Course Note for CPSC532D STAT Learn Theory

Lec ERM, uniform convergence bound.

empirical risk minimization

$\mathrm{ERM}(S) \in \underset{h\in\mathcal{H}}{\mathrm{argmin}}\, L_S(h)$
$= \underset{h\in\mathcal{H}}{\mathrm{argmin}}\, \frac{1}{m}\sum_{i=1}^{m} \ell(h, z_i)$

let $\hat{h}_S = \mathrm{ERM}(S)$

let $h^* = \arg \inf_{h\in\mathcal{H}} L_D(h)$

how ERM generalize? uniform convergence ≤ 0 as $\hat{h}_S \in \mathrm{argmin}\, L_S(h)$ Hoeffding inequality

$$\underbrace{L_D(\hat{h}_S) - L_D(h^*)}_{\text{estimation error}} = \underbrace{L_D(\hat{h}_S) - L_S(\hat{h}_S)}_{\text{how } \hat{h}_S \text{ overfits}} + \underbrace{L_S(\hat{h}_S) - L_S(h^*)}_{\text{how } \hat{h}_S \text{ fits } S} + \underbrace{L_S(h^*) - L_D(h^*)}_{\text{how } h^* \text{ underfits}}$$

$$\underbrace{L_D(\hat{h}_S) - L_{bayes}}_{\text{excess error}} = \underbrace{L_D(\hat{h}_S) - \inf_{h\in\mathcal{H}} L_D(h)}_{\text{estimation error}} + \underbrace{\inf_{h\in\mathcal{H}} L_D(h) - L_{bayes}}_{\text{approximation error}}$$

	estimation error	approximation error
increase $\lvert\mathcal{H}\rvert$	$\uparrow$	$\downarrow$
increase m	$\searrow 0$ (ideally)	$-$

concentration inequalities

bound the deviation of independent random variables from their expectations

$t_i = \ell(h, z_i) - \mathbb{E}_z\, \ell(h, z)$ $\qquad L_D(h) = \mathbb{E}_{S\sim D^m} L_S(h)$

large law number: $m\to\infty$ $\frac{1}{m}\sum_{i=1}^{m} t_i \to 0$

central limit theorem: $\frac{1}{m}\sum_{i=1}^{m} t_i \xrightarrow{d} \mathcal{N}(0, \sigma^2/m)$ $\quad \sigma^2 = \mathrm{Var}(t_i)$

Hoeffding: Let $x_1, \dots, x_m$ be independent with:

- $\mathbb{E}\, x_i = \mu$
- $\Pr(a \le x_i \le b) = 1$

Let $\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$, then

① $\Pr\left(\bar{x} \le \mu + (b-a)\sqrt{\frac{\log 1/\delta}{2m}}\right) \ge 1-\delta$ $\qquad \Pr\left(\bar{x} > \mu + (b-a)\sqrt{\frac{\log 1/\delta}{2m}}\right) < \delta$

Let $Y_i = -x_i$. Then $\Pr\left(\bar{Y} \le -\mu + (b-a)\sqrt{\frac{\log 1/\delta}{2m}}\right) \ge 1-\delta$

② $\Pr\left(\bar{x} \ge \mu - (b-a)\sqrt{\frac{\log 1/\delta}{2m}}\right) \ge 1-\delta$ $\qquad \Pr\left(\bar{x} < \mu - (b-a)\sqrt{\frac{\log 1/\delta}{2m}}\right) < \delta$

$P(A+B) < P(A) + P(B)$

③ $\Pr\left(\lvert\bar{x} - \mu\rvert \le (b-a)\sqrt{\frac{\log 2/\delta}{2m}}\right) \ge 1-\delta$ $\qquad \Pr\left(\lvert\bar{x} - \mu\rvert > (b-a)\sqrt{\frac{\log 2/\delta}{2m}}\right) < \delta$

use Hoeffding to bound $L_S(h^*) - L_D(h^*)$

$\Rightarrow \Pr\left(L_S(h^*) - L_D(h^*) \geq (b-a)\sqrt{\frac{\log 1/\delta}{2m}}\right) \leq \delta$

