MAT280 ($\overset{\text{first}}{\text{M}}$L) - Lecture 1 - 4/4/17

Michael Wolf - Prof. at TUM

Mathematic Foundations of Machine Learning

Not about software/coding/implementation

Grading = 1/3 participation → in lecture, comments on lecture notes online
2/3 research paper presentation (just a paper) → 2/3 person team

What is Machine Learning? CS ⊇ AI ⊇ ML ⊇ Deep learning

Aim: Data → models/algorithm/program

used when: large amt. of data available, target fn. too difficult to implement directly.

Main branches - ★ Supervised ML - training data w/ labels as input
focus of course — outputs program to predict labels of unseen neighbors

Unsupervised ML: input data, output model/pattern in data

Supervised is predictive, Unsupervised in descriptive

Representation: (structure of programs)
- Decision trees
- Neural networks
- K-Nearest Neighbors
- SVMs

optimizable structures

Course Topics:
- Learning Theory
- Neural Networks
- SVMs & Kernel methods
- . . . (if time)

I. Learning Theory (supervised ML)

I.I. Statistical framework

input - "training data" $S = ((x_1, y_1), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$

output - "hypothesis" $h : \mathcal{X} \Rightarrow \mathcal{Y}$

ML algorithm $A : \bigcup_{n \in \mathbb{N}} (\mathcal{X} \times \mathcal{Y})^n \longrightarrow \mathcal{Y}^{\mathcal{X}}, \quad S \longmapsto h$

$\mathrm{range}(A) =: \mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ ↖ could work for any size training set $\forall n \in \mathbb{N}$

assumptions: (x_i, y_i) values of random i.i.d. vars. (X_i, Y_i)
(not always valid IRL) distributed according to some prob. measure P over $\mathcal{X} \times \mathcal{Y}$

Denote expectation value wrt. P as $\mathbb{E}$. Exp. value of S wrt. P^n as $\mathbb{E}_S$.

Goal: Find a "good" h with "loss function" $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$

minimize the "risk": $\mathcal{R}(h) := \int_{\mathcal{X} \times \mathcal{Y}} L(y, h(x)) \, dP(x, y)$

↑ dont know this distributi

Challenge: P is unknown.

Regression: $\mathcal{Y}$ is continuous. If $\mathcal{Y} = \mathbb{R}$, the most common loss fn. is

$$L(y, y') := |y - y'|^2 \quad \text{(quadratic loss)}$$

Then risk is $\mathcal{R}(h) = \mathbb{E}\left[|Y - h(X)|^2 \right]$ (mean squared error

Classification: $\mathcal{Y}$ is discrete. h called "classifier") choice may al be det. by appli

Most Common loss fn: "0-1 loss" $L(y, y') := 1 - \delta_{yy'}$

Risk $\mathcal{R}(h) = \mathbb{P}[h(X) \neq Y] = \mathbb{E}[\mathbb{1}_{h(X) \neq Y}]$

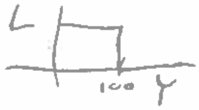
↑ indicator fn.

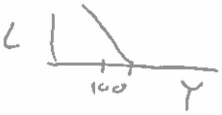
B. classification [illegible] $\mathcal{Y} = \{-1, 1\}$

MAT280 - Lecture 1 Continued - 4/4/17

Choice of loss fn is determined by problem goals / optimizer constraints

↳ E.g. step fn. for goal of $Y \geq 100$, make convex for optimizer





1.2 Error Decomposition

- Prior knowledge is encoded in $\mathcal{F}$ (target hyp. space - det. by model choice)
- Can't minimize risk, minimize empirical risk - ERM

$$\hat{R}(h) := \frac{1}{n} \sum_{i=1}^{n} L(y_i, h(x_i))$$

Errors:

$$R_{\mathcal{F}} := \inf_{h \in \mathcal{F}} R(h) \quad , \quad R_{Y^X} := \inf_{h \in Y^X} R(h)$$

↑ best we can do w/ model choice

↑ best we can do in principle

$$R(h) - R_{Y^X} = \underbrace{(R(h) - R_{\mathcal{F}})}_{\text{estimation error}} + \underbrace{(R_{\mathcal{F}} - R_{Y^X})}_{\text{approximation error}}$$

← independent of S
← how model choice fits

Do ERM, get $\hat{h} \in \mathcal{F}$.

Then $\hat{R}(\hat{h}) \leq \hat{R}(h) \; \forall h \in \mathcal{F}$.

since $R_{\mathcal{F}} = \inf(\text{stuff})$, $-R_{\mathcal{F}} = \sup(-\text{stuff})$

Est. error: $R(\hat{h}) - R_{\mathcal{F}} = R(\hat{h}) - \hat{R}(\hat{h}) + \sup_{h \in \mathcal{F}} (\hat{R}(\hat{h}) - R(h))$

$$\leq 2 \sup_{h \in \mathcal{F}} \underbrace{|R(h) - \hat{R}(h)|}_{\text{generalization error}}$$

d: $R(\hat{h}) - R_{\mathcal{F}} \leq 2 \sup_{h \in \mathcal{F}} |R(h) - \hat{R}(h)|$

↑↑ ERM within $\mathcal{F}$

MAT280 - Lecture 2 - 4/6/17

Ex 1 - Linear Regression: $\mathcal{X} \times \mathcal{Y} = \mathbb{R}^d \times \mathbb{R}$

$$\bar{\mathcal{F}} := \{ h : \mathbb{R}^d \to \mathbb{R} \mid \exists v \in \mathbb{R}^d : h(x) = \langle v, x \rangle \}$$

$$\hat{\mathcal{R}}(v) := \frac{1}{n} \sum_{i=1}^{n} | \langle v, x_i \rangle - y_i |^2$$

$$\nabla \hat{\mathcal{R}}(v) \overset{!}{=} 0 \iff \frac{\partial}{\partial v_k} \hat{\mathcal{R}}(v) = \frac{1}{n} \sum_{i=1}^{n} (\langle v, x_i \rangle - y_i) x_{i,k} = 0 \quad \forall k$$

$\hookrightarrow v = A^{-1} b$ is the ERM $\iff Av = b$ w/ $A := \sum_{i=1}^{n} x_i x_i^T$

$b := \sum_{i=1}^{n} y_i x_i$

Ex. 2 - Polynomial Regression: $\mathcal{X} \times \mathcal{Y} = \mathbb{R} \times \mathbb{R}$, $\bar{\mathcal{F}} := \{ h : \mathbb{R} \to \mathbb{R} \mid \exists a \in \mathbb{R}^{m+1} : h(x) = \sum_{k=0}^{m} a_k x^k \}$

$\Psi : \mathbb{R} \to \mathbb{R}^{m+1}$, $\Psi(x) := (1, x, x^2, \ldots, x^m)$

then $\hat{\mathcal{R}}(v) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \sum_{k=0}^{m} a_k x^k \right)^2 = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \langle a, \Psi(x_i) \rangle \right)^2$

then do gradient & whatnot (may not get explicit ERM)

Can use this trick for any set of linear basis functions.

ERM $\Rightarrow Av = b$ w/ $A = \sum_{i=1}^{n} \Psi(x_i) \Psi(x_i)^T$, $b = \sum_{i=1}^{n} y_i \Psi(x_i)$ ← maybe not invertible

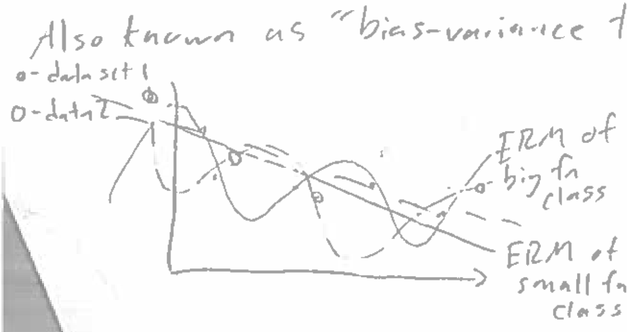
---

Approx. error $:= \mathcal{R}_{\bar{\mathcal{F}}} - \mathcal{R}_{\mathcal{Y}\mathcal{X}}$ ← Inf. risk in fn. class - Inf risk overall fns.

Estimation error $- \mathcal{R}(h) - \mathcal{R}_{\bar{\mathcal{F}}}$ ← risk of hypothesis - inf risk of fn. class

$\lvert\bar{\mathcal{F}}\rvert \nearrow$	$\lvert\bar{\mathcal{F}}\rvert \searrow$
↘	↗
↗	↘

Also known as "bias-variance trade-off"



Do high-deg poly fits on random draws of sample, average will have small/zero bias, but sample-to-sample polynomials vary widely.

Linear fits on random draws will have little variance, but bias stays big.

MAT280 Lecture 2 cont'd

Approaches aiming @ a balanced choice for $\bar{\mathcal{F}}$:

- Split data $\to$ training data $\to$ optimizing hypothesis $\to h_S$
  $\to$ test data $\to$ evaluate performance of h_S
  $\to$ validation data $\to$ tune hyperparameters

- Modification of ERM $\to$ structural risk minimization
  Use $\bar{\mathcal{F}}_1 \subset \bar{\mathcal{F}}_2 \subset \bar{\mathcal{F}}_3 \subset \dots$
  & penalize higher levels

  $\downarrow$

  Regularization – minimize $\hat{\mathcal{R}}(h) + \Omega(h)$

  e.g. "Tikhonov reg." $\Omega(h) = \lambda \|h\|^2$, $\lambda \in \mathbb{R}_+$ chosen by cross-validation.

1.3 PAC Learning (Probably Approximately Correct)

$\hookrightarrow$ introduced by Vailant in 1984

Desirable: find uniform bound on $|\mathcal{R}(h) - \hat{\mathcal{R}}(h)| \leq \dots$

Always have chance of unfair training data! No deterministic bound

Want: $\mathbb{P}_S\left[|\hat{\mathcal{R}}(h) - \mathcal{R}(h)| > \varepsilon\right] < \delta$

Take $\mathcal{X}, \mathcal{Y}$ finite, choose 0-1 loss & assume $\overset{\in\bar{\mathcal{F}}}{f}: \mathcal{X} \to \mathcal{Y}$ determines "true label"

so $S = (x_i, f(x_i))$ and $P(x,y) = \delta_{y, f(x)}\, p(x)$

Lemma: $\forall \varepsilon > 0$, $\forall h \in \mathcal{Y}^{\mathcal{X}}$, $\overset{\mathbb{P}[h(x)\neq f(x)]}{\overset{\shortparallel}{\mathcal{R}(h)}} > \varepsilon \Rightarrow \mathbb{P}_S[\hat{\mathcal{R}}(h) = 0] < e^{-\varepsilon n}{}^{<\text{size of } \delta}$

Pf:

$$\mathbb{P}_S[\hat{\mathcal{R}}(h) = 0] = \mathbb{P}_S[\forall i \in \{1,\dots,n\}: h(x_i) = f(x_i)] \overset{\text{i.i.d.}}{=} \prod_{i=1}^{n} \mathbb{P}[h(x_i) = f(x_i)] \leq (1-\varepsilon)^n$$

$$\leq e^{-\varepsilon n}$$

$\square$

Thm: $\mathbb{P}_S\left[|R(h_S)-\hat{R}(h_S)|>\varepsilon\right]<\delta$ if $n\geq\frac{1}{\varepsilon}\left(\ln|\mathcal{F}|+\ln\frac{1}{\delta}\right)$

under the assumption $\forall S$ $\exists h_S\in\mathcal{F}: \hat{R}(h_S)=0$.

Pf: $\mathbb{P}_S\left[|R(h_S)-\hat{R}(h_S)|>\varepsilon\right]=\mathbb{P}_S\left[R(h_S)>\varepsilon\right]$

$$\leq \mathbb{P}_S\left[\exists h\in\mathcal{F}: R(h)>\varepsilon \wedge \hat{R}(h)=0\right]$$

union bound → $$\leq \sum_{h\in\mathcal{F}:R(h)>\varepsilon}\mathbb{P}_S\left[\hat{R}(h)=0\right]$$

Lemma → $$\leq \sum_{h\in\mathcal{F}:R(h)>\varepsilon} e^{-\varepsilon n} \leq |\mathcal{F}|e^{-\varepsilon n}=\delta$$ solve for n get

MAT480 - Lec 3 - 7/11/14

Assumptions:

- $\mathbb{P}[Y=y \mid X=x] = \delta_{y,f(x)}$
- $\forall S = ((x_i, f(x_i)))_{i=1}^{n} : \exists h_S \in \mathcal{F} : \hat{R}(h_S) = 0$

$$\Bigg\} \Rightarrow \mathbb{P}_S[R(h_S) > \varepsilon] \leq \begin{cases} (1-\varepsilon)^n |\mathcal{F}| \\ \delta \text{ if} \\ n \geq \frac{1}{\varepsilon}\ln\left(\frac{|\mathcal{F}|}{\delta}\right) \end{cases}$$

Rmk: • These $\mathbb{P}$ bounds assume $\mathcal{F}$ is finite.

• If $\mathcal{F} = \mathcal{Y}^{\mathcal{X}}$, then $n > |\mathcal{X}|$ but then there's nothing to generalize to, meaningless!

1.4 No Free Lunch

Thm: Let $\mathcal{X}, \mathcal{Y}$ be finite, $|\mathcal{X}| > n$. $R_f(h) := \mathbb{P}[h(x) \neq f(x)]$ $\quad (\mathcal{F} = \mathcal{Y}^{\mathcal{X}})$

$S \mapsto h_S \in \mathcal{F}$: $\mathbb{E}_f\left[\mathbb{E}_S[R_f(h_S)]\right] \geq \left(1 - \frac{1}{|\mathcal{Y}|}\right)\left(1 - \frac{n}{|\mathcal{X}|}\right)$

if uniform dist. over f and x are x used

Pf:

$$\mathbb{E}_f \mathbb{E}_S[R_f(h_S)] = \frac{1}{|\mathcal{X}|}\mathbb{E}_f \mathbb{E}_S\left[\sum_{x \in \mathcal{X}} \mathbb{1}_{h_S(x) \neq f(x)}\right]$$

$$\geq \frac{1}{|\mathcal{X}|}\mathbb{E}_f \mathbb{E}_S\left[\sum_{x \notin \mathcal{X}_S} \text{''}\right] \quad \text{where } \mathcal{X}_S \subset \mathcal{X} \text{ appearing in } S.$$

$$\vdots$$

$$\geq \frac{1}{|\mathcal{X}|}\sum_{x \notin \mathcal{X}_S}\left(1 - \frac{1}{|\mathcal{Y}|}\right) = \left(1 - \frac{1}{|\mathcal{Y}|}\right)\left(\frac{|\mathcal{X}| - n}{|\mathcal{X}|}\right) = \left(1 - \frac{1}{|\mathcal{Y}|}\right)\left(1 - \frac{n}{|\mathcal{X}|}\right)$$

Random guessing $\rightarrow$ $\left(1 - \frac{1}{|\mathcal{Y}|}\right)$. Sophisticated alg. only $\cdot\left(1 - \frac{n}{|\mathcal{X}|}\right)$ better

this additional factor reflects the fact that training data is already known.

No "better" learners for all data sets – neural nets perform better than decision trees on some data sets but worse on others

– on average, all learners are no better than random guessing!

Need to restrict fn. class $\mathcal{F}$ a priori.

Cardinality $\not\Leftrightarrow$ complexity of $\mathcal{F}$

E.g. $\mathcal{X} = \{x_1, x_2, \ldots\}$, $\infty > |\mathcal{Y}|, \mathcal{Y} \ni 0$, $\mathcal{F} := \bigcup_{n \in \mathbb{N}} \{f : \mathcal{X} \to \mathcal{Y} : \forall m > n : f(x_m) = 0\}$ ← countable but infly complex!

