Part 2 Generalization in Deep Learning

What is Generalization ?

- $\log(1+2+3) = \log 1 + \log 2 + \log 3$
- $\log(1+1.5+5) = \log 1 + \log 1.5 + \log 5$
- $\log(2+2) = \log 2 + \log 2$
- $\log(1+1.25+9) = \log 1 + \log 1.25 + \log 9$

$\boxed{\log(a+b) = \log a + \log b}$???

SLT (Statistical Learning Theory)

prediction $\quad S = \{(x_1, y_1), \cdots, (x_n, y_n)\} \rightarrow \boxed{f: \mathcal{X} \to \mathcal{Y}}$

↳ given observation $x \in \mathcal{X}$, predict the label $y \in \mathcal{Y}$ via $f(x) = \hat{y}$

"good prediction" → loss function, $\ell: \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$

$\ell(\hat{y}, y)$: cost (loss) $\qquad$ Aim: $\ell(f(x), y)$ small

Example. Classification $\mathcal{Y} = \{1, \cdots, k\}$

$$\ell(\hat{y}, y) = \mathbb{1}\{\hat{y} \neq y\} = \begin{cases} 1, & \text{if } \hat{y} \neq y \\ 0, & \text{if } \hat{y} = y \end{cases}$$

Example. Regression $\mathcal{Y} = \mathbb{R}$

$$\ell(\hat{y}, y) = (\hat{y} - y)^2$$

Probabistic Assumption : i.i.d $\quad (x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ $\quad$ test example (x, y) $\xrightarrow{\text{i.i.d}}$ P (over $\mathcal{X} \times \mathcal{Y}$) (unknown)

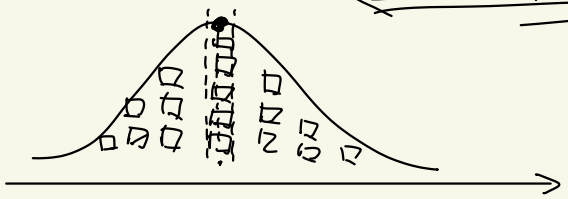
in practice

↳ risk (population loss) $\quad L(f) = \mathbb{E}_{P}[\ell(f(x), y)]$ ↓

e.g. $L(f) = L_{01}(f) = \mathbb{E}[\mathbb{1}\{f(x) \neq y\}] = \Pr(f(x) \neq y)$

In practice.. distribution shift (covariate shift) → domain adaptation $\quad P, Q$ discrepency

Why not to estimate P (or $P_{Y|X}$)



Statistic $\langle$ Induction / Deduction

Goal: unknown quantities related to unknown P.

$\downarrow$

$\checkmark$ empirical data $\rightarrow$ empirical loss: $\boxed{\hat{L}(f) = \frac{1}{n}\sum_{i=1}^{n} l(y_i, f(x_i))}$

Generalization Theory

$x \mapsto \hat{f}_n(x) = \hat{f}_n(x; S)$

random

by LLN, asymptotically $\hat{L}(f) \xrightarrow{P} L(f)$ $(n \to \infty)$

ERM

$(L(f) = \mathbb{E}[\hat{L}(f)])$

$L(\hat{f}_n) = \mathbb{E}_{(x,y)\sim P}[\, l(\hat{f}_n(x), y) \mid S \,]$ random

$\star$ ① $\boxed{\mathbb{E}[L(\hat{f}_n)] \text{ small}}$

② $L(\hat{f}_n)$ small w.h.p

concentration inequality

over training data

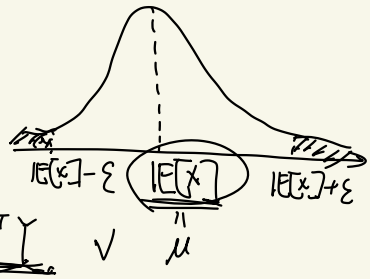
e.g. $\hat{L}(f)$ given fixed f.

$\hat{f}_n = \arg\min_f \hat{L}(f)$ (ERM) $\rightarrow$ random

$\hat{L}(\hat{f}_n)$

$L(\hat{f}_n)$

$L(f)$ given fixed f $\rightarrow$ fixed!



$\hat{f}_n$

- closed-form: $\hat{f}_n(x) = X^T(X^TX)^{-1}X^TY$ $\checkmark$
- ERM: solution to an optimization objection
- Iterative procedure

The Gold Standard:

$f^* = \arg\min_f L(f)$ $\Rightarrow$ $L(f^*) = \inf_f L(f)$ Bayes error

over all measurable prediction rule

Bayes optimal

$\hat{f}_n \longleftrightarrow f^*$

posterior mean

Example. Classification (binary) $\mathcal{Y} = \{0,1\}$

$\eta(x) = \mathbb{E}[Y \mid X = x]$

$f^*(x) = \mathbb{1}\{\eta(x) \geq \frac{1}{2}\}$

Regression $\mathcal{Y} = \mathbb{R}$.

$f^*(x) = \eta(x) = \mathbb{E}[Y \mid X = x]$

$L(f) = \mathbb{E}[(y - f(x))^2] = \mathbb{E}[(y - \eta(x) + \eta(x) - f(x))^2] = \mathbb{E}[(y-\eta(x))^2] + \mathbb{E}[(f(x)-\eta(x))^2]$

$f(x) = \eta(x)$

The big question: is there a way construct a learning algorithm s.t. $\boxed{L(\hat{f}_n) - L(f^*)}$ is small for large enough data?

If $n \to \infty$, $\lim_{n\to\infty} L(\hat{f}_n) = L(f^*)$ a.s (consistency)

$\mathcal{X}$: finite / countable set $\checkmark$

$\mathcal{X}$: infinite. x & y continuous relationship $\checkmark$

$L(\hat{f}_n) - L(f^*)$ NFL (No Free Lunch): without any further assumptions,

$\forall \hat{f}_n$ n. $\varepsilon > 0$. $\exists$ "bad distribution" P s.t. $L(f^*) = 0$.

learning algorithm

$(a_n \to 0 \quad \forall n.\ \mathbb{E}[L(\hat{f}_n)] \geq a_n)$ $\qquad \mathbb{E}[L(\hat{f}_n)] \geq \frac{1}{2} - \varepsilon$.

lower bound

We always have some domain knowledges (prior)

① assumption on distribution P (e.g. margin, smoothness of regression function)

② redefine the goal

in reference set $\mathcal{F}$ of predictors: $\boxed{L(\hat{f}_n) - \inf_{f \in \mathcal{F}} L(f)}$

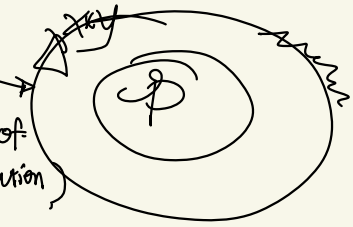
(hypothesis space)

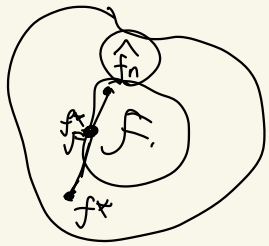
(e.g. linear models, 2-layer NNs, Lipschitz functions, strongly-convex....)

inductive bias

excess risk

"model-specific"





Perceptron → linear classifier

T. n Online Learning.

$(x_1, y_1), \cdots, (x_T, y_T) \in \mathcal{X} \times \{\pm 1\}$.

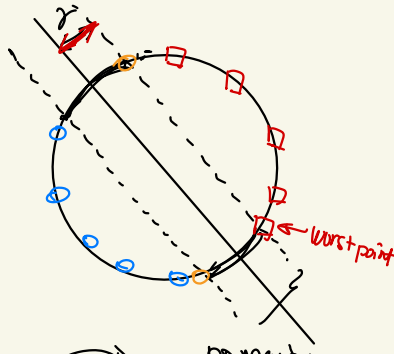
Maintain a hypothesis $w_t \in \mathbb{R}^d$ (initially, $w_1 = 0$)

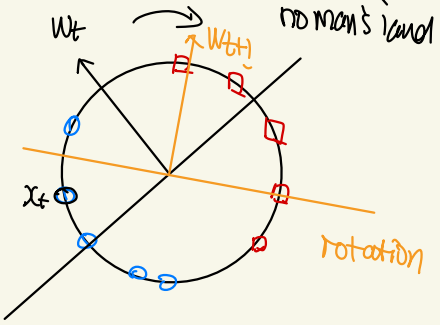
On round t:

- Consider (x_t, y_t)
- Form prediction $\hat{y}_t = \text{sign}(\langle w_t, x_t \rangle)$
- If $\hat{y}_t \neq y_t$. update $w_{t+1} = w_t + y_t x_t$. → wrong

  Else $w_{t+1} = w_t$ → correct

Suppose all data in unit ball $\|x_t\| \leq 1$.





Realizable margin $\gamma = \max_{\|w\|=1} \min_{i \in [T]} \boxed{(y_i \langle w, x_i \rangle)}$ → measures the difficulty of classification on fixed dataset.

best direction

worst point.

Thm (Novikoff '62) Perceptron makes at most $\frac{1}{\gamma^2}$ mistakes (and corrections) on any sequence of examples with margin γ.

↳ deterministic.

non-realizable: $\gamma = 0$

Proof: T iterations # of mistakes m

At round t, a mistake happens

$$\|w_{t+1}\|^2 = \|w_t + y_t x_t\|^2 \le \|w_t\|^2 + \underbrace{2\, y_t \cdot \langle w_t, x_t\rangle}_{\le 0} + 1 \le \|w_t\|^2 + 1,$$

true label, prediction

$$\Rightarrow \|w_T\|^2 \le m.$$

For optimal hyperplane w^*: $\gamma \le \langle w^*, y_t x_t\rangle = \langle w^*, w_{t+1} - w_t\rangle$

$$\Rightarrow m\cdot\gamma \le \langle w^*, w_T\rangle \le \underbrace{\|w^*\|}_{1}\|w_T\| \le \sqrt{m} \Rightarrow m \le \frac{1}{\gamma^2} \quad \square$$

$$\Rightarrow \text{\# of mistakes} = \sum_{t=1}^{T} \mathbb{1}\{y_t\langle w, x_t\rangle \le 0\} \le \frac{D^2}{\gamma^2} \qquad (*)$$

$\gamma = \gamma(x_{1:T}, y_{1:T})$ margin

$D = D(x_{1:T}, y_{1:T}) = \max_t \|x_t\|$

- Do one pass on i.i.d sequence $(X_1, Y_1), \dots, (X_n, Y_n)$ (i.e. $T = n$)

Claim: Expected indicator loss of randomly picked function $x \mapsto \langle w_\tau, x\rangle$, $\tau \sim \text{Uniform}(1, \dots, n)$

X, Y is at most $\frac{1}{n}\cdot \mathbb{E}\left[\frac{D^2}{\gamma^2}\right]$

Proof: $(*)$ taking expectation, $\mathbb{E}_{S,\tau}\left[\frac{1}{n}\sum_{t=1}^{n} \mathbb{1}\{y_t\langle w_t, x_t\rangle \le 0\}\right] \le \mathbb{E}\left[\frac{1}{n}\cdot\frac{D^2}{\gamma^2}\right]$

$$\text{LHS} = \mathbb{E}_\tau\left[\mathbb{E}_S\left[\mathbb{1}\{y_\tau\langle w_\tau, x_\tau\rangle \le 0\}\right]\right]$$

$$= \mathbb{E}_S\left[\mathbb{E}_\tau\left[\mathbb{E}_{X,Y}\left[\mathbb{1}\{Y\cdot\langle w_t, X\rangle \le 0\}\right]\right]\right] = \mathbb{E}_S\left[\mathbb{E}_\tau\left[L_{01}(w_\tau)\right]\right] \quad \square$$

- Do multiple passes: cycle through data $(X_1, Y_1), \dots, (X_n, Y_n) \overset{i.i.d}{\sim} P$ ⟨ parametric / nonparametric

repeatly until no more mistakes $\left(T \le n \times \frac{D^2}{\gamma^2}\right)$

$\Rightarrow$ final hyperplane of perceptron seperates the data perfectly $\boxed{\hat{L}_{01}(w_T) = 0}$

$$\hat{L}_{01}(w) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{Y_i\langle w, X_i\rangle \le 0\}$$

interpolation

ERM with finite-time convergence.

Claim: $\boxed{\mathbb{E}\left[L_{01}(w_T)\right] \le \frac{1}{n+1}\cdot\mathbb{E}\left[\frac{D^2}{\gamma^2}\right]}$

random (wrt. $(X_1, Y_1), \dots, (X_n, Y_n)$)

Proof: $Z = (X, Y)$ $\ell(w, Z) = \mathbb{1}\{Y\cdot\langle w, X\rangle \le 0\}$

$$\mathbb{E}_S\left[\mathbb{E}_Z\left[\ell(w_T, Z)\right]\right] = \mathbb{E}_{S, Z_{n+1}}\left[\frac{1}{n+1}\sum_{t=1}^{n+1} \ell(w^{(-t)}, Z_t)\right]$$

* Bayes Error: $L_{01}(f^*) = 0$

Leave-One-Out is unbiased estimate of expected loss.

$W^{(-t)}$ Perceptron's final hyperplane after cycling through data $Z_1, \cdots, Z_{t-1}, Z_{t+1}, \cdots, Z_{n+1}$

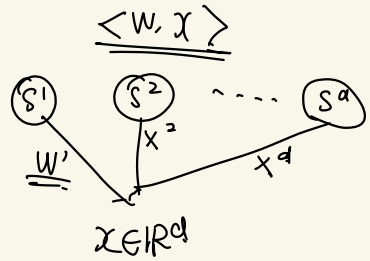
Cycling on $Z_1, \cdots, Z_{n+1}$. Let $i_1, \cdots, i_m$ be indices that Perceptron errs.

$m \le \frac{D^2}{\gamma^2}$

$t \notin \{i_1, \cdots, i_m\}$: correct. $\rightarrow$ Z_t doesn't matter.

$t \in \{i_1, \cdots, i_m\}$: $\ell(W^{(-t)}, Z_t) = 1$ (at most $m \le \frac{D^2}{\gamma^2}$ times) $\square$

* Perceptron : 1-layer NN (no hidden layer) with d neurons



dimension d never appears in the mistake bound! $d = +\infty$

$\boxed{d > n : \text{overparameterization}}$:

any n points can be separated (zero empirical error)

interpolation

$\Rightarrow$ More parameters than data does NOT generally contradicts good performance.

Perfectly fitting the data does NOT generally contradicts good generalization. benign overfitting

$\mathbb{E}[L_{01}(W_T)] - L_{01}(f^*) \le O(\frac{1}{n})$ under assumption: $\mathcal{P}$ linearly separable with margin γ.