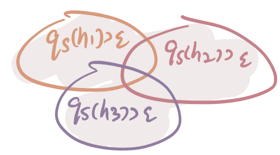
we can't use Hoeffding on $L_D(\hat{h}_S) - L_S(\hat{h}_S)$
$\ell(\hat{h}_S, z_i)$ are not independent of each other
change z_i, will change $\hat{h}_S$

we can't use Hoeffding directly on $L_D(\hat{h}_S) - L_S(\hat{h}_S)$.
because $\ell(\hat{h}_S, z_i)$ is not independant of each other. change z_i will change $\hat{h}_S$, thus affect all other $\ell(\hat{h}_S, z_j)$

For $L_D(\hat{h}_S) - L_S(\hat{h}_S)$, we could use uniform convergence:
bound every $h \in \mathcal{H}$, the gap will not be too big. then it'll be small for $\hat{h}_S$

$$\sup_{h \in \mathcal{H}} L_D(h) - L_S(h) \leq \varepsilon \Rightarrow \forall h \in \mathcal{H} \quad L_D(h) - L_S(h) \leq \varepsilon$$

If $|\mathcal{H}| < \infty$: union bound



Let $q_S(h) = L_D(h) - L_S(h)$

$$\Pr(\exists h \in \mathcal{H}, q_S(h) > \varepsilon) \leq \sum_{h \in \mathcal{H}} \Pr(q_S(h) > \varepsilon)$$

$P(A \cup B) \leq P(A) + P(B)$

Not tight enough because of $P(AB) \geq 0$

now $\ell(h, z)$ is indepedent

$$\Pr(\exists h \in \mathcal{H}, L_D(h) - L_S(h) > \varepsilon) \leq \sum_{h \in \mathcal{H}} \Pr(L_D(h) - L_S(h) > \varepsilon) \leq |\mathcal{H}| \exp\left(\frac{-2m\varepsilon^2}{(b-a)^2}\right)$$

$$\Rightarrow \Pr\left(L_S(h^*) - L_D(h^*) + \sup_{h \in \mathcal{H}} L_D(h) - L_S(h) > 2\varepsilon\right) \leq (1 + |\mathcal{H}|) \exp\left(\frac{-2m\varepsilon^2}{(b-a)^2}\right)$$

so the overall bound would be:

Assume
- $\ell(h, z) \in [a, b] \ \forall z, h$,
- $\mathcal{H}$ finite
- $\hat{h}_S$ is an ERM

$$L_D(\hat{h}_S) - \inf_{h \in \mathcal{H}} L_D(h) \leq L_D(\hat{h}_S) - L_S(\hat{h}_S) + L_S(h^*) - L_D(h^*)$$
$$\leq \sup_{h \in \mathcal{H}} [L_D(h) - L_S(h)] + L_S(h^*) - L_D(h^*)$$

by showing:
- $\forall h \in \mathcal{H}, \ \Pr(L_D(h) - L_S(h) > \varepsilon) \leq \frac{\delta}{|\mathcal{H}|+1}$
- $\Pr(L_S(h^*) - L_D(h^*) > \varepsilon) \leq \frac{\delta}{|\mathcal{H}|+1}$

$$\therefore \Pr\left(L_D(\hat{h}_S) - \min_{h \in \mathcal{H}} L_D(h) \leq 2\varepsilon\right) \geq 1 - \delta \qquad \varepsilon = (b-a)\sqrt{\frac{1}{2m} \log \frac{|\mathcal{H}|+1}{\delta}}$$

$$\Pr\left(L_D(\hat{h}_S) - \min_{h \in \mathcal{H}} L_D(h) \leq \sqrt{\frac{2}{m} \log \frac{|\mathcal{H}|+1}{\delta}}\right) \geq 1 - \delta$$