1.5 Growth function:

Def: $|\mathcal{Y}| < \infty$, $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$. For every $C \subseteq \mathcal{X}$ define

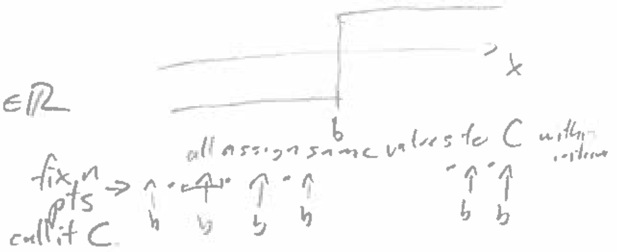
$$\mathcal{F}_C := \{ f \in \mathcal{Y}^C \mid \exists F \in \mathcal{F}\ \forall x \in C : F(x) = f(x) \}$$

~~The "growth function" $\Gamma : \mathbb{N} \to \mathbb{N}$ assigned to $\mathcal{F}$ is $\Gamma(n) := \max_{C \subseteq \mathcal{X} : |C| = n} |\mathcal{F}_C|$~~

Note: $\Gamma(n) \leq |\mathcal{Y}|^n$

Ex: $\mathcal{F} \subseteq \{-1, 1\}^{\mathbb{R}}$, $\mathcal{F} := \{ x \mapsto \mathrm{sgn}[x - b] \}_{b \in \mathbb{R}}$

$\to \Gamma(n) = n + 1$



Lemma: [Hoeffding's ineq. '63] Consider $Z_1, \ldots, Z_n$ real indep. rand. vars. w/ $\mathrm{range}(Z_i) \subseteq [a_i, b_i]$. Then $\forall \varepsilon > 0$,

$$\mathbb{P}\left[\sum_{i=1}^{n} (Z_i - \mathbb{E}[Z_i]) \geq \varepsilon \right] \leq \exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^{n} (a_i - b_i)^2} \right)$$

$$\frac{1}{2} \mathbb{P}\left[\left| \sum_{i=1}^{n} (Z_i - \mathbb{E}[Z_i]) \right| \geq \varepsilon \right] \leq \ldots$$

$$\frac{1}{2} \mathbb{P}\left[\left| \sum_{i=1}^{n} (Z_i - \underset{\text{independent copies of } Z_i}{\underset{\uparrow}{Z_i'}}) \right| \geq \varepsilon \right] \leq \ldots$$

Thm: $|\mathcal{Y}| < \infty$, $\mathrm{range}(L) \subseteq [0, c]$, $\delta \in [0, 1]$.
With prob. at least $1 - \delta$ wrt. repeated sampling of training data of size $n \in \mathbb{N}$:

$$\forall h \in \mathcal{F} : |R(h) - \hat{R}(h)| \leq c \sqrt{\frac{8 \ln(\Gamma(2n) \frac{4}{\delta})}{n}}$$

Pf: S, S' are i.i.d. rand. vars. over $(\mathcal{X} \times \mathcal{Y})^n$ w/ distr. according to P^n.

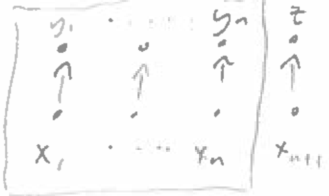
$|R(h) - \hat{R}(h)| > \varepsilon$, $|\hat{R}'(h) - \hat{R}(h)| < \varepsilon/2 \Rightarrow |R(h) - \hat{R}'(h)| > \varepsilon/2$ by Δ-ineq.

$$\mathbb{1}_{|R(h) - \hat{R}(h)| > \varepsilon} \mathbb{1}_{|R(h) - \hat{R}'(h)| < \varepsilon/2} \leq \mathbb{1}_{|\hat{R}'(h) - \hat{R}(h)| > \varepsilon/2}$$

assume $n \geq \frac{4c^2}{\varepsilon^2} \ln 2$. $\mathbb{E}_{S'}\left[\mathbb{1}_{|R(h) - \hat{R}'(h)| < \varepsilon/2} \right] \geq 1 - 2\exp\left[-\frac{\varepsilon^2 n}{2c^2} \right] \geq 1/2$.

MAT 2400 - Lec 4 - 4/13/17

Recall: $\Gamma(n) := \max_{C \subseteq X,\ |C| = n} \{ |\mathcal{F}_C| \}$



need restriction on fn class to use info in ~~box~~ !

Thm: ... $\forall h \in \mathcal{F}: |R(h) - \hat{R}(h)| \le c \sqrt{\frac{8 \ln(\Gamma(2n) \, 4/\delta)}{n}}$

Pf: Last time, showed $\mathbb{P}_S[\exists h \in \mathcal{F}: |R(h) - \hat{R}(h)| > \varepsilon]$

$$\le 2 \mathbb{P}_{SS'}[\exists h \in \mathcal{F}: |\hat{R}'(h) - \hat{R}(h)| > \varepsilon/2]$$

Note: $2\mathbb{P}_{SS'}[\exists h \in \mathcal{F}: |\hat{R}'(h) - \hat{R}(h)| > \varepsilon/2] = 2\mathbb{P}_{SS'}[\exists h \in \mathcal{F}: \frac{1}{n}|\sum_{i=1}^{n} L(y_i, h(x_i)) - L(y_i', h(x_i'))| > \varepsilon/2]$

Mult. by $\underset{\text{unif. dist.}}{\sigma_i} = \{\pm 1\}$, since S & S' are i.i.d, $= 2\mathbb{P}_{SS'}[\exists h \in \mathcal{F}: \frac{1}{n}|\sum_{i=1}^{n} [L(y_i, h(x_i)) - L(y_i', h(x_i'))]\sigma_i| > \varepsilon/2$

$$= 2\mathbb{E}_{SS'}\left[\mathbb{P}_\sigma\left[\exists h \in \mathcal{F}_{|S \cup S'}: \frac{1}{n}\Big|\sum_{i=1}^{n} \cdots\Big| > \varepsilon/2\right]\right]$$

$$\underset{\text{union bound}}{\le} 2\mathbb{E}_{SS'}\left[\sum_{h \in \mathcal{F}_{|S \cup S'}} \mathbb{P}_\sigma\left(\frac{1}{n}\Big|\underbrace{\cdots}_{\sum_{i=1}^{n} z_i}\Big| > \varepsilon/2\right]\right]$$

$\leftarrow$ exp. of $z_i = 0$ b/c of σ_i

$$= 2\mathbb{E}_{SS'}\left[\sum_{h \in \mathcal{F}_{|S \cup S'}} \mathbb{P}_\sigma\left[\Big|\sum_{i=1}^{n} z_i\Big| > n\varepsilon/2\right]\right]$$

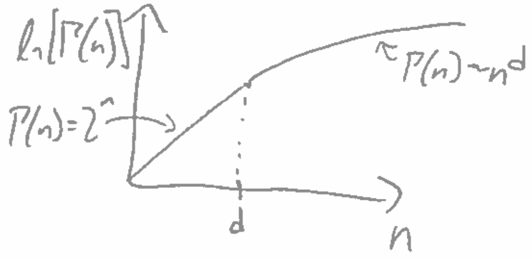
$$\underset{\text{Hoeffding}}{\le} 4 \underbrace{\mathbb{E}_{SS'}[|\mathcal{F}|_{S \cup S'}]}_{\le \Gamma(2n)} \exp\left[-\frac{n\varepsilon^2}{8c^2}\right] =: \delta \;\rightarrow \text{solve for } \varepsilon.$$

... assumption $\Longleftrightarrow$ $\delta \le 2\sqrt{2}\,\Gamma(2n)$ ✓ since $\delta \in (0,1)$ □

Lemma: $\mathcal{F}_1 \subseteq \mathcal{Y}^{\mathcal{X}}$, $\mathcal{F}_2 \subseteq \mathcal{Z}^{\mathcal{Y}}$, $\mathcal{F} := \mathcal{F}_2 \circ \mathcal{F}_1 \Rightarrow \Gamma(n) \le \Gamma_1(n)\Gamma_2(n)$

1.6 VC-Dimension:

Consider $|\mathcal{Y}| = 2$.



Def: The VC (Vapnik-Chervonenkis) dimension of $\bar{f} \subseteq \mathcal{Y}^n$ w/ $|\mathcal{Y}|=2$ is $VCdim(\bar{f}) := \max\{n \in \mathbb{N} : \Gamma(n) = 2^n\}$ (∞ if max DNE)

Ex:

threshold fns. $\Gamma(n) = n+1 \Rightarrow VCdim(\bar{f}) = 1$

Thm: Let $d := VCdim(\bar{f})$. Then $\Gamma(n) \begin{cases} = 2^n & \text{if } n \leq d \\ \leq \left(\frac{en}{d}\right)^d & \text{if } n > d \end{cases}$

Pf: Assume $n > d$ ($n \leq d$ pf. by definition)

WTS: $\forall A \subseteq X$ w/ $|A| = n$. $|\bar{f}_{|A}| \overset{?}{\leq} |\{B \subseteq A \mid \bar{f}_{|B} = y^B\}|$.

pf. by induction over n

$$\leq |\{B \subseteq A \mid |B| \leq d\}|$$

$$= \sum_{i=0}^{d} \binom{n}{i} \leq \sum_{i=0}^{n} \binom{n}{i} \left(\frac{n}{d}\right)^{d-i}$$

$$= \left(\frac{n}{d}\right)^d \left(1 + \frac{d}{n}\right)^n \leq \left(\frac{n}{d}\right)^d e^d.$$

Pick $a \in A$ & define

$\bar{f}' := \{h \in \bar{f}_{|A} \mid \exists g \in \bar{f}_{|A} : h(a) \neq g(a) \wedge (h-g)|_{A \setminus a} = 0\}$.

$\bar{f}_a := \bar{f}'|_{A \setminus a}$.

Note $|\bar{f}_{|A}| = |\bar{f}_{|A \setminus a}| + |\bar{f}_a|$.

$|\bar{f}_{|A \setminus a}| \leq |\{B \subseteq A \mid \bar{f}_{|B} = y^B \wedge a \notin B\}|$ by IH.

$|\bar{f}_a| = |\bar{f}'_{|A \setminus a}| \leq |\{B \subseteq A \setminus a \mid \bar{f}'_{|B} = y^B\}|$ by IH.

$= |\{B \subseteq A \setminus a \mid \bar{f}'_{|B \cup a} = y^{B \cup a}\}|$ by def. of $\bar{f}'$

$= |\{B \subseteq A \mid \bar{f}'_{|B} = y^B \wedge a \in B\}|$

$\leq |\{B \subseteq A \mid \bar{f}_{|B} = y^B \wedge a \in B\}|$

add together
$a \notin B$
cancels/sums
with
$a \in B$

$\Rightarrow |\bar{f}_{|A}| \leq |\{B \subseteq A \mid \bar{f}_B \in y^B\}|$ $\square$

Cor: $(\varepsilon, \delta) \in (0,1]^2$. R = error prob. Then $\forall h \in \bar{f}: |\hat{R}(h) - R(h)| \leq \varepsilon$ holds w/ prob $\geq 1 - \delta$ if $n \geq \frac{32}{\varepsilon^2}\left[d \ln\left(\frac{8d}{\varepsilon^2}\right) + \ln\left(\frac{6}{\delta}\right)\right]$

$\rightsquigarrow$ "$n \geq d/\varepsilon^2$ scaling"

Thm: Let G be an $\mathbb{R}$-vector space of fns. from $\mathcal{X} \to \mathbb{R}$

Then $\mathcal{F} := \{ x \mapsto \operatorname{sgn}[g(x)] \mid g \in G \} \subseteq \{\pm 1\}^{\mathcal{X}}$ satisfies

$$\operatorname{VCdim}(\mathcal{F}) = \dim(G).$$

Pf: let $k = \dim(G) + 1$ & assume $\operatorname{VCdim}(\mathcal{F}) \geq k$.

$\exists \{x_1, \ldots, x_k\} = C \subseteq \mathcal{X}$ s.t. $\mathcal{F}|_C = \{-1, 1\}^C$

$L : G \to \mathbb{R}^k \ : \ L(g) := (g(x_1), \ldots, g(x_k))$

$\dim(\operatorname{range}(L)) \leq \dim(G) \Rightarrow \exists v \neq 0 \in (\operatorname{range}(L))^{\perp} = \ker(L^*)$

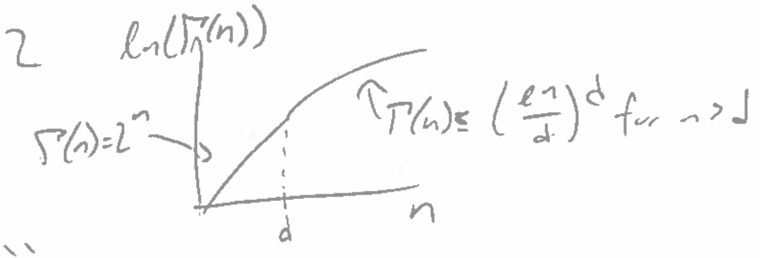
Then $\forall g \in G : 0 = \langle L^*(v), g \rangle = \langle v, L(g) \rangle = \sum_{i=1}^{k} v_i \, g(x_i)$

If $\mathcal{F}|_C = \{-1, 1\}^C$, we can choose g st. $\operatorname{sgn}(g(x_i)) = \operatorname{sgn}(v_i)$

& then $\sum_{i=1}^{k} v_i \, g(x_i) > 0$ ↯. $\square$

MAT280 - Lec 5 - 4/18/17

Recap: $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$, $|\mathcal{Y}| = 2$



"$n \geq \frac{1}{\epsilon^2}\left(d \overset{\text{WTS}}{\ln\left(\frac{d}{\epsilon^2}\right)} + \ln \frac{1}{\delta}\right)$" $\Rightarrow \forall h \in \mathcal{F} : |R(h) - \hat{R}(h)| \leq \epsilon$ w/ prob. $1-\delta$

$\Rightarrow$ If $\mathcal{F} \subseteq$ halfspaces in $\mathbb{R}^m \Rightarrow \operatorname{VCdim}(\mathcal{F}) = m+1$.

Ex: $\mathcal{F}$ is indicator fns of · Euclidean balls in $\mathbb{R}^m \to \operatorname{VCdim}(\mathcal{F}) = m+1$

· axis-aligned boxes in $\mathbb{R}^m \to \operatorname{VCdim}(\mathcal{F}) = 2m$ also no. of parameters

counterex: $\mathcal{F} := \{ x \in \mathbb{R} \mapsto \operatorname{sgn}(\sin(\alpha x)) \mid \alpha \in \mathbb{R} \}$. $\operatorname{VCdim}(\mathcal{F}) = \infty$

pf. idea: Choose set of pts. $A = \{2^{-k} : k = 1, \ldots, n\}$ & show can hit each pt w/ opposite sine signs using arbitrarily large frequency α.

1.7 - Fundamental Theorem of binary classification (qualitative version)

Def: $\mathrm{poly}(\frac{1}{\varepsilon}, \frac{1}{\delta}) :=$ fns of the form $(0,1]^2 \ni (\varepsilon, \delta) \mapsto r(\varepsilon, \delta) \in \mathbb{R}^+$
polynomial in both $\frac{1}{\varepsilon}$ & $\frac{1}{\delta}$.

Thm: Consider $\mathcal{F} \in \{\pm 1\}^{\mathcal{X}}$ $n := |S|$ - training data set, $S \sim P^n$.

