On the Intrinsic Dimensions of Data in Kernel Learning

Rustem Takhanov
Nazarbayev University, Nazarbayev University Research Administration

Abstract

The manifold hypothesis suggests that the generalization performance of machine learning methods improves significantly when the intrinsic dimension of the input distribution's support is low. In the context of Kernel Ridge Regression (KRR), we investigate two alternative notions of intrinsic dimension. The first, denoted d_ϱ, is the upper Minkowski dimension defined with respect to the canonical metric induced by a kernel function K on a domain Ω. The second, denoted d_K, is the effective dimension, derived from the decay rate of Kolmogorov n-widths associated with K on Ω. Given a probability measure μ on Ω, we analyze the relationship between these n-widths and eigenvalues of the integral operator $\phi \mapsto \int_\Omega K(\cdot, x)\phi(x)\, d\mu(x)$. We show that, for a fixed domain Ω, the Kolmogorov n-widths characterize the worst-case eigenvalue decay across all probability measures μ supported on Ω. These eigenvalues are central to understanding the generalization behavior of constrained KRR, enabling us to derive an excess error bound of order $\mathcal{O}(n^{-\frac{2+d_K}{2+2d_K}+\varepsilon})$ for any $\varepsilon > 0$, when the training set size n is large. We also propose an algorithm that estimates upper bounds on the n-widths using only a finite sample from μ. For distributions close to uniform, we prove that ε-accurate upper bounds on all n-widths can be computed with high probability using at most $\mathcal{O}\left(\varepsilon^{-d_\varrho}\log\frac{1}{\varepsilon}\right)$ samples, with fewer required for small n.

Finally, we compute the effective dimension d_K for various fractal sets and present additional numerical experiments. Our results show that, for kernels such as the Laplace kernel, the effective dimension d_K can be significantly smaller than the Minkowski dimension d_ϱ, even though $d_K = d_\varrho$ provably holds on regular domains.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

1 INTRODUCTION

The manifold hypothesis posits that, within high-dimensional spaces, probability distributions tend to concentrate near lower-dimensional structures, suggesting that data primarily lies along smooth manifolds whose intrinsic dimension is significantly smaller than that of the ambient space. It is highly desirable for the sample complexity of a machine learning (ML) algorithm to depend on the intrinsic dimension of the data, rather than the ambient dimension of the space in which the data resides. When sample complexity scales with the ambient dimension instead, the algorithm suffers from the well-known "curse of dimensionality". Fortunately, many widely used algorithms avoid the curse of dimensionality. Notable examples include kernel regression [Bickel and Li, 2007] and k-nearest neighbors (k-NN) regression [Kpotufe, 2011].

A general theory of machine learning algorithms that adapt to intrinsic dimension has yet to be developed. A natural move in this direction is to establish generalization bounds for kernel methods that explicitly reflect the intrinsic dimension of the data. Some results of this kind rely explicitly on the assumption that the data lies on a low-dimensional Riemannian manifold and make full use of the manifold's geometric structure [Ozakin and Gray, 2009, McRae et al., 2020, Cheng and Xie, 2024].

In this paper, we study excess risk bounds for Kernel Ridge Regression. Unlike the previously cited works, we do not impose strong regularity assumptions on the underlying low-dimensional structures. Instead, we employ concepts from geometric measure theory to characterize the support of the data distribution. This approach enables us to handle less regular supports — those that are not only theoretically plausible but also observed in real-world datasets — such as sets with

fractional Hausdorff dimension (e.g., fractals). The main contribution of this paper is the identification of two alternative notions of intrinsic dimension, both defined solely in terms of the support of the data distribution, independent of the distribution itself. The first notion is the upper Minkowski dimension, defined with respect to an appropriate metric. The second is based on the behavior of Kolmogorov n-widths of the support, viewed as a subset of the corresponding reproducing kernel Hilbert space (RKHS). The first definition is proportional to the Hausdorff dimension of the support and does not take into account the global structure of the kernel. In contrast, the second notion — appearing in our excess risk bounds — captures the interaction between the kernel and the support. We argue that this parameter governs the generalization behavior of constrained KRR.

The paper is *organized* as follows. Section 2 introduces the reproducing kernel Hilbert space (RKHS) and the metric structure induced by the kernel K on the domain Ω, which supports the input distribution μ. We then define the Kolmogorov n-widths of Ω under its embedding into the RKHS, following conventional definitions. This framework allows us to define two notions of intrinsic dimension for Ω: the upper Minkowski dimension, denoted d_ϱ, and the effective dimension, denoted d_K. We prove that $d_K \leq d_\varrho$, and, through illustrative examples, justify particular attention to the case where equality holds, i.e., $d_K = d_\varrho$. In Section 3, we show that the Kolmogorov n-width characterizes the worst-case decay rate of the nth eigenvalue of the associated integral operator on $L_2(\Omega, \mu)$. This result may be seen as a counterpart to the classical Ismagilov's theorem in approximation theory and forms the foundation for our main excess risk bound for constrained KRR, presented in Section 4. The bound exhibits the asymptotic rate $\mathcal{O}\big(n^{-\frac{2+d_K}{2+2d_K}+\varepsilon}\big)$ for any $\varepsilon > 0$. In the second, empirically focused part of the paper, we present an algorithm for estimating n-widths from a sample, which may be either drawn from the distribution μ or deterministically constructed. In Section 5, we analyze this algorithm and show that, under the so-called C-uniformness assumption on μ, it produces ε-accurate upper bounds on the n-widths using a sample of size $\mathcal{O}\big(\varepsilon^{-d_\varrho}\log\frac{1}{\varepsilon}\big)$. In Section 6 we report the results of numerical experiments with the calculation of n-widths and effective dimensions d_K for various kernels and domains. Some of these experiments give rise to new conjectures concerning the gap between d_K and d_ϱ for such kernels as the Laplace kernel, given on fractal sets. Proofs of most statements, additional remarks on the tables, and details of the numerical experiments are provided in the Appendix.

Related work. It is well known that the leading term in the Shannon entropy of a discretized n-dimensional random vector is proportional to a parameter that can be interpreted as the information dimension of its support [Rényi, 1959]. Several other classical notions of a distribution's dimension are discussed in standard texts on geometric measure theory, such as [Mattila, 1995]. The concept of effective dimension in the context of kernel methods for supervised learning is well established. A distribution-dependent definition based on the eigenvalues of the kernel operator was introduced in [Zhang, 2002] and further developed in [Mendelson, 2003] and [Caponnetto and De Vito, 2007]. In this paper, we consider two alternative notions of intrinsic dimension. The first is closely related to the definition used by [Hamm and Steinwart, 2022], who employed it to derive an improved excess risk bound for regularized ERMs over RKHSs.

Our second notion of intrinsic dimension is based on the behavior of Kolmogorov n-widths. While n-widths have a long history in approximation theory [Pinkus, 1985], their application in machine learning is a relatively recent development [Steinwart, 2017]. Notably, [Wu and Long, 2022] and [Siegel and Xu, 2024] demonstrated that the asymptotic behavior of n-widths characterizes the gap in approximation power between two-layer neural networks and random feature models [Rahimi and Recht, 2007]. In certain settings — such as the unit sphere $\mathbb{S}^{d-1}$ — the computation of n-widths reduces to the analysis of kernel eigenvalues. For many commonly used kernels, these eigenvalues have been explicitly computed in [Bach, 2017] and [Bietti and Mairal, 2019].

In the second part of the paper, we develop a numerical algorithm for the approximate computation of n-widths from a sample drawn over a given domain. This algorithm is intended to serve as a new numerical tool for studying kernels on domains, analogous to the widely used approach of estimating eigenvalues from empirical kernel matrices [Koltchinskii and Giné, 2000, Williams and Shawe-Taylor, 2002].

Notations. Given $f : \mathbb{N} \to \mathbb{R}$ and $g : \mathbb{N} \to \mathbb{R}_+$ (or, alternatively, given two functions of a small argument $f : \mathbb{R}_+ \to \mathbb{R}, g : \mathbb{R}_+ \to \mathbb{R}_+$), we write $f(x) \ll g(x)$ if there exist constants $c_1, c_2 \in \mathbb{R}_+$ such that for all natural numbers $x > c_2$ (correspondingly, all positive reals $x < c_2$) we have $|f(x)| \leq c_1 g(x)$. We write $f \asymp g$ if $f \ll g$ and $g \ll f$. The space of measurable functions $f : \Omega \to \mathbb{R}$ such that $\int_\Omega |f|^p d\mu < \infty$ is denoted by $L_p(\Omega, \mu)$. Accordingly, $\|f\|_{L_p(\Omega,\mu)} = (\int_\Omega |f|^p d\mu)^{\frac{1}{p}}$. A $d-1$-sphere is the set $\mathbb{S}^{d-1} = \{\mathbf{x} \in \mathbb{R}^d \mid \|\mathbf{x}\| = 1\}$ where $\|\mathbf{x}\|$ denotes the canonical euclidean norm.

2 KERNEL-BASED UPPER METRIC AND EFFECTIVE DIMENSIONS

Let $\Omega \subseteq \mathbb{R}^d$ be a compact set and $K : \Omega \times \Omega \to \mathbb{R}$ a continuous positive-semidefinite kernel. Let $\mathcal{H}_K$ denote the reproducing kernel Hilbert space (RKHS) associated with K. The Mercer kernel K induces a canonical pseudometric on Ω defined by

$$\varrho(x,y) = \|K(x,\cdot) - K(y,\cdot)\|_{\mathcal{H}_K} = \sqrt{K(x,x) + K(y,y) - 2K(x,y)}. \tag{1}$$

This pseudometric, sometimes referred to as the "canonical metric", is widely used in kernel-based analysis [Adler and Taylor, 2007].

Let $B(z,r)$ denote the closed ball of radius r centered at z in (Ω, ϱ). By $\mathcal{N}(\varepsilon, \Omega, \varrho)$ we denote the ε-covering number of (Ω, ϱ), i.e.

$$\mathcal{N}(\varepsilon,\Omega,\varrho) = \min\big\{M \in \mathbb{N} \mid \exists z_1,\cdots,z_M \in \Omega \text{ s.t. } \bigcup_{i=1}^{M} B(z_i,\varepsilon) \supseteq \Omega\big\}. \tag{2}$$

It was shown in [Takhanov, 2024b] that the Dudley-type integral $\int_0^\infty \sqrt{\log \mathcal{N}(\varepsilon,\Omega,\varrho)}d\varepsilon$ plays a central role in bounding the generalization error of kernel methods. The function inverse to $\mathcal{N}(\varepsilon,\Omega,\varrho)$ is denoted by $\varepsilon(n)$, i.e. $\varepsilon(n) = \min\big\{\varepsilon > 0 \mid \mathcal{N}(\varepsilon,\Omega,\varrho) \le n\big\}$.

Then,

$$d_\varrho = \limsup_{\varepsilon\to+0} \frac{\log \mathcal{N}(\varepsilon,\Omega,\varrho)}{\log \frac{1}{\varepsilon}} = \limsup_{n\to+\infty} \frac{\log n}{\log(1/\varepsilon(n))} \tag{3}$$

is called the *kernel-based upper metric dimension* of Ω [Mattila, 1995]. For many kernels of interest (e.g., translation-invariant kernels, polynomial kernels, and neural tangent kernels), one typically has $\varrho(x,y) = \|x-y\|^a(b+o(1)), a, b > 0$ as $\|x-y\| \to 0$, implying $d_\varrho = \frac{d_H}{a}$, where d_H is the upper metric dimension of Ω equipped with the euclidean metric. If Ω is a k-dimensional smooth manifold embedded in $\mathbb{R}^d$, then $d_H = k$. For less regular sets, such as fractals, d_H coincides with the Hausdorff dimension for all interesting domains Ω. For readers unfamiliar with these notions, we refer to the popular book by [Peitgen et al., 2004], which contains numerous illustrations clarifying different definitions of dimension.

Since the map $x \in \Omega \mapsto K(x,\cdot) \in \mathcal{H}_K$ defines an embedding of Ω into $\mathcal{H}_K$, it is natural to study the image of Ω under this mapping as a subset of $\mathcal{H}_K$. The Kolmogorov n-width of this image is defined by

$$w_K(n) = \inf_{L_n} \sup_{x\in\Omega} \inf_{f\in L_n} \|K(x,\cdot) - f\|_{\mathcal{H}_K}, \tag{4}$$

where the infimum is taken over all n-dimensional subspaces L_n of $\mathcal{H}_K$ [Kolmogorov, 1936].

We will call the value

$$d_K = \limsup_{n\to+\infty} \frac{\log n}{\log(1/w_K(n))} \tag{5}$$

the *kernel-based effective dimension* of Ω.

In summary, the metric dimension defined by the kernel-induced pseudometric (cf. Equation (3)) depends on both the domain Ω and the local (typically linear) behavior of the kernel — e.g., its Taylor expansion in the translation-invariant case. The dimension (5) depends on the domain, and also, on the global structure of the kernel. We now state a basic but useful result:

Theorem 1. *We have $d_K \le d_\varrho$.*

Proof. It is sufficient to prove that

$$w_K(n) \le \varepsilon(n),$$

for any $n \in \mathbb{N}$. Indeed, let $\varepsilon = \varepsilon(n)$ and $z_1,\cdots,z_n \in \Omega$ be such that $\bigcup_{i=1}^n B(z_i,\varepsilon) \supseteq \Omega$. Given $x \in \Omega$, we define $i(x) \in \arg\min_{i:1\le i\le n} \varrho(x,z_i)$ and set $c(i(x),x) = 1$, and $c(j,x) = 0$ if $j \ne i(x)$. Let L_n be the span of $\{K(z_i,\cdot)\}_1^n$. Then,

$$w_K(n) \le \sup_{x\in\Omega} \min_{f\in L_n} \|K(x,\cdot) - f\|_{\mathcal{H}_K} \le \sup_{x\in\Omega} \|K(x,\cdot) - \sum_{i=1}^n c(i,x)K(z_i,\cdot)\|_{\mathcal{H}_K} \le \varepsilon.$$

Theorem proved. □

To illustrate these definitions, let us estimate n-widths for certain kernels on regular domains and exactly calculate both d_ϱ and d_K. In particular, we identify cases where the metric dimension and the effective dimension coincide. This situation is quite typical, as the next theorem demonstrates for several important kernels, including the Laplace kernel, when the domain Ω is sufficiently regular.

Theorem 2. *In the following cases we have $d_K = d_\varrho$:*

- *$K(x,y) = e^{-\gamma\|x-y\|^a}$, $a \in (0,1]$ and $\Omega \subseteq \mathbb{R}^d$ is Riemann measurable with non-zero volume;*
- *Matérn kernel: $K(x,y) = k_{\mathrm{M}}(\|x-y\|)$ where $k_{\mathrm{M}}(r) = \frac{2^{1-\nu}}{\Gamma(\nu)}(\frac{\sqrt{2\nu}r}{l})^\nu K_\nu(\frac{\sqrt{2\nu}r}{l})$, $\nu \in (0,1), l > 0$ and K_ν is a modified Bessel function, $\Omega \subseteq \mathbb{R}^d$ is Riemann measurable with non-zero volume;*

- $K(x,y) = e^{-\gamma\|x-y\|^a}$, $a \in (0,1]$ *and* $\Omega = \mathbb{S}^{d-1}$.

One of the key empirical findings of the experimental section (see Section 6) is that even for the mentioned kernels, the relation $d_K < d_\varrho$ often holds when the domain Ω exhibits fractal structure. A possible explanation for this phenomenon can be inferred from Table 1, which compares the kernel-based upper metric and effective dimensions for several kernels on $\mathbb{S}^{d-1}$.

As the table shows, for the Gaussian kernel we have $d_K = 0$. This reflects the fact that Kolmogorov n-widths decay more rapidly for smoother kernels. In this context, kernels for which $d_K = d_\varrho$ can be interpreted as the *least smooth*, since they yield the slowest possible decay rates of n-widths for a given domain geometry ($w_K(n) \asymp n^{-\frac{1}{d_\varrho}}$). The observed gap between d_K and d_ϱ for highly fractional domains suggests that kernel-based methods (even with least smooth kernels) treat fractals in such a way as if they were effectively lower-dimensional than their true metric complexity. In other words, the effective dimension captures a tradeoff between the smoothness of the kernel and the regularity of the domain.

3 n-WIDTHS AND EIGENVALUES

Recall that a kernel K and a Borel probability measure μ on a compact domain Ω together define a compact integral operator $\mathrm{O}_{K,\mu} : L_2(\Omega,\mu) \to L_2(\Omega,\mu)$ given by $\mathrm{O}_{K,\mu}\phi(x) = \int_\Omega K(x,y)\phi(y)d\mu(y)$. Let $\lambda_1(\mathrm{O}_{K,\mu}) \geq \lambda_2(\mathrm{O}_{K,\mu}) \geq \cdots$ denote its eigenvalues, counted with multiplicities and ordered in decreasing order. These eigenvalues play a central role in understanding the generalization ability of kernel methods [Cucker and Smale, 2001, Williamson et al., 2001]. This makes it natural to analyze the relationship between eigenvalues and n-widths. The key intuition is as follows: given the fixed domain Ω, the behaviour of n-widths captures the information on the worst case behaviour of eigenvalues. Our main result formalizing this relationship is stated below.

Theorem 3. *For any probabilistic Borel measure μ on Ω, we have*

$$\lambda_{2n}(\mathrm{O}_{K,\mu}) \leq \frac{w_K(n)^2}{n}.$$

Moreover,

$$\limsup_{n\to+\infty} \sup_{\mu} \frac{n\lambda_n(\mathrm{O}_{K,\mu})}{w_K(n)^2} \geq \frac{1}{e},$$

where the supremum is taken over all Borel probability measures μ supported on Ω.