L3. MARKOV Inequality

to prove Hoeffding

x cannot be too big with high prob $\mathbb{P}(X \geq 0) = 1$

- Markov's inequality: X nonnegative-valued random variable
  $\Pr(X \geq t) \leq \frac{1}{t}\mathbb{E}X$ for all $t > 0$ — is weak because only assume non-neg

Pf $\begin{cases} X \geq 0 \\ X \geq t \text{ when } X \geq t \end{cases}$ $\therefore$ $\underbrace{X \geq t\mathbb{1}(X \geq t)}_{\text{always holds}}$ $\therefore \mathbb{E}X \geq t\mathbb{E}\mathbb{1}(X \geq t)$ (t is scalar) $= t\Pr(X \geq t)$

Let $\delta = \frac{\mathbb{E}X}{t}$ then $\Pr(X \geq \frac{\mathbb{E}X}{\delta}) \leq \delta \Rightarrow \Pr(X \leq \frac{\mathbb{E}X}{\delta}) \geq 1-\delta$

$Y = (X - \mathbb{E}X)^2$

- Chebyshev's inequality: for any X (just random variable), $\Pr(|X - \mathbb{E}X| \geq \varepsilon) \leq \frac{\mathrm{Var}X}{\varepsilon^2}$

Pf $\Pr(|X - \mathbb{E}X| \geq \varepsilon) = \underbrace{\Pr((X-\mathbb{E}X)^2 \geq \varepsilon^2) \leq \frac{1}{\varepsilon^2}\mathbb{E}(X-\mathbb{E}X)^2}_{\text{Markov inequality}} = \frac{\mathrm{Var}X}{\varepsilon^2}$

$\frac{\mathrm{Var}}{\varepsilon^2} = \delta$

with probality at least $1-\delta$, $|X - \mathbb{E}X| \leq \sqrt{\frac{\mathrm{Var}X}{\delta}} = \frac{\mathrm{std}(X)}{\sqrt{\delta}}$

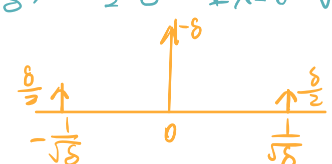
- Consider iid $X_1, X_2, \cdots, X_m$, $\mathbb{E}(X_i) = \mu$, $\mathrm{Var}(X_i) = \sigma^2$
  $\bar{X} = \frac{1}{m}\sum_{i=1}^{m} X_i$ $\quad \mathbb{E}\bar{X} = \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}X_i = \mu$ $\quad \mathrm{Var}(\bar{X}) = \frac{1}{m^2}\cdot m\sigma^2 = \frac{\sigma^2}{m}$

$\Rightarrow$ with $\Pr \geq 1-\delta$, $|\bar{X} - \mu| \leq \frac{\sigma}{\sqrt{m\delta}}$

chebyshev's inequality may not be good enough!
$\frac{1}{\sqrt{m}}$ is fine, but the dependence on δ is the problem $\sqrt{\log\frac{1}{\delta}}$ better than $\frac{1}{\sqrt{\delta}}$

$\Pr(X=0) = 1-\delta$, $\Pr(X = \frac{1}{\sqrt{\delta}}) = \Pr(X = -\frac{1}{\sqrt{\delta}}) = \frac{1}{2}\delta$ $\quad \mathbb{E}X = 0 \quad \mathrm{Var}X = 1$

$\Pr(|\bar{X}| \leq \frac{1}{\sqrt{m\delta}}) \geq 1-\delta$



chebyshev's cares about the worst case

- Chernoff bounds: construct a non-negative random variable
  $\lambda > 0$ arbitrary $\quad Y = e^{\lambda(X-\mu)}$ $\quad \Pr(Y \geq t) \leq \frac{1}{t}\mathbb{E}Y$

