Blessing of Dimensionality in Natural Data

Hong Ye Tan (Supervisor: Prof. Carola-Bibiane Schönlieb)
Joint work with: Subhadip Mukherjee, Junqi Tang
Department of Applied Mathematics and Theoretical Physics, University of Cambridge

UNIVERSITY OF CAMBRIDGE

Background

Problem: *Curse of Dimensionality*

- Dimension is difficult – volumes scale exponentially
- Spurious relations, lack of convergence, … *in theory*
- Evidence to the contrary: *methods still work in practice*

Remedy: *Manifold Assumption*

- Natural datasets exhibit low-dimensional structure
  - E.g. natural images, genomics, human speech
- Manifolds: high-dimensional "surfaces"
  - Made by gluing "charts" together that look Euclidean.
  - "Union of manifolds" hypothesis: many manifolds instead of one

Related Literature

- *Manifold learning*: given points, what is the underlying manifold?
  - Related: kernel methods
- *Learning on a manifold*: given a function that takes values on a manifold, how difficult is it to learn? Sample complexity?
- *Dimensionality reduction*:
  - *Linear:* principal component analysis (PCA), linear determinant analysis (LDA), non-negative matrix factorization (NMF)
  - *Nonlinear:* autoencoders, kernel PCA, Isomap, local linear embedding

Aim

Q. How hard is training with natural data?

- Classical bounds are very loose.
- Given structure on our data, we wish to exploit this for better rates.

Fundamental concept: statistical complexity

- **Trade-off**: approximation power and generalisation: "bias-variance"

Theorem. The sample complexity for a function class with pseudo-dimension n is

$$m_L(\epsilon,\delta) = \frac{128}{\epsilon^2}\left(2n\log\left(\frac{34}{\varepsilon}\right) + \log\left(\frac{16}{\delta}\right)\right).$$

With probability at least $1-\delta$, the generalisation error with m_L samples is at most ε.

We consider approximating a general class of functions that covers many realistic applications and is "not too large".

Sobolev class: $W^{1,\infty}$ is the class of bounded functions with bounded first derivative.

- Think of elements as reconstruction operators, something we want to approximate.
- E.g. deblurring operator, CT reconstruction, speech recognition
- Our results also hold for $W^{1,p},\ p\in[1,\infty]$

Results

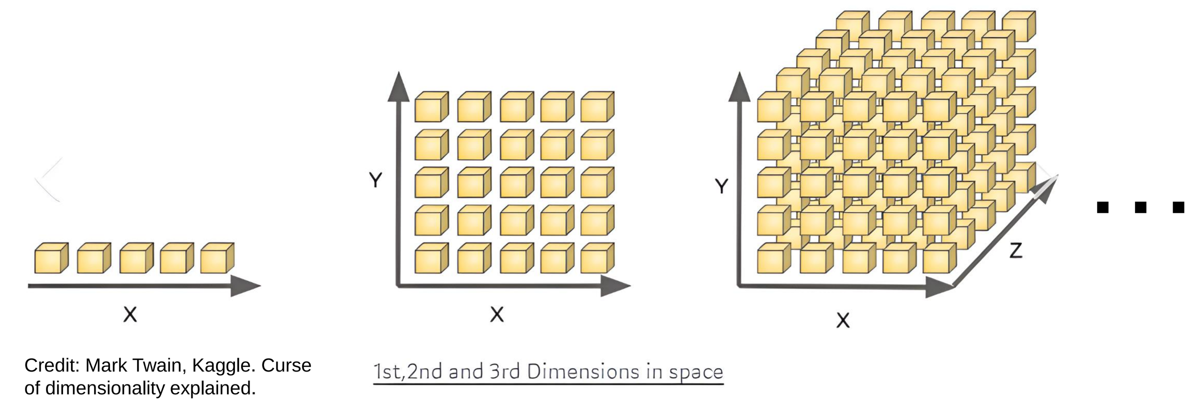
- We lower-bound the optimum approximation power of function classes with given statistical complexity, i.e. **bias**

