Machine Learning

Lecture 17: Generalization 4

Hao Tang

October 28, 2022

Universal approximation

- For every $\epsilon > 0$, given any Lipschitz function $f : [-1,1]^d \to [-1,1]$, there is a network g such that $|g(x) - f(x)| \leq \epsilon$ for any x.
- The number of nodes needed to achieve this is $O(2^d)$.

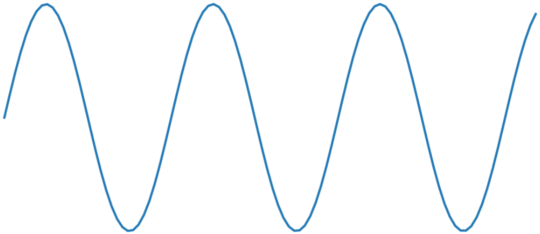
Universal approximation

- Polynomials are universal approximators.
- Decision trees are universal approximators.
- Universal approximation does not explain why neural networks are so "special."

Depth separation

- There exists functions which can be approximated with small depth 3 networks, but cannot be approximated with depth 2 networks without using $O(2^d)$ nodes.
- Functions to show these results tend to oscillate a lot.
- Some believe the results are pathological and do not happen in practice.

VC dimension of a sine function



Universal approximation

- What can be implemented with polynomial number of of nodes?
- Any Turing machine that runs in T operations can be implemented with a neural network of depth $O(T)$ with a total $O(T^2)$ nodes.
- Recall that VC dimension of neural networks is $O(|E| \log |E|)$, where E is the number of edges in the network.

Hardness of optimizing neural networks

- Training a 2-layer 3-node neural network is NP-complete.
- The proof converts instances of an NP-complete problem into data points.
- If we can minimize the loss of the training set, we solve the NP-complete problem.
- Maybe we don't need to solve this exactly?

Hardness of optimizing neural networks

- Approximating ERM is NP hard.
- The loss is not necessarily convex.
- ERM is hard for neural networks.

Optimizing neural networks on random labels

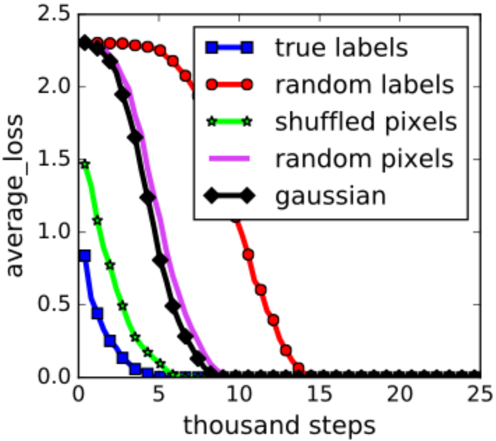


Image credit: (Zhang *et al.*, 2017)

Overparameterization

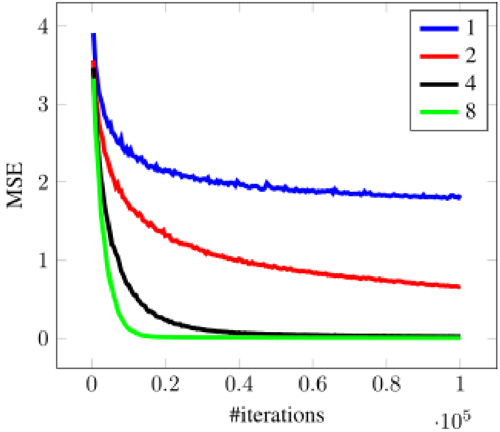


Image credit: (Livni *et al.*, 2014)

Overparameterization

- Overparameterization means using a lot more nodes than the number of points.
- Overparameterization helps optimization.
- Wouldn't the model just memorize the training set?
- Wouldn't the hypothesis class be too large to have good generalization error?

Overparameterization

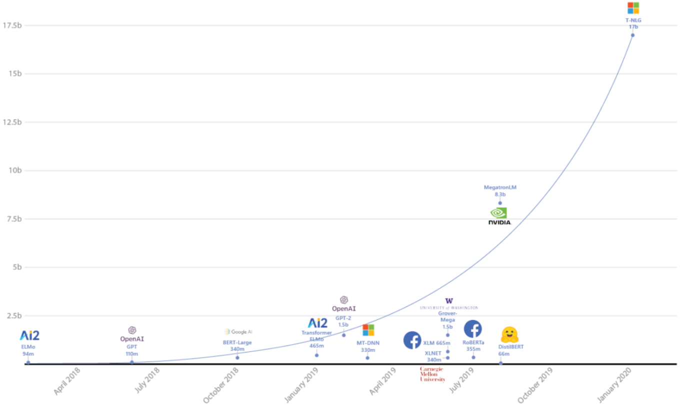


Image credit: (Rosset, 2020)