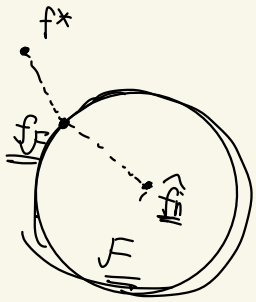
$\Rightarrow$ $\mathbb{E}[L_{01}(W_T)] - L_{01}(w^*)$ with $L_{01}(w^*) - L_{01}(f^*)$ small

linear all

Class of functions $\mathcal{F}$: captures the relationship between X & Y

$$L(\hat{f}_n) - L(f^*) = \underbrace{L(\hat{f}_n) - \inf_{f \in \mathcal{F}} L(f)}_{\text{Estimation error}} + \underbrace{\inf_{f \in \mathcal{F}} L(f) - L(f^*)}_{\text{Approximation error}}$$

(variance) (bias)



$\mathcal{F}$ large : small approximation but large estimation error

n large : approximation $\sim$ estimation error $\downarrow$

$\rightarrow$ trade-off

SRM (Structural Risk Min.)

Model Selection

Method of Sieves

Bias-Variance decomposition

$\mathcal{P}$ $<$ complexity $\uparrow$ (class containing f^* is too large)

estimation error $\uparrow$

$\hookrightarrow$ "biased" solution (e.g. min-norm solution) in hopes of reducing variance.

* bias-variance tradeoff need NOT to be the "approx.-est." form.

Uniform Laws of Large Numbers

$\mathcal{F}$.

ERM: $\hat{f}_n = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \hat{L}(f)$, $\hat{L}(f) = P_n\, \ell(f(X), Y) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(X_i), Y_i)$

expectation of empirical distribution

$f_{\mathcal{F}} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} L(f)$

excess risk: $L(\hat{f}_n) - L(f_{\mathcal{F}})$

$$= \left(L(\hat{f}_n) - \hat{L}(\hat{f}_n)\right) + \left(\hat{L}(\hat{f}_n) - \hat{L}(f_{\mathcal{F}})\right) + \left(\hat{L}(f_{\mathcal{F}}) - L(f_{\mathcal{F}})\right)$$

≤ 0 expectation 0

$$\mathbb{E}[L(\hat{f}_n)] - L(f_{\mathcal{F}}) \leq \mathbb{E}\left[L(\hat{f}_n) - \hat{L}(\hat{f}_n)\right] \leq \mathbb{E}\left[\sup_{f \in \mathcal{F}} |L(f) - \hat{L}(f)|\right].$$

$$L(\hat{f}_n) - L(f_{\mathcal{F}}) \leq 2 \cdot \boxed{\sup_{f \in \mathcal{F}} |L(f) - \hat{L}(f)|}$$

Uniform Convergence.

excess risk (generalization risk)

generalization gap (optimism)

sup → when learning is possible. sample complexity

$\mathcal{F}$ → $\mathcal{G} = \ell \circ \mathcal{F} = \{(x,y) \mapsto \ell(f(x), y) : f \in \mathcal{F}, (x,y) \in \mathcal{Z}\}$, $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$

function class

loss class

$g \in \mathcal{G} : \mathcal{Z} \to [0,1]$

$Z_1, \dots, Z_n, Z \overset{i.i.d}{\sim} P$

$$U = \sup_{g \in \mathcal{G}} \left| \mathbb{E}[g(Z)] - \frac{1}{n}\sum_{i=1}^{n} g(Z_i) \right|$$

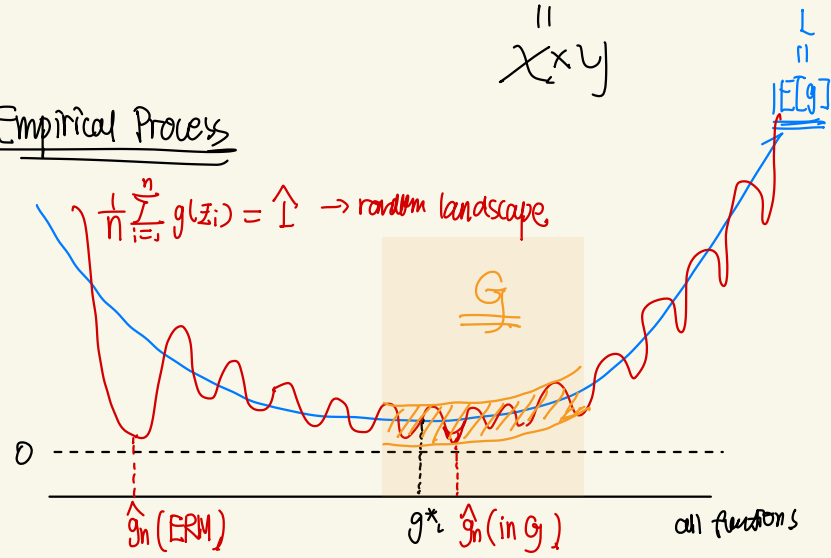
$$= \sup_{g \in \mathcal{G}} |Pg - P_n g|$$

$$= \|P - P_n\|_{\mathcal{G}} \longrightarrow 0: \text{ uniform law of large numbers}$$

empirical process

randomness: $Z_1, \dots, Z_n$.

Empirical Process



def (Glivenko-Cantelli Class) $\mathcal{G}$ is a GC-class for P if $\|P_n - P\|_{\mathcal{G}} \xrightarrow{P} 0$.

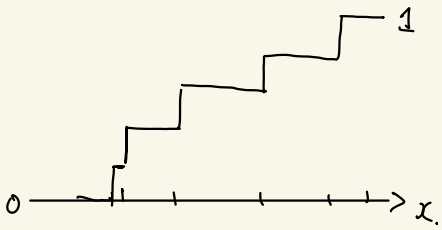
$$\left(\Pr\left(\sup_{g \in \mathcal{G}} |P_n g - Pg| > \varepsilon\right) \to 0\right)$$

$$\|P_n - P\|_{\mathcal{G}} = \sup_{g \in \mathcal{G}} |P_n g - Pg|$$

Thm (Glivenko-Cantelli) $\|F_n - F\|_\infty \xrightarrow{a.s.} 0$

CDF $\quad F_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{Z_i \le t\}$

$Z_1, \dots, Z_n \overset{i.i.d}{\sim} F$



$\|F - G\|_\infty = \sup_t |F(t) - G(t)|$

$\Downarrow$

$\|P_n - P\|_{\mathcal{G}} \xrightarrow{as} 0.$

$\mathcal{G} = \{ Z \mapsto \mathbb{1}\{Z \le \theta\} : \theta \in \mathbb{R} \}$

step function (one-layer NN with one input).

*Not all $\mathcal{G}$ are GC classes

e.g. $\mathcal{G} = \{ \mathbb{1}\{Z \in S\} : S \subseteq \mathbb{R}, |S| < \infty \}$

$Pg = 0, \ \forall g \in \mathcal{G}$

$\sup_{g \in \mathcal{G}} P_n g = 1 \ \forall n$

$P_n g \xrightarrow{as} Pg \ \forall g \in \mathcal{G}$

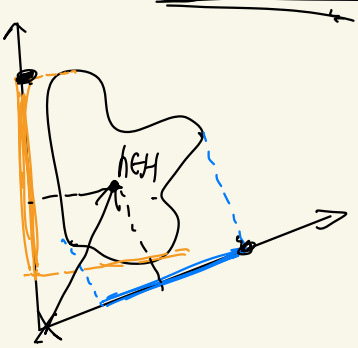
convergence is NOT uniform over $\mathcal{G}$

$\Rightarrow$ $\mathcal{G}$ too large (rich)

- How to decide whether "$\mathcal{G}$ is too large"?

$H \subseteq [-1,1]^n$:

average length of projection onto random direction



Rademacher Process: $\hat{R}_n(h) = \frac{1}{n}\langle \varepsilon, h \rangle$

$\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)$ random direction

$\varepsilon_i \sim \begin{cases} +1, & \text{wp } \frac{1}{2} \\ -1 & \text{wp } \frac{1}{2} \end{cases}$

Rademacher Averages: expected maximum of the process

$\mathbb{E}_\varepsilon \|\hat{R}_n\|_H = \mathbb{E}_\varepsilon \left[\sup_{h \in H} \left| \frac{1}{n} \langle \varepsilon, h \rangle \right| \right]$

Examples. $H = \{h_0\}$ $\quad \mathbb{E}\|\hat{R}_n\|_H = O(n^{-1/2})$

$H = \{-1,1\}^n$ $\quad \mathbb{E}\|\hat{R}_n\|_H = 1.$

$\varepsilon_1^2 + \dots + \varepsilon_n^2 = n.$

$|\varepsilon_1 h_1 + \dots + \varepsilon_n h_n|$

$= |h_0| \, |\varepsilon_1 + \dots + \varepsilon_n| = O(\sqrt{n})$

$\left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i h_i \right| = 1.$

$V + C \cdot \{-1,1\}^n \subseteq H$ $\quad \mathbb{E}\|\hat{R}_n\|_H \ge C - o(1)$

$H = \{-1, 1\}$ $\quad \mathbb{E}\|\hat{R}_n\|_H = O(n^{-1/2})$

H finite: $\mathbb{E}\|\hat{R}_n\|_H \le C \cdot \sqrt{\frac{\log|H|}{n}}$ $\quad$ Massart's Lemma. $\rightarrow$ can be loose!

(since "overlapping" between vectors)

$B_p^n = \{ x \in \mathbb{R}^n : \|x\|_p \le 1 \}.$

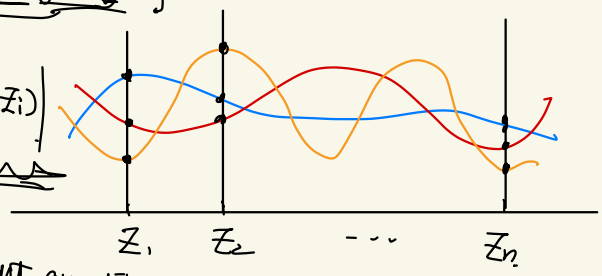
$H = B_2^p \Rightarrow \frac{1}{n} \mathbb{E}\left[\max_{\|g\|_2 \le 1} |\langle \varepsilon, g \rangle| \right] = \frac{1}{n} \mathbb{E}[\|\varepsilon\|_2] = \frac{1}{\sqrt{n}}$

$H = B_1^p \Rightarrow$ Rademacher Averages $\frac{1}{n}$

H: output space of G

$$H = G(Z_1^n) = \{ (g(Z_1), \cdots, g(Z_n)) : g \in G \} \subseteq \mathbb{R}^n$$

Conditionally on Z_1^n: $\mathbb{E}_{\varepsilon} \|\hat{R}_n\|_H = \mathbb{E}_{\varepsilon} \sup_{g \in G} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i g(Z_i) \right|$



data-dependent quantity

Rademacher Complexity: $\mathbb{E}_S \|R_n\|_G = \mathbb{E}_{Z_1^n} \mathbb{E}_{\varepsilon} \|\hat{R}_n\|_{G(Z_1^n)}$

set of functions

subset of $\mathbb{R}^n$

$$\mathbb{E}_{Z_1^n}\left[\mathbb{E}_{\varepsilon}\left[\sup_{g \in G} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i g(Z_i) \right| \right] \right]$$

$$R_n(g) = \frac{1}{n} \sum_{i=1}^n \varepsilon_i g(Z_i)$$

① Symmetrization: bound $\mathbb{E}\|P - P_n\|_G$ via $\mathbb{E}\|R_n\|_G$ ← Rademacher Complexity $\mathcal{R}_n(G)$

② $\|P - P_n\|_G$ and $\|R_n\|_G$ are concentrates about their expectation

Empirical RC

Thm $\forall G$: $\mathbb{E}\|P - P_n\|_G \le 2 \cdot \mathbb{E}\|R_n\|_G$

$G \subseteq [0,1]^Z$: $\frac{1}{2}\mathbb{E}\|R_n\|_G - \sqrt{\frac{\log 2}{2n}} \le \mathbb{E}\|P - P_n\|_G \le 2\,\mathbb{E}\|R_n\|_G$

② $\Rightarrow$ W.P. $\ge 1 - 2\exp(-2\varepsilon^2 n)$: $\mathbb{E}\|P - P_n\|_G - \varepsilon \le \|P - P_n\|_G \le \mathbb{E}\|P - P_n\|_G + \varepsilon$

$\Rightarrow$ $\mathbb{E}\|R_n\|_G \to 0$ iff $\|P - P_n\|_G \xrightarrow{a.s.} 0$

(sub-Gaussianity)

Proof: Symmetrization $Pg \to P_n' g$

$\mathbb{E}_{Z}[g(Z)]$ $\quad$ $\mathbb{E}_{Z_1', \cdots, Z_n'}\left[\frac{1}{n}\sum_{i=1}^n g(Z_i')\right]$

$$\mathbb{E}\|P - P_n\|_G = \mathbb{E}_{Z_1^n}\left[\sup_{g \in G} \left| \mathbb{E}_{Z_1^{n\prime}}\left[\frac{1}{n} \sum_{i=1}^n (g(Z_i') - g(Z_i)) \right] \right| \,\middle|\, Z_1^n \right]$$

$$\le \mathbb{E}_{Z_1^n} \mathbb{E}_{Z_1^{n\prime}}\left[\sup_{g \in G} \left| \frac{1}{n} \sum_{i=1}^n (g(Z_i') - g(Z_i)) \right| \,\middle|\, Z_1^n \right]$$

$\sup A + \sup B \ge \sup(A+B)$

$$= \mathbb{E}_{Z,Z'}\left[\sup_{g\in G}\left|\frac{1}{n}\sum_{i=1}^{n}(g(Z_i')-g(Z_i))\right|\right] = \mathbb{E}\,\|P_n'-P_n\|_G$$

$\varepsilon_i \in \{\pm 1\}$

$$= \mathbb{E}_{Z,Z'}\left[\sup_{g\in G}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\,(g(Z_i')-g(Z_i))\right|\right] \quad \star$$

$Z_i, Z_i' \overset{iid}{\sim} P$

$$\leq \mathbb{E}_{Z_i'}\left[\sup_{g\in G}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i g(Z_i')\right|\right] + \mathbb{E}_{Z_i}\left[\sup_{g\in G}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i g(Z_i)\right|\right]$$

$$= 2\,\underbrace{\mathbb{E}\left[\sup_{g\in G}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i g(Z_i)\right|\right]}_{\text{Rademacher Complexity}} = 2\,\mathbb{E}\,\|R_n\|_G$$

$$\mathbb{E}\,\|R_n\|_G = \mathbb{E}\left[\sup_{g\in G}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\,(g(Z_i)-\mathbb{E}[g(Z_i)]+\mathbb{E}[g(Z_i)])\right|\right]$$

$$\leq \mathbb{E}\left[\sup_{g\in G}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(g(Z_i)-\mathbb{E}[g(Z_i)])\right|\right] + \mathbb{E}\left[\sup_{g\in G}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\,\mathbb{E}[g(Z_i)]\right|\right]$$

$$\leq \mathbb{E}\left[\sup_{g\in G}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(g(Z_i)-g(Z_i'))\right|\right] + \|P\|_G\cdot\mathbb{E}_{\varepsilon_i}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\right|\right]$$

$$= \mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(g(Z_i)-g(Z_i'))\right\|_G + \|P\|_G\cdot\mathbb{E}_{\varepsilon}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\right|\right]$$

$$= \mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}(g(Z_i)-\mathbb{E}[g(Z_i)]+\mathbb{E}[g(Z_i')]-g(Z_i'))\right\|_G + \|P\|_G\cdot\mathbb{E}_{\varepsilon}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\right|\right]$$