$\Pr(e^{\lambda(X-\mu)} \geq e^{\lambda\varepsilon}) \leq e^{-\lambda\varepsilon}\mathbb{E}e^{\lambda(X-\mu)}$
$\Updownarrow$
$= \Pr(\lambda(X-\mu) \geq \lambda\varepsilon)$

centred moment-generating function $M_X(\lambda)$

$e^t = 1 + t + \frac{t^2}{2!} + \frac{t^3}{3!} + \cdots$

$M_X(\lambda) = \mathbb{E} e^{\lambda(X-\mu)} = 1 + \lambda \underbrace{\mathbb{E}[X-\mu]}_{0} + \frac{\lambda^2}{2!}\mathbb{E}[(X-\mu)^2] + \frac{\lambda^3}{3!}\mathbb{E}[(X-\mu)^3] + \cdots$

kth derivative of $M_X(\lambda)$ at $\lambda=0$: $M_X^{(k)}(0) = \mathbb{E}[(X-\mu)^k]$

If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $\mathbb{E} e^{\lambda(X-\mu)} = e^{\frac{1}{2}\lambda^2\sigma^2}$

for $X \sim \mathcal{N}(0,1)$

$$\mathbb{E}_{X\sim\mathcal{N}(0,1)} e^{\lambda X} = \int \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} e^{\lambda x} dx = e^{-\frac{1}{2}\lambda^2} \underbrace{\int \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\lambda)^2} dx}_{1 \quad \mathcal{N}(\lambda,1)} = e^{-\frac{1}{2}\lambda^2}$$

for $Y \sim \mathcal{N}(0, \sigma^2)$

$$\mathbb{E}_{Y\sim\mathcal{N}(0,\sigma^2)} e^{\lambda Y} = \mathbb{E}_{\sigma X\sim\mathcal{N}(0,\sigma^2)} e^{\lambda\sigma X} = \mathbb{E}_{X\sim\mathcal{N}(0,1)} e^{(\lambda\sigma)X} = e^{-\frac{1}{2}\lambda^2\sigma^2}$$

$\Pr(X \geq \mu + \varepsilon) \leq \underbrace{e^{-\lambda\varepsilon} e^{\frac{1}{2}\lambda^2\sigma^2}}$ we can optimize λ to get tighter bound

$\sigma^2\lambda = \varepsilon \rightarrow \lambda = \frac{\varepsilon}{\sigma^2} \rightarrow e^{\frac{1}{2}\frac{\varepsilon^2}{\sigma^2} - \frac{\varepsilon^2}{\sigma^2}} = e^{-\frac{\varepsilon^2}{2\sigma^2}}$

$\Pr(X \geq \mu + \varepsilon) \leq e^{-\frac{\varepsilon^2}{2\sigma^2}}$ $\qquad e^{-\frac{\varepsilon^2}{2\sigma^2}} = \delta$

For $X \sim \mathcal{N}(\mu, \sigma^2)$:

- with probability at least $1-\delta$ $\quad X \leq \mu + \sigma\sqrt{2\log\frac{1}{\delta}}$

If X is $\underbrace{\text{s.t. } \mathbb{E} e^{\lambda(X-\mathbb{E}X)} \leq e^{\frac{1}{2}\lambda^2\sigma^2}}_{SG(\sigma)}$, then $\Pr(X \leq \mathbb{E}X + \sigma\sqrt{2\log\frac{1}{\delta}}) \geq 1-\delta$

$SG(\sigma)$ sub-gaussian with σ (no heavier tails than Gaussian)

σ does not imply the variance of the sub-gaussian X
may have relation but not equal !

