On the Intrinsic Dimensions of Data in Kernel Learning

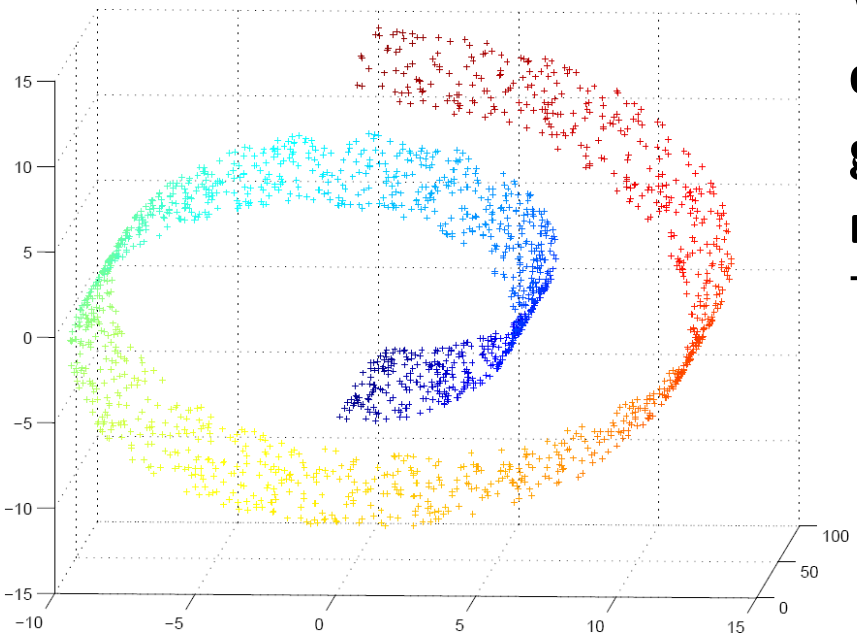
Rustem Takhanov

Nazarbayev University, Astana, Kazakhstan

AISTATS 2026

Motivation

- Many datasets lie on **low-dimensional manifolds** (manifold hypothesis)
- Ideally, the generalization should depend on **intrinsic dimension**, not ambient dimension
- Kernel methods often avoid curse of dimensionality



Which notion of intrinsic dimension controls generallization of kernel methods?

This was addressed before for:

- k-NN regression (Kpotufe, 2011)
- Kernel regression (Bickel and Li, 2007).

Two Notions of Intrinsic Dimension

A domain Ω and a kernel $K : \Omega \times \Omega \to \mathbb{R}$ induces metric

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

and n-covering radius

$$\varepsilon_K(n) = \min \left\{ \varepsilon > 0 \mid \exists z_1, \cdots, z_n \in \Omega \text{ s.t. } \bigcup_{i=1}^{n} B(z_i, \varepsilon) \supseteq \Omega \right\}.$$

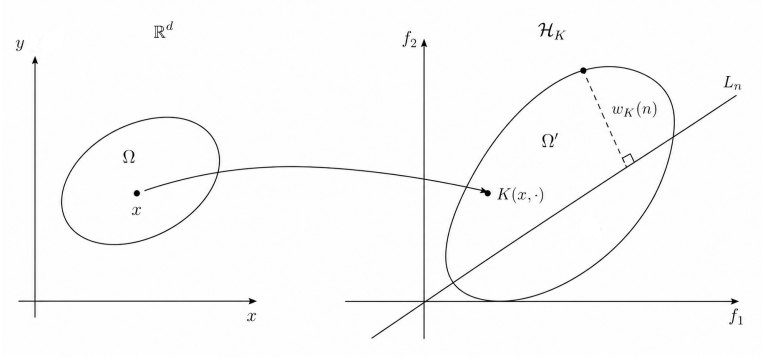
1. Metric Dimension, d_ρ

$$d_\rho = \limsup_{n \to \infty} \frac{\log n}{\log(1/\varepsilon_K(n))}$$

- Based on kernel-induced metric ρ
- Proportional to Hausdorff dimension (which, for smooth manifolds, equals their dimension).

Two Notions of Intrinsic Dimension

The Kolmogorov n-width of the image of Ω under canonical embedding $\Omega \hookrightarrow \mathcal{H}_K$ is defined by



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

where the infimum is taken over all n-dimensional subspaces L_n of $\mathcal{H}_K$

2. Effective Dimension, d_K

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

- Based on Kolmogorov n-widths
- More globally captures interaction: **kernel + domain**.