Massart's Lemma

$$\leq 2\,\mathbb{E}\,\|P_n-P\|_G + 1\cdot\sqrt{\frac{2\log 2}{n}}$$

$$\Rightarrow G\subseteq[0,1]^n:\ \frac{1}{2}\mathbb{E}\,\|R_n\|_G - \sqrt{\frac{\log 2}{2n}} \leq \mathbb{E}\,\|P_n-P\|_G \leq 2\,\mathbb{E}\,\|R_n\|_G \quad \square$$

Rademacher Complexity of Neural Networks

Rademacher Complexity $\mathbb{E}\,\|R_n\|_G = \mathbb{E}_{Z_1^n}\left[\mathbb{E}_{\varepsilon}\left[\sup_{g\in G}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i g(Z_i)\right|\right]\right]$

$$\frac{1}{2}\mathbb{E}\,\|R_n\|_G - \sqrt{\frac{\log 2}{2n}} \leq \mathbb{E}\,\|P-P_n\|_G \leq 2\,\mathbb{E}\,\|R_n\|_G \qquad G\subseteq[0,1]^n$$

prediction problem: $G = \ell\circ\mathcal{F}$ $\begin{cases}\ell: \text{loss function}\\ \mathcal{F}: \text{hypothesis class}\end{cases}$

$$\mathbb{E}\,\|P-P_n\|_{\ell\circ\mathcal{F}} \longleftrightarrow \mathbb{E}\,\|R_n\|_{\ell\circ\mathcal{F}} \qquad \mathbb{E}\,\|R_n\|_{\mathcal{F}}\ ?$$

Classification: $\mathcal{F}\subseteq\{\pm 1\}^{\mathcal{X}}$, $Y\in\{\pm 1\}$ $\quad \ell(y,y') = \mathbb{1}\{y\neq y'\} = \frac{1-yy'}{2}$ ✓

$$\mathbb{E}\|R_n\|_{\ell\circ\mathcal{F}} = \mathbb{E}\Big[\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \ell(y_i, f(x_i))\Big|\Big]$$

$$= \frac{1}{2}\mathbb{E}\Big[\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^n \varepsilon_i + \frac{1}{n}\sum_{i=1}^n -\varepsilon_i y_i f(x_i)\Big|\Big]$$

$$\le O\Big(\frac{1}{\sqrt{n}}\Big) + \frac{1}{2}\mathbb{E}\|R_n\|_{\mathcal{F}}.$$

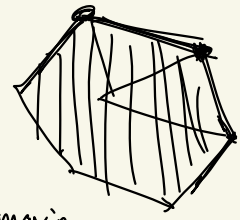
Other loss functions ?

$$\|R_n\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(z_i)\Big|$$

- Subset: $\mathcal{F}\subseteq\mathcal{G} \Rightarrow \|R_n\|_{\mathcal{F}} \le \|R_n\|_{\mathcal{G}}$
- Scaling: $\|R_n\|_{c\cdot\mathcal{F}} = |c|\cdot\|R_n\|_{\mathcal{F}}$
- Union: $\|R_n\|_{\mathcal{F}\cup\mathcal{G}} \le \|R_n\|_{\mathcal{F}} + \|R_n\|_{\mathcal{G}}$
- Plus Constant: $|g(x)|\le 1$. $\big|\mathbb{E}\|R_n\|_{\mathcal{F}+\mathcal{G}} - \mathbb{E}\|R_n\|_{\mathcal{F}}\big| \le \sqrt{\frac{2\log 2}{n}}$
- Convex-Hull: $\|R_n\|_{\mathrm{co}\,\mathcal{F}} = \|R_n\|_{\mathcal{F}}$
- Contraction: $\ell: \mathbb{R}\times\mathbb{R}\to[-1,1]$. $\hat{y}\mapsto \ell(\hat{y}, y)$ 1-Lipschitz $\forall y$

$$\ell\circ\mathcal{F} = \{(x,y)\mapsto \ell(f(x), y)\}$$

$$\Rightarrow \mathbb{E}\|R_n\|_{\ell\circ\mathcal{F}} \le 2\,\mathbb{E}\|R_n\|_{\mathcal{F}} + \frac{c}{\sqrt{n}}$$ (Talagrand's Contraction Lemma)



maximum is being achieved at corners

Deep Networks: deep compositions of non-linear functions

$$h = h_m \circ h_{m-1} \circ \cdots \circ h_1.$$

e.g. Sigmoid Network $h_i: x\mapsto \sigma(W_i x)$ $\sigma(v)_i = \frac{1}{1+\exp(-v_i)}$

ReLU Network: $h_i: x\mapsto r(W_i x)$ $r(v)_i = \max(0, v_i)$

① Rademacher Complexity for Sigmoid Networks

2-layer NN on $\mathbb{R}^d$: $\mathcal{F}_{B,d} = \Big\{x\mapsto \sum_{i=1}^k w_i\,\sigma(v_i^T x) : \|w\|_1\le B,\ \|v_i\|_1\le B,\ k\ge 1\Big\}$

$\|x\|_\infty \le 1$ a.s.

σ: 1-Lipschitz: $|\sigma(a)-\sigma(b)|\le|a-b|$, $\sigma(0)=0$

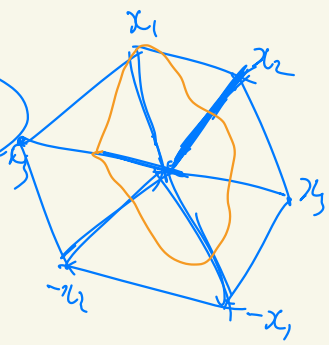
Thm $\mathbb{E}\|R_n\|_{\mathcal{F}_{B,d}} \le 2B^2\sqrt{\frac{2\log 2d}{n}}$.

Proof: $\mathcal{G} = \{(x_1,\cdots,x_d)\mapsto x_j : 1\le j\le d\}$

$\mathcal{V}_B = \{x\mapsto v^T x : \|v\|_1 = \sum_{i=1}^d |v_i| \le B\} = B\cdot\mathrm{co}(\mathcal{G}\cup -\mathcal{G})$

$\mathrm{co}(\mathcal{F}) = \{\sum_{i=1}^k \alpha_i f_i : k\ge 1,\ \alpha_i\ge 0,\ \|\alpha\|_1 = 1,\ f_i\in\mathcal{F}\}$

$\mathcal{F}_{B,d} = \Big\{x\mapsto \sum_{i=1}^k w_i\,\sigma(v_i(x)) : k\ge 1,\ \sum_{i=1}^k |w_i|\le B,\ v_i\in\mathcal{V}_B\Big\}$



$$= B\cdot \mathrm{co}(\sigma\circ \mathcal{V}_B \cup -\sigma\circ \mathcal{V}_B)$$

$$\begin{aligned}
\mathbb{E}\|R_n\|_{\mathcal{F}_{B,d}} &= \mathbb{E}\|R_n\|_{B\cdot \mathrm{co}(\sigma\circ \mathcal{V}_B \cup -\sigma\circ \mathcal{V}_B)} \\
&= B\cdot \mathbb{E}\|R_n\|_{\mathrm{co}(\sigma\circ \mathcal{V}_B \cup -\sigma\circ \mathcal{V}_B)} \\
&= B\cdot \mathbb{E}\|R_n\|_{\sigma\circ \mathcal{V}_B \cup -\sigma\circ \mathcal{V}_B} \\
&\le B\cdot \mathbb{E}\|R_n\|_{\sigma\circ \mathcal{V}_B} + B\cdot \mathbb{E}\|R_n\|_{-\sigma\circ \mathcal{V}_B} \\
&= 2B\cdot \mathbb{E}\|R_n\|_{\sigma\circ \mathcal{V}_B} \\
&\le 2B\cdot \mathbb{E}\|R_n\|_{\mathcal{V}_B} \qquad \text{contraction} \\
&= 2B\,\mathbb{E}\|R_n\|_{B\,\mathrm{co}(G\cup -G)} \\
&= 2B^2\,\mathbb{E}\|R_n\|_{G\cup -G} \qquad |G\cup -G| = 2d \\
&\le 2B^2\cdot\sqrt{\frac{2\log 2d}{n}} \qquad \text{Massart's Lemma}\left(f\in\mathcal{F},\ |f(x)|\le 1;\ \mathbb{E}\|R_n\|_{\mathcal{F}} \le \mathbb{E}\cdot\sqrt{\frac{2\log(2|\mathcal{F}(x_1^n)|)}{n}}\right) \quad \square
\end{aligned}$$

② Rademacher Complexity of ReLU Networks

$\sigma(x) = \max(0,x)$ $\Rightarrow$ positive homogeneity property: $\forall \alpha\ge 0,\ x\in\mathbb{R}$ $\sigma(\alpha x) = \alpha\cdot\sigma(x)$

Thm $(x_1,y_1),\cdots,(x_n,y_n)\in \mathcal{X}\times\{\pm 1\}$, $\|x_i\|\le 1$ $\Rightarrow$ $\mathbb{E}\|R_n\|_{\mathcal{F}_{F,B}} \le \frac{(2B)^L}{\sqrt{n}}$

where $f\in\mathcal{F}_{F,B}$ is an L-layer network $\mathcal{F}_{F,B} = W_L\,\sigma(W_{L-1}\cdots\sigma(W_1 x)\cdots)$

$\|W_i\|_F \le B$

1-Lipschitz · positive homogeneous

Lemma $$\mathbb{E}_\varepsilon\left[\sup_{\substack{f\in\mathcal{F}\\ \|W\|_F\le B}} \frac{1}{n}\left\|\sum_{i=1}^n \varepsilon_i\,\sigma(Wf(x_i))\right\|_2\right] \le 2B\cdot\mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\left\|\sum_{i=1}^n\varepsilon_i f(x_i)\right\|_2\right]$$

$\|W\|_F\le B$

1) Iterating this Lemma $\to (2B)^2$

2) $\frac{1}{n}\left\|\sum_{i=1}^n \varepsilon_i x_i\right\|_2 \le \frac{1}{\sqrt{n}}$, $\left(\|\varepsilon\|_2 = \sqrt{n},\ \|x_i\|_2\le 1\right)$

Proof of Lemma. $W = \begin{pmatrix} -w_1^T- \\ \vdots \\ -w_k^T- \end{pmatrix}$,

$$\left\|\sum_{i=1}^n \varepsilon_i\,\sigma(Wf(x_i))\right\|^2 = \sum_{j=1}^k\left(\sum_{i=1}^n \varepsilon_i\,\sigma(w_j^T f(x_i))\right)^2$$

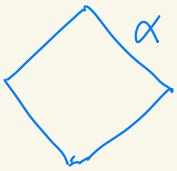
$$= \sum_{j=1}^n \|w_j\|^2\left(\sum_{i=1}^n \varepsilon_i\,\sigma\left(\frac{w_j^T}{\|w_j\|}f(x_i)\right)\right)^2$$

(positive homogeneity) scaling

$$\sup_{\|W\|_F \le B} \sum_{j=1}^{k} \|w_j\|^2 \left(\sum_{i=1}^{n} \varepsilon_i \, \sigma\left(\frac{w_j^T}{\|w_j\|} f(x_i) \right) \right)^3$$

$$= \sup_{\|w_j\|=1} \boxed{\sum_{j=1}^{k} \alpha_j} \left(\sum_{i=1}^{n} \varepsilon_i \, \sigma\left(w_j^T f(x_i) \right) \right)^2$$

$\|\alpha\|_1 \le B^2$



$$= B^2 \cdot \sup_{\|w\|=1} \left(\sum_{i=1}^{n} \varepsilon_i \, \sigma(w^T f(x_i)) \right)^2 \le 4B^2 \cdot \sup_{\|w\|=1} \left(\sum_{i=1}^{n} \varepsilon_i \, w^T f(x_i) \right)^2 \le 4B^2 \sum_{i=1}^{n} \varepsilon_i \, f(x_i)^2 \quad \square$$

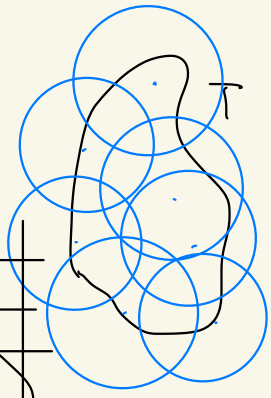
Bounding Rademacher Complexity

① Covering Number

def (covering number) ε-cover of subset T of ametric space (S, d) is a set $\hat{T} \subseteq T$

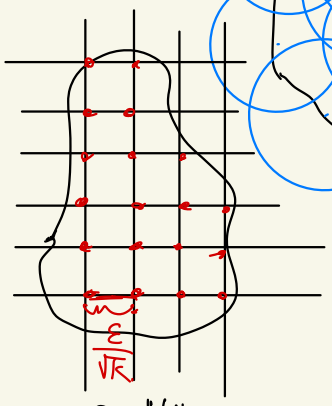
s.t. $\forall t \in T \ \exists \hat{t} \in \hat{T}$ s.t. $d(t, \hat{t}) < \varepsilon$.

