Overfitting and Generalization Performance

March 30, 2021

Main Papers

- *Reconciling modern machine-learning practice and the classical bias-variance trade-off*, Belkin et al
- *To undersand deep learning we need to understand kernel learning*, Belkin et al

Introduction

General Aim

Given training sample

$$(x_1, y_1), ..., (x_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$$

learn a predictor $h_n : \mathbb{R}^d \to \mathbb{R}$ that predicts y given new x.

Empirical Risk Minimization (ERM)

Minimize training risk: $\frac{1}{n} \sum_{i=1}^{n} \ell(h(x_i), y_i)$ given a loss function ℓ.

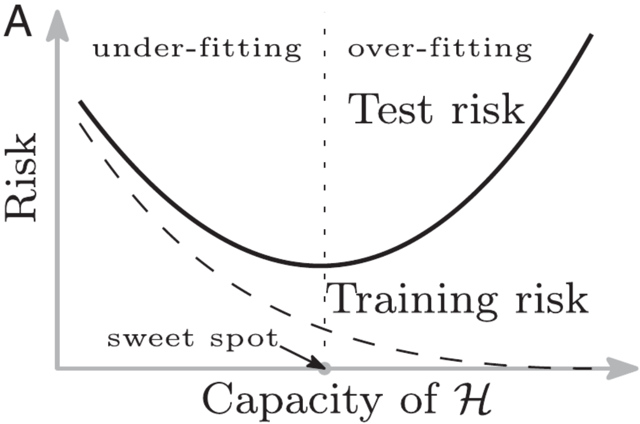
Generalization

- Find h_n that performs well on unseen data.
- Minimize true risk: $h^*(x) = \arg\min_h \mathbb{E}[\ell(h(x), y)]$ where (x, y) drawn independently from P.
- ERM goal: $h^*_{ERM}(x) = \arg\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(h(x_i), y_i)$
- $\mathcal{H}$ is a function class that contains functions approximating h^*.

"Classical" thinking

- Finding a balance between underfitting and overfitting.
- "Bias-Variance Tradeoff"
- 0 training error does not tend to generalize well.
- Control function class $\mathcal{H}$ implicitly or explicitly.

Generalization of performance



Classical curve from bias variance tradeoff.

Modern practice

- Modern ML methods such as large neural networks and other non-linear predictors have very low to no training risk
- NN architectures chosen such that interpolation can be achieved.
- Works even when training data have high levels of noise.

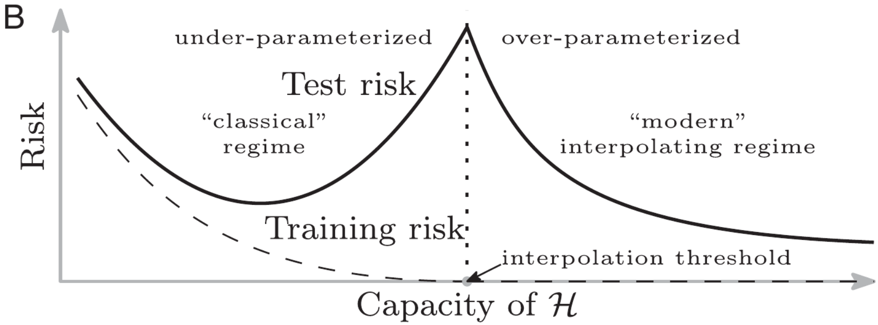
# params	random crop	weight decay	train accuracy	test accuracy
1,649,402	yes	yes	100.0	89.05
	yes	no	100.0	89.31
	no	yes	100.0	86.03
	no	no	100.0	85.75

Understanding Deep Learning Requires Rethinking Generalization, Zhang et al

"Double Descent"

- "Double Descent" curve proposed and empirically observed to some extent.
- Curve the extends beyond the point of interpolation
- Risk decreases beyond this point, typically surpassing performance of classical stopping point.

Double Descent Curve



Explanations on Double Descent

Why should the test risk decrease even when empirical risk stays the same?

- Capacity of function class needs not suit the appropriate inductive bias for the problem.
- By having a larger function class, may find a function that matches the inductive bias better.
- Eg., smoother function, smaller norm, larger margin.

Short introduction to Kernels and RKHS

- For kernel $k : X \times X \to \mathbb{K}$, there exists $\mathbb{K}$-Hilbert space H and map $\psi : X \to H$ such that for all $x_1, x_2 \in X$,

$$k(x_1, x_2) = \langle \psi(x_2), \psi(x_1) \rangle.$$

- ψ is called a feature map of k.
- Gaussian RBF kernel with width γ:

$$k_{\gamma,\mathbb{C}^d}(x, x') := e^{\frac{-\|x-x'\|_2^2}{\gamma^2}}.$$

Let H be a $\mathbb{K}$-Hilbert space, consisting of functions $f : X \to K$.