Let $\mathcal{R}$ be the error prob. Then TFAE:

(1) $\mathrm{VCdim}(\mathcal{F}) < \infty$

(2) $\exists r \in \mathrm{poly}(\frac{1}{\varepsilon}, \frac{1}{\delta})$ s.t. $\forall (\varepsilon, \delta) \in (0,1]^2$ $\forall P$:
$$n \geq r(\varepsilon, \delta) \Rightarrow \mathbb{P}_S[\exists h \in \mathcal{F} : |\hat{\mathcal{R}}(h) - \mathcal{R}(h)| \geq \varepsilon] \leq \delta.$$

(3) $\exists r \in \mathrm{poly}(\frac{1}{\varepsilon}, \frac{1}{\delta})$ & a learner $S \mapsto h_S \in \mathcal{F}$ s.t. $\forall (\varepsilon, \delta) \in (0,1]^2$, $\forall P$:
$$n \geq r(\varepsilon, \delta) \Rightarrow \mathbb{P}[|\mathcal{R}(h_S) - \mathcal{R}_{\mathcal{F}}| \geq \varepsilon] \leq \delta.$$

(4) ... as (3) but w/ $S \mapsto h_S$ is ERM.

Pf: (1) $\Rightarrow$ (2) ✓ by Cor. on. 4(3), (2) $\Rightarrow$ (4) (via Error Decomposition) ✓, (4) $\Rightarrow$ (3) by def. ✓

WTS: (3) $\Rightarrow$ (1) by no free lunch thm & contradiction

Choose $\varepsilon = \delta = 1/4$. Let $n = r(\varepsilon, \delta)$ & supp. $\mathrm{VCdim}(\mathcal{F}) = \infty$.

$\forall N \in \mathbb{N}$ $\exists C \subseteq \mathcal{X} : |C| = N$ s.t. $|\mathcal{F}_{|C}| = 2^N \Rightarrow \mathcal{F}_{|C} = \{\pm 1\}^C$. (by def. of VCdim

From No-Free-Lunch-thm, $\exists f : C \to \{\pm 1\}$ and $P(x,y) := (\mathbb{1}_{x \in C \wedge f(x) = y})/N$
over $\mathcal{X} \times \{-1, 1\}$ wrt. which $\mathbb{E}_S[\mathcal{R}(h_S)] \geq \frac{1}{2}(1 - \frac{n}{N}) = (1 - \frac{1}{|\mathcal{Y}|})(1 - \frac{n}{|C|})$

but $\mathbb{E}_S[\mathcal{R}(h_S)] \leq \mathbb{P}_S[\mathcal{R}(h_S) \geq \varepsilon] + \varepsilon(1 - \mathbb{P}_S[\mathcal{R}(h_S) \geq \varepsilon])$ combine

$\Rightarrow \mathbb{P}_S[\mathcal{R}(h_S) \geq \frac{1}{4}] \geq \frac{1}{3} - \frac{2n}{3N}$ ↯ $\delta = 1/4$ b/c $\mathcal{R}_{\mathcal{F}} = 0$. □

Quantitative version yields "sample complexity" of $\mathcal{F}$:
$$r_{\mathcal{F}}(\varepsilon, \delta) = \Theta\left(\frac{\mathrm{VCdim}(\mathcal{F}) + \ln \frac{1}{\delta}}{\varepsilon^2}\right)$$

Lec 5 7/18 con't'd

1.8 Rademacher complexities:

not nec $\mathbb{Z}$, any set!

Def: Consider a set of real-valued fns. $\mathcal{G} \subseteq \mathbb{R}^{\mathbb{Z}}$ & a vector $z \in \mathbb{Z}^n$. The "empirical Rademacher complexity" of $\mathcal{G}$ w.r.t. z is $\hat{\underline{R}}(\mathcal{G}) := \mathbb{E}_\sigma\left[\frac{1}{n} \sup_{g \in \mathcal{G}} \sum_{i=1}^n \sigma_i g(z_i)\right]$, $\sigma \in \{-1,1\}^n$ uniform "Rademacher var"

If z_i's are i.i.d r.v.s. then the "Rademacher complexity" is defined as

$$\underline{R}_n(\mathcal{G}) := \mathbb{E}(\hat{\underline{R}}(\mathcal{G})).$$

with $g(z) = (g(z_1), \dots, g(z_n))$. $\hat{\underline{R}}(\mathcal{G}) = \frac{1}{n} \mathbb{E}_\sigma \sup_{g \in \mathcal{G}} \langle \sigma, g(z) \rangle$

will be big if $\mathcal{G}$ is "rich enough" for $g(z)$ to align w/ random signs.

Lemma: (McDiarmid's Ineq.)

Let $(z_1, \dots, z_n) =: z$ be indep. rand. vars. w/ values in $\mathbb{Z}$, and $\varphi : \mathbb{Z}^n \to \mathbb{R}$ s.t. $|\varphi(z) - \varphi(z')| \leq c_i$ whenever z & z' differ only in the i^{th} component.

Then $\forall \varepsilon > 0$, $\mathbb{P}[\varphi(z) - \mathbb{E}[\varphi(z)] \geq \varepsilon] \leq \exp\left\{-\frac{2\varepsilon^2}{\sum_{i=1}^n c_i^2}\right\}$.

Lec 6 - 4/20/17

Want to bound risk by empirical risk: $|R(h) - \hat{R}(h)| \le$ "growth fn" $\le$ "VC dim" (for classification)

" $\le$ "Rademacher compl." $\le$ "emp. Rademach"

" $\le$ "L_2-covering number" $\le$ "L_1-covering #"

$\le$ "L_1-packing"

$\le$ "Γ" $\le$ "VC-dim" (beats other VC bound)

Recall: $\mathcal{G} \subseteq \mathbb{R}^{Z}$, $z \in Z^n$, Z = some set (not integers $\mathbb{Z}$)

"Empirical Rad. compl" $\underline{\hat{R}}(\mathcal{G}) := \mathbb{E}_\sigma \left[\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} g(z_i)\sigma_i \right]$

Def: (Rademacher complexity) $\mathbb{E}_Z[\underline{\hat{R}}(\mathcal{G})] =: \underline{R}_n(\mathcal{G})$ ← Rademacher complexity

Lemma: $\mathcal{G} \subseteq [a,b]^Z$. For any $\varepsilon > 0$, prob. measure P^n on Z^n:

$$\mathbb{P}_Z\left[\underline{R}_n(\mathcal{G}) - \underline{\hat{R}}(\mathcal{G}) \ge \varepsilon\right] \le \exp\left[-\frac{2n\varepsilon^2}{(b-a)^2}\right]$$

Pf: $\varphi: Z^n \to \mathbb{R}$, $\varphi(z) := \underline{\hat{R}}(\mathcal{G})$. So $\mathbb{E}_Z[\varphi(z)] = \underline{R}_n(\mathcal{G})$.

Let $z, z' \in Z^n$ differ in only one component. Then for fixed σ_i,

$\sup_{g \in \mathcal{G}} \sum_i \sigma_i g(z_i)$ changes at most by $|a-b|$

Thus $|\varphi(z) - \varphi(z')| = |\underline{\hat{R}}_z(\mathcal{G}) - \underline{\hat{R}}_{z'}(\mathcal{G})| \le \frac{|b-a|}{n}$

Then the result follows from McDiarmid's inequality. □

Thm: $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$, $L: \mathcal{Y} \times \mathcal{Y} \to [0,c]$, $Z := \mathcal{X} \times \mathcal{Y}$.

$\mathcal{G} := \{(x,y) \mapsto L(y, h(x)) \mid h \in \mathcal{F}\} \subseteq [0,c]^Z$. For any $\delta > 0$ and prob measure P on Z, we have w/prob at least $1-\delta$ wrt repeated sampling of $S \in (\mathcal{X} \times \mathcal{Y})^n$, dist. according to P^n:

$$\forall h: \quad R(h) - \hat{R}(h) \le \begin{cases} 2\underline{R}_n(\mathcal{G}) + c\sqrt{\dfrac{\ln 1/\delta}{2n}} \\ 2\underline{\hat{R}}(\mathcal{G}) + 3c\sqrt{\dfrac{\ln 2/\delta}{2n}} \end{cases}$$

using Lemma

Lemma: $\mathcal{F} \subseteq \{-1,1\}^{\mathcal{X}}$, $L = 0\text{-}1$ loss, $S = ((x_i, y_i))_{i=1}^{n}$, $S_{\mathcal{X}} := (x_i)_{i=1}^{n}$

p marginal prob of P on $\mathcal{X}$.

$$\hat{\underline{R}}_S(\mathcal{G}) = \frac{1}{2}\hat{\underline{R}}_{S_{\mathcal{X}}}(\mathcal{F}).$$

Pf: $L(y, h(x)) = \frac{1}{2}(1 - y h(x))$.

$$\hat{\underline{R}}_S(\mathcal{G}) = \mathbb{E}_\sigma\Big[\sup_{h\in\mathcal{F}} \frac{1}{n}\sum \frac{\sigma_i}{2}(1 - y_i h(x_i))\Big]$$

$$\overset{\mathbb{E}_\sigma(\sigma_i)=0}{\underset{=}{\smile}} \mathbb{E}_\sigma\Big[\sup_{h\in\mathcal{F}} \frac{1}{2n}\sum \sigma_i \cdot h(x_i)\Big] \leftarrow \begin{matrix}\sigma_i \sim -\sigma_i \text{ dist. same} \\ \underline{y_i} \in \{-1,1\} \text{ is } \underline{\text{fixed}}.\end{matrix}$$

$$= \frac{1}{2}\hat{\underline{R}}_{S_{\mathcal{X}}}(\mathcal{F}) \qquad \square$$

Properties:

- $\hat{\underline{R}}(c\mathcal{G}) = |c|\hat{\underline{R}}(\mathcal{G})$ for $c \in \mathbb{R}$.
- $\mathcal{G}_1 \subseteq \mathcal{G}_2 \Rightarrow \hat{\underline{R}}(\mathcal{G}_1) \le \hat{\underline{R}}(\mathcal{G}_2)$.
- $\hat{\underline{R}}(\mathcal{G}_1 + \mathcal{G}_2) = \hat{\underline{R}}(\mathcal{G}_1) + \hat{\underline{R}}(\mathcal{G}_2)$.
- $\hat{\underline{R}}(\text{conv. } \mathcal{G}) = \hat{\underline{R}}(\mathcal{G})$. (convex hull) $\text{conv}(\mathcal{G}) := \Big\{ z \mapsto \sum_{i=1}^{n} \lambda_i g_i(z) : \begin{matrix} n \in \mathbb{N}, g_i \in \mathcal{G}, \\ \lambda_i \ge 0, \\ \sum \lambda_i = 1 \end{matrix} \Big\}$
- $\varphi: \mathbb{R} \to \mathbb{R}$: L-Lipschitz $\Rightarrow \hat{\underline{R}}(\varphi \circ \mathcal{G}) \le L \cdot \hat{\underline{R}}(\mathcal{G})$

- $\hat{\underline{R}}(\mathcal{G})$ is NP-hard to compute

1.9 Covering Numbers

Def: (M, d) - pseudometric space (same as metric except $d(x,y) = 0 \not\Rightarrow x = y$)

Let $A, B \subseteq M$, $\varepsilon > 0$. A is an ε-cover of B if $\forall b \in B\ \exists a \in A: d(a,b) \le \varepsilon$.

This is called "internal" if, in addition, $A \subseteq B$.

The "ε-covering number" of B, $N(\varepsilon, B)$ is the smallest cardinality of an ε-cover of B.

Def: $A \subseteq B$ is an "ε-packing" of B if $a, b \in A \wedge a \ne b \Rightarrow d(a,b) > \varepsilon$.

The "ε-packing number" of B, $M(\varepsilon, B)$ is largest cardinality of ε-packing of B.

[illegible]ion: $N(\varepsilon/2, B) \ge M(\varepsilon, B) \ge N_{in}(\varepsilon, B) \ge N(\varepsilon, B)$

Covering/Packing # Example: (norm. balls in $\mathbb{R}^d$) : $\|\cdot\|$ any norm on $\mathbb{R}^d$.

Consider $B_r(x) := \{z \in \mathbb{R}^d \mid \|z - x\| \leq r\}$

Supp. $\{x_1, \dots, x_M\} \subseteq \mathbb{R}^d$ is a max. ε-packing of $B_r(0)$. So $M = M(\varepsilon, B_r(0))$.

Then $i \neq j$, $B_{\varepsilon/2}(x_i) \cap B_{\varepsilon/2}(x_j) = \varphi$, $B_{\varepsilon/2}(x_i) \subseteq B_{r+\varepsilon/2}(0)$.

~~$v := \mathrm{vol}(B_1(0))$. Then~~ $$M(\varepsilon, B_r(0)) \leq \frac{\mathrm{vol}(B_{r+\varepsilon/2}(0))}{\mathrm{vol}(B_{\varepsilon/2}(x_i))} = \frac{(r+\varepsilon/2)^d v}{(\varepsilon/2)^d v} \underset{\varepsilon \leq r}{\overset{\uparrow}{\leq}} \left(\frac{3r}{\varepsilon}\right)^d$$

$\ln M(\varepsilon, B) = \Theta(d \ln 1/\varepsilon)$. essentially same is true when alg. dim d replaced by combinatorial dim - e.g. VC-dim.

Consider $g \in \mathcal{G} \subseteq \mathbb{R}^Z$, $p \in [1, \infty)$, $z \in Z^n$.

Define $\|g\|_{p,z} := \left(\frac{1}{n}\sum_{i=1}^{n} |g(z_i)|^p\right)^{1/p}$ seminorm Note: $\|g\|_{p,z} \leq \|g\|_{q,z}$ if $p \leq q$.

pseudometric $(g_1, g_2) \longmapsto \|g_1 - g_2\|_{p,z}$

Then $M(\varepsilon, \mathcal{G}, \|\cdot\|_{p,z}) \leq M(\varepsilon, \mathcal{G}, \|\cdot\|_{q,z})$ if $p \leq q$.

Lemma: $\mathcal{F} \subseteq \{0,1\}^X$, $d := \mathrm{VCdim}(\mathcal{F})$. (Pf. in Lecture Notes)

$$\underset{\uparrow \text{dep. on } p.}{M}(\varepsilon, \mathcal{F}) \leq \left(\frac{9}{\varepsilon^p} \ln \frac{2e}{\varepsilon^p}\right)^d \Leftarrow \text{bound independent of } n!$$

Lec 7 - 4/25/17

Thm (Dudley's chaining thm) Let $z \in Z^n$. $\mathcal{G} \subseteq \mathbb{R}^Z$ equipped w/ $\|\cdot\|_{2,z}$. $\gamma_0 = \sup_{g \in \mathcal{G}} \|g\|_{2,z}$

Then $$\hat{R}(\mathcal{G}) \leq \frac{12}{\sqrt{n}} \int_0^{\gamma_0} \ln(\underset{\uparrow}{N}(\beta, \mathcal{G}))^{1/2} \, d\beta$$

β-covering number. if ind. of n, this bound $\simeq n^{-1/2}$!

Pf: $\gamma_j := 2^{-j}\gamma_0$, $j \in \mathbb{N}$. Let $G_j \subseteq \mathbb{R}^Z$ be minimal γ_j-cover of $\mathcal{G}$.