Remark 1. *This theorem addresses the question: if the domain Ω is fixed but the distribution μ varies, how poorly can the eigenvalues of $\mathrm{O}_{K,\mu}$ decay? The first inequality shows that the Kolmogorov n-widths provide an upper bound for the eigenvalue decay for any μ. The second inequality confirms that this bound is essentially optimal (up to a constant) and cannot be improved uniformly over all μ, under the condition that domain is fixed and $w_K(n) \asymp w_K(2n)$. It highlights that the Kolmogorov n-widths reflect the worst-case spectral behavior induced by arbitrary distributions.*

The proof of Theorem 3 proceeds in two steps: establishing the upper and lower bounds. A major tool in the first part of the proof is the following remarkable result known as Ismagilov's theorem.

Theorem 4 ([Ismagilov, 1968])**.** *Let μ be a probabilistic Borel, nondegenerate measure on Ω. Let $\lambda_1 \geq \lambda_2 \geq \cdots$ be positive eigenvalues of $\mathrm{O}_{K,\mu}$ (counting multiplicities) with corresponding orthogonal unit eigenvectors $\{\psi_i\}_{i=1}^\infty$. Then,*

$$\sqrt{\sum_{i=n+1}^{\infty} \lambda_i} \leq w_K(n) \leq \sup_{x\in\Omega} \sqrt{\sum_{i=n+1}^{\infty} \lambda_i \psi_i(x)^2}.$$

If the measure μ in Ismagilov's theorem is chosen appropriately — typically as the uniform distribution on Ω — and the corresponding eigenfunctions of the integral operator $\mathrm{O}_{K,\mu}$ grow moderately (e.g., satisfy a bound of the form $\sup_{n\in\mathbb{N},x\in\Omega} \frac{1}{n}\sum_{i=1}^n \psi_i(x)^2 < +\infty$), then we obtain sharp estimates of the form $w_K(n) \asymp \sqrt{\sum_{i=n+1}^\infty \lambda_i}$. In practice, Ismagilov's theorem is most effective when applied with such carefully selected measures μ, which yield well-behaved eigenfunctions and allow accurate estimation of Kolmogorov widths via spectral data.

Proof of the first part of Theorem 3. Let us first assume that μ is non-degenerate on Ω, meaning its support equals Ω. From Ismagilov's theorem we obtain

$$n\lambda_{2n}(\mathrm{O}_{K,\mu}) \leq \sum_{i=n+1}^{\infty} \lambda_i(\mathrm{O}_{K,\mu}) \leq w_K(n)^2.$$

Thus, $\lambda_{2n}(\mathrm{O}_{K,\mu}) \leq \frac{w_K(n)^2}{n}$. Any probabilistic Borel measure μ can be approximated by a non-degenerate $(1-\varepsilon)\mu + \varepsilon\nu$, where ν is a fixed non-degenerate probabilistic Borel measure on Ω. Using Weyl's inequality we have $\lambda_{2n}(\mathrm{O}_{K,(1-\varepsilon)\mu+\varepsilon\nu}) \overset{\varepsilon\to+0}{\to} \lambda_{2n}(\mathrm{O}_{K,\mu})$ and, therefore, $\lambda_{2n}(\mathrm{O}_{K,\mu}) \leq \frac{w_K(n)^2}{n}$ holds for any probabilistic Borel measure μ. This concludes the proof of the first part of the theorem. □

K	$w_K(n)$	d_ϱ	d_K	Using the source
$e^{-\gamma\|x-y\|}$	$\asymp n^{-\frac{1}{2d-2}}$	$2d-2$	$2d-2$	[Geifman et al., 2020]
$\mathrm{NTK}_{\max(0,x)}$	$\asymp n^{-\frac{1}{2d-2}}$	$2d-2$	$2d-2$	[Geifman et al., 2020]
$e^{-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{\sigma^2}}, \sigma > \sqrt{\frac{2}{d}}$	$\searrow$ exp. fast	$d-1$	0	[Minh et al., 2006]
$\mathrm{NNGP}_{\cos}, \mathrm{NNGP}_{\sin}$	$\ll n^{-\frac{1}{2}}$	$d-1$	≤ 2	[Wu and Long, 2022]
$\mathrm{NNGP}_{\max(0,x)^\alpha}, \alpha \geq 0$	$\gg n^{-\frac{1+2\alpha}{2d-2}}$	$\frac{2d-2}{1+\alpha}(?)$	$\geq \frac{2d-2}{1+2\alpha}$	[Wu and Long, 2022]
$\mathrm{NNGP}_{\max(0,x)^\alpha}, \alpha \in \{0,1\}$	$\asymp n^{-\frac{1+2\alpha}{2d-2}}$	$\frac{2d-2}{1+\alpha}$	$\frac{2d-2}{1+2\alpha}$	[Bach, 2017]

Table 1: The Kolmogorov n-widths for various K on $\Omega = \mathbb{S}^{d-1}$

The proof of the second part of Theorem 3 is technical and provided in the Appendix A. Its central idea is to construct a sequence of points $x_1, x_2, \cdots \in \Omega$ such that the corresponding sequence of empirical measures $\mu_m := \frac{1}{m}\sum_{i=1}^m \delta_{x_i}, \quad m \in \mathbb{N}$, induces integral operators O_{K,μ_m} whose eigenvalues exhibit the worst-case decay behavior relative to the Kolmogorov n-widths. Here, δ_x denotes the probabilistic measure concentrated at point x. Specifically, we prove that $\limsup_{n\to\infty} \sup_{m\in\mathbb{N}} \frac{n\lambda_n(\mathrm{O}_{K,\mu_m})}{w_K(n)^2} \geq \frac{1}{e}$.

The construction of the sequence $\{x_i\}_{i=1}^\infty \subset \Omega$ is not only essential to the proof but also serves as the foundation for Algorithm 1, which we present in Section 5. That algorithm provides a practical method for approximating upper bounds on Kolmogorov n-widths. Below, we briefly describe the core idea behind this construction.

Given $X_n = (x_1, \cdots, x_n)$ and $Y_m = (y_1, \cdots, y_m)$, let $K[X_n, Y_m]$ denote the matrix $[K(x_i, y_j)]_{i=1}^n{}_{j=1}^m \in \mathbb{R}^{n\times m}$. For any $n \in \mathbb{N}$ and any $X_n = (x_1, \cdots, x_n) \in \Omega^n$, let us denote $f_n : \Omega^n \to \Omega$ by

$$f_n(X_n) = \arg\max_{x\in\Omega} \det(K[(X_n, x), (X_n, x)]). \tag{6}$$

Given any $x_1 \in \arg\max_{x\in\Omega} K(x,x)$, the collection of functions $\{f_n\}$ induces a sequence $\{x_i\}_1^\infty \subseteq \Omega$ defined by the rule

$$x_{n+1} = f_n(x_1, \cdots, x_n). \tag{7}$$

Recall that the Schur complement $(M \mid A)$ of the block matrix $M = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$, where A, B, C, D are matrices with dimensions $n\times n, n\times m, m\times n, m\times m$ respectively, is defined as the $D - CA^{-1}B$. A key property of the Schur complement is that $\det(M) = \det(A)\det((M|A))$ if $\det(A) \neq 0$ [Horn and Zhang, 2005]. The key lemma for the proof is formulated below.

Lemma 1. *For any $n \in \mathbb{N}$ and any $X_n \in \Omega^n$ such that $\det(K[X_n, X_n]) > 0$, $X_{n+1} = (X_n, f_n(X_n))$ satisfies*

$$(K[X_{n+1}, X_{n+1}] \mid K[X_n, X_n]) \geq w_K(n)^2.$$

4 AN APPLICATION TO ERM

Let $\mathcal{Z} = (\Omega \times \mathbb{R}, \mathcal{B}(\Omega\times\mathbb{R}), P)$ be a probability space, where $\mathcal{B}(\Omega\times\mathbb{R})$ is a family of Borel subsets of $\Omega\times\mathbb{R}$. We denote the probability function over inputs by μ, i.e. $\mu(B) = P(B\times\mathbb{R})$. Let $\mathcal{Z}^n$ denote the Cartesian product of n copies of $\mathcal{Z}$ and $\mathcal{T} = \{(X_i, Y_i)\}_{i=1}^n$ denote a random variable distributed according to $\mathcal{Z}^n$. The hypothesis class is defined by $\mathcal{F} = B_{\mathcal{H}_K}$ where $B_{\mathcal{H}_K}$ is a unit ball of $\mathcal{H}_K$ centered at origin. We are also given a loss function $l : \mathbb{R}\times\mathbb{R} \to \mathbb{R}_+$. We consider the empirical risk minimization (ERM) algorithm for regression, i.e. the algorithm that, given $\mathcal{T}$, selects a function $\hat{f} \in B_{\mathcal{H}_K}$ such that

$$\hat{f} = \arg\min_{f\in B_{\mathcal{H}_K}} \frac{1}{n}\sum_{i=1}^n l(f(X_i), Y_i).$$

The optimal regression function in $B_{\mathcal{H}_K}$, minimizing the true risk, is denoted by f^*, i.e. $f^* = \arg\min_{f\in B_{\mathcal{H}_K}} \mathbb{E}[l(f(X), Y)]$.

Following the framework of [Bartlett et al., 2002], we assume that the loss function satisfies conditions: (1) f^* is well-defined for any probability function P; (2) there is $L > 0$ such that $|l(y_1, y) - l(y_2, y)| \leq L|y_1 - y_2|$; (3) there is $B > 0$ such that $B(\mathbb{E}[l(f(X), Y)] - \mathbb{E}[l(f^*(X), Y)]) \geq \|f - f^*\|^2_{L_2(\Omega,\mu)}$ for any $f \in B_{\mathcal{H}_K}$. Note that for the square loss case, we have $B = 1$ due to convexity of $B_{\mathcal{H}_K}$. Our main result provides a high-probability bound on the excess risk and is stated below.

Theorem 5. *Let $\varepsilon > 0$ and $N_\varepsilon = \min\{N \in \mathbb{N} \mid \sup_{i>N} w_K(i) i^{\frac{1}{d_K+\varepsilon}} \leq 1\}$. For any $x > 0$, with probability at least $1 - e^{-x}$ (over randomness in $\mathcal{T}$), we*

have

$$\mathbb{E}[l(\hat{f}(X),Y)] - \mathbb{E}[l(f^*(X),Y)] \leq$$

$$1995B^{\frac{1}{1+d_K+\varepsilon}}L^{1-\frac{1}{1+d_K+\varepsilon}}n^{-\frac{2+d_K+\varepsilon}{2(1+d_K+\varepsilon)}} + \frac{(11L+27B)x}{n},$$

provided that $n \geq \frac{B^2(3+d_K+\varepsilon)N_\varepsilon^{\frac{2(1+d_K+\varepsilon)}{d_K+\varepsilon}}}{2L^2}$.

Note that excess risk rate $\mathcal{O}(\frac{1}{n})$ represents the best possible convergence achievable in well-specified parametric settings or under strong regularity assumptions [Klochkov and Zhivotovskiy, 2021]. For kernel methods, a typical and often minimax-optimal convergence rate is $\mathcal{O}(\frac{1}{\sqrt{n}})$. As the latter theorem shows, $d_K = 0$, which we have for, e.g., the Gaussian kernel, leads to the excess risk rate arbitrarily close to $\mathcal{O}(\frac{1}{n})$. The larger the dimension d_K the closer we approach to $\mathcal{O}(\frac{1}{\sqrt{n}})$ rate. For kernels such as the Laplace kernel, we have $d_K = d_\varrho$ on regular domains, and the excess risk convergence behaves like $\mathcal{O}(n^{-\frac{2+d_\varrho}{2+2d_\varrho}+\varepsilon})$ which aligns with the manifold hypothesis.

Analogously, Theorem 3 allows to apply Kolmogorov n-widths to other kernel methods, if their statistical properties are defined by the eigenvalue decay of the associated integral operator. E.g., one could estimate the effective dimension in the sense of [Caponnetto and De Vito, 2007] or the Signal Capture Threshold [Jacot et al., 2020], and then derive implications for unconstrained KRR.

5 THE ESTIMATION OF THE n-WIDTH'S UPPER BOUND

In the previous section, we demonstrated that the Kolmogorov n-width, $w_K(n)$, plays a fundamental role in understanding how the geometry of the domain Ω influences the generalization ability of kernel ridge regression. Given the domain Ω and the kernel K, let us now address the problem of the estimation of an upper bound on $w_K(n)$. We begin by describing an algorithm under the assumption that we can efficiently perform function maximization over Ω. Afterwards, we extend the approach to a more practical setting in which we are given only a finite set of points $\tilde{\Omega} \subset \Omega$, either computed directly or sampled from some distribution over Ω.

The Algorithm 1 is inspired by the construction of a sequence $\{x_t\}$ defined by the greedy rule (7), which arises naturally in the proof of Theorem 3 (see also Lemma 1). The key idea behind that proof is that, for any $X_t = (x_1, \cdots, x_t) \in \Omega^t$, the quantity $\sup_{x\in\Omega}(K[(X_t,x),(X_t,x)] \mid K[X_t,X_t])$ serves as a valid upper bound on the squared Kolmogorov width $w_K(t)^2$. Based on this, we compute a sequence of values $\{w_t\}_{t=0}^{T-1}$ such that $w_K(t) \leq w_t$ for $t = 0, 1, \cdots, T-1$. Note that $\forall x \in \Omega, (K[(X_{t-1},x),(X_{t-1},x)] \mid K[X_{t-1},X_{t-1}]) = 0$ implies $\mathcal{H}_K \subseteq \operatorname{span}(\{K(x_t,\cdot)\}_{t=1}^{t-1})$. Thus, Algorithm 1 is well-defined if $T \geq \dim(\mathcal{H}_K)$, that is we have $\det(K[X_{t-1},X_{t-1}]) \neq 0$ at each step. In all our applications, we have $\dim(\mathcal{H}_K) = +\infty$.

Algorithm 1 The empirical n-width upper bound algorithm. Parameter: $T \in \mathbb{N}$

Input: $\tilde{\Omega}$. Either $\tilde{\Omega} = \Omega$ (ideal case) or $\tilde{\Omega} \subset \Omega$ is finite (empirical case)
$x_1 \longleftarrow \arg\max_{x\in\tilde{\Omega}} K(x,x)$
$w_0 \longleftarrow K(x_1,x_1)^{1/2}$
$X_1 = \{x_1\}$
for $t = 2, \cdots, T$ **do**
  $x_t \longleftarrow \arg\max_{x\in\tilde{\Omega}} K(x,x) - K[X_{t-1},x]^\top K[X_{t-1},X_{t-1}]^{-1}K[X_{t-1},x]$
  $w_{t-1} \longleftarrow (K(x_t,x_t) - K[X_{t-1},x_t]^\top K[X_{t-1},X_{t-1}]^{-1}K[X_{t-1},x_t])^{1/2}$
  $X_t = X_{t-1} \cup \{x_t\}$
end for
Output: $\{w_t\}_{t=0}^{T-1}$

We now consider the setting where $\tilde{\Omega} \subset \Omega$ is a finite set, and all maximization steps in Algorithm 1 are performed over $\tilde{\Omega}$. We begin with the case in which $\tilde{\Omega}$ is computed deterministically and serves as a finite approximation of the entire domain Ω. Since the maximization is now restricted to a subset of Ω, we no longer can guarantee $\sup_{x\in\tilde{\Omega}}(K[(X_t,x),(X_t,x)] \mid K[X_t,X_t]) \geq w_K(t)^2$ unless the sample $\tilde{\Omega}$ is sufficiently representative of Ω.

Any sequence $\{x_t\}_{t=1}^\infty \subseteq \Omega$ defines a sequence of functions $\{S_t : \Omega \to \mathbb{R}\}_{t=0}^\infty$ by $S_0(x) = K(x,x)$ and $S_t(x) = K(x,x) - K[X_t,x]^\top K[X_t,X_t]^{-1}K[X_t,x]$ if $\det(K[X_t,X_t]) \neq 0$ where $X_t = (x_1, \cdots, x_t)$. If $\det(K[X_t,X_t]) = 0$, we define $S_t(x) = 0$. The key observation underlying our analysis of Algorithm 1 is the following lemma, which states that functions in $\{\sqrt{S_t}\}$ are all 1-Lipschitz w.r.t. the metric ϱ.

Lemma 2. *For any* $\{x_t\}_{t=1}^\infty \subseteq \Omega$ *and any* $n \in \mathbb{N}\cup\{0\}$, *we have* $\sqrt{S_n(x)} - \sqrt{S_n(y)} \leq \varrho(x,y)$.

As the following theorem shows, a slightly adjusted output of the empirical Algorithm 1 can serve as a valid upper bound on the Kolmogorov n-widths, provided that the finite set $\tilde{\Omega}$ is ε-dense in (Ω,ϱ).

Theorem 6. *Let* $\{w_t\}_{t=0}^{T-1}$ *be the sequence of approximate upper bounds produced by Algorithm 1 using the finite set* $\tilde{\Omega} = \Omega_1 \subset \Omega$. *Let us assume that* Ω_1 *is an*

ε-net in (Ω, ϱ). Then, we have

$$w_t \geq w_K(t) - \varepsilon, t = 0, \cdots, T-1.$$

Proof. Let $\{x_t\}_{t=1}^T \subseteq \Omega_1$ be the sequence of elements computed in Algorithm 1, given the input $\tilde{\Omega} = \Omega_1$. This sequence defines the sequence of functions $\{S_t\}_{t=0}^T$ and, by construction of the algorithm, we have $w_{t-1} = \sqrt{S_{t-1}(x_t)} = \sqrt{\max_{x\in\Omega_1} S_{t-1}(x)}$.