- Function $k : X \times X \to \mathbb{K}$ is a reproducing kernel of H if
  $f(x) = \langle f, k(\cdot, x)\rangle \ \forall f \in H, \ \forall x \in X$
  and $k(\cdot, x) \in H \ \forall x \in X$.
- Dirac functional $\delta_x : H \to \mathbb{K}$, $\delta_x(f) := f(x)$ is continuous.
  H is called a reproducing kernel Hilbert space.
- Reproducing kernels are kernels.

Random Fourier Features

Let the function class $\mathcal{H}_\infty$ be the Reproducing Kernel Hilbert Space (RKHS) corresponding to the Gaussian kernel.
We consider the following non-linear parametric model:

Random Fourier Features (RFF)

Random Features for Large-Scale Kernel Machines (Rahimi et al)

Let the function class $\mathcal{H}_N$ consist of functions $h_n : \mathbb{R}^d \to \mathbb{C}$ of the form:

$$h(\cdot) = \sum_{k=1}^{N} a_k \phi(\cdot, v_k)$$

where $\phi(\cdot, v_k) := e^{i\langle v_k, \cdot \rangle}$ vectors $v_1, ..., v_N$ are sampled independently from the standard normal distribution in $\mathbb{R}^d$.

H_N has N parameters in $\mathbb{C}$, $\{ a_1, ..., a_N \}$.
As $N \to \infty$, H_N becomes a closer approximation to $\mathcal{H}_\infty$

Empirical Evidence - Learning Procedure

- Given training sample $(x_1, y_1), ..., (x_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$.
- Minimize empirical risk: $\frac{1}{n}\sum_{j=1}^{n}(h(x_j) - y_j)^2$ for $h \in \mathcal{H}_N$.
- When minimizer not unique ($N > n$), choose the minimizer with the coefficients $(a_1, ..., a_N)$ that have the smallest ℓ_2 norm.
- Let this predictor be: $h_{n,N} \in \mathcal{H}_N$.

Min Norm RKHS solution $h_{n,\infty}$

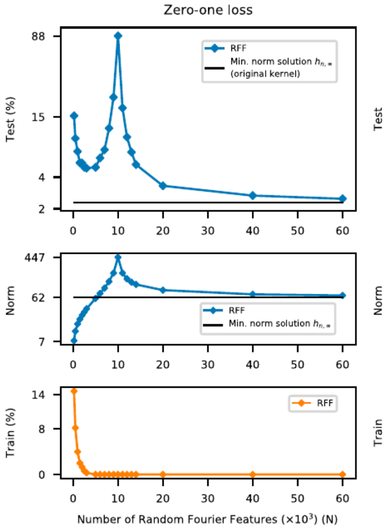
- $f^* = \arg\min_{f \in \mathcal{H}_\infty, f(x_i)=y_i} \|f\|_{\mathcal{H}_\infty}$
- $f^*(\cdot) = \sum_{i=1}^n \alpha_i k(\cdot, x_i)$ (Representer theorem)
- Since f^* interpolates, $(\alpha_1, \ldots, \alpha_n)^\mathsf{T} = K^{-1}(y_1, \ldots, y_n)^\mathsf{T}$
- $\|f\|^2_{\mathcal{H}_\infty} = \sum_{i,j} \alpha_i \alpha_j K(x_i, x_j)$

Approximation theorem

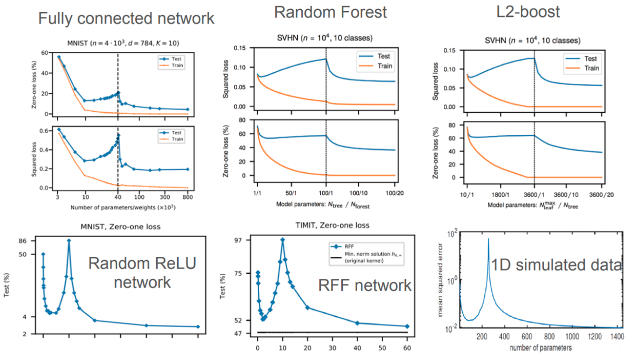
Fix $h^* \in \mathcal{H}_\infty$. Let $(x_1, y_1), \ldots, (x_n, y_n)$ be i.i.d. random variables where x_i drawn randomly from a compact cube $\Omega \subset \mathbb{R}^d$, $y_i = h^*(x_i)\,\forall i$. There exists $A, B > 0$ such that for any interpolating $h \in \mathcal{H}_\infty$ with high probability

$$\sup_{x \in \Omega} |h(x) - h^*(x)| < A e^{-B(n/\log n)^{1/d}} (\|h^*\|_{\mathcal{H}_\infty} + \|h\|_{\mathcal{H}_\infty})$$

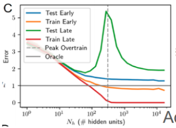
Empirical Evidence - Results



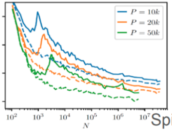
Empirical Evidence - Other Examples



[B., Hsu, Ma, Mandal, 18]



Advani, Saxe, 2017



Spigler, et al, 2018