Theorem. The statistical complexity of $W^{1,\infty}$ depends only on the dimensionality of the underlying manifold. In particular, the best approximation of $W^{1,\infty}$ with a function class of pseudo-dimension at most n scales with the intrinsic dimension of the data d:

$$\rho_n(W^{1,\infty}) \gtrsim (n+\log n)^{-1/d}$$

- This bound does ***not*** depend on the ambient data dimension! Only on the intrinsic dimension and properties of the manifold. ☆
- Matches existing bounds when data lies in a d-dimensional space.
- This bound is over optimal function classes with a given pseudo-dimension. It provides ***best-case*** bounds for uniform approximation, e.g. ReLU networks must have at least this width/depth/parameters to approximate this class.

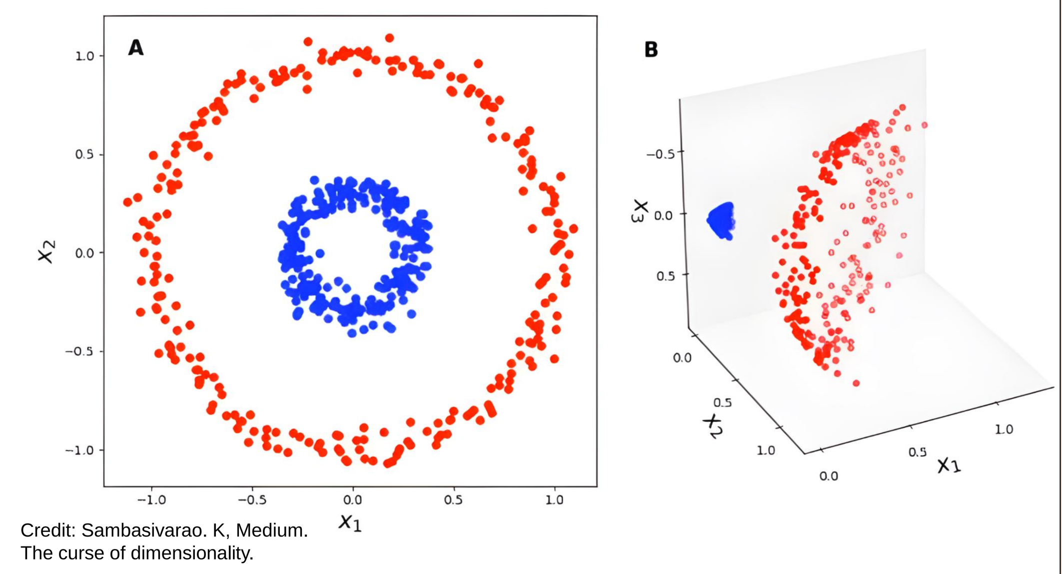
Figures



Breaks down in high dimensions:

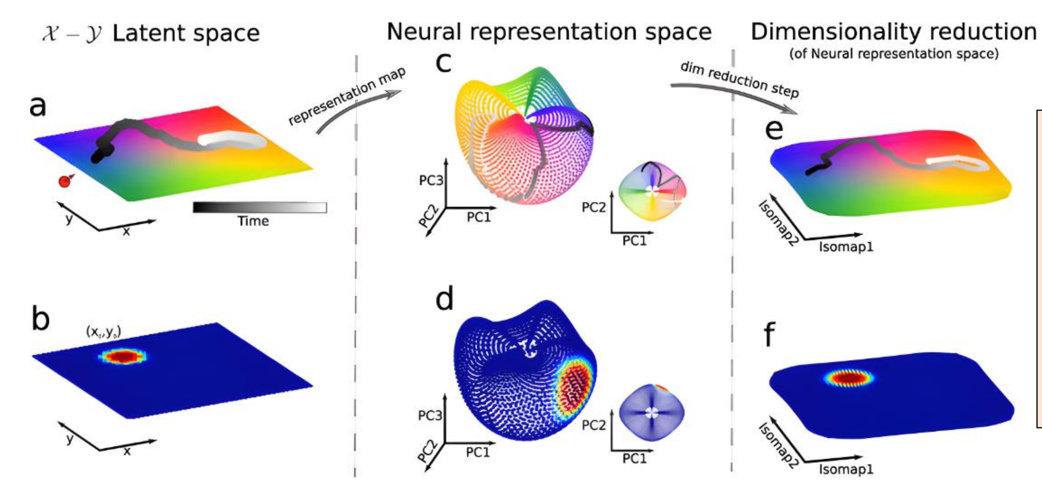
- Nearest neighbour
- k-NN
- Sampling
- Optimisation
- Anomaly detection

