Data classes for which DNNs can overcome the curse of dimensionality.

Theories of Deep Learning: C6.5, Video 4

Prof. Jared Tanner
Mathematical Institute
University of Oxford

University of Oxford
Mathematical Institute

Oxford Mathematics

DNNs as function approximators

Functions act as classifiers and other machine learning tasks

UNIVERSITY OF OXFORD
Mathematical Institute

Classification of inputs $x \in \mathbb{R}^n$ to c classes denoted by $\{e_i\}_{i=1}^{c}$, is modelled by a function $H(x)$ for which $H(x) = e_i$ for all x in class i where $e_i(\ell) = 1$ for $i = \ell$ and 0 otherwise.

- Network architectures are able to approximate any function (Cybenko (89') and Hornik (90')).
- The compositional nature of DNNs result in an exponential expressivity only obtained by exponentially wide shallow NNs.
- Telgarsky 15' give a precise example of the aforementioned for ReLU activation
- Yarotsky 16' develop local exponential approximation bounds using polynomial approximation and $\mathcal{O}(\log(1/\epsilon))$ depth.

For $\sigma(x) = \max(x, 0)$ let $f(x) = h_3(x) = \sigma(2\sigma(x) - 4\sigma(x - 1/2))$ and iterate this 2-layer network k times to obtain a $2k$-layer network $f^k(x) = f(f(\cdots(f(x)\cdots))$ with the property that it is piecewise linear with change in slope at $x_i = i2^{-k}$ for $i = 0, 1, \ldots, 2^k$ and moreover takes on the values $f^k(x_i) = 0$ for i even and $f^k(x_i) = 1$ for i odd.

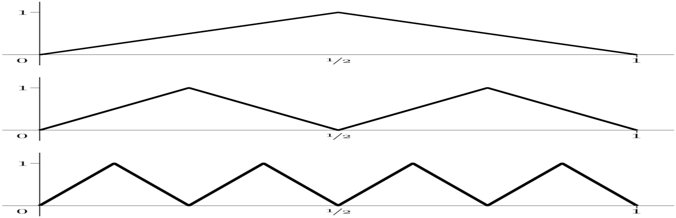


Figure 2: f_m, f_m^2, and f_m^3.

The Sobolev norm is similar to that of functions with $m-1$ derivatives that are Lipschitz continuous $C^{m-1}([0,1]^d)$ excluding sets of measure zero.

$$\|f\|_{W_m^\infty}([0,1]^d) = \max_{|s|\leq m}\text{esssupp}_{x\in[0,1]^d}|D^s f(x)|.$$

Define the unit ball of functions in $W_m^\infty([0,1]^d)$ as

$$F_{m,d} = \left\{ f \in W_m^\infty([0,1]^d) : \|f\|_{W_m^\infty}([0,1]^d) \leq 1 \right\}.$$

Theorem (Yarotsky 16')

For any d, m and $\epsilon \in (0,1)$, there is a ReLU network with depth at most $c(1+\ln(1/\epsilon))$ and at most $c\epsilon^{-d/m}(1+\log(1/\epsilon))$ weights (width $\mathcal{O}(\epsilon^{-d/m})$), for c a function of d, m, that can approximate any function from $F_{d,m}$ within absolute error ϵ.

Curse of dimensionality (Yarotsky 16')

Exponential in d/m growth in number of weights.

Yarotsky 16' results show exponential approximation in depth, but the overall number of weights is $\mathcal{O}(\epsilon^{-d/m})$. Recall

$$\|f\|_{W_m^\infty}([0,1]^d) = \max_{|s|\leq m} \text{esssupp}_{x\in[0,1]^d} |D^s f(x)|.$$

Theorem (Yarotsky 16')

For any d, m and $\epsilon \in (0,1)$, there is a ReLU network with depth at most $c(1+\ln(1/\epsilon))$ and at most $c\epsilon^{-d/m}(1+\log(1/\epsilon))$ weights (width $\mathcal{O}(\epsilon^{-d/m})$), for c a function of d, m, that can approximate any function from $F_{d,m}$ within absolute error ϵ.

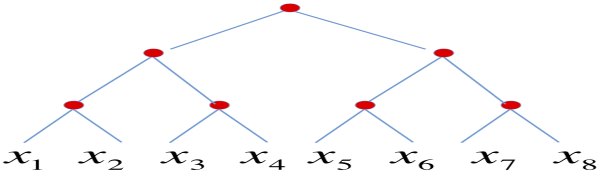
https://arxiv.org/pdf/1610.01145.pdf

To avoid curse of dimensionality need $m \sim d$ or more structure in the function F to be approximated; e.g. compositional structure.