Let $\tilde{x} \in \arg\max_{x\in\Omega} S_{t-1}(x)$. By assumption, there exists $\tilde{\tilde{x}} \in \Omega_1$ such that $\varrho(\tilde{x}, \tilde{\tilde{x}}) \leq \varepsilon$. From Lemma 2 we conclude $\sqrt{S_{t-1}(\tilde{\tilde{x}})} \geq \sqrt{S_{t-1}(\tilde{x})} - \varepsilon$, which implies $\sqrt{\max_{x\in\Omega_1} S_{t-1}(x)} \geq \sqrt{\max_{x\in\Omega} S_{t-1}(x)} - \varepsilon$. Therefore,

$$w_{t-1} + \varepsilon \geq \max_{x\in\Omega} \sqrt{S_{t-1}(x)} =$$

$$\max_{x\in\Omega} \min_{c_{1:t-1}} \|K(x,\cdot) - \sum_{i=1}^{t-1} c_i K(x_i,\cdot)\|_{\mathcal{H}_K} \geq w_K(t-1).$$

This completes the proof. □

Sampling points from a distribution μ whose support is the full domain Ω is another practically important scenario. That is, we assume $\tilde{\Omega} = \{Z_1, \cdots, Z_N\}$ and $Z_1, \cdots, Z_N \sim^{\text{iid}} \mu$. If μ is the uniform distribution over Ω, then with high probability the sample $\tilde{\Omega}$ can serve as a good approximation of Ω. We now introduce a generalization of this natural uniformity assumption.

Definition 1. *Let $C > 0$. The measure μ is called C-uniform over Ω if there is $\varepsilon_0 > 0$ such that $\mu(B(x,\varepsilon)) \geq C\varepsilon^{d_\varrho}$ for any $x \in \Omega$ and $\varepsilon \in (0, \varepsilon_0)$.*

Remark 2. *In geometric measure theory, the quantity*

$$\Theta_*(x,\mu) = \liminf_{\varepsilon\to 0+} (2\varepsilon)^{-d_\varrho} \mu(B(x,\varepsilon))$$

is known as the lower d_ϱ-density of μ at the point x [Mattila, 1995]. This notion generalizes the concept of a probability density function to sets Ω of potentially fractal nature, and to distributions supported on such sets. In the case of regular domains — e.g., compact sets with smooth boundaries — and continuous measures μ, this density behaves proportionally to the standard probability density function (provided that $\varrho(x,y) = \|x-y\|^a(b+o(1)), a, b > 0$ as $\|x-y\| \to 0$). So, in those cases, the condition in Definition 1 is equivalent to requiring a positive lower bound on the pdf of μ.

The following theorem shows how the difference between the ideal setting ($\tilde{\Omega} = \Omega$) and the empirical setting ($\tilde{\Omega} = \{Z_1, \cdots, Z_N\}$) vanishes as the sample size N increases. To ensure that Algorithm 1 is almost surely well-defined, we additionally assume that for $X_1, \cdots, X_n \sim^{\text{iid}} \mu$ the probability of the event $\det([K(X_i, X_j)]_{i,j=1}^n) = 0$ is zero.

Fractal	d_ϱ	d_K^{emp}
Cantor set	1.2618	1.2415
Weierstrass function	3.0	2.7052
Sierpiński carpet	3.7855	3.2896
Menger sponge	5.4536	4.2506
Lorenz attractor	4.12	3.2839

Table 2: Comparison of the Laplace kernel-based upper metric dimension and the empirical upper bound on the effective dimension, computed using Algorithm 1, for several well-known fractal sets. The upper metric dimension corresponds to twice the Hausdorff dimension.

Theorem 7. *Suppose that the measure μ is C-uniform over Ω and $Z_1, \cdots, Z_N \sim^{\text{iid}} \mu$. Let $\{w_t\}_{t=0}^{T-1}$ be an output of Algorithm 1 for $\tilde{\Omega} = \{Z_1, \cdots, Z_N\}$. For $\delta \in (0,1)$, we have*

$$\mathbb{P}[w_t \geq w_K(t) - 2\varepsilon, t = 0, \cdots, T-1] \geq 1 - \delta,$$

provided that $N \geq \frac{\log \mathcal{N}(\varepsilon,\Omega,\varrho) + \log\frac{1}{\delta}}{C\varepsilon^{d_\varrho}}$.

Remark 3. *As $\varepsilon \to +0$, $\mathcal{N}(\varepsilon,\Omega,\varrho) \ll (\frac{1}{\varepsilon})^{d_\varrho+\zeta}$ for any $\zeta > 0$. Thus, we need a sample size of $N = \frac{(d_\varrho+\zeta)\log\frac{1}{\varepsilon} + \log\frac{1}{\delta} + O(1)}{C\varepsilon^{d_\varrho}}$ to have for $t = 0, \cdots, T-1$, $w_t \geq w_K(t) - 2\varepsilon$ with probability at least $1-\delta$. The Appendix C.1 contains a proof showing that the condition $T \leq \frac{NC\varepsilon^{d_\varrho} + \log\delta}{\log N}$ is sufficient to ensure that $w_t \geq w_K(t) - \varepsilon$ holds for all $t = 0, \ldots, T-1$ for a target confidence level $1-\delta$. This result implies that fewer samples are required for smaller values of t.*

In Appendix C.2, we present two implementations of Algorithm 1: one for the ideal setting ($\tilde{\Omega} = \Omega$), where optimization over the infinite domain $\tilde{\Omega}$ is replaced by a non-convex optimization procedure, and another for the finite-sample case $\tilde{\Omega} = \{Z_1, \ldots, Z_N\}$. We experimented with both implementations, and the corresponding results are reported in Appendices D.3, D.4, and D.5.

6 EXPERIMENTS

Experiments with fractals. Our first set of experiments focused on the Laplace kernel, which is closely related — both theoretically and empirically — to the neural tangent kernel (NTK) associated with ReLU activation functions [Chen and Xu, 2021]. A remarkable property of the Laplace kernel is that $d_\varrho = d_K$ for any Riemann measurable $\Omega \subseteq \mathbb{R}^d$ with non-zero volume and for $\Omega = \mathbb{S}^{d-1}$. One might conjecture that this equality extends to less regular subsets of $\mathbb{R}^d$, such as fractals — but our results show this is not the case.

We carried out a series of numerical experiments estimating upper bounds on d_K for various well-known fractal sets whose Hausdorff dimensions are explicitly

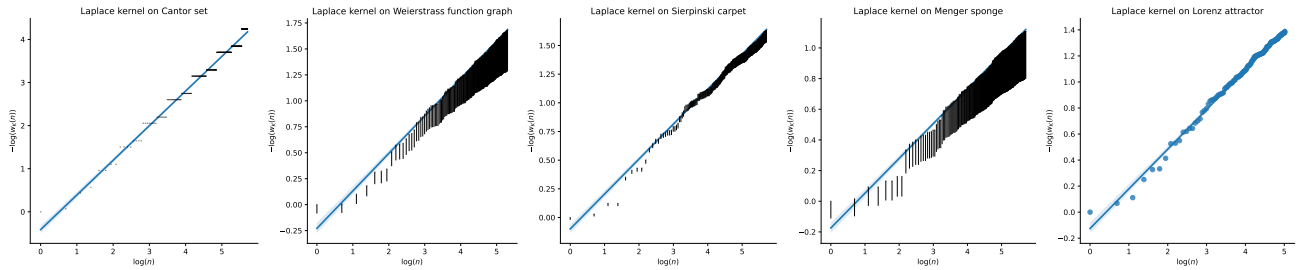

Figure 1: Plot of $-\log(w_n)$ versus $\log n$, where w_n is the output of Algorithm 1 for the Laplace kernel on various fractal sets. Black vertical bars represent uncertainty intervals of the form $[-\log(w_n+\varepsilon), -\log(w_n)]$, where ε is the maximum possible deviation from the true value of $w_K(n)$, estimated using Theorem 6 and known geometric properties of each fractal. For the Lorenz attractor, ε is not reliably estimable. The slope of the fitted line, computed using the RANSAC algorithm [Fischler and Bolles, 1987], is defined as inversely proportional to the empirically estimated effective dimension d_K^{emp}.

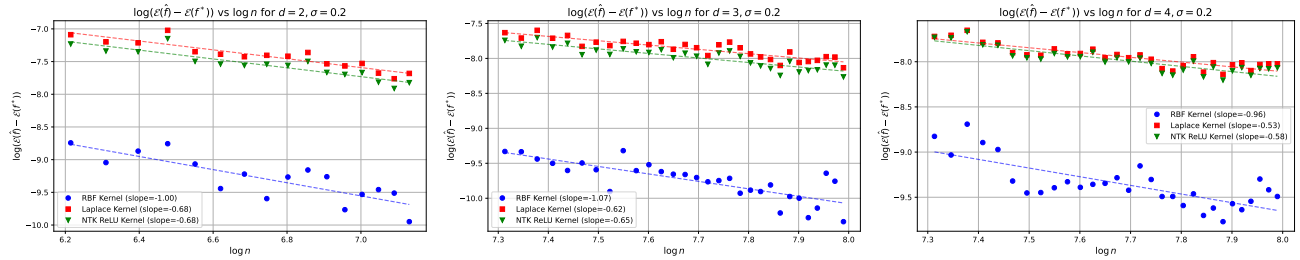

Figure 2: Scatter plots of log-excess risk against log-sample size, with fitted linear regression lines, for $d = 2, 3, 4$.

known. As shown in Table 2 and plots 1, the computed upper bounds on d_K are consistently smaller than the corresponding kernel-based metric dimensions d_ϱ. This strongly suggests that, for typical fractal sets, the inequality $d_K < d_\varrho$ holds.

Verification of decay rates of Theorem 5. According to Theorem 5, the excess risk of constrained KRR satisfies

$$\mathcal{E}(\hat{f}) - \mathcal{E}(f^*) = \mathcal{O}(n^{-\frac{2+d_K+\varepsilon}{2(1+d_K+\varepsilon)}}) \quad \text{as } n \to \infty.$$

To empirically verify this rate, we performed numerical experiments for the empirical risk minimization problem with the squared loss $\ell(y, y') = (y - y')^2$ over the hypothesis class $B_{\mathcal{H}_K}$.

We considered three kernels: (a) Gaussian: $k(x, y) = e^{-0.1\|x-y\|^2}$; (b) Laplace: $k(x, y) = e^{-\|x-y\|}$; (c) ReLU NTK: $k_{\text{NTK}}(x, y) = r(u)$, where $r(u) = \frac{1}{\pi}\big[(2u + 1)(\pi - \arccos u) + \sqrt{1-u^2}\big] + 1$ with $u = \langle x, y\rangle$. We set $\Omega = \mathbb{S}^{d-1}$ and chose the input distribution to be the rotation-invariant measure on $\mathbb{S}^{d-1}$. For each input X, the output Y was drawn independently from $0.2\text{U}([-1, 1])$ where $\text{U}([-1, 1])$ denotes the uniform distribution. This corresponds to a pure-noise target with optimal regression function $f^* \equiv 0 \in B_{\mathcal{H}_K}$. From Table 1, $d_K = 0$ for the Gaussian kernel and $d_K = 2d-2$ for the Laplace and ReLU NTK kernels. Scatter plots of log-excess-risk versus $\log n$ (with fitted linear regression lines) for $d = 2, 3, 4$ are shown in Figure 2.

Theorem 5 predicts that, for large n, the slopes of these plots should be bounded above by: -1 for the Gaussian kernel; $-\frac{2}{3}$ $(d = 2)$, -0.6 $(d = 3)$, $-\frac{4}{7}$ $(d = 4)$ for the Laplace and ReLU NTK kernels. The experimental results confirm these bounds, with all kernels exhibiting decay rates close to theoretical.

Details of the experimental setup, additional results for NTK kernels, and their discussion are provided in the Appendix.

7 CONCLUSIONS AND FUTURE RESEARCH

Unlike the eigenvalues of the kernel operator, Kolmogorov n-widths depend solely on the domain and the kernel, not on the underlying distribution. Their role in quantifying the separation between the approximation power of neural networks and random feature models is well established [Wu and Long, 2022]. Our excess risk bound reveals that n-widths also govern the generalization behavior of KRR. A fundamental trade-off emerges: slower decay of n-widths implies greater approximation capacity but weaker generalization, while faster decay leads to stronger general-

ization but reduced approximation power. Thus, n-widths serve as a meaningful complexity measure, and a deeper investigation of their properties forms a natural direction for future work.

Another promising direction for future research is to investigate kernels for which the effective dimension d_K coincides with the Minkowski dimension d_ϱ on regular domains, and to understand the extent to which this equality breaks down on irregular or fractal domains. Notably, this question is particularly relevant for ReLU-based neural networks, whose associated NTKs belong to this class. Gaining insight into how the equality $d_K = d_\varrho$ breaks down could help clarify the types of domains on which ReLU networks are likely to generalize effectively.

Acknowledgments

This research has been funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP27510283), PI R. Takhanov.

References

[Adler and Taylor, 2007] Adler, R. J. and Taylor, J. E. (2007). *Gaussian Fields*, pages 7–48. Springer New York, New York, NY.

[Bach, 2017] Bach, F. (2017). Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53.

[Bartlett et al., 2002] Bartlett, P. L., Bousquet, O., and Mendelson, S. (2002). Localized rademacher complexities. In Kivinen, J. and Sloan, R. H., editors, *Computational Learning Theory*, pages 44–58, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Bickel and Li, 2007] Bickel, P. J. and Li, B. (2007). Local polynomial regression on unknown manifolds. *Lecture Notes-Monograph Series*, 54:177–186.

[Bietti and Mairal, 2019] Bietti, A. and Mairal, J. (2019). *On the inductive bias of neural tangent kernels*. Curran Associates Inc., Red Hook, NY, USA.

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

[Chen and Xu, 2021] Chen, L. and Xu, S. (2021). Deep neural tangent kernel and laplace kernel have the same {rkhs}. In *International Conference on Learning Representations*.

[Cheng and Xie, 2024] Cheng, X. and Xie, Y. (2024). Kernel two-sample tests for manifold data. *Bernoulli*, 30(4):2572 – 2597.

[Cucker and Smale, 2001] Cucker, F. and Smale, S. (2001). On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39:1–49.

[Fischler and Bolles, 1987] Fischler, M. A. and Bolles, R. C. (1987). Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. In Fischler, M. A. and Firschein, O., editors, *Readings in Computer Vision*, pages 726–740. Morgan Kaufmann, San Francisco (CA).

[Geifman et al., 2020] Geifman, A., Yadav, A., Kasten, Y., Galun, M., Jacobs, D., and Ronen, B. (2020). On the similarity between the laplace and neural tangent kernels. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1451–1461. Curran Associates, Inc.

[Hamm and Steinwart, 2022] Hamm, T. and Steinwart, I. (2022). Intrinsic dimension adaptive partitioning for kernel methods. *SIAM Journal on Mathematics of Data Science*, 4(2):721–749.

[Han et al., 2022] Han, I., Zandieh, A., Lee, J., Novak, R., Xiao, L., and Karbasi, A. (2022). Fast neural kernel embeddings for general activations. In *Advances in Neural Information Processing Systems*.

[Horn and Zhang, 2005] Horn, R. A. and Zhang, F. (2005). *Basic Properties of the Schur Complement*, pages 17–46. Springer US, Boston, MA.

[Ismagilov, 1968] Ismagilov, R. S. (1968). On n-dimensional diameters of compacts in a hilbert space. *Functional Analysis and Its Applications*, 2(2):125–132.

[Jacot et al., 2020] Jacot, A., Şimşek, B., Spadaro, F., Hongler, C., and Gabriel, F. (2020). Kernel alignment risk estimator: risk prediction from training data. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

[Jacot et al., 2018] Jacot, A., Gabriel, F., and Hongler, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 8580–8589, Red Hook, NY, USA. Curran Associates Inc.

[Klochkov and Zhivotovskiy, 2021] Klochkov, Y. and Zhivotovskiy, N. (2021). Stability and deviation optimal risk bounds with convergence rate $o(1/n)$. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*.

[Kolmogorov, 1936] Kolmogorov, A. (1936). Über die beste annäherung von funktionen einer gegebenen funktionenklasse. *Annals of Mathematics*, 37(1):107–110.

[Koltchinskii and Giné, 2000] Koltchinskii, V. and Giné, E. (2000). Random matrix approximation of spectra of integral operators. *Bernoulli*, 6(1):113 – 167.

[Kpotufe, 2011] Kpotufe, S. (2011). k-nn regression adapts to local intrinsic dimension. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, NIPS'11, page 729–737, Red Hook, NY, USA. Curran Associates Inc.

[Mattila, 1995] Mattila, P. (1995). *Geometry of Sets and Measures in Euclidean Spaces: Fractals and Rectifiability*. Cambridge Studies in Advanced Mathematics. Cambridge University Press.

[McRae et al., 2020] McRae, A., Romberg, J., and Davenport, M. (2020). Sample complexity and effective dimension for regression on manifolds. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12993–13004. Curran Associates, Inc.

[Mendelson, 2002] Mendelson, S. (2002). Geometric parameters of kernel machines. In Kivinen, J. and Sloan, R. H., editors, *Computational Learning Theory*, pages 29–43, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Mendelson, 2003] Mendelson, S. (2003). On the performance of kernel classes. *J. Mach. Learn. Res.*, 4(null):759–771.

[Minh et al., 2006] Minh, H. Q., Niyogi, P., and Yao, Y. (2006). Mercer's theorem, feature maps, and smoothing. In Lugosi, G. and Simon, H. U., editors, *Learning Theory*, pages 154–168, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Novak et al., 2020] Novak, R., Xiao, L., Hron, J., Lee, J., Alemi, A. A., Sohl-Dickstein, J., and Schoenholz, S. S. (2020). Neural tangents: Fast and easy infinite neural networks in python. In *International Conference on Learning Representations*.

