UNIVERSITY OF CAMBRIDGE

Blessing of dimensionality using low-dimensional data

Hong Ye Tan

Joint work with: Carola-Bibiane Schönlieb, Subhadip Mukherjee, Junqi Tang

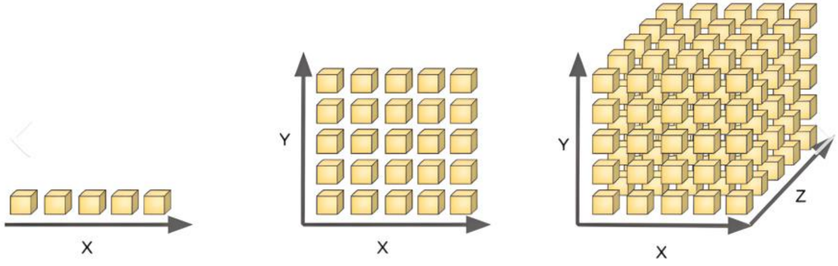
arXiv:2408.06996

GSK.ai PhD Symposium

24/09/24

Background: Dimensionality

- **Dimensionality** = difficulty
- **Canonical example**: sampling
  - Complexity scales exponentially with dimension
  - Optimisation, nearest neighbours, metric learning, data visualisation…

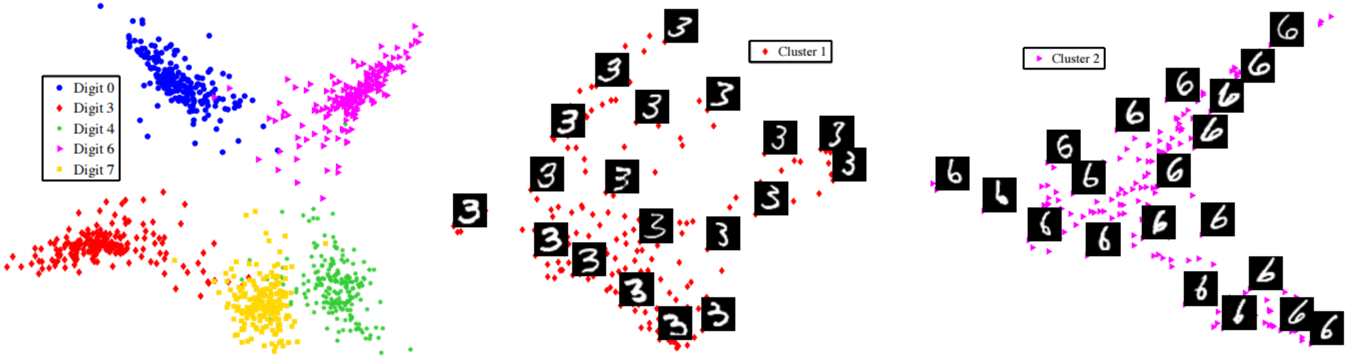


1st,2nd and 3rd Dimensions in space

- The curse can be slightly lifted depending on additional constraints
  - Optimising *non-convex* function: exponential in dimension (random search)
  - *Convex and bounded*: linear (centre of gravity)
  - *Convex and smooth*: bounds using e.g. Lipschitz constant

Background: Manifolds

- **Not all hope is lost!**
- **Manifold Hypothesis:** Natural data is intrinsically low-dimensional.
  - ➢ Original motivation: statistical tools still work when we have high dimensions and relatively little data, when they are not supposed to
  - ➢ a.k.a. blessing of dimensionality, concentration of measure.
  - ➢ Then dimensionality reduction methods work and give us nice visualisations



Project goal

Generalisation ← **Low Statistical Complexity** ← **Useful Functions**

- **Intuition:** in lower dimensions, it should be easier to approximate, say, Lipschitz continuous functions with a simple class of functions.
- We want to determine how difficult it is to approximate a:
  - **Sufficiently useful** class of functions
  - with a simple class of functions

Simple?
– notion of statistical complexity

⚠ Purely theoretical

Function class complexity

Maths

- Sobolev space $\mathcal{W}^{1,p}(M)$
  L^p functions with L^p derivative
  - ➢ Infinite dimensional, pretty nasty
  - ➢ We need to approximate (a bounded version) with simpler functions
- "Pseudo-dimension"
  - Measure of statistical complexity

Definition 2.1. *Let $\mathcal{H}$ be a class of real-valued functions with domain $\mathcal{X}$. Let $X_n = \{x_1, ..., x_n\} \subset \mathcal{X}$, and consider a collection of reals $s_1, ..., s_n \subset \mathbb{R}^n$. When evaluated at each x_i, a function $h \in \mathcal{H}$ will lie on one side[2] of the corresponding x_i, i.e. $\mathrm{sign}(h(x_i) - s_i) = \pm 1$. The vector of such sides $(\mathrm{sign}(h(x_i) - s_i))_{i=1}^n$ is thus an element of $\{\pm 1\}^n$.*

We say that $\mathcal{H}$ P-shatters *X_n if there exist reals $s_1, ..., s_n$ such that all possible sign combinations are obtained, i.e.,*

$$\{(\mathrm{sign}(h(x_i) - s_i))_i \mid h \in \mathcal{H}\} = \{\pm 1\}^n.$$

The pseudo-dimension *$\dim_p(\mathcal{H})$ is the cardinality of the largest set that is P-shattered:*

$$\dim_p(\mathcal{H}) = \sup\{n \in \mathbb{N} \mid \exists \{x_1, ..., x_n\} \subset \mathcal{X} \text{ that is P-shattered by } \mathcal{H}\}. \quad (1)$$

