Mathematics of Deep Learning

Course Logistics

Lectures Th 4:55 → 6:55 (5 min breaks)

Recitation Mondays @ 3pm – 4pm
Carlos Domingo (PhD student).

Office Hours Fridays at 9am → 10:30 ish.

Evaluation: Final Project 90%
Participation 10% (OH, Campus Wire, Class).

- Final Project: In-depth survey on a selected topic.
  NOT a Research Project. By groups of 2, 4
- List of topics are in the website (tentative)
- Abstract: due mid-semester

Bibliography: Two major references:
- ML from first principles by Francis Bach
- Lecture Notes by Matus Telgarski

Main Objective of this course

- When and why can neural nets learn in high-dimension.

- Focus on the simplest learning setup: supervised learning.
- Focus on the simplest class of NN: shallow nets.

>50%
- Study role of architecture (depth)
- Study generative modeling.

- Discuss old & new results.

Lecture 1: The curse of dimensionality; NNs and approximation.

Basic Supervised Learning Setup

→ Given data $\{(x_i, y_i)\}_{i=1 \dots n}$ with $x_i \in \mathcal{X}$ (high-dim), $y_i \in \mathcal{Y}$ (label / target space)

estimate a mapping $f: \mathcal{X} \to \mathcal{Y}$ that

generalises on future data ? "glorified" interpolation.

→ IID assumption: data is drawn iid from an underlying distribution ν on $\mathcal{X} \times \mathcal{Y}$. $(x_i, y_i) \overset{iid}{\sim} \nu$

→ Pointwise loss $\ell: \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$

Given $f: X \to Y$, this defines

→ $R(f) := \mathbb{E}_{x,y}[\ell(f(x), y)]$ "population risk".

→ $\hat{R}(f) := \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i), y_i)$ "empirical risk".

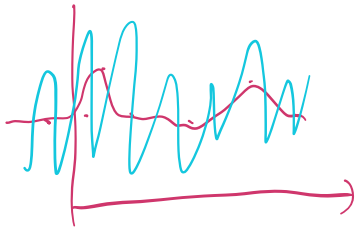
$\mathbb{E}\hat{R} = R$ ($\hat{R}$ is an unbiased estimator of R).

→ Hypothesis Space $\mathcal{F} = \{ f: X \to Y \}$. Assume

$\mathcal{F}$ is a normed space (Banach): for each $f \in \mathcal{F}$

we can assign a "complexity" measure $\begin{pmatrix} \gamma(f) \\ \|f\|_{\mathcal{F}} \end{pmatrix}$

$\mathcal{F}_\delta = \{ f \in \mathcal{F};\ \|f\|_{\mathcal{F}} \leq \delta \}$.



Examples of complexity:

→ (Euclidean) norm of the weights in a NN

→ Number of parameters

→ Number of gradient-descent iterations for models in $\mathcal{F}$ reached by gradient-descent.

→ Empirical Risk Minimisation (ERM)

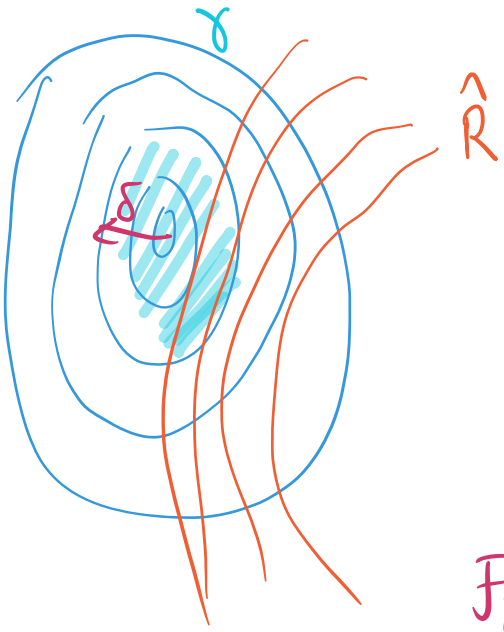
we search for small empirical risk with small complexity

(C) $\min_{\gamma(f) \leq \delta} \hat{R}(f)$ (P) $\min_{f \in \mathcal{F}} \hat{R}(f) + \lambda\, \gamma(f)$

(I) $\min \gamma(f)$

s.t. $f(x_i) = y_i \iff \hat{R}(f) = 0$

"interpolant" or "memorization" regime.



$$\mathcal{F}_\delta = \{ f \in \mathcal{F};\ \gamma(f) \le \delta \}$$

→ Basic Decomposition of Risk

- let $\hat{f} \in \mathcal{F}_\delta$ produced by an arbitrary algorithm.