Overparameterization

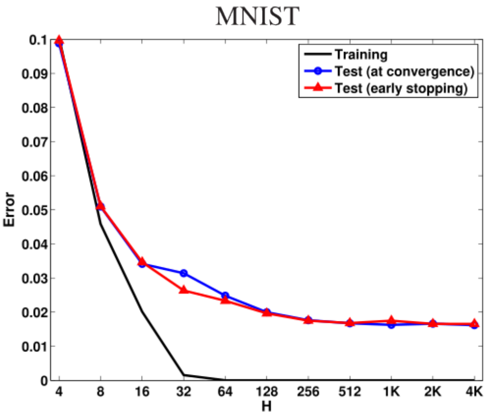
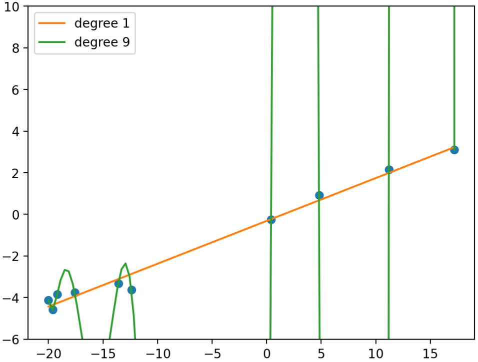


Image credit: (Neyshabur *et al.*, 2014)

Interpolation

- Fitting a data set to training error zero is called interpolation.
- Why doesn't interpolation overfit?

Interpolation



Interpolation

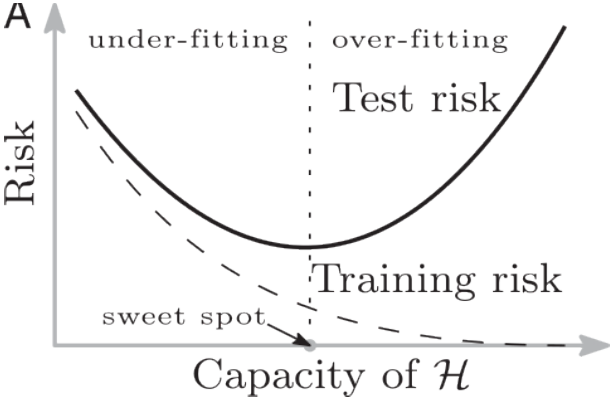


Image credit: (Belkin *et al.*, 2019)

Interpolation

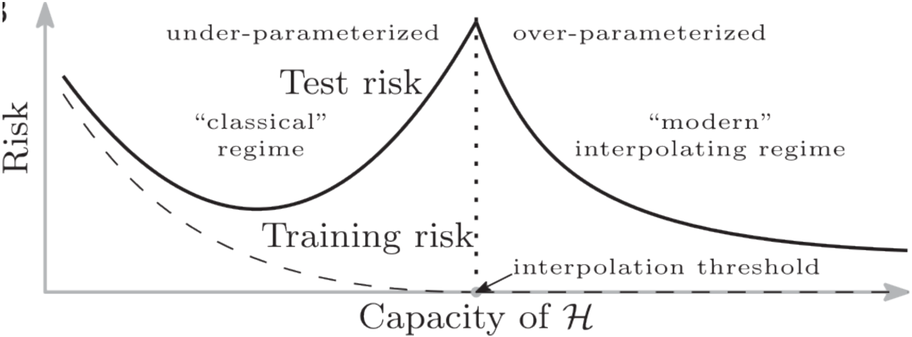


Image credit: (Belkin *et al.*, 2019)

Interpolation

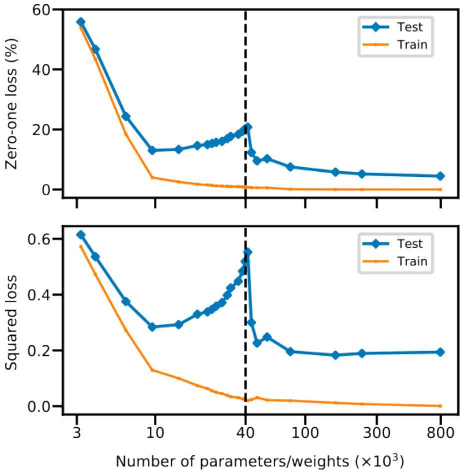


Image credit: (Belkin *et al.*, 2019)

Overfitting

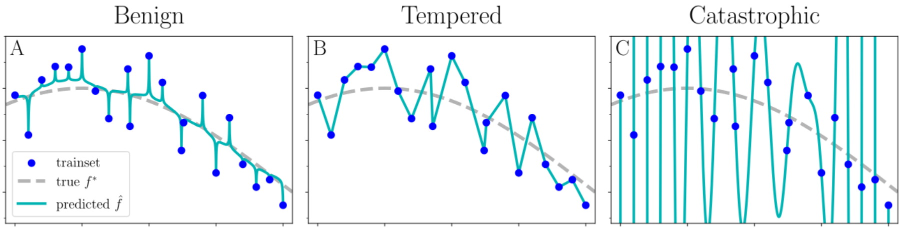


Image credit: (Mallinar *et al.*, 2022)

In practice

- Always start with the training error.
- Always start with ERM.
- Why is the training error not close to zero?
- Regularize