- sub-Gaussian : $X \sim SG(\sigma)$

random variable X with mean $\mu = \mathbb{E}[X]$ is sub-Gaussian with σ

(i) if $\mathbb{E}[e^{\lambda(X-\mu)}]$ exists and for all $\lambda \in \mathbb{R}$, $\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\frac{1}{2}\lambda^2\sigma^2}$

if $\sigma_1 < \sigma_2$ then $X \sim SG(\sigma_1) \Rightarrow X \sim SG(\sigma_2)$ $\qquad SG(\sigma_1) \subseteq SG(\sigma_2)$

(ii) If $X \in SG(\sigma)$, $aX \in SG(|a|\sigma)$ for $\forall a \in \mathbb{R}$

$M_X(\lambda) \leq e^{\frac{1}{2}\lambda^2\sigma^2} \Rightarrow \mathbb{E} e^{\lambda(aX - a\mathbb{E}X)} = \mathbb{E} e^{a\lambda(X-\mathbb{E}X)} \leq e^{\frac{1}{2}(a\lambda)^2\sigma^2} = e^{\frac{1}{2}\lambda^2(|a|\sigma)^2}$

(iii) If $X_1 \in SG(\sigma_1)$, $X_2 \in SG(\sigma_2)$ are independent $\quad X_1 + X_2 \in SG(\sqrt{\sigma_1^2 + \sigma_2^2})$

$\mathbb{E} e^{\lambda(X_1+X_2-\mathbb{E}X_1-\mathbb{E}X_2)} = \mathbb{E} e^{\lambda(X_1-\mathbb{E}X_1)} \mathbb{E} e^{\lambda(X_2-\mathbb{E}X_2)} \leq e^{\frac{1}{2}\lambda^2\sigma_1^2} \cdot e^{\frac{1}{2}\lambda^2\sigma_2^2}$

we can use the above properties to construct a set of X

Hoeffding's lemma :

a real-valued random variable bounded in $[a,b]$ is $SG(\frac{b-a}{2})$

i.e. $Pr(a \le X \le b) = 1$, $X \in SG(\frac{b-a}{2})$ see pf in notes 3.

- Hoeffding: If $X_1, X_2, \dots X_m$ iid with $\mathbb{E}X_i = \mu$, $Pr(a \le X_i \le b) = 1$

  then $Pr(\frac{1}{m}\sum_{i=1}^m X_i > \mu + (b-a)\sqrt{\frac{1}{2m}\log\frac{1}{\delta}}) \le \delta$

  $\frac{1}{m}\sum_{i=1}^m x_i \in SG(\frac{1}{m}\sqrt{m(\frac{b-a}{2})^2}) = SG(\frac{b-a}{2\sqrt{m}})$

  with prob at least $1-\delta$ $\frac{1}{m}\sum_{i=1}^m X_i \le \mu + \frac{b-a}{2\sqrt{m}}\sqrt{2\log\frac{1}{\delta}} = \mu + (b-a)\sqrt{\frac{\log\frac{1}{\delta}}{2m}}$

Lec 4. PAC Learning, infinite $\mathcal{H}$

learning algorithm from sample to $\mathcal{H}$

Defn: an algorithm A is agnostically PAC learns $\mathcal{H}$ with loss ℓ

if there is a function $m(\varepsilon, \delta)$ sample complexity function

s.t. for any D, for any $\varepsilon, \delta \in (0,1)$

if $S \sim D^m$ with $m \ge m(\varepsilon, \delta)$

then $Pr(L_D(A(S)) \le \inf_{h \in \mathcal{H}} L_D(h) + \varepsilon) \ge 1-\delta$

solve for m will get sample complexity

efficient : polynomial run time

Def $\mathcal{H}$ is agnostically PAC learnable if $\exists A$ that agnostically PAC learn $\mathcal{H}$

agnostic PAC is the worst case, $\mathcal{H}$ has nothing to do with data distribution.

an $m(\varepsilon, \delta)$ should work for any D

PAC learnable does not show how learn quickly in terms of m.