ε-covering number: $\mathcal{N}(\varepsilon, T, d) = \min\{ |\hat{T}| : \hat{T} \text{ is an } \varepsilon\text{-cover of } T \}$



$T \subseteq \mathbb{R}^k$: $\mathcal{N}(\varepsilon, T, k) = \Theta\left(\frac{1}{\varepsilon^k}\right)$

Example: $([0,1]^k, L_\infty)$



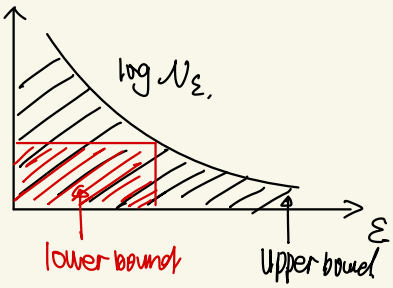
$\sup_{t \in T} \|t\|_d = D$

Thm (Dudley) $\forall \mathcal{F}, X_1, \dots, X_n$

$$\underbrace{\mathbb{E}_\varepsilon \|\hat{R}_n\|_{\mathcal{F}(X_1^n)}}_{\text{empirical}} = \mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(x_i) \right| \right]$$

$$\le C \cdot \int_0^\infty \sqrt{\frac{\log \mathcal{N}(\varepsilon, \mathcal{F}(X_1^n), l_2)}{n}} \, d\varepsilon$$

$$\mathbb{E}_\varepsilon \|\hat{R}_n\|_{\mathcal{F}(X_1^n)} \ge \frac{1}{C \log n} \cdot \sup_{\alpha > 0} \alpha \cdot \sqrt{\frac{\log \mathcal{N}(\alpha, \mathcal{F}(X_1^n), l_2)}{n}}$$



Lemma $\mathcal{F} = \{ f(\theta, \cdot) : \theta \in \Theta \} = f \circ \Theta$

$\|\cdot\|_\Theta$, $\|\cdot\|_F$. $\theta \mapsto f(\theta, \cdot)$ is L-Lipshitz

$\|f(\theta, \cdot) - f(\theta', \cdot)\|_F \le L \cdot \|\theta - \theta'\|_\Theta$

$\Rightarrow \mathcal{N}(\varepsilon, \mathcal{F}, \|\cdot\|_F) \le \mathcal{N}\left(\frac{\varepsilon}{L}, \Theta, \|\cdot\|_\Theta\right)$

Thm (Bartlett, 2017) $\|X_i\|_2 \le C$. $\mathcal{F} = \{ f_\theta : \|W_i^T\|_{2,1} \le b_i, \ \|W_i\|_{2,2} = \max_{x : \|x\|_2 = 1} \|Wx\|_2 \le \kappa_i \}$

$$\mathbb{E}_\varepsilon \|\hat{R}_n\|_{\mathcal{F}(X_1^n)} \le \frac{C}{\sqrt{n}} \cdot \left(\prod_{i=1}^{r} \kappa_i \right) \left(\sum_{i=1}^{r} \frac{b_i^{2/3}}{\kappa_i^{2/3}} \right)^{\frac{3}{2}}$$

② Growth-Function & VC-Dimension

$\mathcal{F} \subseteq \{\pm 1\}^{\mathcal{X}}$. $Y \in \{\pm 1\}$ $\quad \ell(y, y') = \mathbb{1}\{y \neq y'\} = \frac{1 - yy'}{2}$

$\ell \circ \mathcal{F}$ $\qquad \mathbb{E}\|R_n\|_{\ell\circ\mathcal{F}} \le \frac{1}{2} \cdot \mathbb{E}\|R_n\|_{\mathcal{F}} + O\left(\frac{1}{\sqrt{n}}\right)$

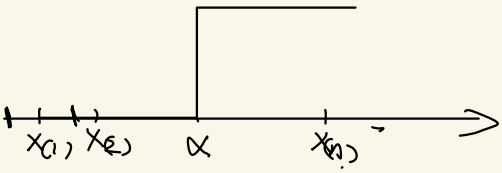
$L(\hat{f}_n) - L(f_{\mathcal{F}})$

$\Rightarrow \sup_{f \in \mathcal{F}} |L(f) - \hat{L}_n(f)|$

Output Space: $\mathcal{F}(X_1^n) = \{(f(X_1), \cdots, f(X_n)) : f \in \mathcal{F}\} \subseteq \{-1, +1\}^n$

Example. $\mathcal{F} = \{x \mapsto \mathbb{1}\{x \le \alpha\} : \alpha \in \mathbb{R}\}$

$|\mathcal{F}(X_1^n)| \le n+1 \Rightarrow \mathbb{E}\|R_n\|_{\mathcal{F}} \le \sqrt{\frac{2\log(n+1)}{n}}$



Example. $\mathcal{F}$ is parameterized by k bits

$\mathcal{F} = \{x \mapsto f(x, \theta) : \theta \in \{0,1\}^k\}$. $\quad f : \mathcal{X} \times \{0,1\}^k \to [0,1]$

$|\mathcal{F}(X_1^n)| \le 2^k \Rightarrow \mathbb{E}\|R_n\|_{\mathcal{F}} \le \sqrt{\frac{2k \cdot \log 2}{n}}$

- How expressive is this class ?

def (Growth function) $\mathcal{F} \subseteq [0,1]^{\mathcal{X}}$. $\quad \Pi_{\mathcal{F}}(n) = \max\{|\mathcal{F}(X_1^n)| : X_1, \cdots, X_n \in \mathcal{X}\}$

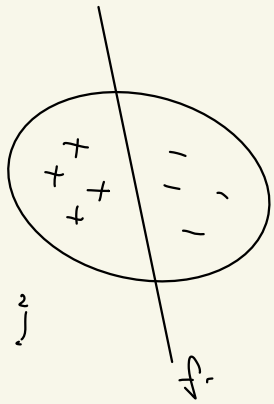
Massart's Lemma: $\mathbb{E}\|R_n\|_{\mathcal{F}} \le \sqrt{\frac{2\log(2\Pi_{\mathcal{F}}(n))}{n}}$

best case

$f(X_i) \in \{\pm 1\}$

$\mathbb{E}\|P - P_n\|_{\mathcal{F}} = O\left(\sqrt{\frac{2\log(2\Pi_{\mathcal{F}}(n))}{n}}\right)$

- $\Pi_{\mathcal{F}}(n) \le |\mathcal{F}|$. $\quad \lim_{n\to\infty} \Pi_{\mathcal{F}}(n) = |\mathcal{F}|$
- $\Pi_{\mathcal{F}}(n) \le 2^n$ $\qquad$ no useful bound



def (VC-dimension) $d_{VC}(\mathcal{F}) = \max\{n : \Pi_{\mathcal{F}}(n) = 2^n\}$

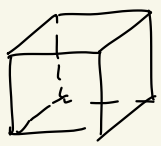
Vapnik Chervonenkis

$= \max\{n : \text{some } x_1, \cdots, x_n \in \mathcal{X} \text{ is shattered by } \mathcal{F}\}$

$|\mathcal{F}(X_1^n)| = 2^n$

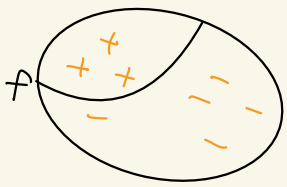
$\|$

$|\{f(X_1), \cdots, f(X_n) : f \in \mathcal{F}\}|$ (shattering)



$\{-1, +1\}^n$

full-projection



$\mathcal{F}$ complex (expressive)

$d_{VC}(\mathcal{F}) \uparrow$

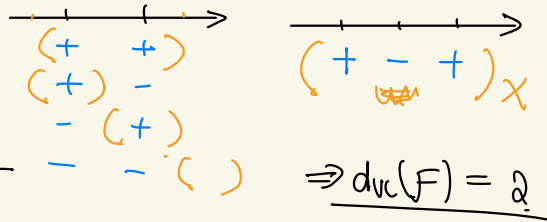
* Regression: fat-shattering

Thm (Sauer-Shelah, Vapnik-Chervonenkis) $d_{VC}(\mathcal{F}) \le d$

$$\Rightarrow \Pi_{\mathcal{F}}(n) \le \sum_{i=0}^{d} \binom{n}{i}$$

Example. $\mathcal{F} = \{ x \mapsto \mathbb{1}\{x \in (a,b)\} : a, b \in \mathbb{R} \}$

$n \ge d$: $\Pi_{\mathcal{F}}(n) \le \left(\frac{en}{d}\right)^d$



$\Rightarrow \Pi_{\mathcal{F}}(n) \begin{cases} = 2^n & n \le d \quad (\text{"overparameterization"}) \\ \le \left(\frac{e}{d}\right)^d \cdot n^d & n > d \end{cases}$

$\Rightarrow \mathrm{dvc}(\mathcal{F}) = 2$

$\Rightarrow \mathrm{dvc}(\mathcal{F}) \le d$, $n \ge d$: $\sup_P \mathbb{E} \| P - P_n \|_{\mathcal{F}} = O\left(\sqrt{\frac{d \cdot \log(n/d)}{n}}\right) \to 0 \quad (n \to \infty)$ d finite

arbitrary distribution, $\Rightarrow$ VC dimension is distribution-free !

Actually, we'll have some knowledge about the distribution $\mathcal{P}$ $\Rightarrow$ VC looser bound.

* Rademacher Complexity : data-dependent (distribution-dependent)

- Uniform convergence (uniformly over probability distribution) $\Leftrightarrow$ finite VC dimension !

$f \in \mathcal{F}$, $P \in \mathcal{P}$

Parameterized Families : $\mathcal{F} = \{ x \mapsto f(x, \theta) : \theta \in \mathbb{R}^p \}$ $\quad f : \mathcal{X} \times \mathbb{R}^p \to \{\pm 1\}$

$\underline{x}$: compute $\theta \mapsto f(x, \theta)$ using $\le t$ operations

$\downarrow$

Computation tree max depth

arithmetic : $+, -, \times, \div$

comparisons : $>, =, <$

output ± 1 !

Thm (Goldberg & Jerrum, 1994) $\mathrm{dvc}(\mathcal{F}) \le 4p(t+2)$

Neural Networks : $f : \mathcal{X} \times \mathbb{R}^p \to \{\pm 1\}$ $\quad L$ layers $\quad k$ non-linear computation unit :

① Piecewise constant (linear threshold unit) ; $\mathrm{dvc}(\mathcal{F}) = \tilde{\Theta}(p)$ (Baum & Haussler, 1989)

② Piecewise linear (ReLU) : $\mathrm{dvc}(\mathcal{F}) = \tilde{\Theta}(pL)$ (Bartlett, Harvey, et al. 2017)

③ Piecewise polynomial : $\mathrm{dvc}(\mathcal{F}) = \tilde{O}(pL^2)$ (Bartlett, Maiorov, Meir, 1998)

④ Sigmoid : $\mathrm{dvc}(\mathcal{F}) = \tilde{O}(p^2 k^2)$ (Karpinsky, MacIntyre, 1996)

VC dimension : # of parameter

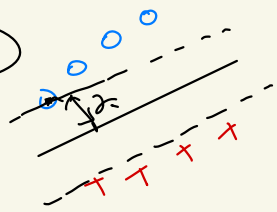
Large Margin Classifier

$\min \frac{1}{2}\|w\|^2$

$\Updownarrow$

$\max \frac{1}{\|w\|} = \gamma$

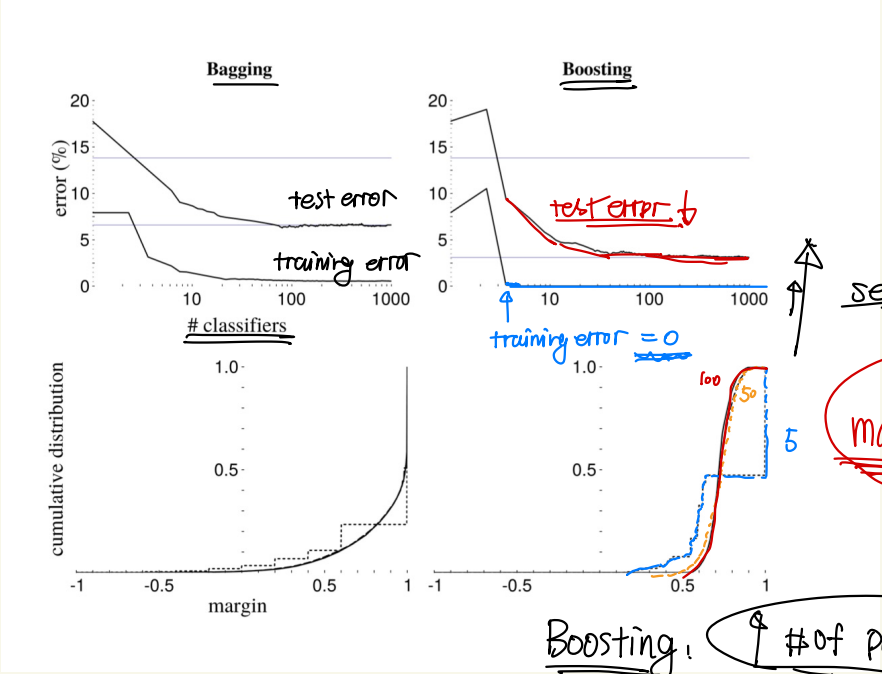
max-margin

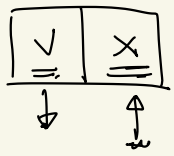


Perceptron : $\gamma = \max_{\|w\|_2 = 1} \min_{i \in [n]} \langle y_i \langle w, x_i \rangle \rangle$

Margin Theory $\to$ SVM, AdaBoost

Ensemble Methods





Weight $e^{-\eta \cdot y_i f(x_i)}$

sequentially

margin ↑

Boosting: ↑ # of params ⇓ ↑ # of classifier ⇓ training time ↑ ⟺ error (performance)

Observation: test error does NOT go up!
(even if we add more parameters)

Classification

$f: \mathcal{X} \to \mathbb{R}$

prediction of x: $\hat{y} = \text{sign}(f(x)) \in \{-1, +1\}$ ⇒ $\begin{cases} y f(x) > 0 & : \text{correct} \\ y f(x) \le 0 & : \text{mistake.} \end{cases}$

$y f(x)$: margin of f on x

Larger margin → more confident correct classification

minimize $\sum_{i=1}^{n} (f(x_i) - y_i)^2$, $\sum_{i=1}^{n} \exp(-y_i f(x_i))$ → encourages large margin

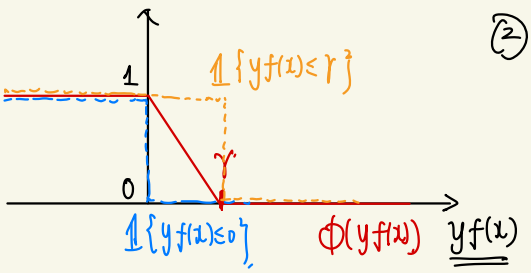
$\ell: \hat{y} \mapsto \ell(\hat{y}, y)$ 1-Lipschitz. $\forall y$. $\quad \mathbb{E}\|R_n\|_{\ell \circ \mathcal{F}} \le 2\ell\, \mathbb{E}\|R_n\|_{\mathcal{F}} + \frac{c}{\sqrt{n}}$

$\mathbb{1}\{\hat{y} \ne y\} = \frac{1 - \hat{y} y}{2}$ ⟶ $\mathbb{E}\|R_n\|_{\ell \circ \text{sign}(\mathcal{F})} \le 2 \cdot \mathbb{E}\|R_n\|_{\text{sign}(\mathcal{F})} + \frac{c}{\sqrt{n}}$

e.g. real values in $\mathcal{F}$ is very small but $\text{sign}(\mathcal{F})$ might be complicated

Surrogate loss

① $\frac{1}{\gamma}$-Lipschitz: $\phi(\alpha) = \begin{cases} 1, & \text{if } \alpha \le 0 \\ 1 - \frac{\alpha}{\gamma}, & \text{if } 0 < \alpha < \gamma \\ 0, & \text{if } \alpha \ge \gamma \end{cases}$



② $\underbrace{\mathbb{1}\{Y f(X) \le 0\}}_{\text{classification 0/1 loss}} \le \phi(Y f(X)) \le \underbrace{\mathbb{1}\{Y f(X) \le \gamma\}}_{\text{large margin loss}}$

$P\phi(Yf(X)) \overset{\text{Lipschitz}}{\Longleftrightarrow} P_n \phi(Yf(X))$ ⇒

$P\,\mathbb{1}\{Yf(X) \le 0\} \le P\,\phi(Yf(X))$

$P_n \phi(Yf(X)) \le P_n \mathbb{1}\{Yf(X) \le \gamma\}$

w.h.p $\forall f \in \mathcal{F}$: $L_{01}(f) = P\,\mathbb{1}\{Yf(X) \le 0\}$

$$\le P\,\phi(Yf(X))$$

$$\le P_n\phi(Yf(X)) + \frac{c}{\gamma}\mathbb{E}\|R_n\|_{\mathcal{F}} + O\left(\frac{1}{\sqrt{n}}\right) \quad \text{(Rademacher)}$$

$$\le \boxed{P_n\,\mathbb{1}\{Yf(X) \le \gamma\}} + \frac{c}{\gamma}\cdot\mathbb{E}\|R_n\|_{\mathcal{F}} + O\left(\frac{1}{\sqrt{n}}\right)$$

$\gamma \uparrow$ $L_{01}(f) \downarrow$ large margin

by using surrogate loss classification $\to$ regression.