$|G_j| = N(\gamma_j, \mathcal{G})$. $\forall g \in \mathcal{G}\ \exists g_j \in G_j$ s.t. $\|g - g_j\|_{2,z} \leq \gamma_j$. Note: $g_0 := 0$ by def. of

For $m \in \mathbb{N}$
$\mathcal{G} \ni g = g - g_m + \sum_{j=1}^{m} (g_j - g_{j-1})$

$$\hat{R}(\mathcal{G}) = \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{g \in \mathcal{G}} \sum_{i=1}^{n} \sigma_i \left(g(z_i) - g_m(z_i) + \sum_{j=1}^{m} g_j(z_i) - g_{j-1}(z_i)\right)\right]$$

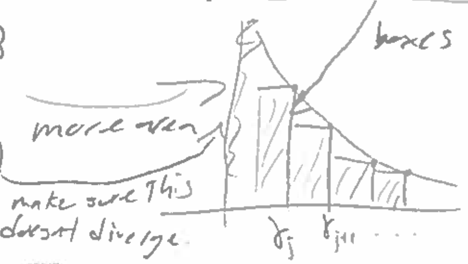
$$\leq \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{g \in \mathcal{G}} \sum_{i=1}^{n} \sigma_i (g(z_i) - g_m(z_i))\right] + \frac{1}{n}\sum_{j=1}^{m} \mathbb{E}_\sigma \sup_{g \in \mathcal{G}} \sum_{i=1}^{n} \sigma_i (g_j(z_i) - g_{j-1}(z_i))$$ $\leftarrow$ treat as inner products

By Cauchy-Schwarz, $\frac{1}{n}\mathbb{E}_\sigma\left[\sup_{g\in\mathcal{G}}\sum_{i=1}^{n}\sigma_i(g(z_i)-g_m(z_i))\right] \le \|\sigma\|_{2,z}\|g-g_m\|_{2,z} = 1\cdot\gamma_m$

By Massart's Lemma, $\frac{1}{n}\sum_{j=1}^{m}\mathbb{E}_\sigma\left[\sup_{g\in\mathcal{G}}\sum_{i=1}^{n}\sigma_i(g_j(z_i)-g_{j-1}(z_i))\right] \le \frac{12}{\sqrt{n}}\sum_{j=1}^{m}\underbrace{(\gamma_j-\gamma_{j+1})\sqrt{\ln(N(\gamma_j,\mathcal{G}))}}$

Hence $\hat{\mathcal{R}}(\mathcal{G}) \le \gamma_m + \frac{12}{\sqrt{n}}\int_{\gamma_{m+1}}^{\gamma_0}\sqrt{\ln(N(\beta,\mathcal{G}))}\,d\beta$

as $m\to\infty$, $\gamma_m\to 0$. (be careful N behaves well in this limit)



$$\Rightarrow \hat{\mathcal{R}}(\mathcal{G}) \le \frac{12}{\sqrt{n}}\int_0^{\gamma_0}\ln(N(\beta,\mathcal{G}))^{1/2}\,d\beta. \quad \square$$

Cor:
To apply this, note $N(\beta,\mathcal{G}) \le M(\beta,\mathcal{G})$ - covering # upper bd by packing #,
then use Lemma $M(\varepsilon,\mathcal{F}) \le \left(\frac{9}{\varepsilon^2}\ln\frac{2e}{\varepsilon^2}\right)^d$

use $\ln x \le \frac{x}{e}$ $\Rightarrow$ " $\le \left(\frac{18}{\varepsilon^4}\right)^d$ & plug into Dudley's

$$\hat{\mathcal{R}}(\mathcal{G}) \le \frac{12}{\sqrt{n}}\int_0^{\gamma_0}\ln\left(\left(\frac{18}{\beta^4}\right)^d\right)^{1/2}d\beta = 12\sqrt{\frac{d}{n}}\int_0^{\gamma_0}\underbrace{\cdots d\beta}_{\text{some constant}}$$

$$\Rightarrow \hat{\mathcal{R}}(\mathcal{G}) \le \underset{\text{ideal?}}{31}\underline{\sqrt{d/n}} \Longleftarrow \text{best known scaling!}$$

Recall: $\mathcal{R}(h)-\hat{\mathcal{R}}(h) \le \hat{\mathcal{R}}(\mathcal{G}) + \mathcal{O}(\frac{1}{\sqrt{n}})$

So now $\mathcal{R}(h)-\hat{\mathcal{R}}(h) \le \mathcal{O}(\sqrt{d/n}) + \mathcal{O}(\frac{1}{\sqrt{n}})$! Best bound.

1.10 Algorithmic Stability: So far, have only seen learning algorithm in its range, $\mathcal{F}$.

- Small changes in input → small changes in hypothesis ⇒ learning alg. stable.
- E.g. linear regression = stable, high deg poly reg = probably not stable
- Lack of stability = sign of overfitting

Def: A learning alg. $S\mapsto h_S$ is "uniformly stable" w/ rate $\varepsilon:\mathbb{N}\to\mathbb{R}$ wrt loss fn L if $\forall n\in\mathbb{N}$, $S\in(\mathcal{X},\mathcal{Y})^n$, $i\in\{1,\dots,n\}$, $(x',y')\in(\mathcal{X}\times\mathcal{Y})$:

$$L(y_i, h_{S^i}(x_i)) - L(y_i, h_S(x_i)) \le \varepsilon(n)$$

w/ $S^i = S$ where i^{th} component is replaced by (x',y').

] very strong, implies next def.

Def: A learning alg. is "on-average stable" w/ rate $\varepsilon:\mathbb{N}\to\mathbb{R}$ if $\forall P$ prob meas. on $\mathcal{X}\times\mathcal{Y}$:

$$\mathbb{E}_{S\sim P^n}\mathbb{E}_{(x_i',y_i')\sim P}\mathbb{E}_{i\sim\text{unif}}\left[L(\dots)-L(\dots)\right] \le \varepsilon(n).$$ w/ all same other assumptions.

Thm: If $S \mapsto h_S$ is ε-on average stable, then

$$\mathbb{E}_S\left[R(h_S) - \hat{R}(h_S)\right] \le \varepsilon(n).$$

on-average stable
$\Rightarrow$ generalizes well

Pf: Relies heavily on i.i.d. assumption for S & (x', y') ($\sim P^n$ and $\sim P$)

$$\mathbb{E}_S\left[R(h_S)\right] = \mathbb{E}_S \, \mathbb{E}_{(x',y')}\left[L(y', h_S(x'))\right]$$

$$= \mathbb{E}_i \, \mathbb{E}_S \, \mathbb{E}_{(x',y')}\left[L(y_i, h_{Si}(x_i))\right] \quad \leftarrow \text{Obs. 1.}$$

$$\mathbb{E}_S\left[\hat{R}(h_S)\right] = \mathbb{E}_S \, \mathbb{E}_i\left[L(y_i, h_S(x_i))\right]$$

$$= \mathbb{E}_S \, \mathbb{E}_{(x',y')} \, \mathbb{E}_i\left[\text{''}\right] \quad \leftarrow \text{Obs. 2}$$

$$\Rightarrow \mathbb{E}_S\left[R(h_S) - \hat{R}(h_S)\right] = \mathbb{E}_S \, E_{(x',y')} \, \mathbb{E}_i\left[L(y_i, h_{Si}(x_i)) - L(y_i, h_S(x_i))\right] \le \varepsilon(n)$$

∎

(We will see that the alg. $S \mapsto \operatorname{argmin}_h \hat{R}(h) + \lambda \|h\|^2$ is unif. stable
Regularization (Tikinov here) adds stability by adding strict convexity.)

Def: f is α-strongly convex if

: $x \mapsto f(x) - \frac{\alpha}{2}\|x\|^2$ is convex

$\Updownarrow$

: $f(\lambda x + (1-\lambda) y) \le \lambda f(x) + (1-\lambda) f(y) - \frac{\alpha}{2} \lambda (1-\lambda) \|x - y\|^2$

Lec 8 - 4/27/17

Thm: Assume range $L \subseteq [-c, c]$ & $S \mapsto h_S$ is unif. stable w/ rate ε_1.
$\forall \varepsilon > 0$ $\forall P$:

$$\mathbb{P}_S\left[|\hat{R}(h_S) - R(h_S)| \geq \varepsilon + \varepsilon_1(n) \right] \leq 2\exp\left[-\frac{n\varepsilon^2}{2(n\varepsilon_1(n) + c)^2} \right]$$

PAC bound

Pf: Let $\varphi(S) := \hat{R}(h_S) - R(h_S)$. By assumption, $|\mathbb{E}[\varphi(S)]| < \varepsilon_1(n)$.
Then $|\varphi(S)| \geq \varepsilon + |\mathbb{E}[\varphi(S)]| \Rightarrow |\varphi(S) - \mathbb{E}[\varphi(S)]| \geq \varepsilon$.

$$\mathbb{P}\left[|\hat{R}(h_S) - R(h_S)| \geq \varepsilon + \varepsilon_1(n) \right] \leq \mathbb{P}\left[|\varphi(S) - \mathbb{E}[\varphi(S)]| \geq \varepsilon \right] \leq 2\exp\left[-\frac{2\varepsilon^2}{n\nu^2} \right]$$ by McDiarmid's

where ν s.t. $\nu \geq |\varphi(S) - \varphi(S^i)|$.

$$|\varphi(S) - \varphi(S^i)| \leq \underbrace{\frac{1}{n}\sum_{j \neq i} |L(y_j, h_S(x_j)) - L(y_j, h_{S^i}(x_j))|}_{\leq \text{ by unif. stable}} + \frac{2c}{n} + |R(h_S) - R(h_{S^i})|$$

$$\leq \varepsilon_1(n) + \frac{2c}{n} + \underbrace{|\mathbb{E}_{(x,y)}[L(y, h_S(x)) - L(y, h_{S^i}(x))]|}_{\leq \varepsilon_1(n)} \leq 2\left(\varepsilon_1(n) + \frac{c}{n}\right) =: \nu.$$

□

Def: Φ is an α-strongly convex function ($\alpha > 0$) if $h \mapsto \Phi(h) - \frac{\alpha}{2}\langle h, h \rangle$ is convex and/or if $\lambda\Phi(h) + (1-\lambda)\Phi(g) \geq \Phi(\lambda h + (1-\lambda)g) + \frac{\alpha}{2}\lambda(1-\lambda)\|h - g\|^2$ $\forall h, g \in \mathcal{F}$ $\forall \lambda \in [0,1]$.

Lemma: If $\Phi: \mathcal{F} \to \mathbb{R}$ is α-strongly convex & attains its minimum at h, then $\forall g \in \mathcal{F}$,

$$\Phi(g) \geq \Phi(h) + \frac{\alpha}{2}\|h - g\|^2.$$

Pf: h is minimizer, so $\Phi(h) \leq \Phi(\lambda h + (1-\lambda)g)$

$$\Rightarrow \Phi(h) + \frac{\alpha}{2}\lambda(1-\lambda)\|h-g\|^2 \leq \lambda\Phi(h) + (1-\lambda)\Phi(g)$$

$$(1-\lambda)\Phi(h) + \frac{\alpha}{2}\lambda(1-\lambda)\|h-g\|^2 \leq (1-\lambda)\Phi(g)$$

$$\Rightarrow \Phi(h) + \frac{\alpha}{2}\lambda\|h-g\|^2 \leq \Phi(g)$$ set $\lambda = 1$. □

Thm: $\lambda > 0$. Let $\mathcal{F}$ be a convex subset of an inner product space.
Assume $h \mapsto L(y, h(x))$ is convex & ℓ-Lipschitz $\forall x, y$.
Then if $S \mapsto h_S$ minimizes $f_S(h) := \hat{R}(h) + \lambda\langle h, h \rangle$,
$S \mapsto h_S$ is unif. stable w/ rate $\varepsilon(n) = \frac{2\ell^2}{\lambda n}$.

Pf. (of Regularization ⇒ Unif Stability): Let $h := h_S$, $h' := h_{S^i}$ ← minimizers of $f_S(h)$ & $f_{S^i}(h)$

$$f_S(h') - f_S(h) = \hat{R}_S(h') - \hat{R}_S(h) + \lambda(\|h'\|^2 - \|h\|^2)$$
$$= \hat{R}_{S^i}(h') - \hat{R}_{S^i}(h) + \lambda(\|h'\|^2 - \|h\|^2)$$
$$+ \frac{1}{n}\left[L(y_i, h'(x_i)) - L(y_i, h(x_i)) + L(y_i', h(x_i')) - L(y_i', h'(x_i'))\right]$$

---

h' min. $f_{S'}$ $\longrightarrow$ $\le \frac{1}{n}[L\cdots] \underset{\substack{\uparrow \\ \text{by Lipschitz}}}{\le} \frac{2\ell}{n}\|h-h'\|$

Moreover, $\lambda\|h-h'\|^2 \le f_S(h') - f_S(h)$ since h minimizes f_S which is 2λ-strongly convex

Put them together, get $\|h-h'\| \le \frac{2\ell}{n\lambda}$

Def. of unif stable: $L(y_i, h'(x_i)) - L(y_i, h(x_i)) \overset{\text{Lipschitz}}{\le} \ell\|h-h'\| \le \frac{2\ell^2}{\lambda n}$. $\square$

Thm: Let $h^* := \underset{h\in\mathcal{F}}{\operatorname{argmin}}\, R(h)$.

If $S \mapsto h_S$ is reg. ERM w/ reg-param. λ & ℓ-Lipschitz loss then

$$\mathbb{E}_S[R(h_S)] \le R(h^*) + \lambda\|h^*\|^2 + \frac{2\ell^2}{\lambda n}.$$

Pf: $\mathbb{E}_S[\hat{R}(h_S)] \le \mathbb{E}_S[\hat{R}(h_S) + \lambda\|h_S\|^2] \underset{\substack{\uparrow \\ \text{by } h_S \text{ reg-ERM.}}}{\le} \mathbb{E}_S[\hat{R}(h^*) + \lambda\|h^*\|^2] = R(h^*) + \lambda\|h^*\|^2.$

$$\mathbb{E}_S[R(h_S)] = \underbrace{\mathbb{E}_S[\hat{R}(h_S)]} + \mathbb{E}_S[R(h_S) - \hat{R}(h_S)]$$
$$\le R(h^*) + \lambda\|h^*\|^2 + \frac{2\ell^2}{\lambda n} \quad \text{← by unif. stable.}$$

$\square$

Future Topics: · PAC-Bayesian · include a priori information about data dist/model

· Ensemble methods – ADA Boost

Next week – start neural networks (for at least 3 weeks)

Lec 4 - 5/2/17

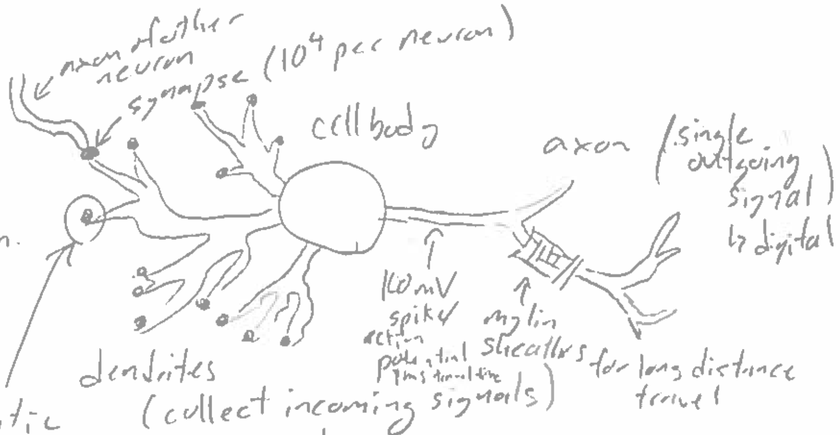
II. Neural Networks

II.1 Biological NNs :

10^{11} neurons in human brain.

grey matter - mostly neurons

white matter - mostly mylin-covered axons


axon of other neuron
synapse (10^4 per neuron)
cell body
axon (single outgoing signal) ↳ digital
100mV spike / action potential 1ms travel time
mylin sheaths for long distance travel
dendrites (collect incoming signals) ~ analog

synaptic weights vary/trainable

- 6 horizontal layers
- cortical columns (~100 neurons)

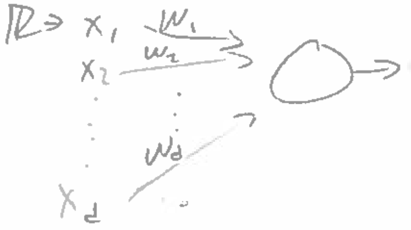
computer	brain
20 W/laptop	20W
# 10^{10} transistors	# 10^{11} neurons
connectedness: ~3	~10^4
clockrate GHz	100 Hz - not universal → constraints logical depth
deterministic	stochastic
2D	3D

Relevant for Model

- many dendritic inputs single axon output
- analog in / digital out
- synaptic variability

Rosenblatt's Perceptron '58

activation fn. $\sigma : \mathbb{R} \to \mathbb{R}$, for perceptron $\sigma(z) := \mathbb{1}_{z \geq 0} = \text{sgn}(z)$.