$$R(\hat{f}) - \inf_{f \in \mathcal{F}} R(f) = R(\hat{f}) - \inf_{f \in \mathcal{F}_\delta} R(f) + \inf_{f \in \mathcal{F}_\delta} R(f) - \inf_{f \in \mathcal{F}} R(f)$$

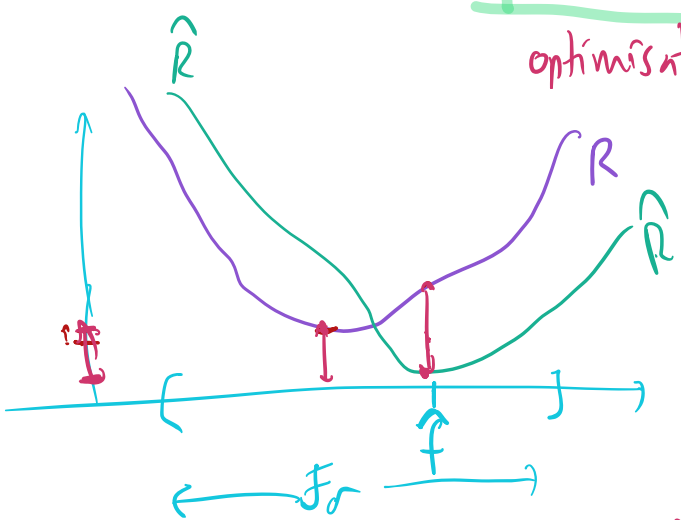
$$= R(\hat{f}) - \hat{R}(\hat{f}) + \hat{R}(\hat{f}) - \inf_{f \in \mathcal{F}_\delta} \hat{R}(f) + \inf_{f \in \mathcal{F}_\delta} \hat{R}(f) - \inf_{f \in \mathcal{F}_\delta} R(f) + \mathcal{E}_{appr.}$$

$R(\hat{f}) - \hat{R}(\hat{f})$: $\le \sup_{\mathcal{F}_\delta} |R - \hat{R}|$

$\hat{R}(\hat{f}) - \inf_{f \in \mathcal{F}_\delta} \hat{R}(f)$: optimisation error.

$\inf_{f \in \mathcal{F}_\delta} \hat{R}(f) - \inf_{f \in \mathcal{F}_\delta} R(f)$: $\le \sup_{f \in \mathcal{F}_\delta} |R(f) - \hat{R}(f)|$

$\mathcal{E}_{appr.}$: approximation error.



$$\left| \inf_{\mathcal{F}_\delta} \hat{R} - \inf_{\mathcal{F}_\delta} R \right| \le \sup_{f \in \mathcal{F}_\delta} |R(f) - \hat{R}(f)|$$

$f^* = \operatorname{argmin} R$

suppose $\inf_{\mathcal{F}_\delta} \hat{R} < \inf_{\mathcal{F}_\delta} R$

$$\left| \inf_{\mathcal{F}_\delta} \hat{R} - \inf_{\mathcal{F}_\delta} R \right| \leq \left| \hat{R}(f^*) - R(f) \right|$$

$$\leq \sup_{f \in \mathcal{F}_\delta} \left| R(f) - \hat{R}(f) \right|$$

$$\leq \boxed{2 \sup_{f \in \mathcal{F}_\delta} \left| R(f) - \hat{R}(f) \right|} + \varepsilon_{opt} + \varepsilon_{appr.}$$

(uniform) statistical error.

"Generalisation error"$(\delta) \leq$ stat error(δ) + opt error(δ) + approx error(δ)

As $\delta \to 0$,	$\downarrow$	?	$\uparrow$
As $\delta \to \infty$	$\uparrow$	?	$\downarrow$

→ Statistical Error. $\sup_{f \in \mathcal{F}_\delta} \left| R(f) - \hat{R}(f) \right|$ Random quantity.

It depends on two basic quantities:

Complexity: δ and n: number of datapoints.

Fix f. $\hat{R}(f) - R(f) =$

$$= \frac{1}{n} \sum_{i=1}^{n} \left[\underbrace{\ell(f(x_i), y_i)}_{\text{Random variable.}} - \mathbb{E}_{x,y}\, \ell(f(x), y) \right]$$

$\frac{1}{n}\sum_{i=1}^{n}(Z_i - \mathbb{E}(Z_i))$ $\quad Z_1 \dots Z_n$ are i.i.d.