Example. Linear Classifiers $\mathcal{F} = \{x \mapsto \langle w, x\rangle : \|w\| \le 1\}$

$$\mathbb{E}\|R_n\|_{\mathcal{F}} \lesssim n^{-\frac{1}{2}} \Rightarrow L_{01}(f) \lesssim \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{Y_i f(X_i) \le \gamma\} + \frac{1}{\gamma}\frac{1}{\sqrt{n}}$$

In general: for any surrogate loss ϕ

$\exists$ transform ψ: 0/1 loss $\Leftrightarrow$ ϕ-loss excess risk

$0 + \frac{1}{\gamma^2}\frac{1}{n}$ Perceptron (linearly-separable)

Thm 1. $\forall P, f$: $\psi(L_{01}(f) - L_{01}^*) \le L_\phi(f) - L_\phi^*$.

2. $|\mathcal{X}| \ge 2$, $\varepsilon > 0$, $\theta \in [0,1]$. $\exists P, f$: $L_{01}(f) - L_{01}^* = \theta$.

$$\psi(\theta) \le L_{01}(f) - L_{01}^* \le \psi(\theta) + \varepsilon$$

"Calibration"

Generalization loss function doesn't vary too quickly

Trade-off: fit to the training data (margin), complexity (size of parameter) # of parameters

$$\Pr(\text{sign}(f(X)) \ne Y) \le \underbrace{\frac{1}{n}\sum_{i=1}^{n}\phi(Y_i f(X_i))}_{\text{fit}} + \underbrace{P_n(f)}_{\text{complexity}}$$

Even if the training set is classified correctly, it might be worthwhile to increase the complexity, to improve its loss functions.

Beyond ERM

* large-margin classification for 2-layer NN.

$$\mathcal{F}_{B,d} = \left\{x \mapsto \sum_{i=1}^{K} W_i\,\sigma(v_i^T x) : \|W\|_1 \le B,\ \|v_i\|_1 \le B,\ K \ge 1,\ x \in \mathbb{R}^d\right\}$$

1-Lipschitz function

w.h.p. $\forall f \in \mathcal{F}_{B,d}$: $L_{01}(f) \le P_n\,\mathbb{1}\{yf(x) \le \gamma\} + \tilde{O}\left(\frac{B^3}{\gamma^2}\sqrt{\frac{\log d}{n}}\right)$

$\mathbb{E}[L(\hat{f}_n) - \hat{L}(\hat{f}_n)] \downarrow$ $\quad \hat{f}_n = \arg\min_f \hat{L}(f) \rightarrow$ ERM

$\hat{f}_n$: certain properties $\quad L(\hat{f}_n) \le \hat{L}(\hat{f}_n) + \cdots + \cdots$ w.h.p.

(a-posteriori bound)

① Compression Bound

$S = \{Z_1, \cdots, Z_n\}$ $\quad \hat{f}_n = \hat{f}_n[S]$

$\mathcal{G} = \{(x,y) \mapsto \ell(f(x), y) : f \in \mathcal{F}\}$. $\hat{g}_n(\cdot) = \ell(\hat{f}_n(\cdot), \cdot) = \hat{g}_n[S](\cdot) \in [a,b]$

Compression function $C_k(S) \sqsubseteq S$ $\quad \hat{f}_n[S] = \hat{f}_k[C_k(S)]$

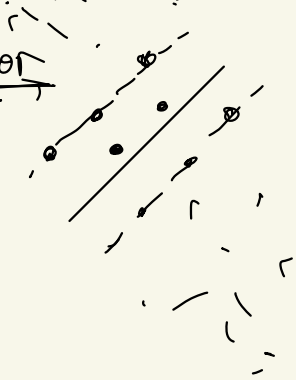
$|C_k(S)| = k \quad |S| = n$

"expressive examples" $\quad$ k sparse ($k \ll n$)

the learning alg. produces the same function when given S or its subset $C_k(S)$

e.g. SVM $\quad (x_1, y_1) \cdots (x_n, y_n)$

Support Vector



$$L(\hat{f}_n) - \hat{L}(\hat{f}_n) = \mathbb{E}_{Z\sim P}[\hat{g}_n(Z)] - \frac{1}{n}\sum_{i=1}^n \hat{g}_n(Z_i)$$

$$= \mathbb{E}[\hat{g}_k[C_k(S)](Z)] - \frac{1}{n}\sum_{i=1}^n \hat{g}_k[C_k(S)](Z_i)$$

$$\le \max_{\substack{I \subseteq \{1,\cdots,n\} \\ |I| \le k}} \left\{ \mathbb{E}[\hat{g}_k[S_I](Z)] - \boxed{\frac{1}{n}\sum_{i=1}^n \hat{g}_k[S_I](Z_i)} \right\}$$

only depends on k out of n sample $\quad$ mostly out of sample

i.i.d $W = \{Z_1', \cdots, Z_k'\}$

$$\le \max_{\substack{I \subseteq \{1,\cdots,n\} \\ |I| = k}} \left\{ \overbrace{\mathbb{E}[\hat{g}_k[S_I](Z)] - \frac{1}{n}\sum_{Z' \in S'} \hat{g}_k[S_I](Z')}^{X_i} \right\} + \frac{k(b-a)}{n}$$

$S' = (S - S_I) \cup W$ $\quad \frac{1}{n}\sum_i \hat{g}_k[S_I](Z_i') - (Z_i)$

$\le \frac{1}{n} \cdot k \cdot (b-a)$ $\quad W \quad S_I$

mean: 0

st: $O\left(\sqrt{\frac{b-a}{n}}\right)$

X_i with $\mathbb{E}[X_i] = 0$. $\mathrm{Var}[X_i] = O\left(\frac{b-a}{n}\right)$ $\quad \log\binom{n}{k} \le k \cdot \log\frac{en}{k}$

maximal inequality: $\mathbb{E}[\max_{i \in I} X_i] \le r \cdot \sqrt{2\log\binom{n}{k}}$

sub-Gaussianity $\quad \sqrt{\frac{b-a}{n}}$

$$\underbrace{\mathbb{E}\left[L(\hat{f}_n) - \hat{L}(\hat{f}_n)\right]}_{\text{limit the bias}} = C\cdot\sqrt{\frac{(b-a)\,k\cdot\log en/k}{n}} + \frac{k(b-a)}{n} \le O\left(\sqrt{\frac{k\log n}{n}}\right)$$

if $K = o\left(\frac{n}{\log n}\right)$

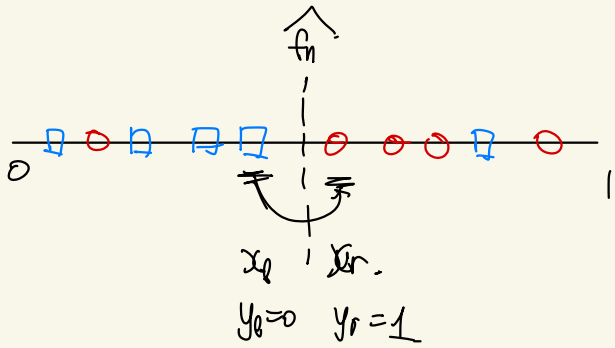
Example. Classification on 1D

$\mathcal{X} = [0,1]$

$\mathcal{Y} = \{0,1\}$

$\mathcal{F} = \{f_\theta : f_\theta(x) = \mathbb{1}(x \ge \theta) : \theta \in [0,1]\}$

$\ell(f_\theta(x), y) = \mathbb{1}\{f_\theta(x) \ne y\}$



$k=2$: $\hat{f}_n[S] = \hat{f}_2[(x_l, 0), (x_r, 1)]$

▷ Compression of size d for hyperplanes

▷ Compression of size $1/\gamma^2$ for margin case.

② Algorithmic Stability

Compression: $\hat{f}_n$ depends only on a subset of k datapoints

Stability: $\hat{f}_n$ does NOT depend on any of the datapoints too strongly

$$\mathbb{E}_S[L(\hat{f}_n)] = \mathbb{E}_{Z_1,\dots,Z_n,Z}\left[\hat{g}_n[\underbrace{Z_1,\dots,Z_n}_{\text{train}}](\underbrace{Z}_{\text{test}})\right]$$

$Z \leftrightarrow Z_1$

$$= \mathbb{E}_{Z_1,\dots,Z_n,Z}\left[\hat{g}_n[Z, Z_2,\dots,Z_n](Z_1)\right]$$

$$\mathbb{E}_S[\hat{L}(\hat{f}_n)] = \mathbb{E}_{Z_1,\dots,Z_n}\left[\frac{1}{n}\sum_{i=1}^{n}\hat{g}_n[Z_1,\dots,Z_n](Z_i)\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{Z_1,\dots,Z_n}\left[\hat{g}_n[Z_1,\dots,Z_n](Z_i)\right]$$

$$= \mathbb{E}_{Z_1,\dots,Z_n}\left[\hat{g}_n[Z_1,\dots,Z_n](Z_1)\right]$$

β-close

$$\underbrace{\mathbb{E}\left[L(\hat{f}_n) - \hat{L}(\hat{f}_n)\right]}_{\text{small}} = \mathbb{E}_{Z_1,\dots,Z_n,Z}\left[\hat{g}_n[Z, Z_2,\dots,Z_n](Z_1) - \hat{g}_n[Z_1,\dots,Z_n](Z_1)\right] \le \beta$$

stability "does not change much"

Note: Stability analysis only makes sensens for those methods that explicitly or implicitly minimize empirical loss.

What about $\hat{f}_n = f_0 \triangleq c$

Uniform Stability

β-uniformly-stable $\quad \forall i \in [n],\ z_1, \cdots, z_n,\ z, z'$

$$\left| \hat{g}_n[S](z) - \hat{g}_n[S^{i,z'}](z) \right| \le \beta$$

$$S^{i,z'} = \{z_1, \cdots, z_{i-1}, z', z_{i+1}, \cdots, z_n\}$$

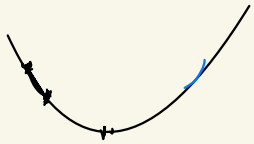
Challenge: it's unclear whether a β-uniformly-stable ERM (or near-ERM) exists

Example. Kernel Ridge Regression

$$\hat{f}_n = \underset{f \in \mathcal{H}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda \|f\|_K^2$$

in RKHS $\mathcal{H}$ w.r.t. kernel K $\qquad$ strongly-convex

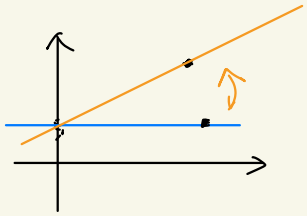
$\|\cdot\|_2^2 \quad \sigma = 2$



Assume $K(x,x) \le \kappa^2, \ \forall x$

Lemma KRR is β-uniformly stable with $\beta = O\left(\frac{1}{\lambda n}\right)$

* Linear Regression not stable



Proof: σ-strongly convex ϕ:

$$\forall u, v \in W: \ \phi(u) \ge \phi(v) + \langle \nabla \phi(v), u - v \rangle + \frac{\sigma}{2} \|u - v\|^2$$

Suppose ϕ, ϕ' σ-strongly-convex

$$\nabla \phi(w) = \nabla \phi'(w') = 0 \Rightarrow \phi(w') \ge \phi(w) + \frac{\sigma}{2} \|w - w'\|^2$$

$$\phi'(w) \ge \phi'(w') + \frac{\sigma}{2} \|w - w'\|^2$$

$$\Rightarrow \overset{2\lambda}{\sigma} \cdot \|w - w'\|^2 \le \left(\phi(w') - \phi'(w')\right) + \left(\phi'(w) - \phi(w)\right)$$

$$\phi(f) = \frac{1}{n} \sum_{i \in S} (f(x_i) - y_i)^2 + \lambda \cdot \|f\|_K^2$$

$$\phi'(f) = \frac{1}{n} \sum_{i \in S'} (f(x_i) - y_i)^2 + \lambda \cdot \|f\|_K^2$$

S, S' only differ in one element: $(x_i, y_i) \to (x_i', y_i')$

$\hat{f}_n = \operatorname{argmin} \phi$

$\hat{f}_n' = \operatorname{argmin} \phi'$

$$\phi(\hat{f}_n') - \phi'(\hat{f}_n') = \frac{1}{n}\left((\hat{f}_n'(x_i) - y_i)^2 - (\hat{f}_n'(x_i') - y_i')^2\right)$$

$$\phi'(\hat{f}_n) - \phi(\hat{f}_n) = \frac{1}{n}\left((\hat{f}_n(x_i') - y_i')^2 - (\hat{f}_n(x_i) - y_i)^2\right)$$

$$\Rightarrow \|\hat{f}_n - \hat{f}_n'\|_K^2 \le \frac{1}{2\lambda n}\left((\hat{f}_n'(x_i) - y_i)^2 - (\hat{f}_n'(x_i') - y_i')^2 + (\hat{f}_n(x_i') - y_i')^2 - (\hat{f}_n(x_i) - y_i)^2\right)$$

$$\le \frac{1}{2\lambda n}\cdot(4C+4)\,\|\hat f_n-\hat f_n'\|_\infty \quad \frac{(\hat f_n'(x_i)-\hat f_n(x_i))\,(\hat f_n'(x_i)+\hat f_n(x_i)-2y_i)}{}$$

$$\le \|\hat f_n'-\hat f_n\|_\infty\cdot(2C+2) \qquad |y_i|\le 1,\ |\hat f_n(x)|\le C,\ |\hat f_n'(x)|\le C$$

$\forall x$: $f(x)=\langle f, K_x\rangle \le \|f\|_K\cdot\|K_x\|_K$

$$=\|f\|_K\cdot\sqrt{\langle K_x,K_x\rangle} = \|f\|_K\cdot\sqrt{K(x,x)} \le K\cdot\|f\|_K \Rightarrow \|f\|_\infty\le K\cdot\|f\|_K$$

$$\|\hat f_n-\hat f_n'\|_K^2 \le \underbrace{\frac{4C+4}{2\lambda n}}_{C'}\|\hat f_n-\hat f_n'\|_\infty \le \frac{K\cdot C'}{2\lambda n}\|\hat f_n-\hat f_n'\|_K$$

$$\Rightarrow \|\hat f_n-\hat f_n'\|_K \le O\left(\frac{1}{\lambda n}\right)$$

$\forall (x_i,y_i)$:

$$(\hat f_n(x_i)-y_i)^2-(\hat f_n'(x_i)-y_i)^2 \le C'\|\hat f_n-\hat f_n'\|_\infty \le K\cdot C'\|f_n-f_n'\|_K \le O\left(\frac{1}{\lambda n}\right) \quad \square$$