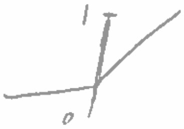


$$\sigma\Big(\underset{\substack{\uparrow \\ \text{bias/} \\ \text{threshold}}}{w_0} + \sum_i \underset{\substack{\uparrow \\ \text{weights}}}{w_i} x_i\Big)$$

other activation fns

$\sigma(z) = \frac{1}{1+e^{-z}}$ - "Logistic sigmoid"

tanh z



★ rectified linear unit (ReLU)

$\sigma(z) = \max\{0, z\}$

almost exclusively used now

Networks

architecture - graph (V, E)

vertices V = artificial neurons

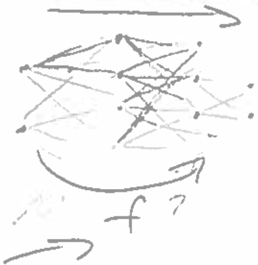
edges E = connections btwn V w/ weights $w_{ij} \in \mathbb{R}$

$i, j \in V$ and $(i,j) \in E$

+ bias per vertex

↳ graph is weighted + directed + acyclic (feed-forward)

often we consider multi-layered Networks/graphs



Feed-Forward Multilayer Acyclic Neural Network.

what kind of fns. can we represent by this model?

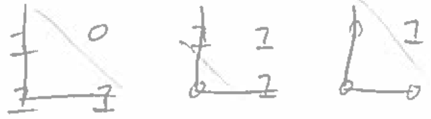
Representations & Approximations

single perceptron: $f(x) := \sigma\left(w_0 + \underbrace{\sum_i w_i x_i}_{\langle w, x \rangle}\right)$, $f : \mathbb{R}^d \to \mathbb{R}$

Constant on hyperplanes → can use σ-step fn. to use this model to separat space via a hyperplane.

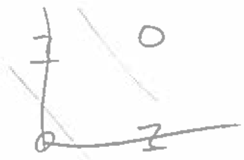
↳ can use on linearly (via hyperplane) separable data for binary classificat

NAND, OR, AND : $\mathbb{R}^2 \to \mathbb{R}$ ← can we rep. these via a perceptron?



← can separate via hyperplane. so yes! → can build any boolean fn. via composition of these

XOR - requires composition of above perceptions
- cant be done by single perceptron since data isnT linearly separable.

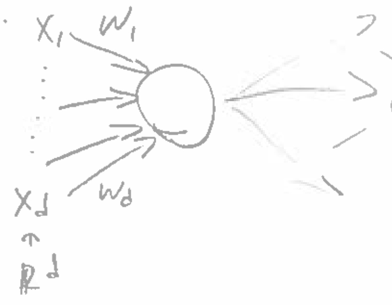


email: mwolf@math.ucdavis.edu

for project group 1-3 people.

Lec 10 - 5/4/17

Neural Network:

Node: 

some output

$\sigma(w_0 + \sum_{i=1}^{d} w_i x_i)$

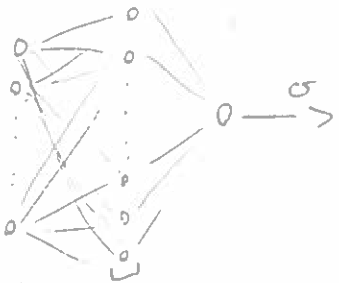
$\sigma: \mathbb{R} \to \mathbb{R}$ activation func.

Thm: Every $f: \{0,1\}^d \to \{0,1\}$ can be represented by a feed-forward NN w/ a single hidden layer containing at most 2^d neurons if $\sigma(z) := \mathbb{I}_{z \geq 0}$ is used.

Pf: If $a, b \in \{0,1\}$ then $2ab - a - b \leq 0$ with "=" iff $a = b$.

Thus $\mathbb{I}_{x=u} = \sigma\left(\sum_{i=1}^{d} 2x_i u_i - x_i - u_i\right)$. $A := f^{-1}(\{1\})$

$$f(x) = \sigma(-1 + \sum_{u \in A} \mathbb{I}_{x=u}) = \sigma\left(-1 + \sum_{u \in A} \sigma\left(\sum_{i=1}^{d} 2x_i u_i - x_i - u_i\right)\right)$$



A ← at most 2^d since $A = f^{-1}(\{1\})$ so at most $A = \{0,1\}^d$. □

Assume input space is $\mathcal{X} = \mathbb{R}^d$, output space $\mathcal{Y} = \{0,1\}$

Consider $\sigma(z) = \mathbb{I}_{z \geq 0}$. Then every indiv. neuron j is characterized by a halfspace $\mathcal{H}_j \subseteq \mathbb{R}^d$ (looking at the hidden layer $j = 1, \ldots, m$)

$$\overset{\mathbb{R}^d}{\overset{\in}{x}} \longmapsto f(x) = \sigma\left(w_0 + \sum_{j=1}^{m} w_j \mathbb{I}_{x \in \mathcal{H}_j}\right)$$

$$\mathcal{A} := \{A \subseteq \{1, \ldots, m\} \mid \sum_{j \in A} w_j \geq -w_0\}$$ ← combinations of neurons in hidden layer who "fire" w/ final output "fire"

$$\mathbb{R}^d \supseteq f^{-1}(\{1\}) = \bigcup_{A \in \mathcal{A}} \underbrace{\bigcap_{j \in A} \mathcal{H}_j}_{\text{convex polyregion for fixed } A \text{ (intersection of half-planes)}}$$

Thm (Zaslavsky's): (Given n-hyperplanes in $\mathbb{R}^d$, how many regions can we make?)

Let $h_1, \dots h_n \subseteq \mathbb{R}^d$ by hyperplanes. The number N of connected components of $\mathbb{R}^d \setminus \bigcup_{j=1}^{n} h_j$ is $N \leq \sum_{i=0}^{d} \binom{n}{i} \leq \left(\frac{en}{d}\right)^d$

$\uparrow$ ~~tight bound~~, often the exact #.

Pf: Shift hyperplanes away from origin, each can be characterized w/ a single vector.

$h := \{x \in \mathbb{R}^d \mid \langle w, x \rangle = 1\}$, $\mathcal{X}$ is the set of all such hyperplanes.

$\updownarrow$

$w \in \mathbb{R}^d$

$\mathcal{F} := \{h \mapsto \text{sgn}[g(h)] \mid g \in \mathcal{G}\}$. $\mathcal{G} = \{h \mapsto \langle x, w \rangle - 1 \mid x \in \mathbb{R}^d\} \subseteq \mathbb{R}^{\mathcal{X}}$.

$\cap$ $\{-1, 1\}^{\mathcal{X}}$

$\uparrow$ affine space w/ $\dim(\mathcal{G}) = d$.

Obs: VC dim$(\mathcal{F}) = d$.

Now assume $A \subseteq \mathcal{X}$ w/ $|A| = n$. separates $\mathbb{R}^d$ into N regions.

Then $\mathcal{F}|_A$ contains at least N different fns.

Thus $N \leq \Gamma(n) \leq \sum_{i=0}^{d} \binom{n}{i}$. □

$\uparrow$ growth fn of $\mathcal{F}$

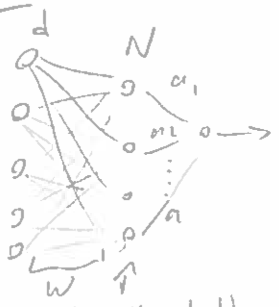
but still need more neurons anyway to prev...

problem: too many neurons = memorization → overfitting

Thm: Let $A = \{x_1, \dots, x_N\} \subseteq \mathbb{R}^d$ and $f: A \to \mathbb{R}$.

There is a feed-forward NN that implements $F: \mathbb{R}^d \to \mathbb{R}$ w/ a single hidden layer containing N neurons (& $2N+d$ parameters) so that $F|_A = f$. We can use $\sigma(z) = \max\{0, z\}$ in the hidden layer & $\sigma(z) = z$ for the output layer.

Pf: $F: \mathbb{R}^d \to \mathbb{R}$, $F(x) = \sum_{j=1}^{N} a_j \max\{0, \langle w, x \rangle - v_j\}$, $a, v \in \mathbb{R}^N$, $w \in \mathbb{R}^d$



v_j = threshold bias

All first-layer connections have shared weights w.

Let $M_{ij} = \max\{0, \langle w, x_i \rangle - v_j\}$

Then $F(x) = Ma$. If M^{-1} exists,

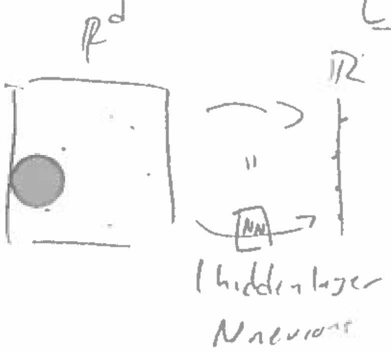
$\Rightarrow a_j = (M^{-1})_{ji} f(x_i)$ solves the problem exactly.

Assuming x_i distinct, $\exists\, w \in \mathbb{R}^d$ s.t. $\langle w, x_i \rangle = z_i$ distinct as well.

Reorder z's & pick v_i's s.t. $v_1 < z_1 < v_2 < z_2 < \dots$

invertible! □

Lec 11 - 5/9/17



Thm: Let $\sigma \in C(\mathbb{R})$. The set $\mathcal{F}_\sigma$ of fns. representable by a NN w/ a single hidden layer & activation function σ is dense in $C(\mathbb{R})$ w.r.t. the topology of unif. conv. on compact sets iff σ is not a polynomial.

pf (sketch): Assume $\sigma \in C^\infty(\mathbb{R})$. If σ is not a poly. there is a $z \in \mathbb{R}$ s.t. $\sigma^{(k)}(z) \neq 0 \quad \forall k \in \mathbb{N}$.

$$\frac{\sigma((\lambda+\delta)x+z) - \sigma(\lambda x + z)}{\delta} \text{ is in } \mathcal{F}_\sigma \text{ if } \delta \neq 0.$$

$$\frac{d}{d\lambda}\sigma(\lambda x + z)\Big|_{\lambda=0} = x\underbrace{\sigma^{(1)}(z)}_{\neq 0}$$, Thus $f(x) = x$ is in closure of $\mathcal{F}_\sigma$.

Higher derivatives yield $x \mapsto x^k \underbrace{\sigma^{(k)}(z)}_{\neq 0}$ => monomials $\in$ closure of $\mathcal{F}_\sigma$

Hence all polynomials $\in$ closure of $\mathcal{F}_\sigma$. $\overset{\text{Weierstrass}}{\Longrightarrow}$ $\mathcal{F}_\sigma$ dense in $C(\mathbb{R})$ (since polys are.) □

Lemma: Let $U \in \mathbb{R}^d$ compact. Then $\mathcal{E} := \operatorname{Span}\{f : U \to \mathbb{R} \mid f(x) = \exp[\sum_{i=1}^d w_i x_i],\ w_i \in \mathbb{R}\}$. is dense in $(C(K), \|\cdot\|_\infty)$.

Pf: Stone-Weierstrass states $\mathcal{E}$ is dense if (i) $\mathcal{E}$ is an algebra. ✓

to get $f=1$, plug in w_i's $= 0$ $\rightarrow$ (ii) $\mathcal{E}$ contains a non-zero const fn.

use $w = x-y$ $\rightarrow$ (iii) $\forall x,y \in K : x \neq y$, $\exists f \in \mathcal{E}$ s.t. $f(x) \neq f(y)$.

$\Rightarrow f(x-y) = e^{\|x-y\|^2} \neq 1 \Rightarrow f(x) \neq f(y)$.

" $f(x)/f(y)$.

□

Thm $d, d' \in \mathbb{N}$, $K \subseteq \mathbb{R}^d$ compact, $\sigma \in C(\mathbb{R})$ nonpolynomial. Then the set of fns. representable by a NN w/ a single hidden layer that uses σ as an act. fn. is dense in the set of cont's $f : U \subseteq \mathbb{R}^d \to \mathbb{R}^{d'}$.

Pf: WLOG, $d'=1$. $K \subseteq \mathbb{R}^d$ is compact.

$\forall \varepsilon > 0$, $\exists n \in \mathbb{N}$, $v_1, \dots, v_n \in \mathbb{R}^d$, $s \in \{\pm 1\}^n$ s.t. $g : \mathbb{R}^d \to \mathbb{R}$ } from Lemma

$g(x) := \sum_{i=1}^{n} s_i e^{v_i \cdot x}$ satisfies $\|g - f\|_\infty \le \varepsilon/2$.

$K_1 := \bigcup_{i=1}^{n} \{v_i \cdot x \mid x \in K\} \subseteq \mathbb{R}$ compact. Then $\exists a_j, b_j, w_j$ s.t.

[illegible] these 2

$$\sup_{y \in K_1} \left| e^y - \sum_{j=1}^{L} a_j \sigma(w_j y - b_j) \right| \le \varepsilon / 2n$$

$$\Big\| f - \underbrace{\sum_{j=1}^{L} \sum_{i=1}^{n} s_i a_j \sigma(w_j v_i \cdot x - b_j)}_{\text{expressable by 1-layer NN.}} \Big\|_\infty \le \Big\| f - \sum_{i=1}^{n} s_i e^{v_i \cdot x} \Big\|_\infty + \sum_{i=1}^{n} \Big\| e^y - \sum_{j=1}^{L} a_j \sigma(w_j y \Big.$$

$$\le \varepsilon/2 + n \, \varepsilon/2n = \varepsilon. \quad \square$$

Kolmogorov's superposition thm: $\forall n \in \mathbb{N}$ $\exists \varphi_j \in C([0,1])$, $j \in \{0, 1, \dots 2n\}$

$\exists \lambda \in \mathbb{R}_+^n$ s.t. for every $f \in C([0,1]^n, \mathbb{R})$ $\exists \phi \in C([0,1])$ s.t.

$$f(x_1, \dots, x_n) = \sum_{j=0}^{2n} \phi\Big(\sum_{k=1}^{n} \lambda_k \varphi_j(x_k) \Big)$$

not of practic[illegible]

can interpret as a 2-hidden layer NN w/ exact #nodes + activation fns.

- 1st hidden layer has $n(2n+1)$ neurons w/ φ_j's as act. fns.
- 2nd hidden layer has $2n+1$ neurons using ϕ
- Output layer uses $\sigma(z) = z$.

Thm: (VC-dim of NNs): For arbitrary $n_0, w \in \mathbb{N}$, fix an architecture of a layered feed-forward NN w/ n_0 inputs, a single output, w parameters (#weights + biases).