[Ozakin and Gray, 2009] Ozakin, A. and Gray, A. (2009). Submanifold density estimation. In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C., and Culotta, A., editors, *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc.

[Peitgen et al., 2004] Peitgen, H., Jürgens, H., and Saupe, D. (2004). *Chaos and Fractals: New Frontiers of Science*. Springer New York.

[Pinkus, 1985] Pinkus, A. (1985). *n-Widths in Approximation Theory*. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Pollard, 1946] Pollard, H. (1946). The representation of e^{-x^λ} as a Laplace integral. *Bulletin of the American Mathematical Society*, 52(10):908 – 910.

[Rahimi and Recht, 2007] Rahimi, A. and Recht, B. (2007). Random features for large-scale kernel machines. In *Proceedings of the 21st International Conference on Neural Information Processing Systems*, NIPS'07, page 1177–1184, Red Hook, NY, USA. Curran Associates Inc.

[Rasmussen and Williams, 2005] Rasmussen, C. E. and Williams, C. K. I. (2005). *Gaussian Processes for Machine Learning*. The MIT Press.

[Rényi, 1959] Rényi, A. (1959). On the dimension and entropy of probability distributions. *Acta Mathematica Academiae Scientiarum Hungarica*, 10(1):193–215.

[Siegel and Xu, 2024] Siegel, J. W. and Xu, J. (2024). Sharp bounds on the approximation rates, metric entropy, and n-widths of shallow neural networks. *Foundations of Computational Mathematics*, 24(2):481–537.

[Steinwart, 2017] Steinwart, I. (2017). A short note on the comparison of interpolation widths, entropy numbers, and kolmogorov widths. *Journal of Approximation Theory*, 215:13–27.

[Takhanov, 2023] Takhanov, R. (2023). On the speed of uniform convergence in mercer's theorem. *Journal of Mathematical Analysis and Applications*, 518(2):126718.

[Takhanov, 2024a] Takhanov, R. (2024a). Multi-layer random features and the approximation power of neural networks. In Kiyavash, N. and Mooij, J. M., editors, *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pages 3299–3322. PMLR.

[Takhanov, 2024b] Takhanov, R. (2024b). Non-asymptotic spectral bounds on the ε-entropy of kernel classes.

[Widom, 1963] Widom, H. (1963). Asymptotic behavior of the eigenvalues of certain integral equations. *Transactions of the American Mathematical Society*, 109(2):278–295.

[Williams and Shawe-Taylor, 2002] Williams, C. and Shawe-Taylor, J. (2002). The stability of kernel principal components analysis and its relation to the process eigenspectrum. In Becker, S., Thrun, S., and Obermayer, K., editors, *Advances in Neural Information Processing Systems*, volume 15. MIT Press.

[Williamson et al., 2001] Williamson, R., Smola, A., and Scholkopf, B. (2001). Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators. *IEEE Transactions on Information Theory*, 47(6):2516–2532.

[Wu and Long, 2022] Wu, L. and Long, J. (2022). A spectral-based analysis of the separation between two-layer neural networks and linear methods. *J. Mach. Learn. Res.*, 23(1).

[Zhang, 2002] Zhang, T. (2002). Effective dimension and generalization of kernel learning. In *Proceedings of the 16th International Conference on Neural Information Processing Systems*, NIPS'02, page 471–478, Cambridge, MA, USA. MIT Press.

Checklist

The checklist follows the references. For each question, choose your answer from the three possible options: Yes, No, Not Applicable. You are encouraged to include a justification to your answer, either by referencing the appropriate section of your paper or providing a brief inline description (1-2 sentences). Please do not modify the questions. Note that the Checklist section does not count towards the page limit. Not including the checklist in the first submission won't result in desk rejection, although in such case we will ask you to upload it during the author response period and include it in camera ready (if accepted).

In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Materials

A Proofs for Section 3

Proof of Lemma 1. Let $L_n \subseteq \mathcal{H}_K$ be the span of $\{K(x_i,\cdot)\}_1^n$. The squared distance between $K(x,\cdot)$ and L_n is

$$\inf_{f\in L_n} \|K(x,\cdot)-f\|^2_{\mathcal{H}_K} = \min_{[c_i]_1^n} \|K(x,\cdot)-\sum_{i=1}^n c_i K(x_i,\cdot)\|^2_{\mathcal{H}_K} =$$
$$\min_{[c_i]_1^n} K(x,x) - 2\sum_{i=1}^n K(x,x_i)c_i + \sum_{i,j=1}^n K(x_i,x_j)c_ic_j.$$

The latter minimum is attained at $[c_i]_1^n = K[X_n,X_n]^{-1}K[X_n,x]$ and is equal to

$$K(x,x) - K[X_n,x]^\top K[X_n,X_n]^{-1}K[X_n,x],$$

i.e. to the Schur complement $(K[(X_n,x),(X_n,x)] \mid K[X_n,X_n])$. Thus, by the definition of $w_K(n)$, we have

$$w_K(n)^2 \le \sup_{x\in\Omega}(K[(X_n,x),(X_n,x)] \mid K[X_n,X_n]).$$

Since $(K[(X_n,x),(X_n,x)] \mid K[X_n,X_n]) = \frac{\det(K[(X_n,x),(X_n,x)])}{\det(K[X_n,X_n])}$ the latter supremum is attained an $x_{n+1} = f_n(X_n)$ and, therefore,

$$w_K(n)^2 \le (K[(X_n,f_n(X_n)),(X_n,f_n(X_n))] \mid K[X_n,X_n]).$$

Lemma proved. □

Lemma 3. *The sequence $\{x_i\}_1^\infty \subseteq \Omega$ defined by (7) satisfies*

$$\prod_{i=1}^n \lambda_i(\mathrm{O}_{K,\mu_n}) \ge \prod_{i=1}^n \frac{w_K(i-1)^2}{ei},$$

where $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$.

Proof. Note that $L_2(\Omega,\mu_n)$ is n-dimensional and eigenvalues of O_{K,μ_n} are equal to eigenvalues of the empirical kernel matrix $[\frac{1}{n}K(x_i,x_j)]_{i,j=1}^n$. Let again denote $(x_1,\cdots,x_n)$ by X_n. Using Lemma 1, for any natural k we obtain

$$\frac{\prod_{i=1}^k \lambda_i(\mathrm{O}_{K,\mu_k})}{\prod_{i=1}^{k-1}\lambda_i(\mathrm{O}_{K,\mu_{k-1}})} = \frac{k^{-k}\prod_{i=1}^k \lambda_i(K[X_k,X_k])}{(k-1)^{-(k-1)}\prod_{i=1}^{k-1}\lambda_i(K[X_{k-1},X_{k-1}])} =$$
$$\frac{1}{k}(1-\frac{1}{k})^{k-1}(K[X_k,X_k] \mid K[X_{k-1},X_{k-1}]) \ge \frac{1}{ek}w_K(k-1)^2.$$

Therefore,

$$\prod_{i=1}^n \lambda_i(\mathrm{O}_{K,\mu_n}) \ge \prod_{k=1}^n \frac{w_K(k-1)^2}{ek}.$$

□

Proof of the second part of Theorem 3. Let us assume the opposite, i.e. $\limsup_{n\to+\infty}\sup_{\mu}\frac{n\lambda_n(\mathrm{O}_{K,\mu})}{w(n)^2}<\frac{1}{e}$, and obtain a contradiction. The latter implies $\lambda_n(\mathrm{O}_{K,\mu})\leq\frac{\alpha(n)w(n)^2}{n}$ for any μ and n, where $\alpha(n)$ is decreasing and $\lim_{n\to+\infty}\alpha(n)<\frac{1}{e}$. Therefore, for any μ and n, we have

$$\prod_{i=1}^{n}\lambda_i(\mathrm{O}_{K,\mu})\leq\prod_{i=1}^{n}\frac{\alpha(i)w(i)^2}{i}.$$

Let us now set $\mu=\mu_n$ where μ_n is defined as in Lemma 3. From that lemma we conclude

$$\prod_{i=1}^{n}\frac{\alpha(i)w(i)^2}{i}\geq\prod_{i=1}^{n}\frac{w_K(i-1)^2}{ei},$$

i.e. $w_K(n)^2\prod_{i=1}^{n}e\alpha(i)\geq w_K(0)^2$. Since $\prod_{i=1}^{n}e\alpha(i)\overset{n\to+\infty}{\to}0$ and $w_K(n)^2\overset{n\to+\infty}{\to}0$, we obtain the contradiction $w_K(0)^2\leq 0$. □

B Proofs for Section 4

In this section, the probability measure μ is fixed and eigenvalues $\{\lambda_i(\mathrm{O}_{K,\mu})\}$ are simply denoted by $\{\lambda_i\}$. Given $S=(X_1,\cdots,X_n)\sim\mu^n$, the empirical local Rademacher compexity is defined by

$$R_n[\mathcal{F}](S,r)=\mathbb{E}_\sigma[\sup_{f\in\mathcal{F},\|f\|^2_{L_2(\Omega,\mu)}\leq r}\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(X_i)],$$

where $\sigma=(\sigma_1,\cdots,\sigma_n)$ are independent and uniformly distributed over $\{-1,1\}$ (also, independent from S). Then, $R_n[\mathcal{F}](r)=\mathbb{E}_S\big[R_n[\mathcal{F}](S,r)\big]$ is called the local Rademacher compexity. It was shown in [Mendelson, 2002] that

$$R_n[B_{\mathcal{H}_K}](r)\leq\big(\frac{2}{n}\sum_{i=1}^{\infty}\min(\lambda_i,r)\big)^{\frac{1}{2}},\tag{8}$$

for every $r>0$.

Recall that a function $\psi:[0,\infty)\to[0,\infty)$ is called sub-root if it is nonnegative, nondecreasing and if $r\to\frac{\psi(r)}{\sqrt{r}}$ is nonincreasing for $r>0$.

Lemma 4. *For any $\varepsilon>0$, there exists a sub-root function ψ_ε and $N_\varepsilon\in\mathbb{N}$ such that $\forall r>0,\big(\frac{2}{n}\sum_{i=1}^{\infty}\min(\lambda_i,r)\big)^{\frac{1}{2}}\leq\psi_\varepsilon(r)$ and for any natural $N\geq 2N_\varepsilon$ and $r_N=(N/2)^{-1-\frac{2}{d_K+\varepsilon}}$ we have*

$$\psi_\varepsilon(r_N)\leq\sqrt{2(3+d_K+\varepsilon)}r_N^{\frac{1}{2+d_K+\varepsilon}}n^{-\frac{1}{2}}.$$

Proof. Since $d_K=\limsup_{i\to+\infty}\frac{\log i}{\log(1/w_K(i))}$ and $\lim_{i\to+\infty}w_K(i)=0$, for any $\varepsilon>0$ there is $N_\varepsilon\in\mathbb{N}$ such that $\frac{\log i}{\log(1/w_K(i))}<d_K+\varepsilon$ and $w_K(i)<1$ whenever $i>N_\varepsilon$. Thus, $w_K(i)<i^{-\frac{1}{d_K+\varepsilon}}$ for $i>N_\varepsilon$. The corollary 3 gives us $\lambda_i\leq\frac{w_K(\lfloor i/2\rfloor)^2}{(i-1)/2}$. Thus, for $i>2N_\varepsilon+1$ we obtain $\lambda_i\leq\frac{((i-1)/2)^{-\frac{2}{d_K+\varepsilon}}}{(i-1)/2}\leq((i-1)/2)^{-1-\frac{2}{d_K+\varepsilon}}$.

The RHS of the inequality (8) can be bounded as

$$\big(\frac{2}{n}\sum_{i=1}^{\infty}\min(\lambda_i,r)\big)^{\frac{1}{2}}\leq\left(\frac{2Nr}{n}+\frac{2}{n}\sum_{i=N+1}^{\infty}((i-1)/2)^{-1-\frac{2}{d_K+\varepsilon}}\right)^{\frac{1}{2}},$$

for any $r>0$ and natural $N\geq 2N_\varepsilon+1$. Since $\sum_{i=N+1}^{\infty}((i-1)/2)^{-1-\frac{2}{d_K+\varepsilon}}<2\int_{(N-1)/2}^{\infty}x^{-1-\frac{2}{d_K+\varepsilon}}dx=2^{\frac{2}{d_K+\varepsilon}}(d_K+\varepsilon)(N-1)^{-\frac{2}{d_K+\varepsilon}}$, we have

$$\big(\frac{2}{n}\sum_{i=1}^{\infty}\min(\lambda_i,r)\big)^{\frac{1}{2}}\leq\psi_\varepsilon(r),$$

where $\psi_\varepsilon(r) = \min_{N\in\mathbb{N}\cap[2N_\varepsilon+1,\infty)} \left(\frac{2Nr}{n} + \frac{2^{1+\frac{2}{d_K+\varepsilon}}}{n}(d_K+\varepsilon)(N-1)^{-\frac{2}{d_K+\varepsilon}}\right)^{\frac{1}{2}}$.

Let us prove that $\psi_\varepsilon(r)$ is sub-root. The fact that it is nonnegative is obvious. Let us prove that $r \to \frac{\psi_\varepsilon(r)}{\sqrt{r}}$ is nonincreasing. Indeed, for any $r_1 > r_2 > 0$ we have

$$\frac{2N}{n} + \frac{2^{1+\frac{2}{d_K+\varepsilon}}}{r_1 n}(d_K+\varepsilon)(N-1)^{-\frac{2}{d_K+\varepsilon}} \le \frac{2N}{n} + \frac{2^{1+\frac{2}{d_K+\varepsilon}}}{r_2 n}(d_K+\varepsilon)(N-1)^{-\frac{2}{d_K+\varepsilon}},$$

for any natural $N \ge 2N_\varepsilon + 1$. Therefore, $\frac{\psi_\varepsilon(r_1)}{\sqrt{r_1}} \le \frac{\psi_\varepsilon(r_2)}{\sqrt{r_2}}$. Analogously, it is proved that $\psi_\varepsilon(r_1) \ge \psi_\varepsilon(r_2)$.

Let us set $r_{N-1} = 2^{1+\frac{2}{d_K+\varepsilon}}(N-1)^{-1-\frac{2}{d_K+\varepsilon}}$ for some natural $N \ge 2N_\varepsilon + 1$. By construction,

$$\psi(r_{N-1}) \le \left(\frac{2^{1+\frac{2}{d_K+\varepsilon}}}{n}\left(\frac{2}{N-1} + 2 + d_K + \varepsilon\right)(N-1)^{-\frac{2}{d_K+\varepsilon}}\right)^{\frac{1}{2}} \le$$
$$\sqrt{2(3+d_K+\varepsilon)}\, r_{N-1}^{\frac{1}{2+d_K+\varepsilon}} n^{-\frac{1}{2}}.$$

□

The following theorem is a direct consequence of Corollary 5.3. from [Bartlett et al., 2002].

Theorem 8 ([Bartlett et al., 2002])**.** *Let $\max_{x\in\Omega} K(x,x) \le 1$ and $l : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ be a loss function, that satisfies condition (1)-(3). Let $\hat{f} = \arg\min_{f\in B_{\mathcal{H}_K}} \sum_{i=1}^n l(f(X_i), Y_i)$ and $f^* = \arg\min_{f\in B_{\mathcal{H}_K}} \mathbb{E}[l(f(X), Y)]$. Assume ψ is a sub-root function for which $\psi(r) \ge BL\cdot R_n[2B_{\mathcal{H}_K}](\frac{r}{L^2})$. Then, for any $x > 0$ and any $r \ge \psi(r)$, with probability at least $1 - e^{-x}$,*

$$\mathbb{E}[l(\hat{f}(X), Y)] - \mathbb{E}[l(f^*(X), Y)] \le 705\frac{r}{B} + \frac{(11L+27B)x}{n}.$$

Proof of Theorem 5. Let us fix $\varepsilon > 0$ and consider a sub-root function ψ_ε whose existence is guaranteed by Lemma 4. Using the inequality (8), we have

$$BL\cdot R_n[2B_{\mathcal{H}_K}](\frac{r}{L^2}) \le BL\Big(\frac{2}{n}\sum_{i=1}^\infty \min(4\lambda_i, \frac{r}{L^2})\Big)^{\frac{1}{2}} \le 2BL\psi_\varepsilon(\frac{r}{4L^2}).$$

Therefore, let us define $\psi(r) = 2BL\psi_\varepsilon(\frac{r}{4L^2})$. By construction, ψ is a sub-root function that satisfies $\psi(r) \ge BL\cdot R_n[2B_{\mathcal{H}_K}](\frac{r}{L^2})$. According to Lemma 4, for any natural $N \ge 2N_\varepsilon$ and $r_N = (N/2)^{-1-\frac{2}{d_K+\varepsilon}}$ we have

$$2BL\psi_\varepsilon(r_N) \le 2BL\sqrt{2(3+d_K+\varepsilon)}\, r_N^{\frac{1}{2+d_K+\varepsilon}} n^{-\frac{1}{2}}.$$

Let us assume $\tilde{r} = 4L^2 r_N$ and find such a natural $N \ge 2N_\varepsilon$ that $2BL\sqrt{2(3+d_K+\varepsilon)}\, r_N^{\frac{1}{2+d_K+\varepsilon}} n^{-\frac{1}{2}} \le \tilde{r}$. In other words,

$$2BL\sqrt{2(3+d_K+\varepsilon)}\, r_N^{\frac{1}{2+d_K+\varepsilon}} n^{-\frac{1}{2}} \le 4L^2 r_N,$$

or

$$\frac{B}{2L}\sqrt{2(3+d_K+\varepsilon)} n^{-\frac{1}{2}} \le r_N^{\frac{1+d_K+\varepsilon}{2+d_K+\varepsilon}} = (N/2)^{-\frac{1+d_K+\varepsilon}{d_K+\varepsilon}}.$$

If $2\left(\frac{2nL^2}{B^2(3+d_K+\varepsilon)}\right)^{\frac{d_K+\varepsilon}{2(1+d_K+\varepsilon)}} \ge 2N_\varepsilon$, we can set $N = \lceil 2\left(\frac{2nL^2}{B^2(3+d_K+\varepsilon)}\right)^{\frac{d_K+\varepsilon}{2(1+d_K+\varepsilon)}} \rceil$. Thus,

$$\psi(\tilde{r}) \le \tilde{r} = 4L^2 r_N \le 4L^2\left(\frac{2\left(\frac{2nL^2}{B^2(3+d_K+\varepsilon)}\right)^{\frac{d_K+\varepsilon}{2(1+d_K+\varepsilon)}}}{2}\right)^{-1-\frac{2}{d_K+\varepsilon}} \le$$
$$4L^2\left(\frac{2nL^2}{B^2(3+d_K+\varepsilon)}\right)^{-\frac{2+d_K+\varepsilon}{2(1+d_K+\varepsilon)}}.$$

From Theorem 8 we obtain

$$\mathbb{E}[l(\hat{f}(X),Y)] - \mathbb{E}[l(f^*(X),Y)] \le 705\frac{\tilde{r}}{B} + \frac{(11L+27B)x}{n} \le$$
$$CB^{\frac{1}{1+d_K+\varepsilon}} L^{1-\frac{1}{1+d_K+\varepsilon}} n^{-\frac{2+d_K+\varepsilon}{2(1+d_K+\varepsilon)}} + \frac{(11L+27B)x}{n}.$$

with probability at least $1-e^{-x}$, where $C = 705 \cdot 2^{1.5} < 1995$. □

C Proof for Section 5

Proof of Lemma 2. Let us denote $S_0(x,y) = K(x,y)$ and $S_n(x,y) = K(x,y) - K[X_n,x]^\top K[X_n,X_n]^{-1}K[X_n,y]$ for any $n \in \mathbb{N}$. Note that $S_i(x) = S_i(x,x)$.

