-d} \lesssim C t^{-\frac{2s}{d}+1}\,.$$

Now we can generalize to differnt γ by rescaling the space $\mathcal{X}$ throw the map $R_\gamma(x) = \frac{x}{\gamma}$ and with the same consideration we did in the previous case it holds that

$$\sup_{x\in\mathcal{X}} \|(\mathrm{I}-P_t)k_{\gamma,x}\|^2_{\mathcal{H}_\gamma} \leq \sup_{\|f\|_{\mathcal{H}_\gamma}\leq 1} \|(\mathrm{I}-P_t)f\|^2_{L_\infty(\mathcal{X})} \leq C\left(\frac{h_t}{\gamma}\right)^{2s-d} \lesssim C\gamma^{d-2s} t^{-\frac{2s}{d}+1}\,.$$

since the infinity norm is invariant on the rescaling.
Finally we obtain that

$$\mathcal{N}_\gamma(\lambda) \leq t + \lambda^{-1} \sup_{x\in\mathcal{X}} \|(\mathrm{I} - P_t)k_{\gamma,x}\|^2_{\mathcal{H}_\gamma} \leq t + \lambda^{-1} C\gamma^{d-2s} t^{-\frac{2s}{d}+1} \quad \forall t \in \mathbb{N}$$

and optimizing over t, hence taking $t = \text{round}\left[C^{d/2s}\lambda^{-d/2s}\gamma^{-\frac{d}{2s}(2s-d)}\right]$ we obtain

$$\mathcal{N}(\lambda) \leq 2C^{d/2s}\lambda^{-d/2s}\gamma^{-\frac{d}{2s}(2s-d)} + 1$$

and conclude by taking denoting $c = 2C^{d/2s}$. □

Appendix I Separation distance

Under Assumption 1 we study a probabilistic lower bound on the separation distance

$$q_\infty = \frac{1}{2}\min_{i\neq j} \|x_i - x_j\|_\infty \ .$$

Lemma 5 (Bound on the separation distance)**.** *Under Assumption 1, let $\delta > 0$ then with probability greater than $1 - \delta$*

$$q_\infty \geq \frac{\sqrt[d]{\delta \ \mathrm{vol}\,(\mathcal{X})}}{4} \cdot \frac{1}{n^{2/d}}$$

Proof. We have that

$$\begin{aligned}
\mathbb{P}\left[\frac{1}{2}\min_{i\neq j}\|x_i - x_j\|_\infty \leq t\right] &= \mathbb{P}\left[\exists i < j \mid \|x_i - x_j\|_\infty \leq 2t\right] \\
&\leq \mathbb{P}\left[\bigcup_{i<j}\{\|x_i - x_j\|_\infty \leq 2t\}\right] \\
&\leq \frac{n(n-1)}{2}\mathbb{P}\left[\|x_1 - x_2\|_\infty \leq 2t\right] \\
&\leq n^2 \ \frac{(4t)^d}{\mathrm{vol}(\mathcal{X})}
\end{aligned}$$

since

$$\begin{aligned}
\mathbb{P}\left[\|x_1 - x_2\|_\infty \leq 2t\right] &= \int_{\mathcal{X}} \mathbb{P}\left[\|x_1 - x_2\|_\infty \leq 2t\right] \ \mathrm{d}\rho_{\mathcal{X}}(x_2) \\
&= \int_{\mathcal{X}} \mathbb{P}\left[x_1 \in B_\infty(x_2, 2t)\right] \ \mathrm{d}\rho_{\mathcal{X}}(x_2) \\
&= \int_{\mathcal{X}} \mathrm{vol}\left(B_\infty(x_2, 2t) \cap \mathcal{X}\right) \ \mathrm{d}\rho_{\mathcal{X}}(x_2) \\
&\leq \frac{\mathrm{vol}\left(B_\infty(x_2, 2t)\right)}{\mathrm{vol}\left(\mathcal{X}\right)} \\
&= \frac{(4t)^d}{\mathrm{vol}\left(\mathcal{X}\right)}
\end{aligned} \tag{29}$$

where $B_\infty(x,t)$ denotes the ball centered at x and radius t with respect to the sup-norm in $\mathbb{R}^d$. Hence

$$\mathbb{P}\left[q_\infty \leq t\right] \leq n^2 \, \frac{(4t)^d}{\mathrm{vol}(\mathcal{X})} \tag{30}$$

Fix $\delta > 0$, then with probability at least $1-\delta$

$$q_\infty \geq \frac{\sqrt[d]{\delta \, \mathrm{vol}\left(\mathcal{X}\right)}}{4} \, n^{-2/d} \, . \tag{31}$$

□

Appendix J Complementary Lemmas

Next lemma is a standard concentration inequalities (see Lemma 4 in [16] for details).

Lemma 6. *Let $0 < \delta < 1/2$. It holds with probability at least $1-\delta$:*

$$\|T_\gamma - T_{\gamma,\mathbf{x}}\| \leq \|T_\gamma - T_{\gamma,\mathbf{x}}\|_{HS} \leq \frac{3\kappa_s}{\sqrt{n}} \log \frac{2}{\delta} \, .$$

Here, $\|\cdot\|_{HS}$ denotes the Hilbert-Schmidt norm.