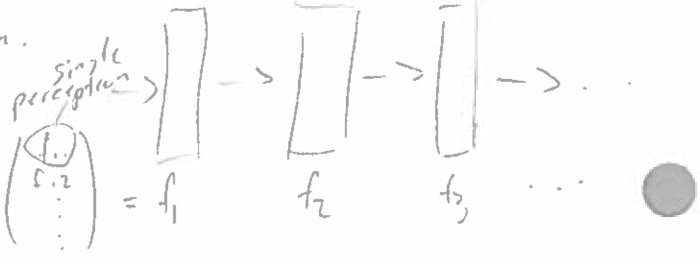
$\mathfrak{F} \subseteq \{-1, 1\}^{\mathbb{R}^{n_0}}$ that can be implemented by such a NN using $\sigma = \mathrm{sgn}$.

Then $\mathrm{VCdim}(\mathfrak{F}) \le 2w \log_2 (ew)$.

Pf(Sketch): Uses composition property of growth fn.

$$\Gamma_{\mathfrak{F}}(n) \le \prod_{i=1}^{m} \Gamma_{f_i}(n) \le \prod_{i=1}^{m} \prod_{j=1}^{n_i} \Gamma_{f_{ij}}(n)$$

$m = \#$ layers $\quad$ $n_i = \#$nodes in layer i $\quad$ ← perceptron



$$\le \prod_{i=1}^{m} \prod_{j=1}^{n_i} \left(\frac{en}{w_{ij}} \right)^{w_{ij}} \le (en)^w$$

$\frac{1}{\cdot} \le 1$

Then use growth fn → VCdim bounds $\quad \square$

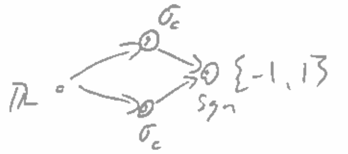
Lec 12 - 5/11/17

Recall: VC dimension tells us how much data is required in binary classification training for good generalization guarantee.

Last time: w/ activation fn $\_\!\!\rceil\!\!\!\_$, $\text{VCdim}(\bar{\mathcal{F}}) = \mathcal{O}(w \log_2 w)$
where w = # of weight parameters in NN.

Recall: $\text{VCdim}\left(\{0,1\}^{\{0,1\}^n}\right) = 2^n$ $\Rightarrow$ NN has to grow exp. in n.

Consider



w/ σ_c:

$\text{VCdim}(\bar{\mathcal{F}}) = \infty$

Linear combs + sgn allow NN to amplify crazy properties of σ_c that are unobvious

Ex: $\sigma_c(z) = \frac{1}{1+e^{-z}} + cz^3 e^{-z^2} \sin(z)$, $c \geq 0$ ↙ this class has VCdim = ∞.

Consider $\bar{\mathcal{F}} = \{f : \mathbb{R}_+ \to \mathbb{R} \mid \exists \alpha \in \mathbb{R} : f(x) = \text{sgn}\sin(\alpha x)\}$. WTS $\bar{\mathcal{F}} \subseteq \bar{\mathcal{F}}_{NN}$

let $f \in \bar{\mathcal{F}}_{NN}$ s.t. $f(x) = \text{sgn}\left[\sigma_c(\alpha x) + \sigma_c(-\alpha x) - 1\right]$

$$= \text{sgn}\left[\underbrace{2c(\alpha x)^3 e^{-\alpha^2 x^2}}_{>0} \sin(\alpha x)\right] = \text{sgn}\left[\sin(\alpha x)\right]. \qquad \square$$

Thus we must require more specific properties on the σ's!

Prop: [Warren] Let $\{p_1, \ldots, p_m\}$ be polys in k variables of degree $\leq d$ w/ coeffs $\in \mathbb{R}$.
Let $\gamma(k,d,m)$ = max # of connected components of $\mathbb{R}^k \setminus \bigcup_{i=1}^{m} p_i^{-1}(\{0\})$
Then $\gamma(k,d,m) \leq (4edm/k)^k$ if $m \geq k$.

Thm: Consider s atomic predicates, each of which is a poly. inequalities of deg. $\leq d$ in $m+k$ vars.
↳ truth $x \mapsto \{0,1\}$

Consider $\Psi : \mathbb{R}^m \times \mathbb{R}^k \to \{0,1\}$ Boolean comb of those predicates. (and/or/xor/etc.)

Define $\bar{\mathcal{F}} := \{\Psi(\cdot, w) \mid w \in \mathbb{R}^k\} \subseteq \{0,1\}^{\mathbb{R}^m}$

Then (i) $\Gamma(n) \leq \gamma(k, d, 2ns)$
(ii) $\text{VCdim}(\bar{\mathcal{F}}) \leq 2k \log_2(8eds)$

Later, show $\text{VCdim}(\bar{\mathcal{F}}_{NN})$ is $\mathcal{O}(w \log w)$ → fixed # layers
$\left.\begin{array}{l}\mathcal{O}(wL\log w) \\ \Omega(wL\log\frac{w}{L})\end{array}\right\}$ p-w linear w/ L layers
↑ lower bd

Pf: WLOG, All poly ineqs. compared to 0 e.g. $p \geq 0$, $p \leq 0$.

Estimate $|\bar{f}|_A|$ where $|A| = n$, $A \subseteq \mathbb{R}^m$.

For $a \in A$, $\Psi_a(w) := \Psi(a, w)$. $P_a :=$ polys in Ψ_a.

Let $P := \bigcup_{a \in A} P_a$, then $|P| \leq n|P_a| \leq ns$.

Thus $\text{VCdim}(\bar{f}_{NN}) = \mathcal{O}(w$

Note $|\bar{f}|_A| \leq |\{Q \in \{-1, 0, 1\}^P \mid Q(p) = \text{sgn}_0(p(w)), w \in \mathbb{R}^k\}|$

$\hookrightarrow$ 0 map

Consider $P' := \{p + \varepsilon \mid p \in P\} \cup \{p - \varepsilon \mid p \in P\}$, $\varepsilon > 0$

Then $|P'| \leq 2|P| \leq 2ns$. Now things on bdry $(Q(x) = 0)$ than cells in sit inside cells in P'.

Thus $|\bar{f}|_A| \leq \gamma(k, d, 2ns)$ by def. of γ from Warren's

$\Rightarrow \Gamma(n) \leq$ " . //.

Then (ii) follows from Warren's prop. $\square$

Now $\text{VC dim}(\bar{f}) = \mathcal{O}(tk)$

Adding $(x \mapsto e^x, +, *, /, -)$ $\Rightarrow$ $\text{VCdim}(\bar{f}) = \mathcal{O}(t^2k^2)$ by similar proof.

What about $\text{VCdim}(\bar{f}_{NN})$? $\rightarrow$ show $\mathcal{O}(wN)$

Thm: Let $N = \#$neurons, $m = \#$inputs, $w = \#$parameters (weights + biases), feed-forward NN

$\sigma =$ piecewise-poly w/ p pieces & degree d output uses $\sigma(z) = \mathbb{1}_{z \geq 0}$.

Then $\text{VCdim}(\bar{f}_{NN}) \leq 2w[N \log_2 p + \log_2(16e \max\{\delta + 1, 2d^\delta\})]$

w/ δ = depth of network (# of layers).

Pf: Regard entire network as $\Psi : \mathbb{R}^m \times \mathbb{R}^w \to \{0, 1\}$. Label neurons forwards $i = 1, \ldots, N$.

$\delta(i)$ = depth of i^{th} neuron

a_i = output of i^{th} neuron (before σ)

Define $I : \mathbb{R} \to \{1, \ldots, p\}$, $I^{-1}(\{j\})$ interval corr. to j^{th} piece of poly.

a_1 is quadratic on $\mathbb{R}^m \times \mathbb{R}^w$. $I(a_1)$ can be det. by p ineqs. quadratic on $\mathbb{R}^m \times \mathbb{R}^w$.

$\deg(i)$ of a_i over $\mathbb{R}^m \times \mathbb{R}^w$ is $d^{\delta(i)} + \sum_{j=0}^{\delta(i)} d^j$. Then $I(a_i)$ can be det. by p ineqs. on $\mathbb{R}^m \times \mathbb{R}^w$ w/ deg

$\Rightarrow p^i$ predicates poly of deg $\leq \deg(i)$. In total, have $\sum_{j=1}^{i} p^j$ polypreds. to det. $I(a_i)$.

Last step, predicate is $\mathbb{1}_{a_N \geq 0}$ $\to$ 1 ineq. Add em all up, get $\leq 2p^N$ poly. predicates of deg.

$\Rightarrow \deg = \max\{\delta + 1, 2d^\delta\}$. Then use Thm to get VCdim bound. $\square$

Lec 13 - 5/16/17

Recall:

perceptron - VC dim n

$\sigma(\underline{v} + \sum_{i=1}^{n} w_i x_i)$

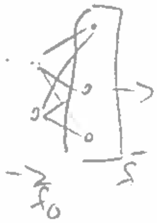
NN - composition of perceptrons has VC dim $O(w \log w)$

Thm: $a, b \in \mathbb{R}$, $\tilde{\sigma}: \mathbb{R} \to \mathbb{R}$, L-lipschitz. Assume $\mathcal{F}_0 \subseteq \mathbb{R}^{\mathcal{X}}$ includes zero-fn.

Define $\mathcal{F} := \{ x \mapsto \tilde{\sigma}(v + \sum_{j=1}^{m} w_j f_j(x)) \mid |v| \le a, \|w\|_1 \le b, f_i \in \mathcal{F}_0 \} \subseteq \mathbb{R}^{\mathcal{X}}$.

wrt. $x \in \mathcal{X}^n$, $\hat{\mathcal{R}}(\mathcal{F}) \le L\left(\frac{a}{\sqrt{n}} + 2b\hat{\mathcal{R}}(\mathcal{F}_0)\right)$

derive $\hat{\mathcal{R}}$ layer-wise



Pf: $\hat{\mathcal{R}}(\mathcal{F}) \le \frac{L}{n} \mathbb{E}_\sigma \left[\sup_{v, w, f_i} \sum_{i=1}^{n} \sigma_i \underbrace{\left(v + \sum_{j=1}^{n} w_j f_j(x_i)\right)} \right]$

$\sum_j w_j f_j \in b \operatorname{conv}\{\mathcal{F}_0 - \mathcal{F}_0\} =: \mathcal{G}_1$

$\{x \mapsto v \mid |v| \le a\} =: \mathcal{G}_2$

$\Rightarrow \hat{\mathcal{R}}(\mathcal{F}) \le L\,\hat{\mathcal{R}}(\mathcal{G}_1 + \mathcal{G}_2) = L(\hat{\mathcal{R}}(\mathcal{G}_1) + \hat{\mathcal{R}}(\mathcal{G}_2)) \le L\left(\frac{a}{\sqrt{n}} + 2b\hat{\mathcal{R}}(\mathcal{F})\right)$

$\hat{\mathcal{R}}(\mathcal{G}_1) = b\,\hat{\mathcal{R}}(\operatorname{conv}\{\mathcal{F}_0 - \mathcal{F}_0\}) = b\,\hat{\mathcal{R}}(\mathcal{F}_0 - \mathcal{F}_0) = b(\hat{\mathcal{R}}(\mathcal{F}_0) + \hat{\mathcal{R}}(-\mathcal{F}_0)) = 2b\hat{R}(f_0)$.

$\hat{\mathcal{R}}(\mathcal{G}_2) \le \frac{a}{n} \mathbb{E}_\sigma[|z|] \le \frac{a}{n} \mathbb{E}_\sigma[|z|^2]^{1/2} = \frac{a}{n}\sqrt{n} = \frac{a}{\sqrt{n}}$.

w/ $z := \sum_{i=1}^{n} \sigma_i$

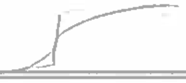
□

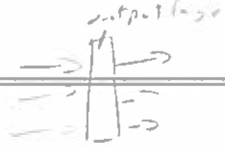
Rmk: Note that $\|w\|_1 \le b$ plays a large role in bounding $\hat{\mathcal{R}}(\mathcal{F})$! keeping weight values low in each layer $\Rightarrow$ generalization.

Training NNs.

Consider Classification: common choices:

- # of neurons in output layer = # of classes (say $m \in \mathbb{N}$)





- softmax activation fn at output

$$\sigma : \mathbb{R}^n \to \mathbb{R}^m \quad \sigma(z)_j := e^{z_j} \Big/ \sum_{k=1}^{n} e^{z_k}$$

- loss fn: cross-entropy (= neg. log. likelihood)

$$L(h(x), \underset{\{1,\dots,m\}}{\underset{\cap}{y}}) = -\sum_{c=1}^{m} \mathbb{1}_{y=c} \ln[h(x)_c]$$

Empirical risk: $\hat{R}(h) = \frac{1}{n} \sum_{i=1}^{n} L(h(x_i), y_i)$ + "regularizer"

suppose h depends on $w_1, \dots, w_N \in \mathbb{R}$, we minimize $\hat{R}(h)$ via gradient descent

$$\mathbb{R}^N \ni w^{(t+1)} = w^{(t)} - \underset{\hat{0}}{\underset{\vee}{\alpha_t}} \nabla_w \hat{R}(h)$$

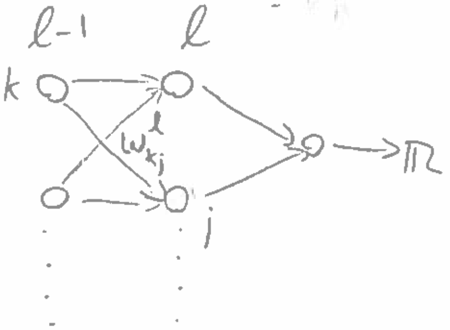
Naively: # NN evaluations for a single step $\sim nN \to$ bad! can get $nN \gtrsim \underbrace{2\text{or}6}_{\text{constant!}}$.

- Reduce n via Stochastic optimization
- Reduce N via backpropagation

Lec13 continued – 5/16/17

Backpropagation: (Automatic differentiation)

Bookkeeping: Layers $\ell \in \{0, \dots, m\}$. $N_\ell = \#$ neurons in ℓ^{th} layer.



$\mathbb{R} \ni w_{jk}^\ell :=$ weight from k^{th} neuron in layer $\ell-1$ to the j^{th} neuron in layer ℓ.

b_j^ℓ : threshold value for j^{th} neuron, layer ℓ.

x_j^ℓ = output of j^{th} neuron, layer ℓ.

$$x_\ell = \sigma(\underbrace{w^\ell x^{\ell-1} + b^\ell}_{z^\ell}).$$

$f : \dots \to \mathbb{R}$ ← would like $f\ \mathbb{R}^N \to \mathbb{R} \to \nabla_w f$

Define: $$\delta_j^\ell := \frac{\partial f}{\partial z_j^\ell} = \sum_k \frac{\partial f}{\partial z_k^{\ell+1}} \cdot \frac{\partial z_k^{\ell+1}}{\partial z_j^\ell} = \sum_k \delta_k^{\ell+1} \cdot w_{kj}^{\ell+1} \cdot \sigma'(z_j^\ell)$$

Compute recursively (backwards) from $\ell = m$. forward to get x_j^ℓ, backward to get δ_j^ℓ.

Thus we only need to run the NN twice for a single computation of $\nabla_w f$

Note: $$\frac{\partial f}{d b_j^\ell} = \sum_k \frac{\partial f}{\partial z_k^\ell} \frac{\partial z_k^\ell}{\partial b_j^\ell} = \sum_k \delta_k^\ell \cdot \underbrace{\delta_{kj}}_{\text{Kronecker-}\delta} = \delta_j^\ell.$$

$$\frac{\partial f}{d w_{jk}^\ell} = \sum_i \underbrace{\frac{\partial f}{\partial z_i^\ell}}_{\delta_i^\ell} \cdot \underbrace{\frac{\partial z_i^\ell}{\partial w_{jk}^\ell}}_{\text{gives } \delta_{ij}} = \delta_j^\ell x_k^{\ell-1}.$$

} comp. of all gradient pieces w/ only 2 runs of NN.