Z $\quad \sigma^2 = \mathbb{E}\left((Z-\mathbb{E}Z)^2\right)$ $\quad \mathbb{E}Z = m$

$$\text{Var}\left[\frac{1}{n}\sum_{i=1}^{n}(Z_i - m)\right] = \frac{\sigma^2}{n}$$

$$\Rightarrow |R(f) - \hat{R}(f)| \sim \frac{\text{std}(\ell(f(X),Y))}{\sqrt{n}}$$ Monte-Carlo Estimate.

non-asymptotic; we can quantify with tail probability Bounds.

→ From point-wise control (fixing f before looking at the data) to uniform control.

Requires concentration tools from empirical process (e.g. Rademacher complexity); but final bound

$$\varepsilon_{stat} \sim \frac{g(\mathcal{F}, \delta)}{\sqrt{n}}$$

The Curse of Dimensionality

Q: How do approximation and statistical errors behave as the input dimension increases?

Statistical curse

We observe $\{(X_i, f^*(X_i))\}_{i=1\dots n}$ $\quad X_i \sim N(0, I_d)$

f^* is the target function.

Q: How many samples are needed to estimate f^* up to error ε, ie $\mathbb{E}_X |\hat{f}(x) - f^*(x)|^2 \leq \varepsilon$?

↳ Suppose first that f^* is linear; $f^*(x) = \langle x, \theta^* \rangle$

$\theta^* \in \mathbb{R}^d$, $x \in \mathbb{R}^d$

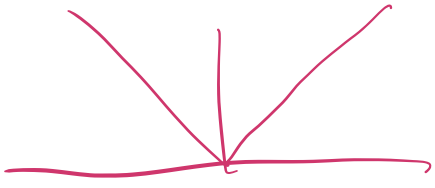
$$F = \{ f : \mathbb{R}^d \to \mathbb{R} \; ; \; f(x) = \langle x, \theta \rangle \} \cong \mathbb{R}^d$$

A: $n = d$ (we just need to solve a linear system of equations).

Remark: $f^* = \varphi(\langle x, \theta^* \rangle)$ φ even and smooth.

A: $n = d+1$ is sufficient for exact recovery.

$\varphi(t) = |t|$



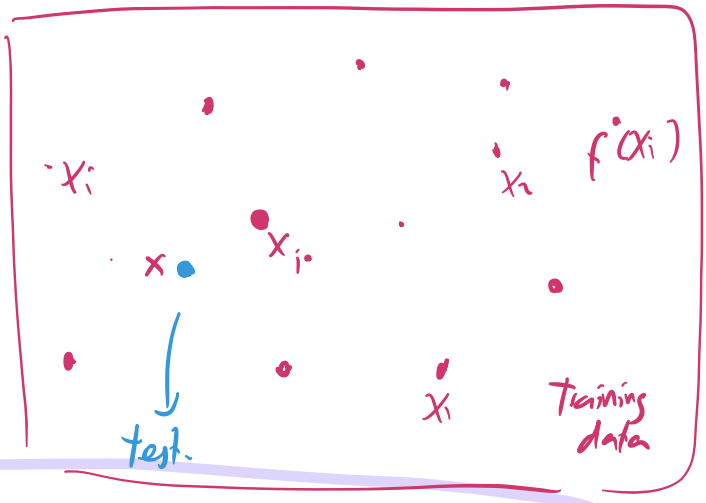
↳ Suppose now that f^* is only locally linear, ie f^* is β- Lipschitz: $|f(x) - f(x')| \leq \beta \cdot \|x - x'\|$

$F = \{ f : \mathbb{R}^d \to \mathbb{R} \; ; \; f$ is Lipschitz + Bounded$\}$

(now $\mathcal{F}$ is a Banach space; $\|f\| = Lip(f) + \|f\|_\infty$.

$$Lip(f) = \inf \{ \beta ; |f(x) - f(x')| \le \beta \|x - x'\| \}$$

→



$$\hat{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \{ Lip(f) ; f(x_i) = f^*(x_i) \ \forall i \}$$