{ Agnostically PAC learning : work for any distribution ERM on finite $\mathcal{H}$

PAC learning is weaker : work only for realizable distribution

Def (PAC-Learn) $\mathcal{A}$ PAC-Learns $\mathcal{H}$ if

$\exists\, m: (0,1)^2 \to \mathbb{N}$ s.t.

for any (ε, δ), for any realizable $\mathcal{D}$, if $m \geq m(\varepsilon, \delta)$

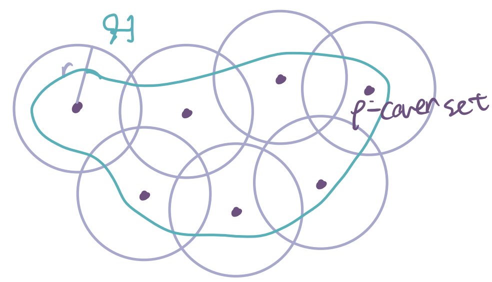
$$\Pr_{S\sim\mathcal{D}^m}\left(L_{\mathcal{D}}(\mathcal{A}(S)) \leq \varepsilon\right) \geq 1-\delta$$

Def (realizable). for a non-negative loss $\ell(h, z) \geq 0$.

$\mathcal{D}$ is realizable by $\mathcal{H}$ if there exists an $h^* \in \mathcal{H}$ s.t. $L_{\mathcal{D}}(h^*) = 0$

we already know finite case, what about infinite $\mathcal{H}$?

idea of covering number



still: uniform convergence

$$\sup_{h\in\mathcal{H}} L_{\mathcal{D}}(h) - L_S(h)$$

$$= \sup_{h\in\mathcal{H}} L_{\mathcal{D}}(h) - L_{\mathcal{D}}(h_j) + L_{\mathcal{D}}(h_j) - L_S(h_j) + L_S(h_j) - L_S(h) \quad \forall h_j$$

$$\leq \underbrace{\sup_{h\in\mathcal{H}} [L_{\mathcal{D}}(h) - L_{\mathcal{D}}(h_j)]}_{\text{Lipschitz}} + \underbrace{\max_{h_j\in T} [L_{\mathcal{D}}(h_j) - L_S(h_j)]}_{\text{union bound on size of } T} + \sup_{h\in\mathcal{H}} [L_S(h_j) - L_S(h)]$$

finite

as ρ decrease, the set to cover points become tighter

T is a ρ-cover for $\mathcal{H}$ of size $N(\mathcal{H}, \rho)$

specific case: Logistic regression

$\mathcal{Z} = \mathcal{X}\times\mathcal{Y}$ $\quad \mathcal{X} = \mathbb{R}^d$ $\quad \mathcal{Y} = \{-1, 1\}$

$\mathcal{H} = \{x \mapsto w\cdot x : w\in\mathbb{R}^d, \underline{\|w\| \leq B}\}$ equivalent to L_2 reg

$\ell_{\log}(h, (x,y)) = \ell_y(h(x)) = \log(1+\exp(-y\,h(x)))$

use absolute value to make analysis easier, but results looser

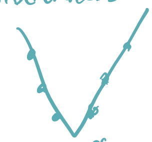
$$|L_{\mathcal{D}}(\underset{\text{weight } w}{h_w}) - L_{\mathcal{D}}(\underset{\text{weight } v}{h_v})| = \left|\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \ell_{\log}(h_w, (x,y)) - \ell_{\log}(h_v, (x,y))\right|$$

$\le \mathbb{E}\,| l_y(h_w(x)) - l_y(h_v(x)) |$ l_y is G-Lipschitz

$\le G\,\mathbb{E}\,| h_w(x) - h_v(x) |$

$G=1$ for l_y

input close → output close
smoothness



G-Lipschitz: l_y is G-Lips if

$$| l_y(\hat{y}_1) - l_y(\hat{y}_2) | \le G\,|\hat{y}_1 - \hat{y}_2| \quad \forall \hat{y}_1, \hat{y}_2$$

Euclidean distance (can be other distance if particular specifed)

