Blessing of Dimensionality

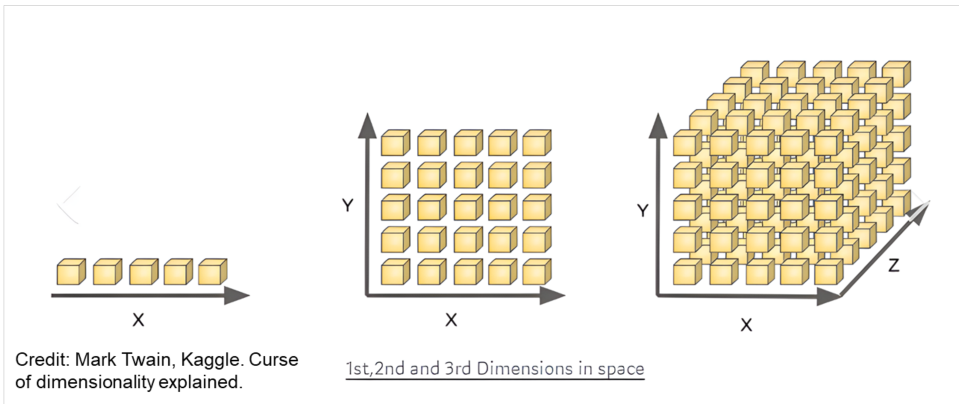
(for approximating Sobolev classes on a manifold)

Hong Ye Tan
Joint work with: Subhadip Mukherjee, Junqi Tang, Carola-Bibiane Schönlieb
Department of Mathematics, UCLA

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*



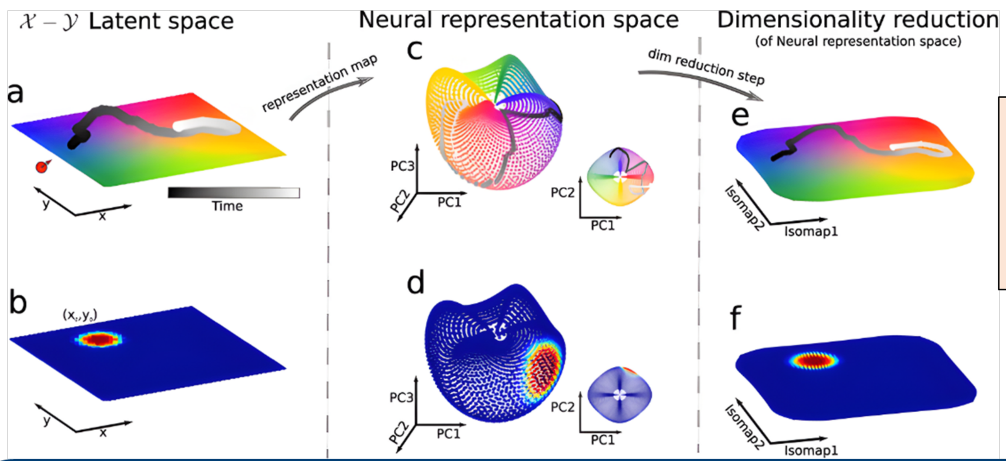
Breaks down in high dimensions:

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

…

Remedy: *Manifold Assumption*

- Natural datasets exhibit low-dimensional structure
  - E.g. natural images, genomics, human speech
- Manifolds: high-dimensional "surfaces"



VAEs, GANs, use a latent space which *assumes a smaller intrinsic dimension.*

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 generalization: "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 generalization 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 $W^{1,p}, p\in[1,\infty]$ or $W^{k,\infty}$

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:

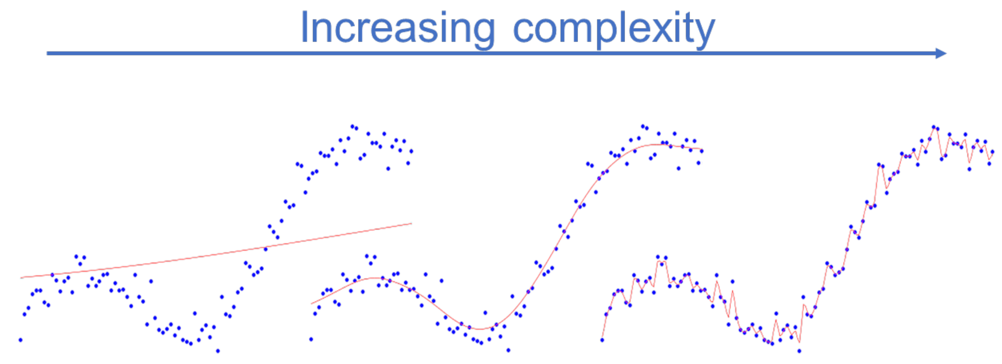
$$\rho_n(W^{1,\infty}) \gtrsim (n+\log n)^{-1/d}$$