Local Methods

Idea: to predict y at a given x, look up in the dataset those Y_i for which X_i is "close" to x.

local similarity

Bias-Variance Decomposition.

$$\mathbb{E}[L(\hat f_n)]-L(f^*)$$

estimation-approximation w.r.t. $\mathcal{F}$:

$$(\mathbb{E}[L(\hat f_n)]-L(f_{\mathcal{F}}))+(L(f_{\mathcal{F}})-L(f^*))$$

fix $\mathcal{F}$ ✗

fix $\hat f_n$ ✓ tunable parameter.

squared loss $\boxed{f^*(x)=\mathbb{E}[Y\mid X=x]}$

$$\mathbb{E}[L(\hat f_n)]-L(f^*) = \mathbb{E}[(\hat f_n(X)-Y)^2]-\mathbb{E}[(f^*(X)-Y)^2]$$

$$=\mathbb{E}[(\hat f_n(X)-f^*(X)+f^*(X)-Y)^2]-\mathbb{E}[(f^*(X)-Y)^2]$$

$$=\mathbb{E}_{S,X}[(\hat f_n(X)-f^*(X))^2]$$

cross-term vanishes

$$=\mathbb{E}_S\int_X(\hat f_n(x)-f^*(x))^2P\,dx$$

$$=\int_X \underbrace{\mathbb{E}_S[(\hat f_n(x)-f^*(x))^2]}_{\text{fixed } x}P\,dx$$

$$\|f-g\|_{L_2(P)}^2=\int_X(f(x)-g(x))^2P(dx)$$

integral is a measure of distance between two function

$$\mathbb{E}_S \|\hat{f}_n - f^*\|^2 = \mathbb{E} \| \hat{f}_n - \mathbb{E}_{Y_{1:n}}[\hat{f}_n] + \mathbb{E}_{Y_{1:n}}[\hat{f}_n] - f^* \|^2.$$

condition on $X_{1:n}$.

variance $\rightarrow$ $= \mathbb{E}_S \|\hat{f}_n - \mathbb{E}_{Y_{1:n}}[\hat{f}_n]\|^2$

bias^2 $\rightarrow$ $+ \mathbb{E}_S \| \mathbb{E}_{Y_{1:n}}[\hat{f}_n] - f^* \|^2$

cross $\mathbb{E}_{X_{1:n}, Y_{1:n}}\left[\hat{f}_n - \mathbb{E}_{Y_{1:n}}[\hat{f}_n] \right] = 0$

eg. $\hat{f}_n$ smoother. variance $\downarrow$ bias $\uparrow$ bias-variance tradeoff

① KNN (K-Nearest-Neighbors)

$(X_1, Y_1), \cdots, (X_n, Y_n) \overset{i.i.d}{\sim} P.$ $\hat{f}_n(x) = \frac{1}{K} \sum_{i=1}^{K} Y_{(i)}$

$(X_{(1)}, Y_{(1)}), \cdots, (X_{(n)}, Y_{(n)})$ sorted (by $\|X_i - x\|$)

Variance: Given x, $\hat{f}_n(x) - \mathbb{E}_{Y_{1:n}}[\hat{f}_n(x)] = \frac{1}{K} \sum_{i=1}^{K} \left(Y_{(i)} - f^*(X_{(i)}) \right) = O\left(\frac{1}{\sqrt{K}}\right)$

Bayes Optimal

$\Rightarrow$ Variance $= O\left(\frac{1}{K}\right)$

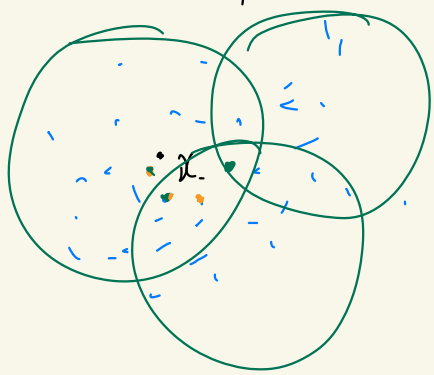
Bias: Given x. $\mathbb{E}_{Y_{1:n}}[\hat{f}_n(x)] - f^*(x) = \frac{1}{K} \sum_{i=1}^{K} \left(f^*(X_{(i)}) - f^*(x) \right)$

f^*: 1-Lipschitz; $\left(\frac{1}{K} \sum_{i=1}^{K} \left(f^*(X_{(i)}) - f^*(x) \right) \right)^2 \leq \frac{1}{K} \sum_{i=1}^{K} \|X_{(i)} - x\|^2$

random fixed

how close the closet K random points are to x.

Assume $\frac{n}{K} \in \mathbb{Z}$. $\rightarrow$ divide the (unsorted) dataset into K blocks at random.



X^i: closet point to x in i-th block

$\{X^1, \cdots, X^K\}$ $\{X_{(1)}, \cdots, X_{(K)}\}$

$$\frac{1}{K} \sum_{i=1}^{K} \|X_{(i)} - x\|^2 \leq \frac{1}{K} \sum_{i=1}^{K} \|X^i - x\|^2.$$

$$\mathbb{E}_{X_1, \cdots, X_n}\left[\frac{1}{K} \sum_{i=1}^{K} \|X_{(i)} - x\|^2 \right] \leq \mathbb{E}_{X_1, \cdots, X_n}\left[\frac{1}{K} \sum_{i=1}^{K} \|X^i - x\| \right]^2 = \mathbb{E}[\|X^1 - x\|^2]$$

If support of X is bounded. $d \geq 3$: $\mathbb{E}[\|X^1 - x\|^2] \lesssim n^{-2/d}$. $\left(\frac{1}{n}\right)^{\frac{1}{d}}$ if d large.

$\Rightarrow$ closet neighbor of a random point X to be no $n^{-1/d}$ away

Bias $\leq O\left(\left(\frac{n}{K}\right)^{-\frac{2}{d}} \right)$

bias-variance: $\frac{1}{K} + \left(\frac{K}{n}\right)^{\frac{2}{d}}$ trade-off

$$\boxed{k^* \sim n^{\frac{2}{2+d}}} \Rightarrow \text{risk: } n^{-\frac{2}{2+d}} \quad \forall x$$

$$\Rightarrow \mathbb{E}[\|\hat{f}_n - f^*\|^2] \lesssim n^{-\frac{2}{2+d}}$$

Truly non-parametric!

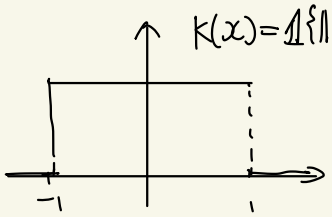
totally different with the analysis by Rademacher Complexity

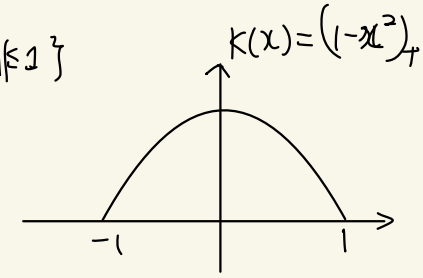
② Kernel Regression : Nadaraya-Watson

$K: \mathbb{R}^d \to \mathbb{R}_{\geq 0}$

$K_h(x) = K\left(\frac{x}{h}\right)$

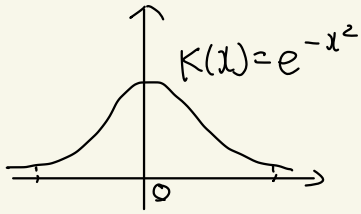
$K_h(x - x') = 0$ if $\|x - x'\| \geq h$.





Assume K is zero outside the unit Euclidean ball.

"h" bandwidth : tunable parameter



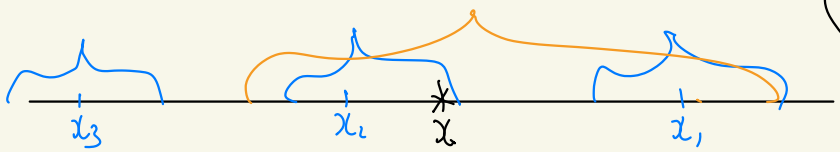
Assume $K(x) > c \cdot \mathbb{1}\{\|x\| \leq 1\}$ for some $c > 0$.

Weighting-function

Nadaraya-Watson estimator: $\hat{f}_n(x) = \sum_{i=1}^n Y_i \cdot W_i(x)$

$$W_i(x) = \frac{K_h(x - X_i)}{\sum_{i=1}^n K_h(x - X_i)}$$

"kernel smoothing"



Bias: Given x: $\mathbb{E}_{Y_{1:n}}[\hat{f}_n(x)] = \mathbb{E}_{Y_{1:n}}\left[\sum_{i=1}^n Y_i W_i(x) \,\middle|\, X_{1:n}\right]$

$= \sum_{i=1}^n f^*(X_i) \cdot W_i(x)$

$K\left(\frac{x}{h}\right)$: simulated annealing

knowledge distillation.

$$\Rightarrow \mathbb{E}_{Y_{1:n}}[\hat{f}_n(x)] - f^*(x) = \sum_{i=1}^n \left(f^*(X_i) - f^*(x)\right) \cdot W_i(x)$$

f^*: 1-Lipschitz. K_h is zero outside the h-radius ball

weight

$$\mathbb{E}_{Y_{1:n}}[\hat{f}_n(x)] - f^*(x) \leq \sum_{i=1}^n \|X_i - x\| \cdot W_i(x) \leq h \cdot \sum_{i=1}^n W_i(x) = h$$

$\Rightarrow$ Bias$^2 \leq h^2$.

Variance, $\hat{f}_n(x) - \mathbb{E}_{Y_{1:n}}[\hat{f}_n(x)] = \sum_{i=1}^n \left(Y_i - \underbrace{f^*(X_i)}_{\mathbb{E}[Y_i | X_i]}\right) W_i(x)$

$$\Rightarrow \mathbb{E}_{X_{1:n}}\left[\sum_{i=1}^n (Y_i - f^*(X_i))^2 W_i(x)^2\right]$$

constant $\leq B$

cross-term 0.

$$B \cdot n \cdot \mathbb{E}\left[\frac{K_h(x-X_1)^2}{\left(\sum_{i=1}^{n} K_h(x-X_i)\right)^2}\right] \to \frac{O(h^d)}{(nh^d)^2} \quad \text{w.h.p.}$$

$$\Rightarrow \text{Variance} = O\left(\frac{1}{nh^d}\right)$$

$$\text{Bias}^2 + \text{Variance}: \; h^2 + \frac{1}{nh^d}. \quad h^* \sim n^{-\frac{1}{2+d}} \Rightarrow h^{*2} + \frac{1}{nh^{*d}} = n^{-\frac{2}{2+d}}.$$

large d

Interpolation

$\ell_{\gamma,F} = \mathbb{1}\{y_i f(x_i) \le \gamma\}$ empirical

VCdim : # of params

Recall: 1. (Vapnik-Chervonenkis, 1971) $\mathbb{E}[\ell_F] \le \hat{\mathbb{E}}[\ell_F] + O\left(\sqrt{\frac{d_{VC}(F)}{n}}\right)$

tradeoff

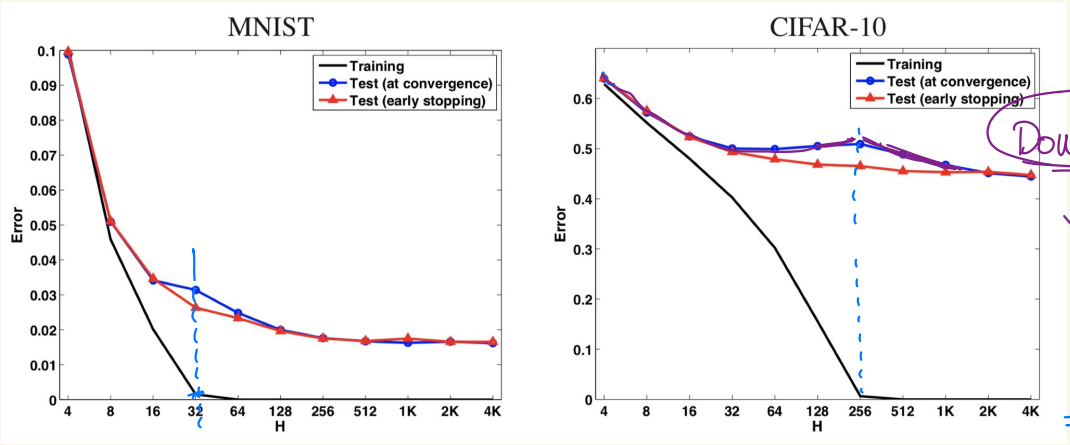
2. Large-Margin Classifier: $\mathbb{E}[\ell_{01,F}] \le \hat{\mathbb{E}}[\ell_{\gamma,F}] + O\left(\frac{B^3}{\gamma^2}\sqrt{\frac{\log d}{n}}\right)$

Classification → Regression

3. Local Methods → $\mathbb{E}_S \|\hat{f}_n - f^*\|^2 = \mathbb{E}_S \|\hat{f}_n - \mathbb{E}_{Y_{1:n}}[\hat{f}_n]\|^2 + \mathbb{E}_S \|\mathbb{E}_{Y_{1:n}}[\hat{f}_n] - f^*\|^2$.

NN

Kernel Smoothing

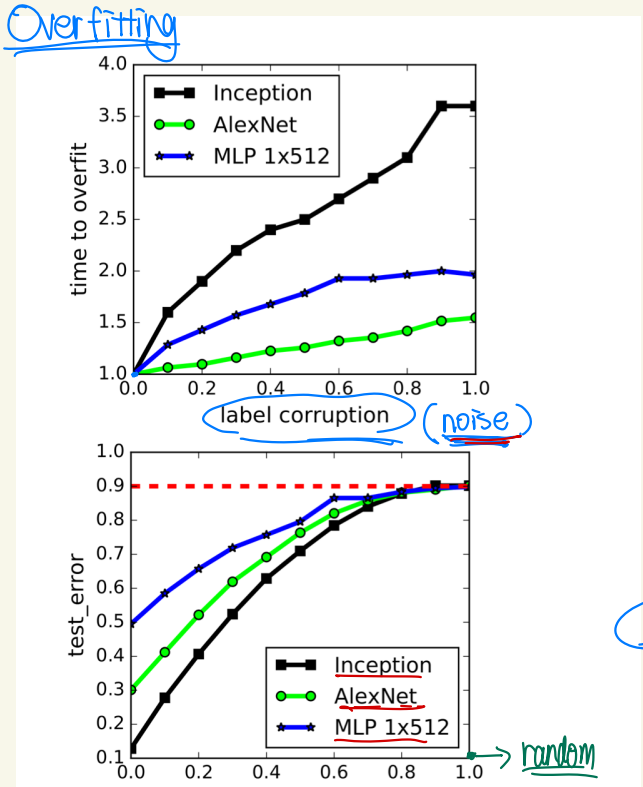


Double Descent

spiking

# of params

Overfitting



- Deep learning can be trained to zero training error (for regression loss)
- ... with near state-of-the-art performance
- ... even for noisy problems ☆
- No trade-off between fit to the training data and model complexity
- $\boxed{\text{Benign Overfitting}}$

interpolation ≠ fitting

Classification, regression

overparameterization ✓

overfitting noisy data ???