British Coastline

- The British coastline
  - ➢ Very grainy around the edges
  - ➢ We need to measure this with a coarse set of tools: rulers
  - ➢ See: coastline paradox
- How long your rulers are

Why statistical complexity? Sample complexity

Maths

- Low complexity
  lower sample complexity

The number of samples required to be within ϵ of the true risk with probability at least $1-\delta$ is bounded above by:

$$m_L(\epsilon,\delta) \leq \frac{128}{\epsilon^2}\left(2\dim_P(\mathcal{H})\log\left(\frac{34}{\varepsilon}\right)+\log\left(\frac{16}{\delta}\right)\right)$$

- Directly related to better generalisation
  - ➢ Simpler description is less prone to error

British Coastline

- Longer rulers, fewer points needed to measure the coast

A. Rae, https://www.statsmapsnpix.com/2016/08/how-long-is-coastline-of-great-britain.html

Approximation complexity

Maths

- Approximation with bounded pseudo-dimension functions

  - Given a maximum pseudo-dimension, the non-linear width of a function class F is the best approximation with such limited pseudo-dimension classes

Definition 2.4. *Let $\mathcal{F}$ be a normed space of functions. Given two subsets $F_1, F_2 \subset \mathcal{F}$, the* (asymmetric) Hausdorff distance *between the two subsets is the largest distance of an element of F_1 to its closest element in F_2:*

$$\operatorname{dist}(F_1, F_2; \mathcal{F}) = \sup_{f_1 \in F_1} \inf_{f_2 \in F_2} \|f_1 - f_2\|_{\mathcal{F}}. \tag{3}$$

For a subset $F \subset \mathcal{F}$, the nonlinear n-width *is given by the optimal (asymmetric) Hausdorff distance between F and $\mathcal{H}^n$, infimized over classes $\mathcal{H}^n$ in $\mathcal{F}$ with $\dim_p(\mathcal{H}^n) \leq n$:*

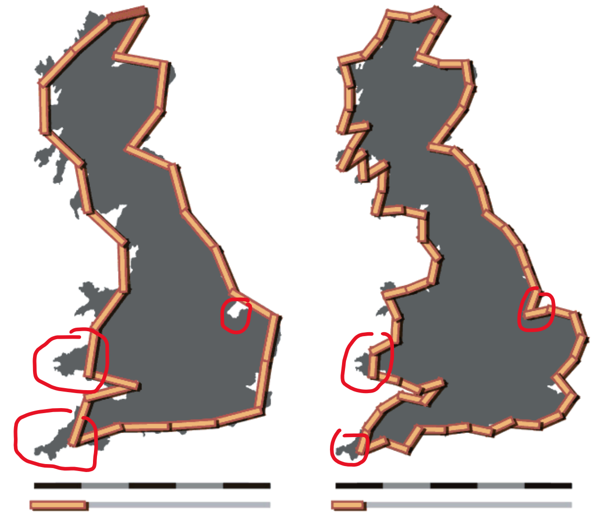
$$\rho_n(F, \mathcal{F}) = \inf_{\mathcal{H}^n} \operatorname{dist}(F, \mathcal{H}^n; \mathcal{F}) = \inf_{\mathcal{H}^n} \sup_{f \in F} \inf_{h \in \mathcal{H}^n} \|f - h\|_{\mathcal{F}}. \tag{4}$$

- *Nonlinear width:* approximation error
  Denoted ρ_n, depends on max complexity n

British Coastline

- Approximation with long rulers

100km rulers 50km rulers



Length: ~2800km ~3400km

- Error from using long rulers

Main result

Standard assumptions

Theorem 3.1. *Let (M, g) be a d-dimensional compact (separable) Riemannian manifold without boundary. From compactness, there exist real constants such that:*

1. *The Ricci curvature satisfies* $\mathrm{Ric} \geq -(d-1)K$, *where* $K > 0$;
2. $\mathrm{inj}(M) > 0$ *is the injectivity radius;*

Moreover, for any $1 \leq p, q \leq +\infty$, *the nonlinear width of* $W^{1,p}(1)$ *satisfies the lower bound for sufficiently large* n:

$$\rho_n(W^{1,p}(1), L^q(M, \mathrm{vol}_M)) \geq C(d, K, \mathrm{vol}(M), p, q)(n + \log n)^{-1/d}. \qquad (13)$$

In particular, the constant is independent of any latent dimension that (M, g) *may be embedded in.*

- Proof is constructive
  - Construct a particular subset of $\mathcal{W}^{1,p}(M)$ that is hard to approximate using low-complexity function classes
  - Uses approximation theory, differential geometry, functional analysis

Classical comparison + interpretation

Us: $$\rho_n(W^{1,p}(1;M), L^q) \gtrsim (n + \log n)^{-1/d}$$ d dimensional manifold

Related to curvature

Unavailable on manifolds

Classical: $$\rho_n(W^{k,p}(1;[0,1]^D), L^q) \gtrsim n^{-k/D}$$ D dimensional hypercube

- We lower bound the optimal approximation error in terms of the allowed statistical complexity n and the intrinsic dimension d
- As the statistical complexity increases, this error decreases
- The rate at which this error decreases is dependent only on the intrinsic dimension (and properties of the manifold)

Summary

- We demonstrate that function approximation is provably difficult, but *only in the intrinsic dimension*.
  - ➢ Increasing statistical complexity to approximate a function class yields slower rates as the dimension increases.
  - ➢ Our result is completely independent of any embedding dimensions
- **Implications**: Well-structured data is crucial for good training
- Open questions:
  - ➢ Extensions to higher Sobolev derivatives (requires better manifold approximation techniques)
  - ➢ Other manifold properties, e.g. reach, condition, that may have better bounds

arXiv:2408.06996