This technique is useful for network/circuit fns. from $\mathbb{R}^N \to \mathbb{R}^M$.

Can compute $\nabla_w f$ w/ same complexity/#comps. as f itself] can write a computational graph w/ $O(N)$ nodes

Thus $w \mapsto \langle \nabla f(w), v \rangle$ is linear in N.

[illegible]. $\nabla^2 f(w) v$ is computable in same time as f.

Lec 14 - 5/18/17

Work for one training step NN → use backprop to reduce to $\underline{n}$

Recall: Computing $\nabla_w \hat{R}(h_w) = \frac{1}{n} \sum_{i=1}^{n} \nabla_w L(h_w(x_i), y_i) + \text{"reg"}$

We can use Stochastic gradient descent to reduce n to 1.

Expectation value of random single gradient eval is full gradient.

Pick x_i at random & compute $\nabla_w L(h_w(x_i, y_i)) + \dots$

& step in that (neg.) direction, on <u>average</u> following true $\nabla_w \hat{R}(h_w)$

Why use the gradient at all?

Consider diff. f: $f(\underline{x}+v) \approx f(\underline{x}) + \langle \nabla f(\underline{x}), v \rangle$ ← Minimize?

Take $\operatorname{arginf}\{\langle \nabla f(\underline{x}), v \rangle : \|v\|=1\} = \begin{cases} -\nabla f(\underline{x}) & \text{if } \|\cdot\|_2 \\ \dfrac{-\rho^{-1} \nabla f(\underline{x})}{\|\rho^{-1} \nabla f(x)\|} & \text{if } \|\cdot\| = \langle v, \rho v \rangle^{1/2} \end{cases}$ as step direction $\rho > 0$ (matrix)

Could use new "direction" $\prod_{i=1}^{\delta} \|w_i\|_1$ ← depth # for better generalization

↑ vector of all weights in layer i

Gradient Descent: Randomly choose datapoint to eval ∇f. $x_t \in \mathbb{R}^d$

$$x_{t+1} = x_t - \alpha \nabla f(x_t), \quad \alpha > 0$$

Lemma: $f \in C^1(\mathbb{R}^d)$ w/ ∇f L-Lipschitz. Then $\forall x, y \in \mathbb{R}^d$,

$$|f(x) - f(y) - \langle \nabla f(x), y-x \rangle| \leq \frac{L}{2} \|y-x\|^2$$

← every point $f(x)$ is bounded by a lower + upper quadratic

Pf: $f(x) - f(y) = \int_0^1 \langle \nabla f(x+t(y-x)), y-x \rangle \, dt$

$|f(x) - f(y) - \langle \nabla f(x), y-x \rangle| \leq \int_0^1 |\langle \nabla f(x+t(y-x)) - \nabla f(x), y-x \rangle| \, dt$

$\overset{\text{Cauchy-Schwarz}}{\leq} \int_0^1 \|\nabla f(x+t(y-x)) - \nabla f(x)\|_2 \|y-x\|_2 \, dt$

$\leq \int_0^1 tL \|y-x\|^2 \, dt = \frac{L}{2} \|y-x\|^2 \quad \square$

Lec 14 continued 5/18/17

Use this lemma for GD:

Set $x = x_t$, $y = x_{t+1}$: $f(x_t) - f(x_{t+1}) \geq \alpha\left(1 - \frac{\alpha L}{2}\right) \|\nabla f(x_t)\|^2$ (*)

w/ $x_{t+1} = x_t - \alpha \nabla f(x_t)$ $\quad > 0$ if $0 < \alpha < 2/L$

& max if $\alpha = 1/L$.

Thm: $f \in C^1$, ∇f - L-Lipschitz, $\alpha \in (0, 2/L)$.

(i) $f(x_{t+1}) \leq f(x_t)$ unless $\nabla f(x_t) = 0$.

(ii) If f is bounded from below, then $\lim_{t\to\infty} \|\nabla f(x_t)\| = 0$.

$$\& \min_{t<T} \|\nabla f(x_t)\|^2 \leq \frac{\alpha(2-\alpha L)}{2T} \underset{\underset{\alpha = 1/L}{\uparrow}}{=} \frac{1}{2LT}.$$

Pf: We just showed (i). WTS (ii).

Consider $\sum_{t=0}^{T-1}$ (*) $\xrightarrow{\text{telescoping}}$ $f(x_0) - f(x_T) \geq \alpha\left(1 - \frac{\alpha L}{2}\right) \sum_{t=0}^{T-1} \|\nabla f(x_t)\|^2$

Can lower bound by min. of sum elements · #steps:

$$f(x_0) - f(x_T) \geq \alpha\left(1 - \frac{\alpha L}{2}\right) \cdot T \cdot \min_{t<T} \|\nabla f(x_t)\|^2$$

Then use simple algebra, trivial. □

Lemma: $f \in C^1(\mathbb{R}^n)$ & μ-strongly convex, ∇f - L-Lipschitz, then

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{\mu L}{\mu + L} \|x - y\|_2^2 + \frac{1}{\mu + L} \|\nabla f(x) - \nabla f(y)\|_2^2$$

Pf: $\phi(x) := f(x) - \frac{\mu}{2}\|x\|^2$ is convex (literally the def. of f μ-strongly convex).

& $\nabla \phi(x) = \nabla f(x) - \mu x$ is $(L - \mu)$-Lipschitz. ... □

Thm: $f \in C^1(\mathbb{R}^n)$, μ-strongly convex, ∇f - L-Lipschitz, $\alpha \in \left(0, \frac{2}{L+\mu}\right]$.

$$\|x_t - x^*\|^2 \leq \left(1 - \frac{2\alpha\mu L}{\mu + L}\right)^t \|x_0 - x^*\|_2^2$$

$$\underset{\underset{\alpha = \frac{1}{\mu + L}}{\uparrow}}{\leq} \frac{L}{2}\left(\frac{\kappa - 1}{\kappa + 1}\right)^{2t} \|x_0 - x^*\|_2^2$$

(linear convergence w/ quadratic rate?)

where $\kappa = \frac{L}{\mu}$ "condition number"

Pf: $\|x_{t+1} - x^*\|_2^2 = \|x_t - \alpha \nabla f(x_t) - x^*\|_2^2$

$= \|x_t - x^*\|_2^2 + \alpha^2 \|\nabla f(x_t)\|^2 - 2\alpha \langle \nabla f(x_t), x_t - x^* \rangle$

lemma & $\nabla f(x^*) = 0$ →
$$\leq \left(1 - \frac{2\alpha \mu L}{\mu + L}\right)\|x_t - x^*\|_2^2 + \underbrace{\alpha\left(\alpha - \frac{2}{\mu + L}\right)}_{\leq 0 \text{ if } \alpha \in (0, \frac{2}{L+\mu})} \|\nabla f(x_t)\|_2^2$$

$\leq$ // // $+ 0$

Then use telescoping to get $\|x_t - x\|^2 \leq \left(1 - \frac{2\alpha\mu L}{\mu + L}\right)^t \|x_0 - x^*\|^2$. □

Thus, if we want to be ε-close to optimum, take $\mathcal{O}(\log 1/\varepsilon)$ steps.
Only have this order work if we have (i) strong convexity of f
(ii) L-smoothness of f (∇f - L-Lipschitz)
(iii) exact gradient computable.

We will show that relaxing all 3 assumptions brings us to $\mathcal{O}(1/\varepsilon^2)$ step.

Thm: Assume $f: \mathbb{R}^d \to \mathbb{R}$ convex. $\forall x \in \mathbb{R}^d$, let $g_1(x), \ldots, g_n(x)$ i.i.d rvs w/ values in $\mathbb{R}^d$ s.t. $\mathbb{E}[g_i(x)] = \nabla f(x)$ ← a subgradient

f
many
subgradients
lie below graph
at x

Assume $\mathbb{E}[\|g_i(x)\|^2] \leq G^2$

Then consider $x_{t+1} := x_t - \alpha g_{t+1}(x_t)$, $t = 0, \ldots, T-1$, $\bar{x} := \frac{1}{T}\sum_{t=0}^{T-1} x_t$.

Then $\mathbb{E}[f(x)] - \underbrace{f(x^*)}_{\min} \leq \frac{2\|x - x^*\| G}{\sqrt{T}}$.

Lec 15 - 5/23/17

Polyak '63: $\frac{1}{2}\|\nabla f(x)\|^2 \geq \mu(f(x) - \underset{\substack{\uparrow \\ \text{global} \\ \text{min.}}}{f(x^*)}) \quad \forall x \in \mathbb{R}, \ \mu > 0$

Lemma: If $f \in C^1(\mathbb{R}^d)$ is μ-strongly convex, the above holds.

Pf: $\phi(x) := f(x) - \frac{\mu}{2}\|x\|^2$ is convex by def. of f μ-strongly convex

$\Rightarrow \phi(y) \geq \phi(x) + \langle \nabla\phi(x), y - x\rangle \quad \forall x, y$

$\Rightarrow f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|y - x\|^2$

minimize wrt. y $\quad f(x^*) \geq f(x) - \frac{1}{2\mu}\|\nabla f(x)\|^2$ $\quad \square$

Rmk: $x \mapsto |x|$ is convex but doesn't satisfy Polyak's ineq.

$x \mapsto x^2 + 3(\sin x)^2$ satisfies Polyak w/ $\mu = \frac{1}{32}$ but is not convex $\rightarrow$

↳ Polyak's ineq. independent of convexity.

Gradient Descent

$x_{t+1} = x_t - \alpha \nabla f(x_t)$

SGD: $-g(x_t)$ where $\mathbb{E}[g(x)] = \nabla f(x)$

SGD: Fixed step size will not let us stay at min, but rather in an α-nbhd of it.

↳ but gets close exp-fast!

Thm: $f \in C^1(\mathbb{R}^d)$, μ-Polyak, L-Lipschitz gradient.

$\forall x \in \mathbb{R}^d$, $g_1, \dots, g_T$ i.i.d. r.v.s w/ values in $\mathbb{R}^d$ s.t. $\mathbb{E}[g_t(x)] = \nabla f(x)$ & $\mathbb{E}[\|g_t(x)\|^2] \leq \gamma^2$.

Consider $x_{t+1} = x_t - \alpha g_t(x_t)$, $\alpha \in [0, \frac{1}{2\mu}]$. Then

$$\mathbb{E}[f(x_T)] - f(x^*) \leq (1 - 2\mu)^T(f(x_0) - f(x^*)) + \frac{L\alpha\gamma^2}{4\mu}.$$

Pf: $f(y) - f(x) - \langle \nabla f(x), y - x\rangle \leq \frac{L}{2}\|y - x\|^2$ from prev. lecture. Use $y = x_{t+1} = x_t - \alpha g_t(x_t)$, $x = x_t$

↳ $f(x_{t+1}) \leq f(x_t) - \alpha\langle \nabla f(x_t), g_t(x_t)\rangle + \frac{L}{2}\alpha^2\|g_t(x_t)\|^2$

$[f(x_{t+1}) | g_1, \dots, g_{t-1}] \leq f(x_t) - \alpha\|\nabla f(x_t)\|^2 + \frac{L\alpha^2}{2}\mathbb{E}[\|g_t(x_t)\|^2 | g_1, \dots, g_{t-1}]$

$\leq f(x_t) - 2\mu\alpha(f(x_t) - f(x^*)) + \frac{L\alpha^2}{2}\gamma^2.$

$[f(x_{t+1})] - f(x^*) \leq (\mathbb{E}[f(x_t)] - f(x^*))(1 - 2\mu\alpha) + L\alpha^2\gamma^2$ $\rightarrow$ iterate inequality recursively. $\square$

Thm: $C \subseteq \mathbb{R}^d$ compact, convex w/ diam δ ($x, y \in C \Rightarrow \|x-y\| \leq \delta$)
$f: C \to \mathbb{R}$ convex w/ global min. at $x^* \in C$.
$\forall x \in \mathbb{R}^d$, $g_1, \ldots, g_T$ are i.i.d. r.v.s w/ $\mathbb{E}[g_t(x)] = \nabla f(x)$
& $\mathbb{E}[\|g_t\|^2] \leq \gamma^2$. Consider $x_t = P_C(x_{t-1} - \alpha_t g_t(x_{t-1}))$

With $\alpha_t \leq \alpha_{t-1}$, $\bar{x} := \frac{1}{T}\sum_{i=0}^{T-1} x_t$, then

$$\mathbb{E}[f(\bar{x})] \leq f(x^*) + \frac{1}{2T}\left(\gamma^2\|\alpha\|_1 + \frac{\delta^2}{\alpha_T}\right)$$
$$\leq f(x^*) + \sqrt{\frac{2}{T}}\gamma^\delta \text{ for } \alpha_t = \frac{\delta}{\sqrt{2t}\,\gamma}$$

Pf: $\mathbb{E}[\|x_t - x\|^2 | g_1, \ldots, g_{t-1}]$

$$\leq \mathbb{E}[\|x_{t-1} - \alpha_t g_t(x_{t-1}) x^*\|^2 | \ldots]$$

$$= \mathbb{E}[\|x_{t-1} - x^*\|^2 - 2\alpha_t \langle g_t(x_{t-1}), x_{t-1} - x^*\rangle + \alpha_t^2 \|g_t(x_{t-1})\|^2$$

$$\leq \|x_{t-1} - x^*\|^2 - 2\alpha_t \langle \nabla f(x_{t-1}), x_{t-1} - x^*\rangle + \alpha_t^2 \gamma^2$$

$$\overset{\text{convexity}}{\leq} \|x_{t-1} - x^*\|^2 - 2\alpha_t (f(x_{t-1}) - f(x^*)) + \alpha_t^2 \gamma^2$$

$$\to \mathbb{E}[f(x_{t-1})] - f(x^*) \leq \frac{\gamma^2}{2}\alpha_t + \frac{1}{2\alpha_t}\mathbb{E}[\|x_{t-1} - x^*\|^2 - \|y_t - x^*\|^2]$$

$\frac{1}{T}\sum_{t=1}^{T} \ldots$ " $\leq \Sigma$ "

side calc.

$$\sum_{t=1}^{T} \frac{1}{2\alpha_t}(\|x_{t-1} - x^*\|^2 - \|x_t - x^*\|^2) = \frac{1}{2\alpha_1}\underbrace{\|x_0 - x^*\|^2}_{\leq \delta^2} \underbrace{- \frac{1}{2\alpha_T}\|x_T - x^*\|^2}_{\leq 0} + \sum_{t=2}^{T} \underbrace{\left(\frac{1}{2\alpha_t} - \frac{1}{2\alpha_{t-1}}\right)}_{\geq 0} \underbrace{\|x_{t-1} - x^*\|}_{\leq \delta^2}$$

$$\leq \frac{\delta^2}{2\alpha_1} + \frac{\delta^2}{2}\sum_{t=2}^{T} \underbrace{\frac{1}{\alpha_t} - \frac{1}{\alpha_{t-1}}}_{\frac{1}{\alpha_T} - \frac{1}{\alpha_1}} \leq \frac{\delta^2}{2\alpha_T}.$$

$$\to \mathbb{E}[f(\bar{x})] \leq f(x^*) + \frac{1}{2T}\left(\gamma^2\|\alpha\|_1 + \frac{\delta^2}{\alpha_T}\right)$$

$$\alpha_t = \frac{\delta}{\sqrt{2t}}\gamma \longrightarrow \leq f(x^*) + \sqrt{\frac{2}{T}}\gamma\delta.$$

$\square$

Stochastic subgradient + convexity $\to \mathcal{O}(\frac{1}{\varepsilon^2})$ steps (conv. up to ε)
+ Polyak $\to \mathcal{O}(\frac{1}{\varepsilon})$ steps

grad. desc. + Polyak $\to \mathcal{O}(\log 1/\varepsilon)$ steps.