(Zhang, Bengio, Hardt, Recht, Vinyals, 2017 : Understanding Deep Learning Requires Rethinking Generalization)

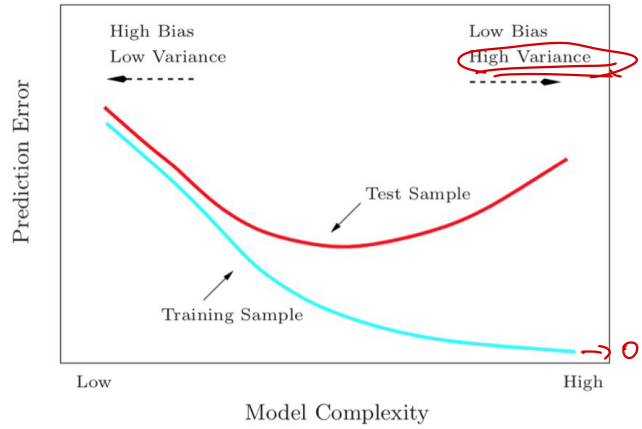
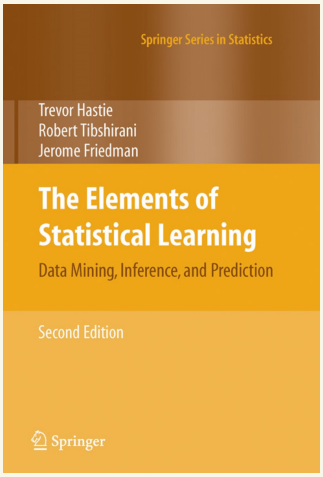


FIGURE 2.11. *Test and training error as a function of model complexity.*

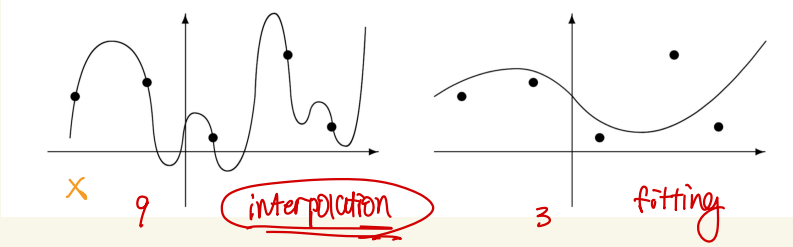
Figure 2.11 shows the typical behavior of the test and training error, as model complexity is varied. The training error tends to decrease whenever we increase the model complexity, that is, whenever we fit the data harder. However with too much fitting, the model adapts itself too closely to the training data, and will not generalize well (i.e., have large test error). In

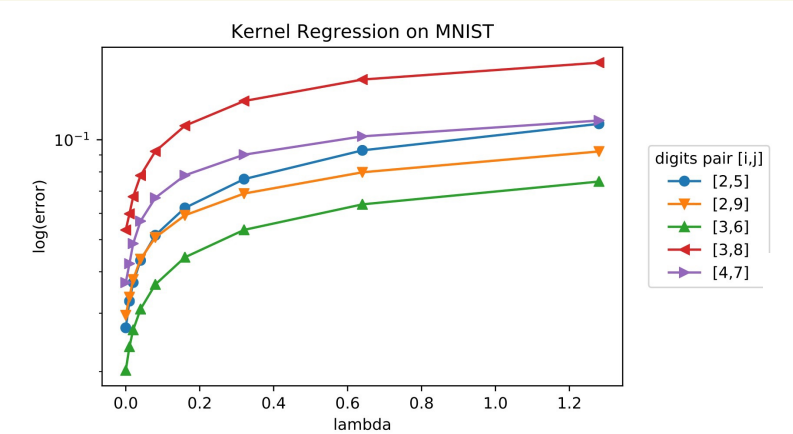


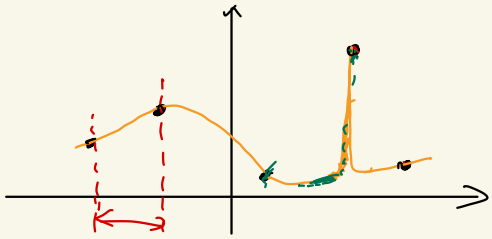
" ... interpolation fits ... are unlikely to predict future data well at all ... "

A "new" statistical phenomenon:

good prediction with zero training error for regression loss! ✓







$$\min_{\beta} \|X\beta - Y\|_2^2 + \lambda \|\beta\|_2^2$$

interpolation.

Overparameterization $\quad \beta^* = (X^T X)^{\dagger} X^T Y$

High-Dimensional Space:

① 1-Nearest-Neighbor: interpolation

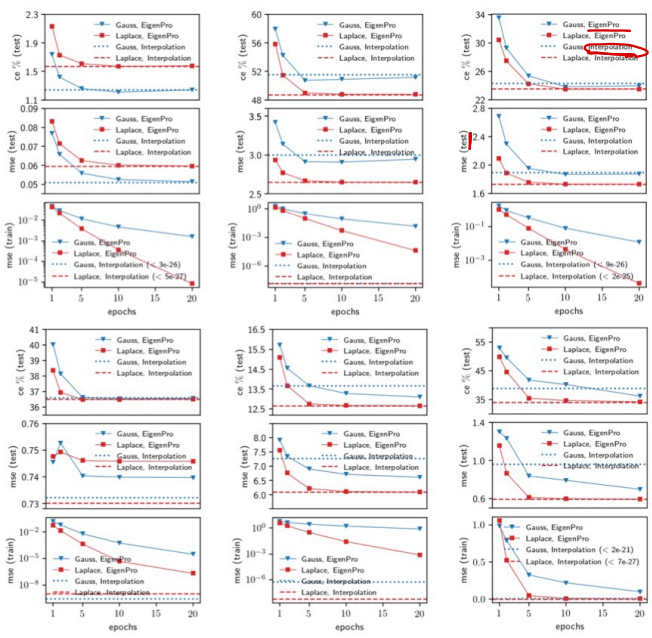
$$\mathbb{E}_S\left[L(\hat{f}_n)\right] - L(f^*)$$

Cover-Hart '67: $\lim_{n\to\infty} \mathbb{E}_S\left[L(\hat{f}_n)\right] \le 2L(f^*)$ ✓

Bayes Optimal → noiseless

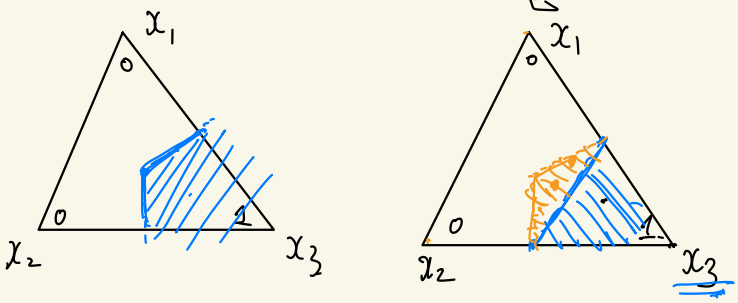
To Understand Deep Learning We Need to Understand Kernel Learning

Mikhail Belkin, Siyuan Ma, Soumik Mandal
Department of Computer Science and Engineering
Ohio State University
{mbelkin, masi}@cse.ohio-state.edu, mandal.32@osu.edu



(d) TIMIT ($5 \cdot 10^4$ subsamples) (e) HINT-S ($2 \cdot 10^4$ subsamples) (f) 20 Newsgroups

Simplex $\quad \limsup_{n\to\infty} \mathbb{E}\|\hat{f}_n - f^*\|^2 \le \frac{2}{d+2} \cdot \mathbb{E}[(f^*(X)-Y)^2]$

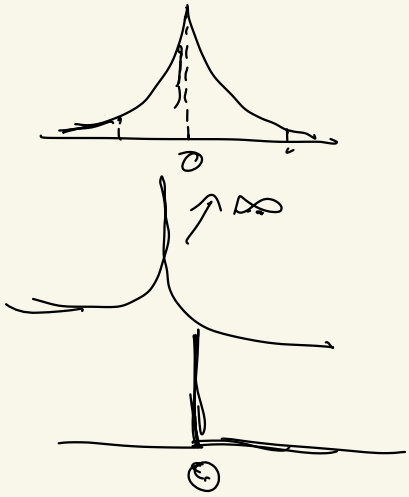


$\hat{f}_n(x) = \sum_{i=1}^{3} W_i \cdot Y_i$

↓

decrease the impact of "noisy" labels

② Kernel Smoothing (Nadaraya-Watson): $K(x) = \min\left\{\frac{1}{\|x\|^\alpha}, \tau\right\}$



$\tau \uparrow \quad \hat{f}_n(X_i) \approx Y_i \quad (W_i(X_i) \uparrow)$

$\tau = \infty \quad \hat{f}_n(X_i) = Y_i$

interpolation → memorize the data.

Benign Overfitting

Overparameterization Regime:

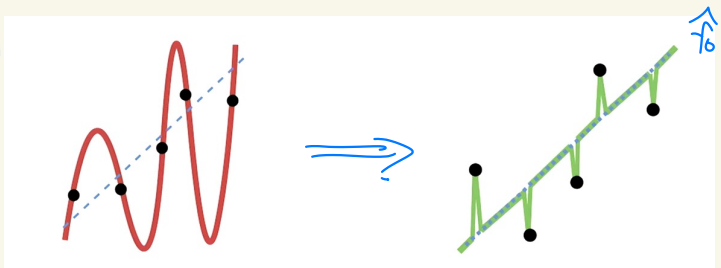
prediction rule $\hat{f} = \hat{f}_0 + \Delta$

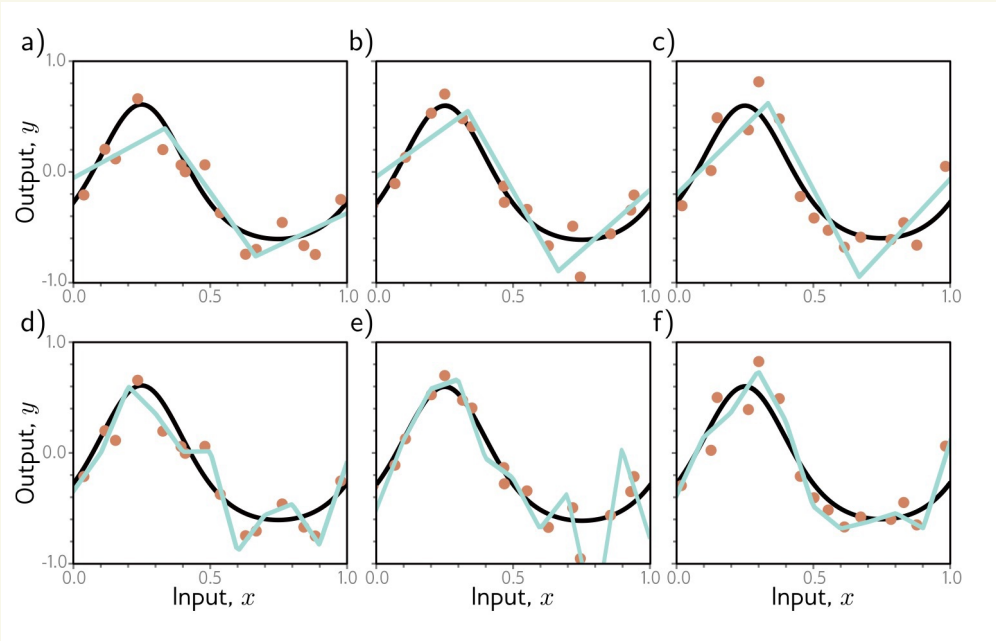
Classical SLT

$\hat{f}_0$: simple component useful for prediction

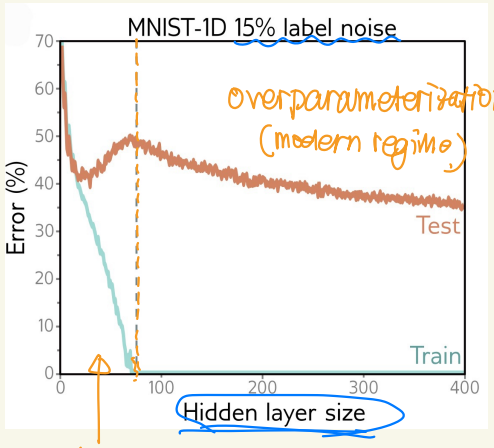
Δ: "spicy" component useful for benign overfitting

prediction ×





Double Descent



overparameterization (modern regime)

underparameterization (regime)

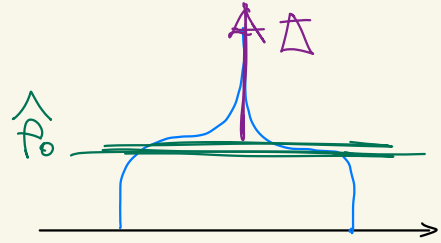
(Understanding Deep Learning)

$\hat{f} = \hat{f}_0 + \Delta$

e.g. (kernel smoothing) $\hat{f}(x) = \sum_{i=1}^{n} y_i \cdot \frac{K_h(x - x_i)}{\sum_{j=1}^{n} K_h(x - x_j)}$, $\quad K_h(x) = \frac{\mathbb{1}\{h\|x\| \le 1\}}{h \cdot \|x\|^\alpha}$.