Next lemma is a concentration result on the product of the regularized empirical covariance and population covariance, see [27] for details.

Lemma 7. *Let $0 < \delta < 1$ and $\lambda \geq \frac{9}{n} \log \frac{n}{\delta}$. It holds with probability greater than $1-\delta$:*

$$\sqrt{\frac{2}{3}} \leq \left\| \left(T_\gamma + \lambda \, \mathrm{I}\right)^{1/2} \left(T_{\gamma,\mathbf{x}} + \lambda \, \mathrm{I}\right)^{-1/2} \right\| \leq \sqrt{2} \, .$$

Lemma 8 (Bound on the trace of the kernel)**.** *Given the kernel functions k_γ defined in* (2) *then there exist a constant κ_s defined by*

$$\kappa_s = 2^{\frac{2s-d-2}{2}} \Gamma(s - d/2)$$

such that

$$\sup_{x,x' \in \mathcal{X}} k_\gamma(x,x') \leq \kappa_s \, .$$

Proof. Recall that

$$\sup_{x,x' \in \mathcal{X}} k_\gamma(x,x') = \sup_{x \in \mathcal{X}} k_\gamma(x,x) = Q_\gamma(0) = Q(0) = 2^{s-1}\Gamma(s) \int_{\mathbb{R}^d} \frac{1}{(1+\|\xi\|^2)^s} \, \mathrm{d}\xi.$$

Note that

$$\begin{aligned} Q(0) &= \lim_{z \to 0} \|z\|^{s-d/2} \, \mathbf{K}_{s-d/2}\left(\|z\|\right) \\ &\leq \|z\|^{s-d/2} \, 2^{s-\frac{d}{2}-1} \Gamma(s-d/2) \, \|z\|^{-(s-d/2)} = \kappa_s \end{aligned}$$

where in the second inequality we use Lemma 5.14 from [34] to bound the modified Bessel function $\mathbf{K}_{s-d/2}$. □

Lemma 9 (Covariance of the noise)**.** *Define*

$$\widehat{\varepsilon} = \mathbf{y} - \widehat{f_\rho}$$

where $\mathbf{y} = (y_1, \dots, y_n)$ *and* $\widehat{f_\rho} = (f_\rho(x_1), \dots, f_\rho(x_n))$*. Under Assumption 2 the random vector* $\widehat{\varepsilon}$ *has zero-mean (conditionally to* $\mathbf{x}$*) and the covariance satisfies*

$$\mathbb{E}_{\mathbf{y}} \left[\widehat{\varepsilon} \otimes \widehat{\varepsilon} \right] \preceq \frac{\sigma^2}{n} \, \mathrm{I}_{\mathbb{R}^n}$$

where $A \preceq B$ *is the partial order induced by the notion of positive operator and* $\mathrm{I}_{\mathbb{R}^n}$ *denotes the identity matrix of size* n*.*

Proof. For every $v \in \mathbb{R}^n$ we can compute

$$\begin{aligned}
\mathbb{E} \left[\langle \widehat{\varepsilon} \otimes \widehat{\varepsilon} \, v, v \rangle_n \mid \mathbf{x} \right] &= \mathbb{E} \left[\langle v, \widehat{\varepsilon} \rangle_n \langle v, \widehat{\varepsilon} \rangle_n \mid \mathbf{x} \right] \\
&= \frac{1}{n^2} \sum_{i,j=1}^{n} v_i v_j \, \mathbb{E} \left[\varepsilon_i \varepsilon_j \mid \mathbf{x} \right] \\
&\leq \frac{1}{n^2} \sum_{i,j=1}^{n} v_i v_j \delta_{i,j} \sigma^2 \\
&= \frac{\sigma^2}{n^2} \sum_{i=1}^{n} v_i^2 \\
&= \frac{\sigma^2}{n} \langle v, v \rangle_n
\end{aligned}$$

where we use that the variables ε_i are independents and the variance is bounded by σ^2. □

Lemma 10 (Lemma 9 in [2] or Lemma 35 in [**?**])**.**
Let the random variables $\varepsilon_1, \dots, \varepsilon_n$*, be conditionally independent given* $\mathbf{x}$ *and conditionally* σ^2*-subgaussian, that is, for all* $\lambda \in \mathbb{R}$

$$\mathbb{E} \left[\exp \left(\lambda \varepsilon_i \right) \mid \mathbf{x} \right] \leq \exp \left(\sigma^2 \lambda^2 / 2 \right) \quad \forall i = 1, \dots, n \, .$$

Suppose that $M \in \mathbb{R}^{n \times n}$ *is a.s. positive semidefinite, conditionally on* $\mathbf{x}$*. Then a.s. on* $\mathbf{x}$*, with conditional probability at least* $1 - \delta$*,*

$$\langle \varepsilon, M\varepsilon \rangle \leq \left(4 \log \left(\frac{1}{\delta} \right) + 2 \right) \sigma^2 \operatorname{tr}(M) \, .$$