…

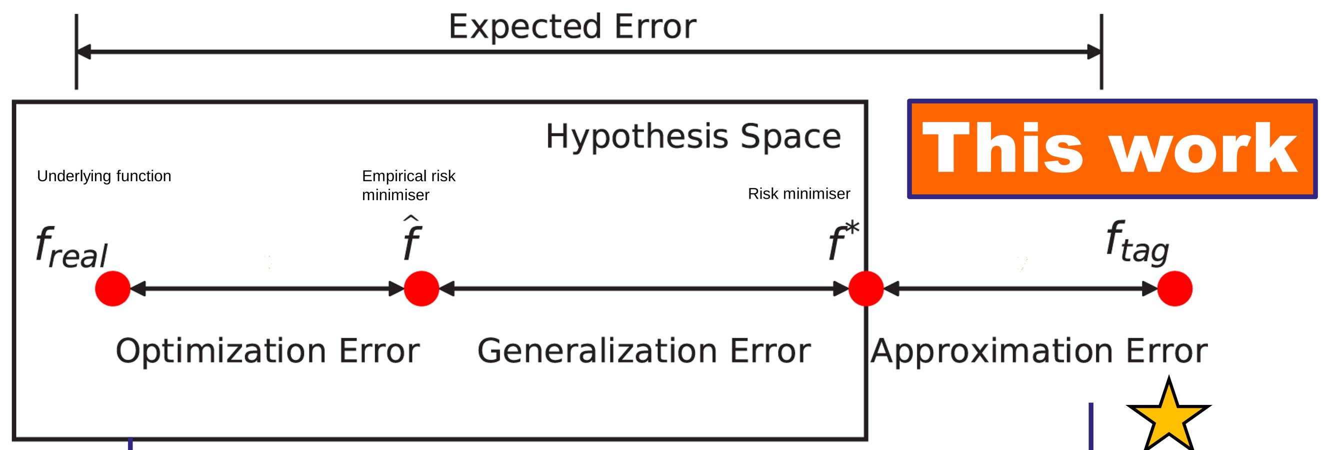


Data is concentrated on a lower-dimensional structure. Here the 3D data lies on two 2D surfaces.

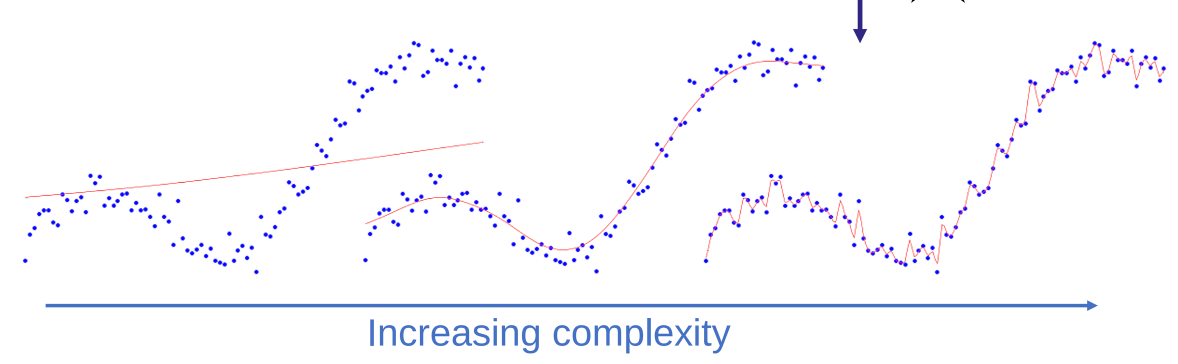
Traditional heuristics: higher dimensions may give rise to simpler description, from circular boundary in 2D to linear boundary in 3D