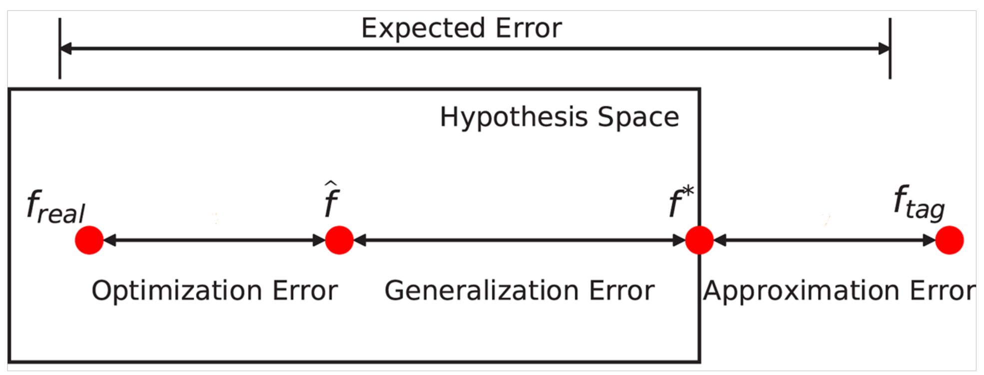
- This bound does ***not*** depend on the ambient data dimension! Only on the intrinsic dimension and properties of the manifold. ☆
- Matches existing bounds when data lies in a d-dimensional space.
- This bound is over optimal function classes with a given pseudo-dimension. It provides ***best-case***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

Where does this fit in?

Test error can be roughly decomposed as

1. Optimization error
   - Cannot perfectly optimize training loss
2. Generalization error
   - Overfitting
3. Approximation error
   - Model is not expressive enough to model the function





Existing results:

- Bigger networks implies better approximation
  - universal approximation

Our result:

- Better approximation *always requires* bigger networks

Model complexity ↑
⇔ Generalization ↓
⇔ Maximum trainability ↑

So what?

(Informal) The following things make training harder.

- More difficult datasets (e.g. MNIST -> ImageNet)
- Adding useless noise dimensions
- "Spiky data": e.g. misclassified training data
- A lower target error. Decreasing the error by a factor of 2 means increasing the complexity by 2^d

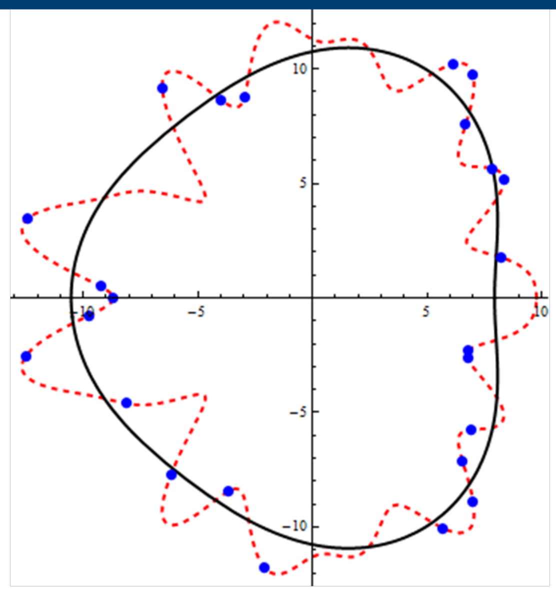
The following things *do not* make training harder.

- Adding some zeros in a new dimension.
- Artificially increasing the resolution of your image
- Data augmentation

Intuitive Examples

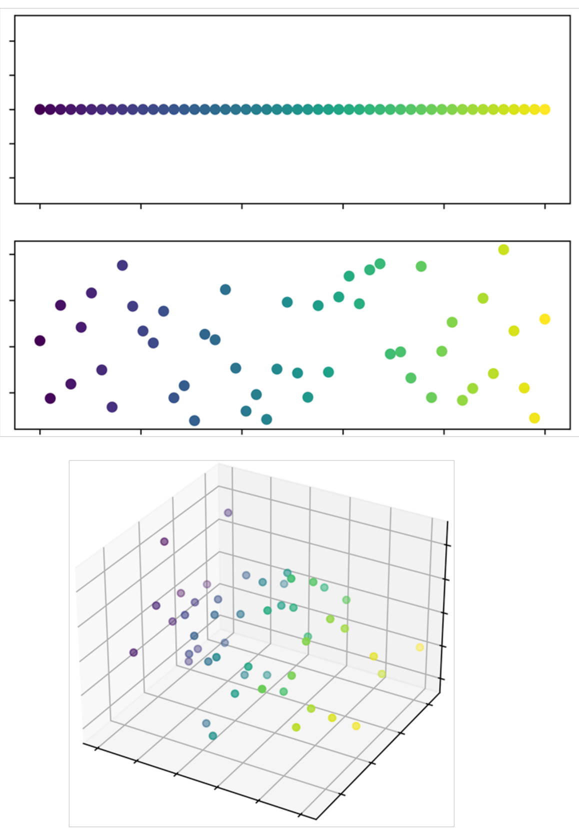
Is it easier to learn the black line or the red dotted line?

- The red line has more curvature and both are 1D. (Specific: need lower bound on Ricci curvature and volume)
- Theory: learning data on the red line is *provably* more difficult. ☆
- **Your data needs structure!**



Which colors are easiest to learn?

- Obviously the 1D one.
- The data is augmented with random noise in orthogonal directions.
- Theory: learning data on the red line is more difficult.
- Exercise: check with some simple neural networks! It gets harder to train.
- ☆ **You should remove confounding variables!**



Conclusions

☆ Natural data is (probably) intrinsically low dimensional.

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

☆ Data analyst jobs are safe.