Compositional structured functions (Poggio et al. 17')

Extending the compositional nature of Yarotsky dimensionally

Consider functions with a binary tree hierarchical structure:



where $x \in \mathbb{R}^8$ and
$f(x) = h_3(h_{21}(h_{11}(x_1, x_2), h_{12}(x_3, x_4)), h_{22}(h_{13}(x_5, x_6), h_{14}(x_7, x_8)))$
Let $W_m^{n,2}$ be the class of all compositional functions $f(\cdot)$ of n variables with binary tree structure and constituent functions $h(\cdot)$ of 2 variables with m bounded derivatives.
`https://arxiv.org/pdf/1611.00740.pdf`

Compositional structured functions (Poggio et al. 17')

Each constituent function is a map from $\mathbb{R}^2 \to \mathbb{R}$

Mathematical Institute

The set $W_m^{n,2}$ of of all compositional functions $f(\cdot)$ of n variables with binary tree structure and constituent functions $h(\cdot)$ of 2 variables with m bounded derivatives can be effectively approximated using a DNN with a rate dictated by the ability to approximate functions $\mathbb{R}^2 \to \mathbb{R}$; e.g. effectively locally $d = 2$.

Theorem (Poggio 17')

Let $f(\cdot) \in W_m^{n,2}$ and consider a DNN with the same binary compositional tree structure and an activation $\sigma(\cdot)$ which is infinitely differentiable, and not a polynomial. The function $f(\cdot)$, can be approximated by ϵ with a number of weights that is $\mathcal{O}\left((n-1)\epsilon^{-2/m}\right)$.

https://arxiv.org/pdf/1611.00740.pdf

Compositional structured functions (Poggio et al. 17')

Compositional functions $W_m^{n,2}$ compared to shallow NNs and Yarotsky(16')

Oxford Mathematical Institute

The set $W_m^{n,2}$ of of all compositional functions $f(\cdot)$ of n variables with binary tree structure are effectively $d = 2$ in the DNN approximation requirements, but are much richer than $d = 2$.

Functions can be approximated within ϵ with a DNN from $\mathcal{O}(ln(1/\epsilon))$ layers with a number of weights:

- $\mathcal{O}(\epsilon^{-d/m})$ for general locally smooth functions (Yarotsky 16'),
- $\mathcal{O}\left((n-1)\epsilon^{-2/m}\right)$ for $f(\cdot) \in W_m^{n,2}$, binary tree structure and constituent functions in $C_m[0,1]^2$.
- $\mathcal{O}(\epsilon^{-d/m})$ for shallow NNs is best possible for $f(\cdot) \in W_m^n$ which have non-binary hierarchical tree structures.

https://arxiv.org/pdf/1611.00740.pdf

Definition of local effective dimensionality (Poggio et al. 17')

Local dimensionality determined by approximation rate ϵ^{-d}.

Mathematical Institute

Definition (Poggio 17')

The effective dimensionality of a function class W is said to be d if for every $\epsilon > 0$, any function within W can be approximated within an accuracy ϵ by a DNN at rate ϵ^{-d}.

In the prior slide we had examples of complex compositional functions with effective dimensionality 2. These could be extended naturally to local *effective dimensionality* d_{eff} and *local smoothness* m_{eff} for rate $\epsilon^{-d_{eff}/m_{eff}}$.

Restriction to a data class decreases d_{eff} and localisation can increase the smoothness m_{eff} substantially.
`https://arxiv.org/pdf/1611.00740.pdf`

Intrinsic dimensionality of sub-manifolds (Hein et al. 05')

MNIST exemplar dataset classes are approximately under 15 dimensional

University of Oxford Mathematical Institute

Estimates of dimensionality within MNIST digit classes using three approaches: the reference below, and two others building on local linear embedding.

Table 7. Number of samples and estimated intrinsic dimensionality of the digits in MNIST.

1	2	3	4	5
7877	6990	7141	6824	6903
8/7/7	13/12/13	14/13/13	13/12/12	12/12/12
6	7	8	9	0
6876	7293	6825	6958	6903
11/11/11	10/10/10	14/13/13	12/11/11	12/11/11

https://icml.cc/Conferences/2005/proceedings/papers/037_Intrinsic_HeinAudibert.pdf

A manifold model can explicitly represent the data through:

$$X = f(CF/\sqrt{d}) \in \mathbb{R}^{p,n}$$

where:

- $F \in \mathbb{R}^{d,n}$ are the d features used to represent the data
- $C \in \mathbb{R}^{p,d}$ combines the $d < n < p$ features
- $f(\cdot)$ is an entrywise locally smooth nonlinear function.

This data model is the same as a generative adversarial network (GAN) and is similar to dictionary learning and subspace clustering models where C is typically sparse.

`https://hal-cea.archives-ouvertes.fr/cea-02529246/document`

Additional approximation theory resources

Review articles and courses elsewhere

Further references for the approximation theory perspective of deep learning include:

- Telgarsky's "Deep Learning Theory" course, lectures 1-11: http://mjt.cs.illinois.edu/courses/dlt-f20/
- Matthew Hirn's "Mathematics of Deep Learning" course: lectures 20-24. https://matthewhirn.com/teaching/spring-2020-cmse-890-002/
- DNN Approximation Theory by Elbrachter et al. (19') https://www.mins.ee.ethz.ch/pubs/files/deep-it-2019.pdf