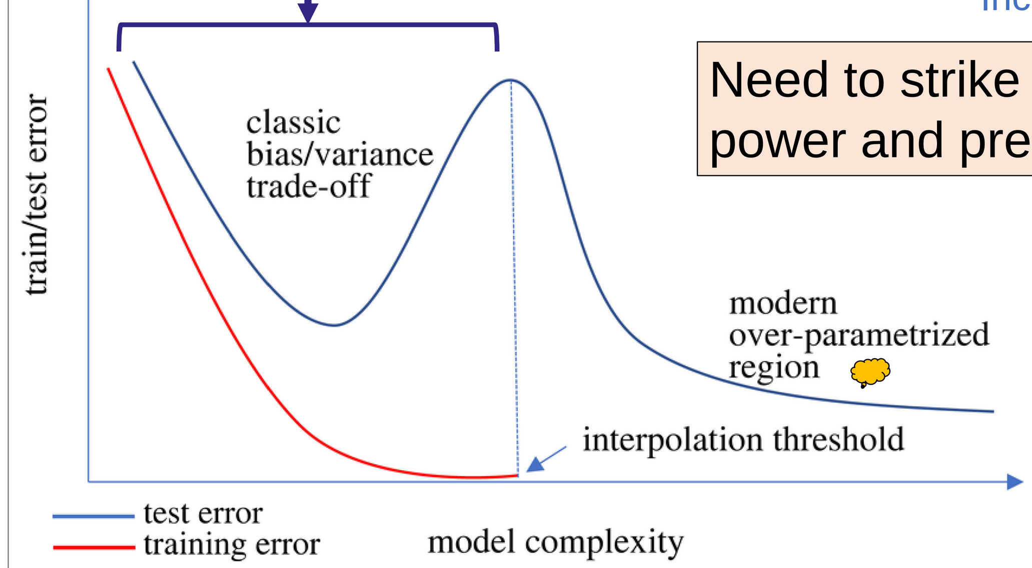
Neural networks can be used for dimensionality reduction. E.g. VAEs, GANs, use a latent space which gives "lower dimension".

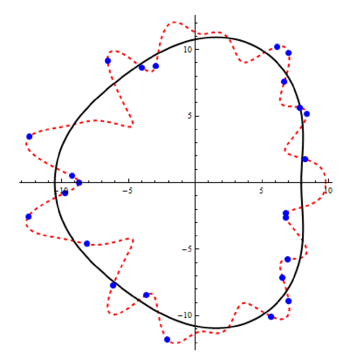


This work ☆



Need to strike a balance between approximation power and preventing overfitting.





Which one is the correct 1D approximant for this 2D data? The simple black line or the complicated dotted red line?

Conclusions

☆ Natural data is intrinsically low dimensional.

☆ We still need to balance approximation power with generalisation. The lower-dimension the data, the better generalisation we get for the same approximation power.

☆ Exploiting known structure can be useful

References

[1] Pope, Zhu, Abdelkader, Goldblum, Goldstein. *The Intrinsic Dimension of Images and Its Impact on Learning.* ICLR 2021
[2] Jin, Lu, Tang, Karniadakis. *Quantifying the generalization error in deep learning in terms of data distribution and neural network smoothness*. Neural Networks, 2020.
[3] Bengio, Courville, Vincent. *Representation Learning: A Review and New Perspectives*. IEEE TPAMI, 2013.

Acknowledgements and Disclosures

HYT is supported by GSK.ai and the Masason Foundation.