The projection of the function $K(x,\cdot)$ onto the span of $\{K(x_1,\cdot),\cdots,K(x_n,\cdot)\}$ in $\mathcal{H}_K$, denoted by $\mathrm{pr}_{\mathrm{span}(X_n)}K(x,\cdot)$, equals

$$[\mathrm{pr}_{\mathrm{span}(X_n)}K(x,\cdot)](y) = K[X_n,x]^\top K[X_n,X_n]^{-1}K[X_n,y].$$

Therefore,

$$[\mathrm{pr}_{\mathrm{span}(X_n)^\perp}K(x,\cdot)](y) = K(x,y) - K[X_n,x]^\top K[X_n,X_n]^{-1}K[X_n,y].$$

and

$$\langle \mathrm{pr}_{\mathrm{span}(X_n)^\perp}K(x,\cdot), \mathrm{pr}_{\mathrm{span}(X_n)^\perp}K(y,\cdot)\rangle_{\mathcal{H}_K} =$$
$$\langle K(x,\cdot) - K(x,X_n)K(X_n,X_n)^{-1}K(X_n,\cdot),$$
$$K(y,\cdot) - K(y,X_n)K(X_n,X_n)^{-1}K(X_n,\cdot)\rangle_{\mathcal{H}_K} =$$
$$K(x,y) - K(x,X_n)K(X_n,X_n)^{-1}K(X_n,y) = S_n(x,y).$$

Let $d_n(x,y) = \sqrt{S_n(x,x) + S_n(y,y) - 2S_n(x,y)}$. Since $\langle \mathrm{pr}_{\mathrm{span}(X_n)^\perp}f, \mathrm{pr}_{\mathrm{span}(X_n)^\perp}g\rangle_{\mathcal{H}_K}$ is a semi-definite bilinear form on $\mathcal{H}_K$, it satisfies the triangle inequality, i.e.

$$d_n(x,y) = \langle \mathrm{pr}_{\mathrm{span}(X_n)^\perp}K(x,\cdot) - K(y,\cdot), \mathrm{pr}_{\mathrm{span}(X_n)^\perp}K(x,\cdot) - K(y,\cdot)\rangle_{\mathcal{H}_K}^{\frac{1}{2}} \ge$$
$$\langle \mathrm{pr}_{\mathrm{span}(X_n)^\perp}K(x,\cdot), \mathrm{pr}_{\mathrm{span}(X_n)^\perp}K(x,\cdot)\rangle_{\mathcal{H}_K}^{\frac{1}{2}} - \langle \mathrm{pr}_{\mathrm{span}(X_n)^\perp}K(y,\cdot), \mathrm{pr}_{\mathrm{span}(X_n)^\perp}K(y,\cdot)\rangle_{\mathcal{H}_K}^{\frac{1}{2}}.$$

Thus,

$$\sqrt{S_n(x,x)} - \sqrt{S_n(y,y)} \le d_n(x,y) \le d_1(x,y) = \varrho(x,y).$$

□

Lemma 5. *Let $Z_1,\cdots,Z_N \sim^{\mathrm{iid}} \mu$ and let*

$$f(\varepsilon) = \min_{x\in\Omega} \mu(B(x,\varepsilon)\cap\Omega).$$

Then, for any $\varepsilon > 0$ we have

$$\mathbb{P}[\{Z_1,\cdots,Z_N\} \text{ is a } 2\varepsilon-\text{net in } (\Omega,\varrho)] \ge 1 - \mathcal{N}(\varepsilon,\Omega,\varrho)e^{-Nf(\varepsilon)}.$$

Proof. Let $\varepsilon > 0$ and $M = \mathcal{N}(\varepsilon,\Omega,\varrho)$. By construction, there are $z_1,z_2,\cdots,z_M \in \Omega$ satisfying $\bigcup_{i=1}^M B(z_i,\varepsilon) \supseteq \Omega$. If $\{Z_1,\cdots,Z_N\} \cap B(z_i,\varepsilon) \neq \emptyset$ for all $i : 1 \le i \le M$, then $\{Z_1,\cdots,Z_N\}$ is the 2ε-net in (Ω,ϱ).

The probability that there is at least one i such that $\{Z_1,\cdots,Z_N\} \cap B(z_i,\varepsilon) = \emptyset$ is not more than $\sum_{i=1}^M (1-\mu(B(z_i,\varepsilon)\cap\Omega))^N$, which does not exceed $M(1-\min_{x\in\Omega}\mu(B(x,\varepsilon)\cap\Omega))^N$. Thus,

$$\mathbb{P}[\{Z_1,\cdots,Z_N\} \text{ is a } 2\varepsilon-\text{net in } (\Omega,\varrho)] \ge$$
$$1 - M(1-f(\varepsilon))^N \ge 1 - Me^{-Nf(\varepsilon)}.$$

Lemma proved. □

Corollary 1. *Suppose that the measure μ is C-uniform over Ω and $Z_1,\cdots,Z_N \sim^{\text{iid}} \mu$. For any sequence $\{x_t\}_{t=1}^{\infty}\subseteq\Omega$ and any $\delta\in(0,1)$, we have*

$$\mathbb{P}[\{Z_1,\cdots,Z_N\} \text{ is a } 2\varepsilon-\text{net in } (\Omega,\varrho)]\geq 1-\delta,$$

provided that $N\geq \frac{\log \mathcal{N}(\varepsilon,\Omega,\varrho)+\log\frac{1}{\delta}}{C\varepsilon^{d_\varrho}}$.

Proof. Since $f(\varepsilon)\geq C\varepsilon^{d_\varrho}$, $\mathcal{N}(\varepsilon,\Omega,\varrho)e^{-Nf(\varepsilon)}\leq \mathcal{N}(\varepsilon,\Omega,\varrho)e^{-NC\varepsilon^{d_\varrho}}$. Then, $\delta\geq \mathcal{N}(\varepsilon,\Omega,\varrho)e^{-NC\varepsilon^{d_\varrho}}$ is equivalent to $N\geq \frac{\log \mathcal{N}(\varepsilon,\Omega,\varrho)+\log\frac{1}{\delta}}{C\varepsilon^{d_\varrho}}$. Therefore

$$\mathbb{P}[\{Z_1,\cdots,Z_N\} \text{ is a } 2\varepsilon-\text{net in } (\Omega,\varrho)]\geq 1-\delta,$$

if $N\geq \frac{\log \mathcal{N}(\varepsilon,\Omega,\varrho)+\log\frac{1}{\delta}}{C\varepsilon^{d_\varrho}}$. □

Proof of Theorem 7. We only to use Corollary 1 in combination with Theorem 6. □

C.1 Sample size for smaller t

Let us call the pair (μ,K) non-degenerate if for $X_1,\cdots,X_n\sim^{\text{iid}}\mu, n\geq 3$ the probability of the event

$$K(X_1,X_1)-K([X_j]_{j=3}^n,X_1)^\top K([X_j]_{j=3}^n,[X_j]_{j=3}^n)^{-1}K([X_j]_{j=3}^n,X_1)=$$
$$K(X_2,X_2)-K([X_j]_{j=3}^n,X_2)^\top K([X_j]_{j=3}^n,[X_j]_{j=3}^n)^{-1}K([X_j]_{j=3}^n,X_2)$$

is zero.

Lemma 6. *Suppose that (μ,K) is non-degenerate, $Z_1,\cdots,Z_N\sim^{\text{iid}}\mu$ and*

$$f(\varepsilon)=\min_{x\in\Omega}\mu(B(x,\varepsilon)\cap\Omega).$$

Let $\{w_t\}_{t=0}^{T-1}$ be an output of Algorithm 1 for $\tilde{\Omega}=\{Z_1,\cdots,Z_N\}$. For any $t\in\{1,\cdots,T\}$ and $\varepsilon>0$, we have

$$\mathbb{P}[w_{t-1}<w_K(t-1)-\varepsilon]\leq \binom{N}{t}e^{-f(\varepsilon)(N-t)}.$$

Proof. Let μ^N be the product measure $\mu\times\cdots\times\mu$ on Ω^N. We will treat elements $\{x_t\}_{t=1}^T$, $\{w_t\}_{t=0}^{T-1}$ generated by the Algorithm 1 as random variables $\Omega^N\to\mathbb{R}$ depending on the random input $\tilde{\Omega}=\{Z_1,\cdots,Z_N\}$ where $Z_1,\cdots,Z_N\sim^{\text{iid}}\mu$.

A sequence of elements $a_1,\cdots,a_t\in\Omega$ defines the sequence of functions $\{S_i^{\mathbf{a}}\}_{i=0}^t$ by $S_i^{\mathbf{a}}(x)=K(x,x)-K([a_j]_{j=1}^i,x)^\top K([a_j]_{j=1}^i,[a_j]_{j=1}^i)^{-1}K([a_j]_{j=1}^i,x)$. Let A_t be a set of tuples $\mathbf{a}=(a_1,\cdots,a_t)\in\Omega^t$ such that $S_{i-1}^{\mathbf{a}}(a_i)\geq S_{i-1}^{\mathbf{a}}(a_j)$ for any $i,j\in\{1,\cdots,t\}$. We denote the event $x_1=a_1,\cdots,x_t=a_t$ by $E_{\mathbf{a}}$. Note that for $E_{\mathbf{a}}=\emptyset$ if $\mathbf{a}\notin A_t$.

Let $P(N,t)$ denote the set of all t-element arrangements of $\{1,\cdots,N\}$. The event $E_{a_1,\cdots,a_t}$ can be represented as

$$E_{a_1,\cdots,a_t}=\bigcup_{(i_1,\cdots,i_t)\in P(N,t)} C_{i_1,\cdots,i_t}^{a_1,\cdots,a_t},$$

where $C_{i_1,\cdots,i_t}^{a_1,\cdots,a_t}$ denotes the event $Z_{i_1}=a_1,\cdots,Z_{i_t}=a_t$, $S_{u-1}^{\mathbf{a}}(a_u)\geq S_{u-1}^{\mathbf{a}}(Z_j)$, $\forall u\in\{1,\cdots,t\}$, $j\in\{1,\cdots,N\}$. Due to symmetry, all probabilities $\mathbb{P}[C_{i_1,\ldots,i_t}^{a_1,\ldots,a_t}]$ are equal for a fixed $\mathbf{a}$. Therefore, for any Borel set $B\subseteq\Omega^t$, we have

$$\mathbb{P}[(x_1,\cdots,x_t)\in B]=\mathbb{P}[\bigcup_{(a_1,\cdots,a_t)\in B}E_{a_1,\cdots,a_t}]\leq |P_t^N|\cdot\mathbb{P}[\bigcup_{(a_1,\cdots,a_t)\in B}C_{1,\cdots,t}^{a_1,\cdots,a_t}].$$

We are specifically interested in the case of

$$B=\{(a_1,\cdots,a_t)\in A_t\mid S_{t-1}^{\mathbf{a}}(a_t)<w_K(t-1)-\varepsilon\},$$

due to $\mathbb{P}[w_{t-1} < w_K(t-1) - \varepsilon] = \mathbb{P}[(x_1, \cdots, x_t) \in B]$. We will estimate $\mathbb{P}[\bigcup_{(a_1,\cdots,a_t)\in B} C_{1,\cdots,t}^{a_1,\cdots,a_t}]$ for this specific choice of B. Let us denote

$$\Omega_{a_1,\cdots,a_t} = \{x \in \Omega \mid S_{u-1}^{\mathbf{a}}(a_u) \geq S_{u-1}^{\mathbf{a}}(x), \forall u \in \{1, \cdots, t\}\}.$$

The event $C_{1,\cdots,t}^{a_1,\cdots,a_t}$, as an element of sigma algebra over Ω^N, satisfies

$$C_{1,\cdots,t}^{a_1,\cdots,a_t} = \{(a_1, \cdots, a_t)\} \times \Omega_{a_1,\cdots,a_t}^{N-t}.$$

By Fubini's theorem, we have

$$\mathbb{P}[(x_1, \cdots, x_t) \in B] \leq \frac{N!}{(N-t)!} \int_B \Big(\int_{\Omega_{\mathbf{a}}^{N-t}} d\mu^{N-t} \Big) d\mu^t(\mathbf{a}) =$$
$$\frac{N!}{(N-t)!} \int_B \mu^{N-t}(\Omega_{\mathbf{a}}^{N-t}) d\mu^t(\mathbf{a}) = \frac{N!}{(N-t)!} \int_B \mu(\Omega_{\mathbf{a}})^{N-t} d\mu^t(\mathbf{a}).$$

Let $a_t^* \in \arg\max_{x\in\Omega} \sqrt{S_{t-1}^{\mathbf{a}}(x)}$. For any $z \in \Omega_{\mathbf{a}}$ we have (a) $\sqrt{S_{t-1}(a_t)} \geq \sqrt{S_{t-1}(z)}$, (b) $w_K(t-1) \leq \sqrt{S_{t-1}(a_t^*)}$. Since $\sqrt{S_{t-1}}$ is 1-Lipschitz (Lemma 2), we have $w_K(t-1) - \sqrt{S_{t-1}(a_t)} \leq \sqrt{S_{t-1}(a_t^*)} - \sqrt{S_{t-1}(z)} \leq \varrho(a_t^*, z)$. Thus, $\sqrt{S_{t-1}(a_t)} < w_K(t-1) - \varepsilon$ implies $\varrho(a_t^*, z) > \varepsilon$. This leads to

$$\mu(\Omega_{\mathbf{a}}) \leq \mu(\{z \in \Omega \mid \varrho(a_t^*, z) > \varepsilon\}) \leq 1 - f(\varepsilon),$$

for any $\mathbf{a} \in B$. Thus, we proved

$$\mathbb{P}[(x_1, \cdots, x_t) \in B] \leq \frac{N!}{(N-t)!}(1 - f(\varepsilon))^{N-t} \int_{A_t} d\mu^t(\mathbf{a}).$$

Due to non-degeneracy condition, $t! \int_{A_t} d\mu^t(\mathbf{a}) = \int_{\Omega^t} d\mu^t(\mathbf{a}) = 1$, and we obtain

$$\mathbb{P}[w_{t-1} < w_K(t-1) - \varepsilon] \leq \binom{N}{t}(1 - f(\varepsilon))^{N-t} \leq \binom{N}{t} e^{-f(\varepsilon)(N-t)}.$$

□

Theorem 9. *Suppose that the measure μ is C-uniform over Ω, (μ, K) is non-degenerate, and $Z_1, \cdots, Z_N \sim^{\text{iid}} \mu$. Let $\{w_t\}_{t=0}^{T-1}$ be an output of Algorithm 1 for $\tilde{\Omega} = \{Z_1, \cdots, Z_N\}$. Let $\delta \in (0,1)$ and $\varepsilon > 0$. If $10 \leq T \leq \frac{NC\varepsilon^{d_\varrho} + \log\delta}{\log N}$, then*

$$w_{t-1} \geq w_K(t-1) - \varepsilon, t = 1, \cdots, T$$

with probability at least $1 - \delta$.