Newton + strong convexity $\to \mathcal{O}(\log\log 1/\varepsilon)$ steps
↳ 2nd-order Hessian inversion method.

Lecture 16 - 5/30/17

Ideas to improve (S)GD: - momentum (introduces short term memory)
- exploit 2nd order info "$D^2 f \cdot v$"
- hypergradient descent
- . . .

Deep Neural Networks

= NN w/ more than a single hidden layer

- until ~2006, almost all NNs in practice were "shallow"
  - more computing power & GPUs now
  - more data
  - more tricks → different initialization
    - → -- activation fn. - sigmoid < tanh < ReLU
    - → training changed

Problem w/ "vanishing gradient in backpropagation ☆

$\delta^{L-1} = W\delta^L$, $\delta^{L-2} = WW'\delta^L$, . . .

eigenvals of W's kill δ in deep nets!

Taking partial deriv. of weight far away from output gives a nearly 0 value → weight updates slowly

☆ Initialization of these far-away weights is crucial.

*Neural networks unpopular in academia since even if it works, nobody understands the representation or what it's doing.

Why and when should one use a deep net?

↳ Representational efficiency

Results that indicate that deep NNs can be more efficient (regarding representation/approximation) than shallow nets.

- 2013 - number of affine regions is larger when using ReLUs.
- 2014 - Bianchini & Scarselli : $f : \mathbb{R}^d \to \mathbb{R}$

$$A := \{x \in \mathbb{R}^d \mid f(x) \geq 0\}$$

Consider $\sum_{i=0}^{d-1} b_i(A)$ $\leq \Theta((\#\text{neurons})^d)$ for shallow NNs

$\geq$ can be $\Omega(2^\delta)$ for depth δ

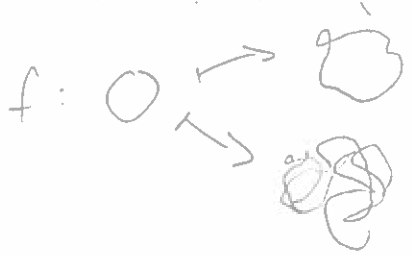
b_i Betti number

↳ number of topological "holes"

- 2016 - Poole et. al. $f : \mathbb{R}^d \to \mathbb{R}^{d'}$, where $d, d' \geq 2$

shallow:

deep:



consider : $\frac{\text{length}}{\text{diam}}$ of image

$\mathbb{E}$ of ↗ w.r.t. randomly chosen network weights (no training)

- Hence decision boundaries can be more complex in deep!

- 2015 - Telgarsky: $\bar{\mathfrak{f}}(m, L) \subseteq \mathbb{R}^{\mathbb{R}}$, fns rep. by NN using ReLU.

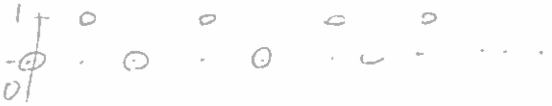
m neurons/layer, L layer

For every $f \in \bar{\mathfrak{f}}(m,L)$, define $\tilde{f}(x) := \mathbb{1}_{f(x) \geq 1/2}$.

Emp. risk of training set S : $\hat{R}(f) = \frac{1}{|S|} \sum_{(x,y) \in S} \mathbb{1}_{\tilde{f}(x) \neq y}$

For $k \in \mathbb{N}$, $n = 2^k$, choose $S := (x_i, y_i)_{i=0}^{n-1}$, with $x_i = \frac{i}{n}$, $y_i = i \bmod 2$

Thm: 1. There is a $h \in \bar{\mathfrak{f}}(2, 2k)$ for which $\hat{R}(h) = 0$.



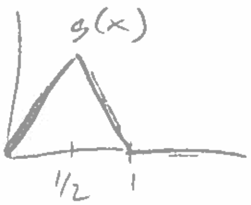
2. If $m, L \in \mathbb{N}$ and $m \leq 2^{(\frac{k-2}{L} - 1)}$, then $\forall f \in \bar{\mathfrak{f}}(m,L)$, $\hat{R}(f) \geq 1/6$.

Rmk: To represent $h \in \bar{\mathfrak{f}}(2,2k)$ w/ $\sim\sqrt{k}$ layers still requires $\sim 2^{\sqrt{k}}$ neurons/layer

Lecture 16 cont'd — 5/30/17

Telgarsky's thm. proof:

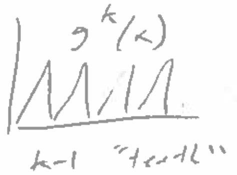
Let $g:\mathbb{R}\to\mathbb{R}$, $g(x) := \begin{cases} 2x, & 0 \le x \le 1/2 \\ 2(1-x), & 1/2 \le x \le 1 \\ 0, & o/w. \end{cases}$



ReLU:
$\sigma(x) := \max\{0, x\}$ $\qquad = \sigma(2\sigma(x) - 4\sigma(x - 1/4))$

Thus $g \in \bar{\mathcal{F}}(2,2)$.

Then $h(x) := g^k(x) = g \circ \cdots \circ g(x) \in \bar{\mathcal{F}}(2,2k)$



This easily represents the training data set S.

1. $\hat{R}(h) = 0$.

2. Every $f \in \bar{\mathcal{F}}(m,L)$ is piecewise linear w/ at most $t = (2m)^L$ pieces

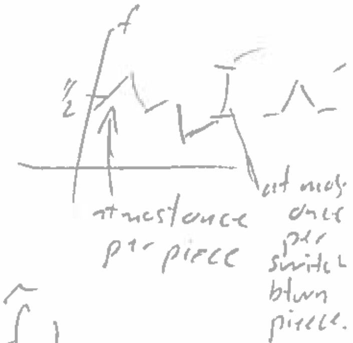
If f_1 has t_1 pieces → $f_1 + f_2$ has ≤ $(t_1 + t_2)$ pieces
& f_2 has t_2 pieces → $f_1 \circ f_2$ has ≤ $t_1 \cdot t_2$ pieces

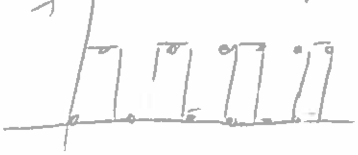
→ graph of f can cross 1/2 at most $2t-1$ times

→ $\tilde{f}$ is piecewise constant w/ at most $2t$ intervals

→ $n - 2t$ points are in intervals containing ≥1 point

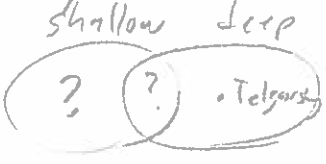
→ out of these, at least 1/3 misclassified





This example shows deep nets can represent fns. much more simply than shallow!

⊗ Doesn't show that there are no fns. more simply represented by shallow than deep

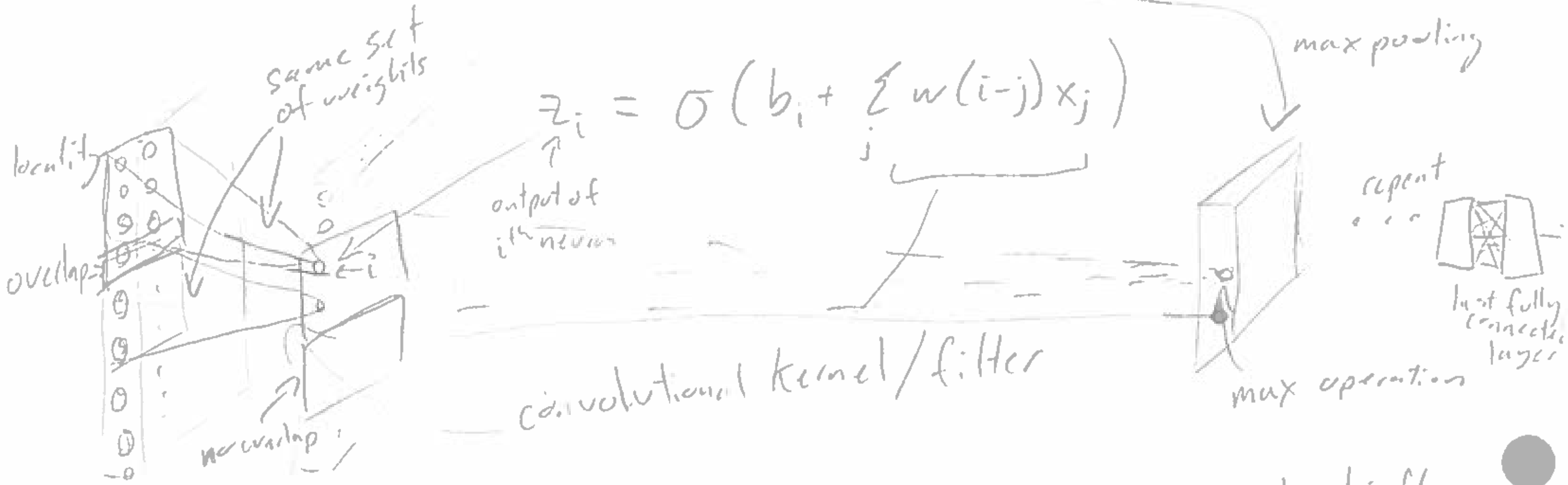


Lecture 17 - 6/1/17

Convolutional Nets

, 2012 ImageNet breakthrough using conv. nets

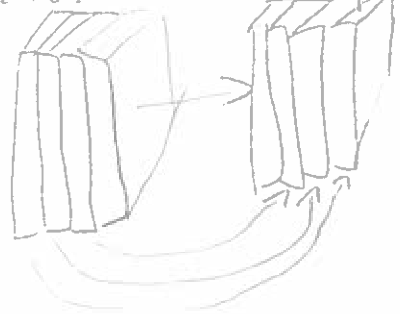
, repeating ingredients - locality ~~- small support of conv. kernel~~
- weight sharing - same conv. kernel used throught layer
  L> translation - invariant
- pooling & down-sampling



$$z_i = \sigma\left(b_i + \sum_j \underbrace{w(i-j)}_{} x_j\right)$$

☒ Convolution reflects structure in data - 2D like image, local influences
L> Translational invariance of network <=> tr. inv. of images - dog is a do
when on left or right.

☒ Conv. is run in parallel!



☒ Depth is usually 10~20 layers

Lecture 17 cont'd – 6/1/17

Optimization – How does all this work?

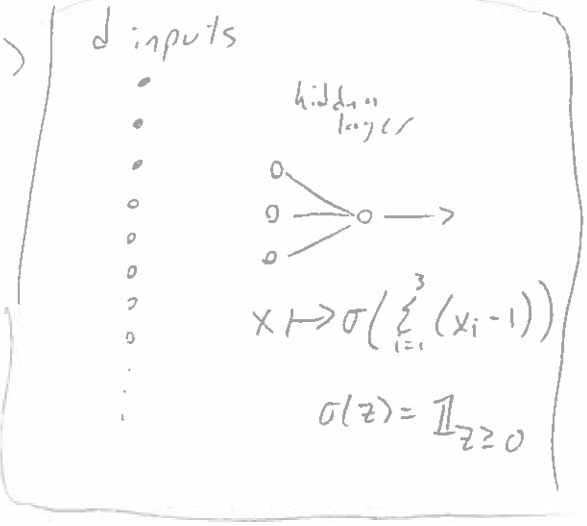
NP-hardness of ERM: graph $G = (\underset{\text{vertices}}{V}, \overset{\text{edges}}{E})$, $V = \{1, \dots, d\}$

assign $S_G := \{(e_i, 0), (e_i + e_j, 1), (0, 1)\}_{i \in V, (i,j) \in E}$

where $e_i \in \{0,1\}^d$ s.t. $(e_i)_k = \delta_{ik}$. ↳ labels vertices 0, labels edges 1

Recall: G is 3-colorable iff $\exists\, \chi : V \to \{1,2,3\}$ s.t. $(i,j) \in E \Rightarrow \chi(i) \neq \chi(j)$

Prop: For any G w/ d vertices, $\exists\, h \in \mathcal{F}_d$ that correctly classifies S_G iff G is 3-colorable. →


d inputs
hidden layer
$x \mapsto \sigma\left(\sum_{i=1}^{3} (x_i - 1)\right)$
$\sigma(z) = \mathbb{1}_{z \geq 0}$

(embedding an NP-hard problem in a simple NN)

Pf: Assume G 3-colorable.

let $\underset{\in\{1,2,3\}}{w_{\ell,}}\underset{\uparrow\text{input}}{{}_{i}} = \begin{cases} -1 & \text{if } \chi(i) = \ell \\ 1 & \text{o/w} \end{cases}$

$$h(x) = 1 \text{ iff } \forall\, \ell \in \{1,2,3\} : \sum_k w_{\ell,k}\, x_k \geq -1/2$$

This classifies S_G correctly since

- $h(0) = 1$
- $h(e_i) = 0$ since if $\chi(i) = \ell$ then $w_{\ell,i} = -1$ so $\sum_k w_{\ell,k} (e_i)_k = -1 < -1/2$.
- $h(e_i + e_j) = 1$ since $\sum_k w_{\ell,k} (e_i + e_j)_k = w_{\ell,i} + w_{\ell,j} \geq 0$ since G is 3-colorable ($\Rightarrow \chi(i) = \ell$ or $\chi(j) = \ell$.

//

Converse?

Converse pf: Assume $h \in \mathcal{F}_d$ that correctly classifies S_G.

$h^{-1}(\{1\}) = H_1 \cap H_2 \cap H_3 := H \ni 0$ (since output is only 1 iff all 3 hidden nodes output 1)

$\forall (i,j) \in E,\ e_i + e_j \in H$

$0,\ e_i + e_j \in H \Rightarrow \frac{e_i + e_j}{2} \in \mathcal{H}$ by convexity.

$\chi(i) := \min\{l \mid e_i \notin \mathcal{H}_l\}$

If $(i,j) \in E$, WTS $\chi(i) \neq \chi(j)$. Assume $(i,j) \in E$ but $\chi(i) = \chi(j) = l$.

Then $e_i, e_j \notin \mathcal{H}_l$ so by convexity $\frac{e_i + e_j}{2} \notin \mathcal{H}_l$ $\lightning$

Thus $(i,j) \in E \Rightarrow \chi(i) \neq \chi(j) \Rightarrow G$ is 3-colorable. □

Hence even a simple NN can embed an NP-hard problem.

But that's a discrete combinatorial problem, what about smooth targets?

NP hardness of classifying stationary points:

Consider $Q \in \mathbb{Z}^{d \times d}$, $f: \mathbb{R}^d \to \mathbb{R}$. $f(x) = \sum_{i,j=1}^{d} Q_{ij} x_i^2 x_j^2$.

At $x = 0$, $\nabla f = 0$ & $\nabla^2 f = 0 \Rightarrow$ stationary pt. but no info on if max, min, saddle

If at $x = 0$ f has local min, then it's a global min.

Suppose $\exists x$ st. $f(x) < 0$, then $\mathbb{R} \ni \lambda \mapsto f(\lambda x) = \lambda^4 \underbrace{f(x)}_{<0} \Rightarrow$ close to x, fn. descends in some directions then it cannot be a local min.

Def: Q is "copositive" iff $\langle z, Qz \rangle \geq 0\ \forall z \in \mathbb{R}^d_+$

The question "Is Q not copositive?" is NP-complete!!