Recall MSE: $\mathbb{E}_{x \sim \nu} |\hat{f}(x) - f^*(x)|^2$

$$|\hat{f}(x) - f^*(x)| \le |\hat{f}(x) - \hat{f}(x_{i^*})| + |\hat{f}(x_{i^*}) - f^*(x_{i^*})| + |f^*(x_{i^*}) - f^*(x)|$$

$$|f^*(x_{i^*}) - f^*(x)| \le \beta \|x_{i^*} - x\|$$

$$|\hat{f}(x_{i^*}) - \hat{f}(x)| \le Lip(\hat{f}) \|x_{i^*} - x\|$$

$$|\hat{f}(x) - f^*(x)| \le 2\beta \|x_{i^*} - x\|$$

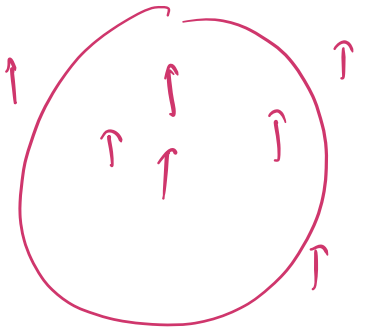
$$\Rightarrow \mathbb{E} |\hat{f}(x) - f^*(x)|^2 \le 4\beta^2 \mathbb{E}_x \|x - x_{i^*}\|^2$$

$x \sim N(0, I)$

$$W_2^2(\nu, \hat{\nu}_n) \simeq n^{-1/d}$$

$$\nu = N(0, I_d)$$

$$\hat{\nu}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i} \quad \text{empirical measure.}$$

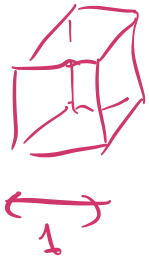


If we want $\|\hat{f} - f^*\|^2 \leq \varepsilon$

$$\varepsilon \simeq n^{-1/d} \rightarrow n \sim \varepsilon^{-d}$$

This sample complexity is cursed by dimension.

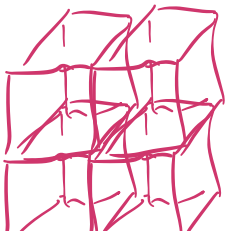
→ Is this necessary? Can we learn without exponential dependency on dimension?



$B = [-1/2, 1/2]^d$.

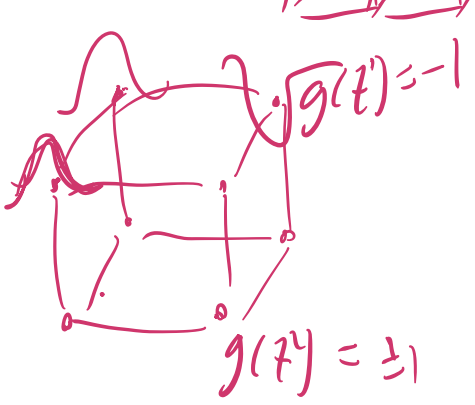
$\psi : B \rightarrow \mathbb{R}$ $\quad \psi(x) = \text{dist}(x, \partial B)$

Fact: ψ is 1-Lipschitz. and supported in B.



2^d copies of B.

$z = z_1 \dots z_d \quad z_i = \pm 1$

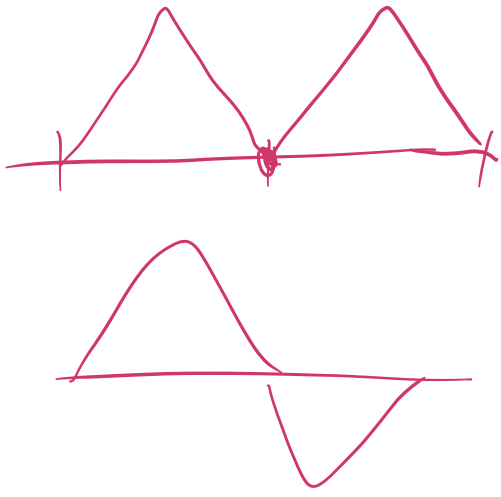


$z \in \{-1, +1\}^d$

$$f^*(x) = \sum_{z \in \{\pm 1\}^d} g(z)\, \varphi(x - z)$$

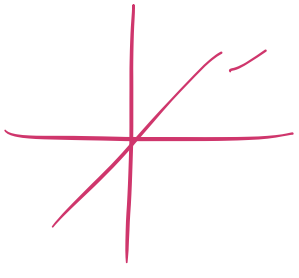
Claim: f^* is also 1- Lipschitz

f_1 f_2



→ If $n << 2^d$ then any estimator will incur in a relative error

$$\mathbb{E}_x \left| \hat{f}(x) - f^*(x) \right|^2 = \Theta(1) \cdot \mathbb{E} |f^*(x)|^2$$

→  most quadrants will be empty of data.

→ To summarize: linear functions (global reg) → $n = d$ (OK)

Lipschitz functions (local reg) → $n \sim \varepsilon^{-d}$

↳ Too hard to learn!

Next lectures → can shallow NNs define "interesting" hypothesis spaces somewhere in between these two extremes?