Proof. Using the previous lemma we obtain

$$\mathbb{P}[w_{t-1} \geq w_K(t-1) - \varepsilon, t = 1, \cdots, T] \geq 1 - \sum_{t=1}^{T} \binom{N}{t} e^{-f(\varepsilon)(N-t)}.$$

The latter sum can be bounded by

$$\sum_{t=1}^{T} \binom{N}{t} e^{-f(\varepsilon)(N-t)} \leq T \binom{N}{T} e^{-f(\varepsilon)(N-T)} \leq T(\frac{Ne^2}{T})^T e^{-f(\varepsilon)N}.$$

Thus, we need to have $T(\frac{Ne^2}{T})^T e^{-f(\varepsilon)N} \leq \delta$ to satisfy $w_{t-1} \geq w_K(t-1) - \varepsilon, t = 1, \cdots, T$ with probability at least $1 - \delta$. This condition is equivalent to

$$-(T-1)\log T + T\log N + 2T \leq f(\varepsilon)N + \log\delta.$$

Since $-(T-1)\log T + 2T \leq 0$ for $T = 10, 11, \cdots$, it is sufficient to satisfy

$$f(\varepsilon)N + \log\delta \geq T\log N,$$

i.e. $T \leq \frac{NC\varepsilon^{d_\varrho} + \log\delta}{\log N}$. □

C.2 Implementation of the Algorithm 1 and its complexity

The finite sample implementation. The actual implementation of the algorithm differs slightly from its formal presentation. Let $|\tilde{\Omega}| = N$. At the $(t+1)$-st iteration, given the inverse matrix $K[X_t, X_t]^{-1}$, the algorithm selects the next point $x_{t+1} \in \tilde{\Omega}$ by maximizing the expression

$$K(x,x) - K[X_t, x]^\top K[X_t, X_t]^{-1} K[X_t, x],$$

which requires $\mathcal{O}(t^2 N)$ arithmetic operations (including multiplications, additions, and comparisons) and $\mathcal{O}(tN)$ kernel evaluations.

The inverse matrix $K[X_t, X_t]^{-1}$ can be updated iteratively using the formula:

$$K[X_t, X_t]^{-1} = \begin{bmatrix} K[X_{t-1}, X_{t-1}]^{-1} + A_t A_t^\top s_t^{-1} & -A_t s_t^{-1} \\ -A_t^\top s_t^{-1} & s_t^{-1} \end{bmatrix},$$

where $A_t = K[X_{t-1}, X_{t-1}]^{-1} K(X_{t-1}, x_t)$ and the Schur complement

$$s_t = K(x_t, x_t) - K[X_{t-1}, x_t]^\top K[X_{t-1}, X_{t-1}]^{-1} K[X_{t-1}, x_t]$$

is equal to $w(t-1)^2$, and therefore, has already been computed at a previous iteration.

Thus, the total computational complexity of the algorithm is bounded by $\mathcal{O}(T^3 N)$ arithmetic operations and $\mathcal{O}(T^2 N)$ kernel evaluations. In practice, the latter often dominates the runtime, especially when kernel evaluations are costly — as is the case with NNGP and NTK kernels (given non-analytically).

The L-BFGS–based implementation. Now let $\tilde{\Omega} = \Omega$. To select the next point $x_{t+1} \in \Omega$, any nonconvex optimization method can be employed. In our experiments, we used the Limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) algorithm to maximize

$$K(x,x) - K[X_t, x]^\top K[X_t, X_t]^{-1} K[X_t, x].$$

The inverse matrix $K[X_t, X_t]^{-1}$ was updated in the same manner as in the finite-sample implementation. The main drawback of this approach is that, at the $(t+1)$-st iteration, the optimized objective is non-convex, so we cannot guarantee that the computed value at the point x_{t+1} is within ε of the true maximum. Consequently, the resulting sequence $\{w(t)\}$ may fail to be ε-close to any upper bound on the actual n-widths. As shown in Figure 4, the L-BFGS–based implementation produces results that are very close to those of the finite-sample approach, while having the advantage of being completely independent of N (which must be large even for $T \approx 300$ in the finite sample setting). We also report cases in which the output sequence $\{w(t)\}$ substantially underestimates the actual n-widths. A possible remedy is to invoke L-BFGS multiple times with independent initializations to obtain a more accurate estimate of the optimal x_{t+1}.

D Additional experiments

D.1 Details of experiments with fractals

In experiments with the Laplace kernel on fractals we used the following values of parameters:

- Cantor set: $N = 2^{15}$, $T = 300$;
- Weierstrass function graph: $N = 2^{24}$, $T = 200$;
- Sierpiński carpet: $N = 8^8$, $T = 200$;
- Menger sponge: $N = 20^5$, $T = 300$;
- Lorenz attractor: $N = 10^6$, $T = 300$.

Our empirical estimation of the effective dimension has two main limitations. First, the relation $w_K(n) \asymp n^{-1/d_K}$ holds only asymptotically, whereas we estimate d_K using n-width approximations for $n \leq 300$. Second, the estimators we employ are upper bounds on the true n-widths and are therefore not unbiased.

To demonstrate that our estimates are nevertheless quite accurate, we also computed lower bounds on the n-widths using Ismagilov's theorem:

$$w_K(n) \geq \sqrt{\sum_{i>n} \lambda_i(\mathcal{O}_{K,\mu})}.$$

We calculated these lower bounds for the first 400 n-widths based on the eigenvalues $\hat{\lambda}_i$ of the empirical kernel matrix $[K(x_i, x_j)]_{i,j=1}^{M}$ with $M = 10{,}000$.

As shown in Figure 3, Ismagilov's lower bounds differ from the upper bounds produced by Algorithm 1 only by a constant factor asymptotically, and their log-log plots have nearly identical slopes. This provides additional evidence that our empirical estimates of the effective dimension are quite accurate.

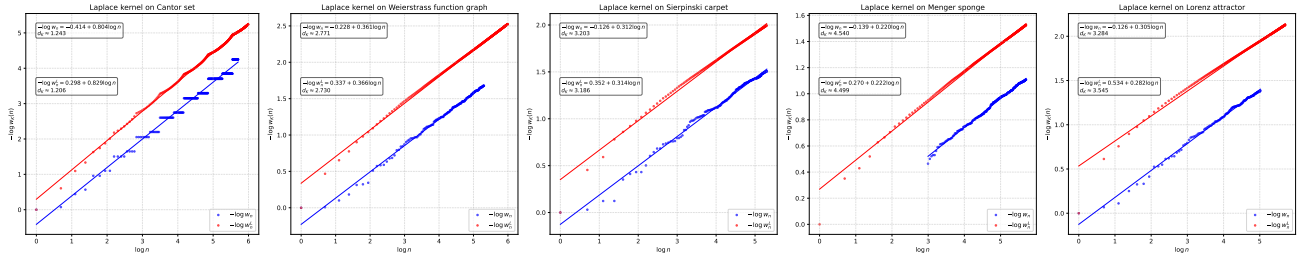

Figure 3: Plot of $-\log(w_n^L)$ (and $-\log(w_n)$) versus $\log n$, where $w_n^L = \sqrt{\sum_{i>n} \hat{\lambda}_i}$ for the Laplace kernel on various fractal sets. The slope of the fitted line is computed using standard linear regression. Estimated effective dimension is defined as inversely proportional to the slope.

D.2 Details of experiments with the excess risk decay rates

In our experiments, we reduced the constrained KRR problem (which involves optimization over $B_{\mathcal{H}_K}$) to an unconstrained KRR with a regularization term $\lambda \|f\|^2_{\mathcal{H}_K}$, denoted by KRR($\lambda$). The optimal value of λ was determined via a half-interval search with 30 iterations, ensuring that the solution f_λ of KRR(λ) satisfies $\|f_\lambda\|_{\mathcal{H}_K} = 1$. This approach is roughly 30 times more expensive than standard KRR, yet remains computationally feasible. For each kernel, we trained $\hat{f}$ on a training set of size n and evaluated the excess risk on a test set of size $n_{\text{test}} = 10000$, averaging the results over 10 independent trials.

D.3 Overestimation and underestimation of n-widths

Algorithm 1 can significantly overestimate the Kolmogorov n-widths. Figure 4 shows the estimated upper bounds on the effective dimension obtained using this algorithm. Notably, a substantial overestimation is observed for the Gaussian kernel. The most plausible explanation is that, in this case, the optimal n-dimensional subspace $L_n \subseteq \mathcal{H}_K$, which minimizes the expression in equation (4), cannot be expressed as the span of n kernel sections of the form $K(x, \cdot)$. This representability condition is essential for the output of Algorithm 1 to accurately approximate the true n-widths.

In the first two plots (corresponding to $a = \frac{1}{2}$ and $a = 1$), we observe that the estimated upper bounds on the effective dimension fall below the true effective dimension as the ambient dimension increases (e.g., for $d = 8$). This underestimation arises from an insufficient sample size — specifically, a small value of the parameter $|\tilde{\Omega}| = N$ in the finite sample implementation of Algorithm 1. As follows from Remark 3 following Theorem 7, achieving uniform ε-accuracy requires a sample size of order $\mathcal{O}\left(\varepsilon^{-\frac{2(d-1)}{a}} \log \frac{1}{\varepsilon}\right)$ for the kernel $e^{-\|x-y\|^a}$ on the domain $\Omega = \mathbb{S}^{d-1}$. Consequently, as the ambient dimension d increases or the kernel becomes less smooth (i.e., a decreases), the required sample size grows rapidly. This explains the observed decline in the accuracy of Algorithm 1 under such conditions.

In experiments with exponential type kernels we used the following values of parameters: $N = 1000000$, $T = 1000$.

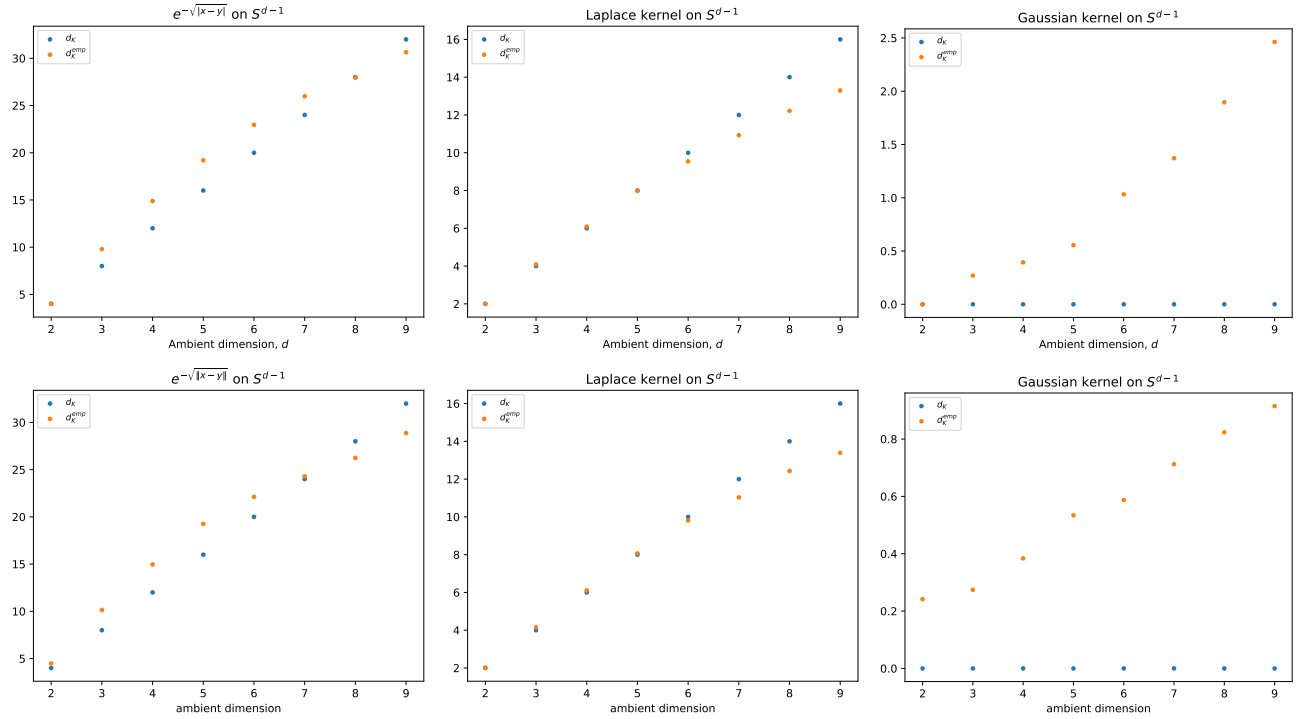

Figure 4: The effective dimension and the empirical upper bound on the effective dimension for exponential type kernels $e^{-\|x-y\|^a}$ on $\Omega = \mathbb{S}^{d-1}$: $a = \frac{1}{2}$, $a = 1$ (the Laplace kernel), $a = 2$ (the Gaussian kernel). The first row displays plots computed with the finite sample implementation of Algorithm 1, whereas the second row displays plots computed using the L-BFGS–based implementation.

D.4 Experiments with analytically given NNGP and NTK kernels

Analytical expressions for the NNGP and NTK kernels are available for the following activation functions: $\mathbb{1}_{x>0}$ (step function), $\max(0, x)$ (ReLU), erf, and $\alpha x\mathbb{1}_{x\le 0} + x\mathbb{1}_{x>0}$ (Leaky ReLU) [Takhanov, 2024a]. For each of these kernels (without bias), defined on the domains $\Omega = \mathbb{S}^{d-1}$ with $2 \le d \le 6$, we computed the empirical effective dimensions d_K^{emp} using the L-BFGS–based implementation of Algorithm 1 with $T = 500$. In every optimization step we invoke L-BFGS 10 or 20 times with independent initializations and output the best point.

After obtaining the empirical n-widths $\{w(t)\}_{t=0}^{T}$, the quantity d_K^{emp} is defined as the reciprocal of a, where a is the slope in the dataset $\{(\log t, -\log w(t))\}_{t=300}^{500}$ estimated via the RANSAC algorithm [Fischler and Bolles, 1987].

Figure 5 shows the computed empirical effective dimensions as functions of d. As the plots indicate, d_K^{emp} is recovered with high accuracy for $2 \le d \le 5$, but is underestimated for $d = 6$. This underestimation arises from the limited value of T in higher dimensions, where the asymptotic relation $w_K(n) \asymp n^{-1/d_K}$ becomes valid only for $n > T$.

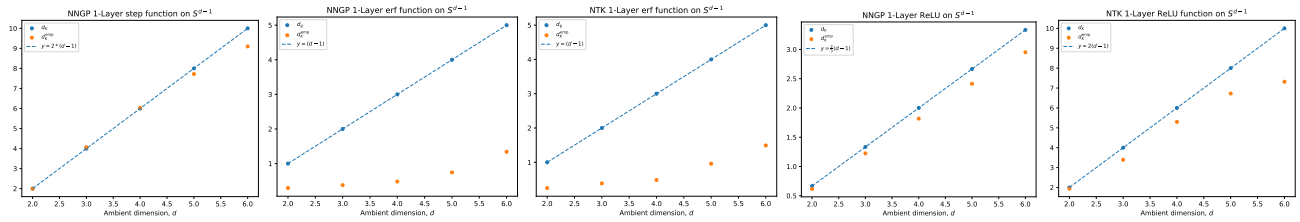

Figure 5: Empirical effective dimensions for infinite-width NNGP and NTK kernels with various activation functions, computed using the L-BFGS–based algorithm. For NNGP ReLU, we have $d_K = \frac{2}{3}(d-1)$, as Table 1 indicates. We do not show plots for the Leaky ReLU activation function ($\alpha < 1$) due to their similarity to the ReLU case.

On some pictures, along with empirical n-widths, we draw a linear plot for d_ϱ. The kernel-based upper Minkowski

dimension d_ϱ can be expressed as $d_\varrho = c\, d_H$, where d_H denotes the Euclidean upper Minkowski dimension (which equals $d-1$ in our setup). We call the coefficient c the *metric dimension factor*. If the kernel K satisfies $\sqrt{K(x,x)+K(y,y)-2K(x,y)} \asymp \|x-y\|^a$, then $c = \frac{1}{a}$. We refer to kernels with $c=1$ as *type I kernels*, and to those with $c=2$ as *type II kernels*.

For all activation functions considered, the corresponding NNGP kernels are of type I. Among NTK kernels, those induced by tanh, sigmoid, erf, and Gaussian activations are of type I, whereas those induced by ReLU and Leaky ReLU activations are of type II.

D.5 Experiments with NNGP and NTK kernels of finite width

Let $\sigma : \mathbb{R} \to \mathbb{R}$ be an activation function. Define the output of a one-hidden-layer neural network as

$$f_1(x \mid \mathbf{w}, W^{(1)}) = \frac{1}{\sqrt{n_1}} \mathbf{w}^\top \sigma(W^{(1)} x),$$

where $W^{(1)} \in \mathbb{R}^{n_1 \times d}$ and $\mathbf{w} \in \mathbb{R}^{n_1}$.

The random 1-layer NNGP kernel of width n_1 on the unit sphere $\mathbb{S}^{d-1}$ is defined as

$$\mathrm{NNGP}_\sigma(x, y \mid W^{(1)}) = \mathbb{E}_{\mathbf{w}\sim\mathcal{N}(0,I_{n_1})} f_1(x \mid \mathbf{w}, W^{(1)}) f_1(y \mid \mathbf{w}, W^{(1)}) = \frac{1}{n_1} \sigma(W^{(1)}x)^\top \sigma(W^{(1)}y),$$

where the entries of $W^{(1)}$ are sampled i.i.d. from $\mathcal{N}(0,1)$. This kernel serves as a Monte Carlo approximation of the infinite-width NNGP kernel (without bias):

$$\mathbb{E}_{\omega\sim\mathcal{N}(0,I_d)}[\sigma(\omega^\top x)\sigma(\omega^\top y)].$$

The random 1-layer NTK kernel (without bias) of width n_1 is defined by the expected inner product of the gradients of the network output with respect to all parameters:

$$\mathrm{NTK}_\sigma(x, y \mid W^{(1)}) = \mathbb{E}_{\mathbf{w}\sim\mathcal{N}(0,I_{n_1})} \left\langle \nabla_{(\mathbf{w},W^{(1)})} f_1(x), \nabla_{(\mathbf{w},W^{(1)})} f_1(y) \right\rangle,$$

with the same initialization for $W^{(1)}$ as above.