Effective dimension cannot be greater than metric

Theorem: $d_K \leq d_\rho$

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

K	$w_K(n)$	d_ϱ	d_K
$e^{-\gamma\|x-y\|}$	$\asymp n^{-\frac{1}{2d-2}}$	$2d-2$	$2d-2$
$\mathrm{NTK}_{\max(0,x)}$	$\asymp n^{-\frac{1}{2d-2}}$	$2d-2$	$2d-2$
$e^{-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{\sigma^2}}, \sigma > \sqrt{\frac{2}{d}}$	$\searrow$ exp. fast	$d-1$	0
$\mathrm{NNGP}_{\cos}, \mathrm{NNGP}_{\sin}$	$\ll n^{-\frac{1}{2}}$	$d-1$	≤ 2
$\mathrm{NNGP}_{\max(0,x)^\alpha}, \alpha \geq 0$	$\gg n^{-\frac{1+2\alpha}{2d-2}}$	$\frac{2d-2}{1+\alpha}$(?)	$\geq \frac{2d-2}{1+2\alpha}$
$\mathrm{NNGP}_{\max(0,x)^\alpha}, \alpha \in \{0,1\}$	$\asymp n^{-\frac{1+2\alpha}{2d-2}}$	$\frac{2d-2}{1+\alpha}$	$\frac{2d-2}{1+2\alpha}$

- Kernel smoothness reduces effective dimension (making it zero for such kernels as the Gaussian kernel).
- Effective dimensions for Laplace and NTK ReLU cases are largest possible, $d_K = 2(d-1) = d_\varrho$.
- Most of the kernel are of intermediate kind: $d_K \in (0, d_\varrho)$.

n-Widths and Eigenvalues

Kernel operator:

$$O_{K,\mu} f(x) = \int_{\Omega} K(x,y) f(y) d\mu(y)$$

Main Result

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

and

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

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

- n-widths characterize **worst-case eigenvalue decay**, provided $w_K(n) \asymp w_K(2n)$
- Independent of distribution μ

Generalization Bound for constrained KRR

The constrained KRR:

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

With high probability

$$\text{Excess Risk} = O\left(n^{-\frac{2+d_K}{2+2d_K}+\varepsilon} \right), \quad \forall \varepsilon > 0.$$

- For $d_K = 0$ (e.g., Gaussian kernel): $\text{Excess Risk} \approx O(1/n)$
- If d_K large: $\text{Excess Risk} \approx O(1/\sqrt{n})$
- d_K controls generalization rate (similarly to VC-dimension)

How to estimate $w_K(n)$ from data?

- $\widetilde{\Omega} = \{Z_1, \cdots, Z_N\} \subseteq \Omega$ — dataset.
- Greedy selection of points:

$$x_t = \arg\max_{x \in \tilde{\Omega}} \left(K[(X_{t-1}, x), (X_{t-1}, x)] \mid K[X_{t-1}, X_{t-1}]\right)$$

  where $X_{t-1} = (x_1, ..., x_{t-1})$ and $(A|B)$ is the Schur complement.
- Let $\{w_t\}_{t=0}^{T-1}$ be the sequence of approximate upper bounds produced by Greedy algorithm using the finite set $\tilde{\Omega}$. Let us assume that $\tilde{\Omega}$ is an ε-net in (Ω, ϱ). Then, we have

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

Sample complexity: if $\widetilde{\Omega} \sim \mu$,

$$N = O\left(\varepsilon^{-d_\rho} \log \frac{1}{\varepsilon}\right)$$

Experiments

- Empirical KRR rates match theory;
- For most of kernels: $d_K < d_\rho$;
- For $K(x, y) = e^{-\gamma\|x-y\|^a}, a \in (0, 1]$, NTK ReLU, Matérn kernel, $\nu \in (0, 1), l > 0$ kernels and regular domains: $d_K = d_\rho$;
- Even for the latter kernels, if the domain Ω is non-regular (e.g. a fractal): $d_K < d_\rho$.

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

Takeaway

Kolmogorov n-widths determine the intrinsic dimension seen by kernels.

- Effective dimension d_K governs generalization rates of kernel methods (for the worst-case distribution over Ω).
- Always $d_K \leq d_\varrho$.
- Kernels can **compress geometric complexity**.
- n-widths can be estimated from data: upper bounds by the Greedy algorithm, and lower bounds from eigenvalues of the empirical kernel matrix via Ismagilov's theorem.
- Open direction: n-widths for NTK and deep kernels.

Open direction:

Understanding n-widths for modern kernels (NTK, deep kernels).