$\hat{f}_0$: standard kernel,

Δ : spiky piece (min $L_2(P)$ norm)



Benign Overfitting in Linear Regression

(2019 Bartlett, Long, Lugosi, Tsigler)

$x \in \mathbb{H}$, $y \in \mathbb{R}$

(x, y) sub-Gaussian, mean zero, well-specified

$$\boxed{\mathbb{E}[y|x] = x^T\theta^*}$$

x satisfies a small ball condition

$\exists\, C>0.\ \Pr\left(\|x\|^2 > C\cdot\mathbb{E}[\|x\|]^2\right) \le \delta$

1) covariance matrix $\Sigma = \mathbb{E}[xx^T] = \sum_i \lambda_i v_i v_i^T$ $\quad (\lambda_1 \ge \lambda_2 \ge \cdots)$

variance on some direction

2) optimal OLS: $\theta^* = \arg\min_\theta \mathbb{E}[(y - x^T\theta)]^2$,

3) variance of noise: $\sigma^2 = \mathbb{E}[(y - x^T\theta^*)]^2$

Data: $X \in \mathbb{H}^n$, $y \in \mathbb{R}^n$

$$X = \begin{pmatrix} - x_1 - \\ - x_2 - \\ \vdots \\ - x_n - \end{pmatrix}, \quad y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

$$\hat{\theta} = (X^TX)^\dagger X^T y$$

min-norm solution

$$\min_{\theta \in \mathbb{H}} \|\theta\|^2 \quad \text{s.t. } \|X\theta - y\|^2 = \min_\theta \|X\beta - y\|^2$$

$$R(\hat{\theta}) = \mathbb{E}[(y - x^T\hat{\theta})^2] - \min_\theta \mathbb{E}[(y - x^T\theta)^2]$$

$$= \mathbb{E}[(y - x^T\hat{\theta})^2 - (y - x^T\theta^*)^2]$$

$$= \mathbb{E}_{(x,y)}[(2y + x^T(\theta^* + \hat{\theta}))\cdot x^T(\theta^* - \hat{\theta})]$$

$$= \mathbb{E}_x[\mathbb{E}_y[(2y - x^T(\theta^* + \hat{\theta}))\, x^T(\theta^* - \hat{\theta}) \,|\, x]]$$

$$= \mathbb{E}_x[(\theta^* - \hat{\theta})^T xx^T(\theta^* - \hat{\theta})] = (\hat{\theta} - \theta^*)^T \boxed{\Sigma} (\hat{\theta} - \theta^*)$$

$$\Sigma = \sum_i \lambda_i v_i v_i^T$$

$\mathbb{E}[y|x] = \theta^{*T}x$.

Σ measures the quality of the estimate.

$$\boxed{\begin{aligned} &\theta_0 = 0. \\ &\dot{\theta}_t = -\nabla_\theta \|X\theta - y\|^2. \end{aligned}}$$

Implicit Bias

Gradient Flow converges to min-norm

Regularization: ridge regression

$$\min \lambda\|\theta\|^2 + \frac{1}{n}\|X\theta - y\|^2 \iff \min \frac{1}{n}\|X\theta - y\|^2 \ \text{s.t. } \|\theta\| \le b \iff \min \|\theta\| \ \text{s.t. } \frac{1}{n}\|X\theta - y\|^2 \le c$$

norm constraint

fit constraint

Overfitting: $c \ll \min_\theta \mathbb{E}[(y - x^T\theta)^2]$

$\downarrow$

0

We consider $\min_\beta \|X\beta - y\|^2 = 0$

$$\hat{\theta} = (X^T X)^{\dagger} X^T y \quad : \quad \min_{\theta \in \mathbb{H}} \|\theta\|^2$$

$$\text{s.t. } \|X\theta - y\|^2 = \min_\beta \|X\beta - y\|^2 = 0 \rightarrow \begin{cases} y_1 = x_1^T \hat{\theta} \\ \vdots \\ x_n = x_n^T \hat{\theta} \end{cases}$$

$$x_1 = x_1^T \theta^* + \varepsilon_1 \qquad \vdots \qquad y_1 = x_n^T \theta^* + \varepsilon_n$$

When can the label noise be hidden in $\hat{\theta}$ w/o. hurting predictive accuracy ?

Thm b.c. any Linear Regression problem $(\theta^*, \sigma^2, \Sigma)$ $\rightarrow \lambda_n > 0$ (we can get a zero-empirical risk)

If $\boxed{k^*} = \min\{k \geq 0 : r_k(\Sigma) \geq bn\}$

effective dimension

① w.h.p : $R(\hat{\theta}) \leq C \cdot \left(\underbrace{\text{bias}(\theta^*, \Sigma, n)}_{\text{bias}} + \underbrace{\sigma^2 \cdot \left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)} \right)}_{\text{variance}} \right)$

(attributes to θ^*) (attributes to noise)

② $X = \Sigma^{\frac{1}{2}} Z$, Z Gaussian $\perp\!\!\!\perp$

θ^* symmetrized : $\mathbb{E}[R(\hat{\theta})] \geq \frac{1}{C}\left(\text{bias}(\theta^*, \Sigma, n) + \sigma^2 \cdot \min\left\{ \frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}, 1 \right\}^2 \right)$

- $\text{bias}(\theta^*, \Sigma, n) = \underbrace{\|\theta^*_{k+1:\infty}\|^2_{\boxed{\Sigma_{k+1:\infty}}}}_{\text{orthogonal subspace}} + \|\theta^*_{1:k}\|^2_{\Sigma^{-1}_{1:k}} \cdot \left(\frac{\sum_{i>k} \lambda_i}{n} \right)^2$

Effective Ranks : $\lambda_1 \geq \lambda_2 \geq \cdots$ eigenvalues of Σ

$k \geq 0$; $\lambda_{k+1} > 0$

$$r_k(\Sigma) = \frac{\sum_{i>k} \lambda_i}{\lambda_{k+1}} \quad \begin{matrix} 1\text{-norm} \\ \infty\text{-norm} \end{matrix}$$

$$R_k(\Sigma) = \frac{\left(\sum_{i>k} \lambda_i\right)^2}{\sum_{i>k} \lambda_i^2} \quad \begin{matrix} 1\text{-norm} \\ 2\text{-norm} \end{matrix}$$

throwing k highest-variance direction ⇓ remaining subspaces

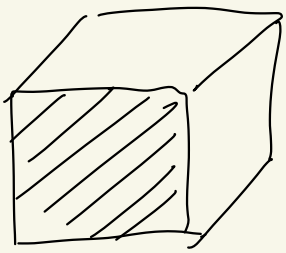
e.g.1 $\Sigma = I_p$. $r_0(I_p) = R_0(I_p) = p$.

e.g.2. $\text{rank}(\Sigma) = p$. $r_0(\Sigma) = \text{rank}(\Sigma) \cdot s(\Sigma)$

$$s(\Sigma) = \frac{\frac{1}{p} \cdot \sum_{i=1}^p \lambda_i}{\lambda_i} \quad \begin{matrix} \rightarrow 1\text{-norm} \\ \rightarrow \infty\text{-norm} \end{matrix}$$

$R_0(\Sigma) = \text{rank}(\Sigma) \cdot S(\Sigma)$.

$$S(\Sigma) = \frac{\left(\frac{1}{p} \sum_{i=1}^p \lambda_i\right)^2}{\frac{1}{p} \sum_{i=1}^p \lambda_i^2} \quad \begin{matrix} \rightarrow 1\text{-norm} \\ \rightarrow 2\text{-norm} \end{matrix}$$

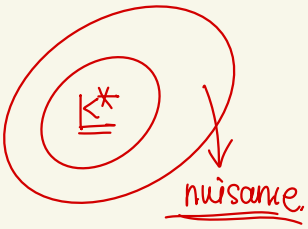


$\hat{f} = \hat{f}_0 + \Delta$

symmetry s. S $\left(\frac{1}{p}, 1\right)$

$\hat{f}_0$: predirection component (k^*-dim subspace corresponding to $\lambda_1, \cdots, \lambda_{k^*}$)

Δ : benign overfitting component (orthogonal subspace)



Δ is benign if $R_{k^*} \gg n$

$\boxed{\Sigma} \rightarrow$ { how the label noise is distributed in $\hat{\theta}$; how errors in $\hat{\theta}$ affect prediction accuracy }

To avoid harming prediction accuracy, the noise "energy" must be distributed across many "unimportant" directions.

$y_i = \theta^* x_i + \varepsilon_i$

$\{k^*\}$ $\{k^*\}^\perp$

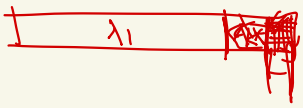
parameter $\uparrow$ $\Rightarrow$ "unimportant" directions $\uparrow$

1) $r_0(\Sigma) \downarrow$: # of non-zero eigenvalues $\gg n$ $\rightarrow$ bias $\downarrow$

2) $r_{k^*}(\Sigma) \uparrow$: # of 'small' eigenvalues $\gg n$

3) $R_{k^*}(\Sigma) \uparrow$: small eigenvalues roughly equal

$\Rightarrow$ variance $\downarrow$



tricky: bias asymptoticity

def $\Sigma_n = \sum_{i=1}^{n} \lambda_i v_i v_i^T$. asymptotically benign if

$$UB = \lim_{n\to\infty} \left(\|\Sigma_n\| \sqrt{\frac{r_0(\Sigma_n)}{n}} + \left(\frac{k_n^*}{n} + \frac{n}{R_{k_n^*}(\Sigma_n)} \right) \right) \rightarrow 0$$

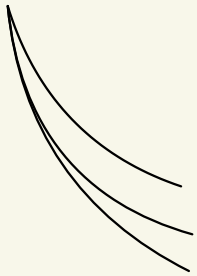
$\lambda_i \downarrow$ exp. $UB > 0$. } infinite dimension

$\lambda_i \downarrow$ poly

$$\lambda_i = \frac{1}{i^\alpha} \cdot \frac{1}{(\log(i+1))^\beta}$$

Σ benign iff $\alpha = 1$, $\beta > 1$:

$\sum_i \lambda_i$ almost diverge



① $k^* = \min\{k : r_k(\Sigma) \geq b \cdot n\} \leq \frac{n}{c_1}$

Bias

XX^T invertible

$\downarrow p$

Σ. $\lambda_{k^*} + \lambda_{k^*+1} + \cdots$

$\downarrow$

ridge regression.

② Noise $\frac{k^*}{n}$.

Variance

$$\hat{\theta} = X(XX^T)^{-1}\varepsilon + \cdots$$

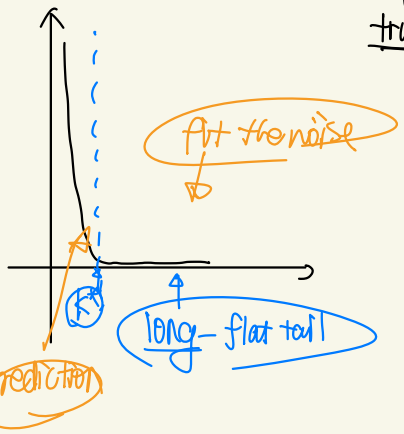
$$R(\hat{\theta}) = (\hat{\theta} - \theta^*)^T \Sigma (\hat{\theta} - \theta^*)$$

$$= \lambda_i \|v_i(\hat{\theta} - \theta^*)\|^2$$

$$\approx \lambda_i \|v_i X^T (XX^T)^{-1} \varepsilon\|^2$$

$$n \cdot \sum_{i > k^*} \lambda_i^2 / p^2 = R_{k^*}(\Sigma)$$

truncation $\infty \rightarrow p_n$ finite dimension



$$\lambda_{k,n} = \begin{cases} e^{-k} + \varepsilon_n & \text{if } k \leq p_n \\ 0 & \text{otherwise} \end{cases}$$

Σ_n is benign $\Leftrightarrow$ $\begin{cases} p_n = \omega(n) \\ \varepsilon_n p_n = o(n) \end{cases}$. $\varepsilon_n p_n = \omega(n e^{-n})$

noise energy

$$R(\hat{\theta}) = O\left(\frac{\varepsilon_n p_n}{n} + \max\left\{ \frac{1}{n}, \frac{n}{p_n} \right\} \right)$$

$\downarrow$ 0 $\downarrow$ 0

quickly converges λ_i + noise in all directions

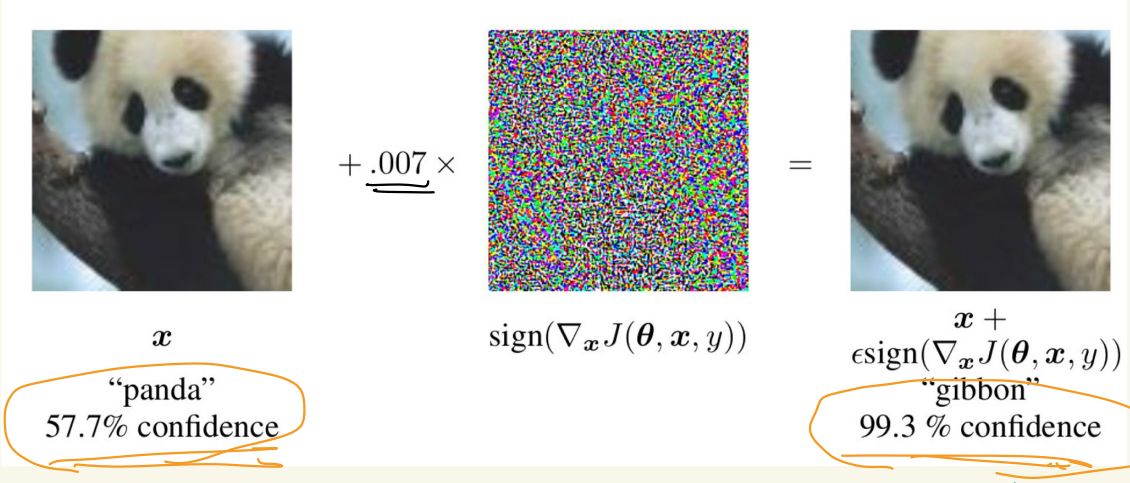
Limitation : Sensitivity

Δ = benign overfitting component spiky $\left\langle \begin{array}{l} \|\Delta\|_{L_\infty} : \text{large} \\ \|\Delta\|_{L_2(P)} : \text{small} \end{array} \right.$

↓

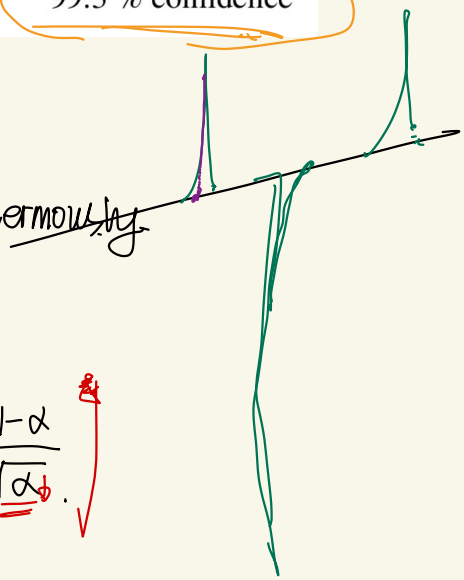
small on avg.

large w.c.



Label noise appears in $\hat{\theta}$: $\exists$ a unit norm Δ,

s.t. perturbing an input x by Δ changes the output enormously

even if $\Delta^T \theta^* = 0$. If $R(\hat{\theta}) \le \alpha$,

$$\left((x+\Delta)^T \hat{\theta} - x^T \hat{\theta}\right)^2 \ge c \cdot \frac{\sigma^2 (n-k^*)}{\sum_{i>k^*} \lambda_i} \ge c \cdot \frac{\sigma^2 \|\theta^*_{1:k^*}\|}{\sqrt{\lambda_1}} \cdot \frac{1-\alpha}{\sqrt{\alpha}}.$$