Now define the 2-layer network:

$$f_2(x \mid \mathbf{w}, W^{(1)}, W^{(2)}) = \frac{1}{\sqrt{n_1}} \mathbf{w}^\top \sigma\Big(\frac{1}{\sqrt{n_1}} W^{(2)} \sigma(W^{(1)} x)\Big),$$

where $W^{(2)} \in \mathbb{R}^{n_1 \times n_1}$ has i.i.d. entries sampled from $\mathcal{N}(0,1)$.

The random 2-layer NNGP kernel (without bias) of width n_1 is given by

$$\mathrm{NNGP}_\sigma(x, y \mid W^{(1)}, W^{(2)}) = \mathbb{E}_{\mathbf{w}\sim\mathcal{N}(0,I_{n_1})} f_2(x \mid \mathbf{w}, W^{(1)}, W^{(2)}) f_2(y \mid \mathbf{w}, W^{(1)}, W^{(2)})$$
$$= \frac{1}{n_1} \sigma\Big(\frac{1}{\sqrt{n_1}} W^{(2)} \sigma(W^{(1)}x)\Big)^\top \sigma\Big(\frac{1}{\sqrt{n_1}} W^{(2)} \sigma(W^{(1)}y)\Big).$$

The random 2-layer NTK kernel (without bias) of width n_1 is defined analogously as

$$\mathrm{NTK}_\sigma(x, y \mid W^{(1)}, W^{(2)}) = \mathbb{E}_{\mathbf{w}\sim\mathcal{N}(0,I_{n_1})} \left\langle \nabla_{(\mathbf{w},W^{(1)},W^{(2)})} f_2(x), \nabla_{(\mathbf{w},W^{(1)},W^{(2)})} f_2(y) \right\rangle.$$

Our definition of NTK kernels is generally consistent with the original formulation [Jacot et al., 2018], with the addition of an explicit expectation over the weights of the final layer. This modification simplifies kernel computation and slightly reduces the variance arising from the random sampling of the remaining weights. We applied the Algorithm 1 to NNGP and NTK kernels of width $n_1 = 1000$ on $\mathbb{S}^{d-1}$ for different activation functions. The parameter values that we used was $N = 500000$, $T = 500$. Results for $n_1 = 1000$, $N = 500000$, $T = 500$ are shown on Figure 6.

For slightly larger parameter settings ($N = 1000000$, $T = 1000$), Algorithm 1 applied to the Laplace kernel began to underestimate d_K starting from dimension $d = 6$. Therefore, we present results only for dimensions up to 6 in the plots. As expected, the empirical upper metric dimension is consistently greater than the empirical effective dimension.

For the infinite-width NTK with ReLU activation, it is known that $d_\varrho = d_K = 2d - 2$, similarly to the Laplace kernel. However, in our experiments, the empirical estimate d_K^{emp} lies significantly below d_ϱ^{emp}. This discrepancy is likely due to two factors: the finite network width ($n_1 = 1000$), and the underestimation of n-widths caused by the limited sample size N.

Based on these observations, we conjecture that for finite-width NTK ReLU kernels, the inequality $d_K < d_\varrho$ holds. As expected, the empirical effective dimensions of NTK ReLU kernels are approximately twice those of the corresponding NNGP ReLU kernels. Results for other activation functions align with known behavior: the n-widths of NTK kernels are typically only slightly larger than those of the corresponding NNGP kernels. These experiments are intended solely to demonstrate the efficiency of Algorithm 1 when applied to NNGP and NTK kernels. From these preliminary experiments, we observe that the main bottleneck in applying Algorithm 1 lies in the difficulty of computing such kernels exactly, due to the absence of closed-form expressions. Even powerful tools such as the `neural-tangents` library [Novak et al., 2020, Han et al., 2022] provide only low-rank approximations (typically of rank 30–40), which leads to a collapse of the computed empirical n-widths. A comprehensive study of the n-widths associated with these kernels is left for future work.

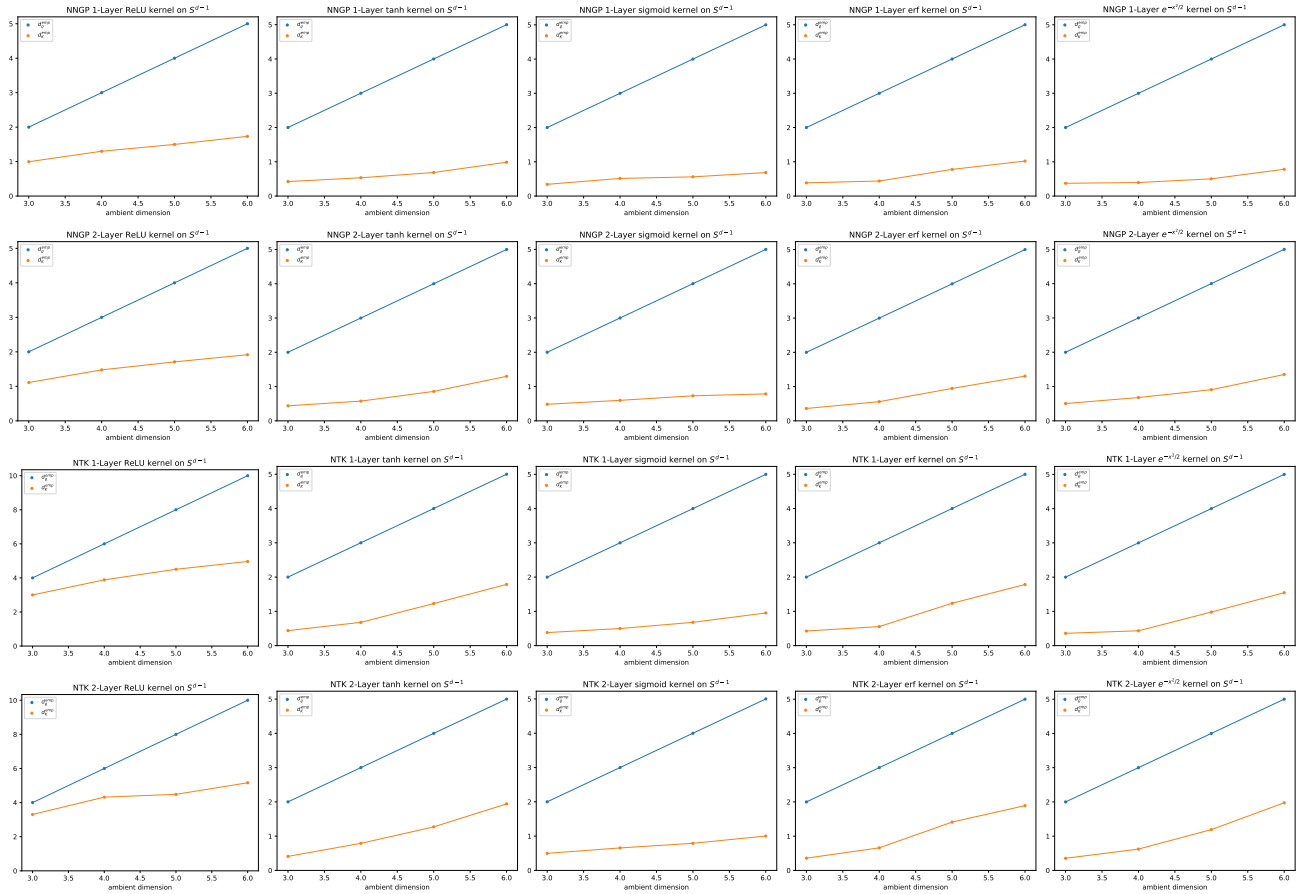

Figure 6: Empirical upper metric dimensions and empirical effective dimensions for NNGP and NTK kernels of width $n_1 = 1000$ on $\mathbb{S}^{d-1}$.

The computational experiments were performed on a system equipped with an Intel Core i9-10900X CPU operating at 3.70 GHz, running Ubuntu 22.04.1 LTS. The software environment included pandas version 2.0.0, numpy version 1.26.0, and seaborn version 0.13.0. Our code is available on github to facilitate the reproducibility of the results.

E Proofs for Section 2: cases of $d_K = d_\varrho$

We begin with a simple lemma that provides a sufficient condition for the equality $d_\varrho = d_K$.

Lemma 7. *To establish $d_K = d_\varrho$, it suffices to construct a Borel probability measure μ on Ω such that $\lambda_n(\mathrm{O}_{K,\mu}) \gg n^{-1-\frac{2}{d_\varrho}}$.*

Proof. From the definition of d_K, for any $\varepsilon > 0$, there exists N_ε such that

$$w_K(n) < n^{-\frac{1}{d_K+\varepsilon}} \quad \text{for all } n > N_\varepsilon.$$

Since $d_K \le d_\varrho$, it follows that

$$w_K(n) < n^{-\frac{1}{d_\varrho+\varepsilon}}.$$

By Ismagilov's theorem, the eigenvalue tail of the integral operator satisfies

$$\sum_{k=n+1}^{\infty} \lambda_k(\mathcal{O}_{K,\mu}) \le w_K(n)^2 \le n^{-\frac{2}{d_\varrho+\varepsilon}}.$$

On the other hand, if we assume $\lambda_n(\mathcal{O}_{K,\mu}) \gg n^{-1-\frac{2}{d_\varrho}}$, then summing the tail gives

$$\sum_{k=n+1}^{\infty} \lambda_k(\mathcal{O}_{K,\mu}) \gg \int_n^{\infty} x^{-1-\frac{2}{d_\varrho}}\, dx \asymp n^{-\frac{2}{d_\varrho}}.$$

Combining the upper and lower bounds,

$$n^{-\frac{2}{d_\varrho}} \ll w_K(n)^2 \le n^{-\frac{2}{d_\varrho+\varepsilon}},$$

which is only possible if $d_K = d_\varrho$. □

We now show that the desired eigenvalue decay holds for regular domains and certain translation-invariant kernels under the uniform measure.

Lemma 8. *Let $\Omega \subseteq \mathbb{R}^d$ be a Riemann measurable set with positive volume, and consider a translation-invariant kernel $K(\mathbf{x},\mathbf{y}) = k(\mathbf{x}-\mathbf{y})$ such that its Fourier transform*

$$\widehat{k}(\boldsymbol{\omega}) = \int_{\mathbb{R}^d} k(\mathbf{x}) e^{-i\boldsymbol{\omega}^\top \mathbf{x}}\, d\mathbf{x}$$

is bounded, nonnegative, and satisfies

$$\liminf_{\|\boldsymbol{\omega}\|\to\infty} \|\boldsymbol{\omega}\|^{d+a} \widehat{k}(\boldsymbol{\omega}) > 0$$

for some $a > 0$. Then,

$$\lambda_n(\mathcal{O}_{K,\mu}) \gg n^{-1-\frac{a}{d}},$$

where μ is the uniform probability measure on Ω.

We will use the following classical result due to Widom.

Theorem 10 ([Widom, 1963])**.** *Let $\Omega \subseteq \mathbb{R}^d$ be Riemann measurable with nonzero volume, μ the uniform distribution on Ω, and $K(\mathbf{x},\mathbf{y}) = k(\mathbf{x}-\mathbf{y})$ with $\widehat{k}$ bounded, nonnegative, and vanishing at infinity. Then*

$$\lambda_n(\mathcal{O}_{K,\mu}) \asymp \frac{1}{\mathrm{vol}(\Omega)}\, \phi_0\left(\frac{(2\pi)^d}{\mathrm{vol}(\Omega)} n\right),$$

where $\phi_0 : [0,\infty) \to [0,\infty)$ is a nonincreasing function equimeasurable with $\widehat{k}$.

This result plays an important role in the theory of translation-invariant kernels, e.g. it can be used to determine the rate of uniform convergence of a truncated Mercer series to the kernel [Takhanov, 2023].

Proof of Lemma 8. Let $\phi_0^{-1}(t) := \sup\{x > 0 \mid \phi_0(x) > t\}$. Since ϕ_0 is equimeasurable with $\widehat{k}$, we have

$$\mathrm{vol}\left(\{\boldsymbol{\omega} \mid \widehat{k}(\boldsymbol{\omega}) > t\}\right) = \phi_0^{-1}(t).$$

From the assumption $\widehat{k}(\boldsymbol{\omega}) \geq c\|\boldsymbol{\omega}\|^{-d-a}$ for $\|\boldsymbol{\omega}\| > C$, it follows that

$$\mathrm{vol}\left(\{\boldsymbol{\omega} \mid \widehat{k}(\boldsymbol{\omega}) > t\}\right) + \omega_d C^d \geq \mathrm{vol}\left(\left\{\boldsymbol{\omega} \mid \frac{c}{\|\boldsymbol{\omega}\|^{d+a}} > t\right\}\right) = \omega_d \left(\frac{c}{t}\right)^{\frac{d}{d+a}},$$

where $\omega_d = \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)}$ is the volume of the unit ball in $\mathbb{R}^d$. Hence,

$$\phi_0^{-1}(t) \geq \omega_d \left(\frac{c}{t}\right)^{\frac{d}{d+a}} - \omega_d C^d \gg t^{-\frac{d}{d+a}} \quad \text{as } t \to 0+.$$

This implies $\phi_0(t) \gg t^{-\frac{d+a}{d}}$ as $t \to \infty$. Substituting into Widom's result gives

$$\lambda_n(\mathcal{O}_{K,\mu}) \gg n^{-\frac{d+a}{d}} = n^{-1-\frac{a}{d}},$$

as claimed. □

E.1 Proof of Theorem 2: the Laplace kernel case

The Fourier transform of $k(\mathbf{x}) = e^{-\gamma\|\mathbf{x}\|}$ is $\widehat{k}(\boldsymbol{\omega}) = c_d \frac{\gamma}{(\gamma^2+\|\boldsymbol{\omega}\|^2)^{\frac{d+1}{2}}}$ where $c_d = 2^d \pi^{\frac{d-1}{2}} \Gamma(\frac{d+1}{2})$. A nonincreasing function ϕ_0 equimeasurable with $\widehat{k}$ satisfies

$$\mathrm{vol}(\{\boldsymbol{\omega} \mid c_d \frac{\gamma}{(\gamma^2 + \|\boldsymbol{\omega}\|^2)^{\frac{d+1}{2}}} > t\}) = \phi_0^{-1}(t).$$

Therefore, $\phi_0^{-1}(t) = \omega_d \left((\frac{c_d \gamma}{t})^{\frac{2}{d+1}} - \gamma^2 \right)^{\frac{d}{2}}$ where $\omega_d = \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)}$. For $t \to +0$ we have $\phi_0^{-1}(t) \asymp t^{-\frac{d}{d+1}}$, and therefore, $\phi_0(t) \asymp t^{-\frac{d+1}{d}}$ as $t \to +\infty$. From Widom's result we obtain $\lambda_n(\mathrm{O}_{K,\mu}) \asymp n^{-\frac{d+1}{d}}$. Since $d_\varrho = 2d$, we obtain $\lambda_n(\mathrm{O}_{K,\mu}) \asymp n^{-1-\frac{2}{d_\varrho}}$. Lemma 7 implies $d_K = d_\varrho$.

E.2 Proof of Theorem 2: the kernel $K(\mathbf{x},\mathbf{y}) = e^{-\gamma\|\mathbf{x}-\mathbf{y}\|^a}$ for $a \in (0,1)$

Let $k(\mathbf{x}) = e^{-\gamma\|\mathbf{x}\|^a}$. In [Pollard, 1946] the following integral representation was derived

$$e^{-r^a} = \int_0^\infty e^{-rt} g_a(t) dt,$$

where $r > 0$, $0 < a < 1$ and g_a is a probability density function of a distribution supported in $[0, +\infty)$ (sometimes called the Lévy stable distribution) which has the following representation:

$$g_a(t) = \frac{1}{\pi} \sum_{n=1}^{\infty} \frac{-\sin(n(a+1)\pi)}{n!} \left(\frac{1}{x}\right)^{an+1} \Gamma(an+1).$$

We will only need the property $g_a(t) \asymp \frac{1}{x^{a+1}}$ as $x \to +\infty$. Since the Fourier transform in d dimensions of $e^{-t\|\mathbf{x}\|}$ is $c_d \frac{t}{(t^2+\|\boldsymbol{\omega}\|^2)^{\frac{d+1}{2}}}$ we conclude

$$\widehat{k}(\boldsymbol{\omega}) = \mathcal{F}[\int_0^\infty e^{-\gamma^{1/a}\|\mathbf{x}\|t} g_a(t) dt] = \int_0^\infty \mathcal{F}[e^{-\gamma^{1/a}\|\mathbf{x}\|t}] g_a(t) dt =$$
$$\int_0^\infty \frac{c_d \gamma^{1/a} t}{(\gamma^{2/a} t^2 + \|\boldsymbol{\omega}\|^2)^{\frac{d+1}{2}}} g_a(t) dt.$$

Since $\int_A^{2A} t^a g_a(t)dt \gg \int_A^{2A} \frac{dt}{t} = \log 2 \gg 1$, we conclude

$$\|\boldsymbol{\omega}\|^{d+a}\widehat{k}(\boldsymbol{\omega}) = c_d\gamma^{1/a}\int_0^\infty \frac{t^{1-a}\|\boldsymbol{\omega}\|^{d+a}}{(\gamma^{2/a}t^2+\|\boldsymbol{\omega}\|^2)^{\frac{d+1}{2}}}t^a g_a(t)dt \geq$$
$$c_d\gamma^{1/a}\int_{\|\boldsymbol{\omega}\|/\gamma^{1/a}}^{2\|\boldsymbol{\omega}\|/\gamma^{1/a}} \frac{t^{1-a}\|\boldsymbol{\omega}\|^{d+a}}{(\gamma^{2/a}t^2+\|\boldsymbol{\omega}\|^2)^{\frac{d+1}{2}}}t^a g_a(t)dt \geq$$
$$c_d\gamma^{1/a}\min_{x\in[\|\boldsymbol{\omega}\|/\gamma^{1/a},2\|\boldsymbol{\omega}\|/\gamma^{1/a}]}\frac{x^{1-a}\|\boldsymbol{\omega}\|^{d+a}}{(\gamma^{2/a}x^2+\|\boldsymbol{\omega}\|^2)^{\frac{d+1}{2}}}\int_{\|\boldsymbol{\omega}\|/\gamma^{1/a}}^{2\|\boldsymbol{\omega}\|/\gamma^{1/a}} t^a g_a(t)dt \gg 1.$$

Thus, $\liminf_{\|\boldsymbol{\omega}\|\to+\infty}\|\boldsymbol{\omega}\|^{d+a}\widehat{k}(\boldsymbol{\omega}) > 0$. Using Lemma 8 we conclude that for any Riemann measurable $\Omega \subseteq \mathbb{R}^d$ and a uniform distribution μ on Ω, we have $\lambda_n(\mathrm{O}_{K,\mu}) \gg n^{-1-\frac{a}{d}}$.

Finally, we show that $d_\varrho = \frac{2d}{a}$. Indeed, observe that the kernel-induced metric satisfies $\varrho(\mathbf{x},\mathbf{y}) = \sqrt{2k(\mathbf{0})-2k(\mathbf{x}-\mathbf{y})} \asymp \|\mathbf{x}-\mathbf{y}\|^{\frac{a}{2}}$. This implies that the entropy number $\varepsilon(n)$ with respect to ϱ scales as $\varepsilon(n) \asymp \varepsilon'(n)^{\frac{a}{2}}$, where $\varepsilon'(n)$ denotes the entropy number under the Euclidean metric. Since the Minkowski dimension of any Riemann measurable set $\Omega \subseteq \mathbb{R}^d$ with nonzero volume is d, we conclude that $d_\varrho = \frac{2d}{a}$. Thus, we have $\lambda_n(\mathrm{O}_{K,\mu}) \gg n^{-1-\frac{2}{d_\varrho}}$. Lemma 7 gives us $d_\varrho = d_K$.

E.3 Proof of Theorem 2: the Matérn kernel case

The Matérn kernel is defined as $k_{\mathrm{M}}(\|\mathbf{x}-\mathbf{y}\|)$ where

$$k_{\mathrm{M}}(r) = \frac{2^{1-\nu}}{\Gamma(\nu)}(\frac{\sqrt{2\nu}r}{l})^\nu K_\nu(\frac{\sqrt{2\nu}r}{l}),$$

with positive parameters ν and l, where K_ν is a modified Bessel function [Rasmussen and Williams, 2005]. Its Fourier transform in d dimensions is

$$\widehat{k_{\mathrm{M}}}(\boldsymbol{\omega}) = \frac{2^d\pi^{\frac{d}{2}}\Gamma(\nu+\frac{d}{2})(2\nu)^\nu}{\Gamma(\nu)l^{2\nu}}\left(\frac{2\nu}{l^2}+4\pi\|\boldsymbol{\omega}\|^2\right)^{-(\nu+\frac{d}{2})}.$$

By construction, $\liminf_{\|\boldsymbol{\omega}\|\to+\infty}\|\boldsymbol{\omega}\|^{d+2\nu}\widehat{k}(\boldsymbol{\omega}) > 0$ if $\nu > 0$. From Lemma 8 we conclude that if $\nu > 0$, then $\lambda_n(\mathrm{O}_{K,\mu}) \gg n^{-1-\frac{2\nu}{d}}$ for a uniform distribution μ on a Riemann measurable domain with non-zero volume $\Omega \subseteq \mathbb{R}^d$.

For $\nu \in (0,1)$, the modified Bessel function satisfies $K_\nu(x) = \frac{\Gamma(\nu)}{2}(\frac{2}{x})^\nu - C_2x^\nu + o(x^\nu)$ for small $x > 0$. Therefore, $\varrho(\mathbf{x},\mathbf{y}) = \sqrt{2k_{\mathrm{M}}(\mathbf{0})-2k_{\mathrm{M}}(\mathbf{x}-\mathbf{y})} \asymp \|\mathbf{x}-\mathbf{y}\|^\nu$. As in the case of exponential type kernels, we have $\varepsilon(n) \asymp \varepsilon'(n)^\nu$, where $\varepsilon'(n)$ is the entropy number w.r.t. to the euclidean metrics. Thus, for any Riemann measurable with non-zero volume $\Omega \subseteq \mathbb{R}^d$, we have $d_\varrho = \frac{d}{\nu}$. This gives us $\lambda_n(\mathrm{O}_{K,\mu}) \gg n^{-1-\frac{2}{d_\varrho}}$ and, using Lemma 7, we obtain $d_\varrho = d_K$.

E.4 Remarks on Table 1

Table 1 presents the Kolmogorov n-widths for zonal kernels defined on the domain $\Omega = \mathbb{S}^{d-1}$, the unit sphere in $\mathbb{R}^d$. Let $\mu_{\mathbb{S}^{d-1}}$ denote the rotation-invariant probability measure on $\mathbb{S}^{d-1}$. Zonal kernels on the sphere admit the following expansion:

$$K(x,y) = \sum_{l=0}^\infty a(l)\sum_{i=1}^{\mathcal{N}(d,l)} Y_{l,i}(x)Y_{l,i}(y),$$

where $\{a(l)\}$ are the eigenvalues of the integral operator $\mathcal{O}_{\mathbb{S}^{d-1},\mu_{\mathbb{S}^{d-1}}}$, and $\{Y_{l,i}\}_{i=1}^{\mathcal{N}(d,l)}$ is an orthonormal basis of the space of real-valued spherical harmonics of degree l. The integer $\mathcal{N}(d,l)$ denotes the dimension of that space.

Using the identity

$$\sum_{i=1}^{\mathcal{N}(d,l)} Y_{l,i}(x)^2 = \mathcal{N}(d,l),$$

and applying Ibragimov's theorem, we obtain the following two-sided estimate:

$$\sqrt{\sum_{l=l_0+1}^{\infty} a(l)\mathcal{N}(d,l)} \le w_K(n) \le \sqrt{\sum_{l=l_0+1}^{\infty} a(l)\mathcal{N}(d,l)},$$

where $n = \sum_{l=0}^{l_0} \mathcal{N}(d,l)$. Hence, we have the explicit expression:

$$w_K(n) = \sqrt{\sum_{l=l_0+1}^{\infty} a(l)\mathcal{N}(d,l)}.$$

This characterization allows for precise asymptotic estimates of $w_K(n)$ for zonal kernels, provided the decay rate of the eigenvalues $\{a(l)\}$ is known.

For the kernel $K(x,y) = e^{-\gamma\|x-y\|}$ and for the NTK-ReLU, it was shown in [Geifman et al., 2020] that $a(l) \asymp l^{-d}$ as $l \to +\infty$. Using $\mathcal{N}(d,l) \asymp l^{d-2}$, we conclude that $w_K(\sum_{l=0}^{l_0} \mathcal{N}(d,l)) \asymp l_0^{-1/2}$ as $l_0 \to +\infty$. Since $n = \sum_{l=0}^{l_0} \mathcal{N}(d,l) \asymp l_0^{d-1}$, we conclude $w_K(n) \asymp n^{-\frac{1}{2d-2}}$. Therefore, $d_K = 2d-2$. The Minkowski dimension of $\mathbb{S}^{d-1}$ in $\mathbb{R}^d$ equals $d-1$. Since $\varrho(x,y) \asymp \|x-y\|^{\frac{1}{2}}$, we conclude $d_\varrho = \frac{d-1}{1/2}$, i.e. $d_\varrho = d_K$.

For the kernel $K(x,y) = e^{-\frac{\|x-y\|^2}{\sigma^2}}$, it was shown in [Minh et al., 2006] that $a(l) \asymp (\frac{2e}{\sigma^2})^l (2l+d-2)^{-(l+\frac{d-1}{2})}$ if $\sigma > (\frac{2}{d})^{\frac{1}{2}}$. Analogously to the previous kernel, this gives $w_K(n) \asymp \sqrt{\int_{l_0}^{\infty} (\frac{2e}{\sigma^2})^x \frac{x^{d-2}dx}{(2x+d-2)^{x+\frac{d-1}{2}}}}$ where $n = \sum_{l=0}^{l_0} \mathcal{N}(d,l)$. That is, $w_K(n) \ll \sqrt{\int_{\mathcal{O}(n^{1/(d-1)})}^{\infty} (\frac{2e}{\sigma^2})^x \frac{x^{d-2}dx}{(2x+d-2)^{x+\frac{d-1}{2}}}}$. The latter expression decays exponentially fast. Therefore, $d_K = 0$.

E.4.1 Neural Network Gaussian Process kernels

Given the activation function $\sigma : \mathbb{R} \to \mathbb{R}$, the Neural Network Gaussian Process (NNGP) kernel without the bias term (in the large width limit), denoted $k_{\mathrm{NNGP},\sigma} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, is defined by

$$k_{\mathrm{NNGP},\sigma}(x,y) = \mathbb{E}_{w\sim\mathcal{N}(\mathbf{0},I_d)}[\sigma(w^\top x)\sigma(w^\top y)],$$

where $\mathcal{N}(\mathbf{0}, I_d)$ denotes the standard Gaussian distribution in d dimensions. The NNGP kernel, restricted to $\mathbb{S}^{d-1}$, is zonal.

The case of smooth activations. If σ is infinitely differentiable and $\max_{x\in\mathbb{R}} |\sigma^{(k)}(x)| \ll k!$, [Wu and Long, 2022] proved that

$$\sum_{l=l_0+1}^{\infty} a(l)\mathcal{N}(d,l) \ll \frac{1}{n},$$

where $n = \sum_{l=0}^{l_0} \mathcal{N}(d,l)$. Therefore,

$$w_K(n) \ll n^{-\frac{1}{2}},$$

and $d_K \le 2$.

As the following lemma demonstrates, for activation functions with a bounded third derivative, the upper metric dimension d_ϱ is typically either $d-1$ or $\frac{d-1}{2}$, depending on the behavior of certain expectations:

Lemma 9. *Let* $\sup_{x\in\mathbb{R}} |\sigma'''(x)| < \infty$. *Then:*

- *If* $\mathbb{E}_{w_1\sim\mathcal{N}(0,1)}\left[\sigma(w_1)\sigma'(w_1)w_1-\sigma''(w_1)\right]\neq 0$*, it follows that* $d_\varrho = d-1$*.*
- *If* $\mathbb{E}_{w_1\sim\mathcal{N}(0,1)}\left[\sigma(w_1)\sigma'(w_1)w_1-\sigma''(w_1)\right]=0$ *and* $\mathbb{E}_{w_1\sim\mathcal{N}(0,1)}\left[\sigma''(w_1)w_1^2-\sigma''(w_1)\right]\neq 0$*, then* $d_\varrho=\frac{d-1}{2}$*.*

Proof. Let $x, y\in\mathbb{S}^{d-1}$ be such that $y = x+\delta x$. By the rotational invariance of the kernel $k_{\mathrm{NNGP},\sigma}(x,y)$, we may assume without loss of generality that $x = e_1 = (1,0,\ldots,0)$. Then the perturbation $\delta x = (\delta x_1,\ldots,\delta x_d)$ satisfies:

$$\delta x_1 = x^\top y - 1, \quad \sum_{i\neq 1}\delta x_i^2 = 2-2x^\top y-(x^\top y-1)^2 = 1-(x^\top y)^2.$$

Define the following constants:

$$\begin{aligned}
C_0 &= \mathbb{E}_{w_1\sim\mathcal{N}(0,1)}[\sigma(w_1)^2],\\
C_1 &= \mathbb{E}_{w_1\sim\mathcal{N}(0,1)}[\sigma(w_1)\sigma'(w_1)w_1],\\
C_2 &= \mathbb{E}_{w_1\sim\mathcal{N}(0,1)}[\sigma''(w_1)w_1^2],\\
D_2 &= \mathbb{E}_{w_1\sim\mathcal{N}(0,1)}[\sigma''(w_1)].
\end{aligned}$$

These will be used to determine the local behavior of the kernel near the diagonal, which governs the scaling of the induced metric. We have

$$\begin{aligned}
&k_{\mathrm{NNGP},\sigma}(x,y) = \mathbb{E}_{w\sim\mathcal{N}(\mathbf{0},I_d)}[\sigma(w^\top x)\sigma(w^\top(x+\delta x))] =\\
&\mathbb{E}_{w\sim\mathcal{N}(\mathbf{0},I_d)}[\sigma(w_1)\sigma(w_1+w^\top\delta x)] =\\
&\mathbb{E}_{w\sim\mathcal{N}(\mathbf{0},I_d)}[\sigma(w_1)(\sigma(w_1)+\sigma'(w_1)w^\top\delta x+\frac{1}{2}\sigma''(w_1)(w^\top\delta x)^2)]+o(\|\delta x\|^2) =\\
&\mathbb{E}_{w_1\sim\mathcal{N}(0,1)}[\sigma(w_1)^2]+\mathbb{E}_{w_1\sim\mathcal{N}(0,1)}[\sigma(w_1)\sigma'(w_1)w_1]\delta x_1+\\
&\frac{1}{2}\mathbb{E}_{w_1\sim\mathcal{N}(0,1)}[\sigma''(w_1)w_1^2]\delta x_1^2+\frac{1}{2}\sum_{i\neq 1}\mathbb{E}_{w_1\sim\mathcal{N}(0,1)}[\sigma''(w_1)]\delta x_i^2+o(\|\delta x\|^2).
\end{aligned}$$

Thus,

$$k_{\mathrm{NNGP},\sigma}(x,y) = C_0+C_1(x^\top y-1)+\frac{1}{2}C_2(x^\top y-1)^2+\frac{1}{2}D_2(1-(x^\top y)^2)+o(\|\delta x\|^2).$$

Note that $1-(x^\top y)^2 = 2(1-(x^\top y))-(1-(x^\top y))^2$. If $C_1-D_2\neq 0$, we have

$$k_{\mathrm{NNGP},\sigma}(x,y) = C_0+(C_1-D_2)(x^\top y-1)+o(\|\delta x\|),$$

and

$$\varrho(x,y)\asymp\sqrt{1-x^\top y}\asymp\|x-y\|.$$

Thus, $d_\varrho = d-1$.

If $C_1-D_2=0$, then

$$k_{\mathrm{NNGP},\sigma}(x,y) = C_0+\frac{1}{2}(C_2-D_2)(x^\top y-1)^2+o(\|\delta x\|^2).$$

When $C_2-D_2\neq 0$, we have

$$\varrho(x,y)\asymp 1-x^\top y\asymp\|x-y\|^2.$$

Thus, $d_\varrho=\frac{d-1}{2}$. □

Example 1. *Let us consider the case of* $\sigma(x)=\cos(x)$. *We have*

$$\mathbb{E}_{w_1\sim\mathcal{N}(0,1)}[\sigma(w_1)\sigma'(w_1)w_1-\sigma''(w_1)]=$$
$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}(-\frac{\sin(2x)x}{2}+\cos(x))e^{-\frac{x^2}{2}}dx\approx 0.47\neq 0.$$

Therefore, $d_\varrho=d-1$. *Analogously, for* $\sigma(x)=\cos(x)$, *we have*

$$\mathbb{E}_{w_1\sim\mathcal{N}(0,1)}[\sigma(w_1)\sigma'(w_1)w_1-\sigma''(w_1)]=$$
$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}(\frac{\sin(2x)x}{2}+\sin(x))e^{-\frac{x^2}{2}}dx\approx 0.13\neq 0,$$

and $d_\varrho=d-1$.

The case of $\sigma(x)=\max(0,x)^\alpha,\alpha\geq 0$**.** For $\alpha\in\{0,1\}$ it was shown in [Bach, 2017] that $a(l)\asymp l^{-d-2\alpha}$. Therefore, $w_K(\sum_{l=0}^{l_0}\mathcal{N}(d,l))\asymp l_0^{-1/2-\alpha}$ as $l_0\to+\infty$ and $w_K(n)\asymp n^{-\frac{1+2\alpha}{2d-2}}$. Thus, $d_K=\frac{2d-2}{1+2\alpha}$.

[Wu and Long, 2022] proved that $w_K(n)\gg n^{-\frac{1+2\alpha}{2d-2}}$ for any $\alpha>0$ and gave some experimental evidence in favour of $w_K(n)\asymp n^{-\frac{1+2\alpha}{2d-2}}$. From this, we can claim $d_K\geq\frac{2d-2}{1+2\alpha}$ for any $\alpha\geq 0$.

For $\alpha=1$, we have $k_{\mathrm{NNGP},\sigma}(x,y)\propto u(\pi-\arccos(u))+\sqrt{1-u^2}$ where $u=x^\top y$. Let $v=1-u=\frac{1}{2}\|x-y\|^2$. Thus, $k_{\mathrm{NNGP},\sigma}(x,y)\propto(1-v)(\pi-\arccos(1-v))+v^{\frac{1}{2}}(2-v)^{\frac{1}{2}}=\pi-\pi v+o(v)$. Therefore, $\varrho(x,y)\asymp\|x-y\|$ and $d_\varrho=d-1$.

For $\alpha=0$, we have $k_{\mathrm{NNGP},\sigma}(x,y)\propto\pi-\arccos(u)$ where $u=x^\top y$. Analogously, we show that $k_{\mathrm{NNGP},\sigma}(x,y)\propto\pi-\sqrt{2v}+o(\sqrt{v})$. Therefore, $\varrho(x,y)\asymp\sqrt{\|x-y\|}$ and $d_\varrho=2(d-1)$.