⇒ if f' exist everywhere, and $\sup |f'(x)| \le G$, f is G-Lips

$$|f(x) - f(x')| = \left| \int_x^{x'} f'(t)\,dt \right| \le \int_x^{x'} |f'(t)|\,dt \le \int_x^{x'} G\,dt = G|x'-x|$$

if f is differetiable the Lips constance is the largest |derivaties| (upper bound of |derivative|)

for the LR: $|h_w(x) - h_v(x)| = |w\cdot x - v\cdot x| = |\langle w-v\rangle \cdot x| \le \|w-v\|\cdot\|x\|$

$l_y(\hat{y}) = \log(1+\exp(-y\hat{y}))$ $l_y'(\hat{y}) = \dfrac{-y\exp(-y\hat{y})}{1+\exp(-y\hat{y})} = \dfrac{-y}{\exp(y\hat{y})+1} = \begin{cases} -\dfrac{1}{e^{\hat{y}}+1}, & y=1 \\ \dfrac{1}{e^{-\hat{y}}+1}, & y=-1 \end{cases}$

$|l_y'(\hat{y})| \le 1$ l_y is 1-Lipschitz

$ax_1 + bx_2 \le \sqrt{a^2+b^2}\sqrt{x_1^2+x_2^2}$ $+2abx_1x_2$ $a^2x_1^2 + b^2x_2^2 + a^2x_2^2 + b^2x_1^2$ $(ax_2 - bx_1)^2$

$|h_w(x) - h_v(x)| = |w\cdot x - v\cdot x| = |\langle w-v\rangle\cdot x| \le \|w-v\|\cdot\|x\|$

Then ① $|L_D(h_w) - L_D(h_v)| \le \underbrace{(\mathbb{E}\|x\|)}_{\text{constant of distribution}} \cdot \underbrace{\|w-v\|}_{\text{distance from } h \text{ to } h_j \text{ at most } \rho}$

$\le C\rho$

Assume $\|x\| \le C$ on $\mathcal{D}$.

② $|L_S(h_w) - L_S(h_v)| \le \frac{1}{m}\sum_{i=1}^m \|x_i\|\,\|w-v\| \le C\rho$

③ $|L_D(h_j) - L_S(h_j)| \le (b-a)\sqrt{\frac{1}{2m}\log\frac{N(B,\rho)}{\delta}}$

$|h_w(x)| = |w\cdot x| \le \|w\|\cdot\|x\| \le BC$

l_y is not bounded, but we assume $\|w\| \le B$ and $\|x\| \le C$ so we can use hoeffdings

$\log(1+\exp(BC)) \le BC+1$ (just simpler)

So $\sup_{h\in H} L_D(h) - L_S(h) \le 2C\rho + (BC+1)\sqrt{\frac{1}{2m}\log\frac{N(B,\rho)}{\delta}}$

what is the covering number?

$N(B,\rho) \le \left(\frac{3B}{\rho}\right)^d$ (proof in notes 4)

$$\sup_{h\in H} L_D(h) - L_S(h) \le 2C\rho + (BC+1)\sqrt{\frac{1}{2m}\left(\log\frac{1}{\delta} + d\log\frac{3B}{\rho}\right)}$$

how to choose ρ to minimize the right term?

Let $\rho = \frac{B}{\sqrt{m}}$ (roughly optimal)

$$\le \frac{2CB}{\sqrt{m}} + (BC+1)\sqrt{\frac{\log\frac{1}{\delta} + \frac{d}{2}\log(9m)}{2m}}$$

"c" dominate by $O_p\left(\sqrt{\frac{\log m}{m}}\right)$

Big Op notation means some random variable is stochastically bounded (Hoeffding)

$$\Pr\left(\sup_{h\in\mathcal{H}} L_D(h)-L_S(h) > \frac{2BC}{\sqrt{m}} + (BC+1)\sqrt{\frac{\log 1/\delta + \frac{d}{2}\log 9m}{2m}}\right) < \delta$$

- the sample complexity depends on the input distribution $\mathbb{E}\|X\| \le c$
  × not allowed in agnostically PAC learning
  So this bound doesn't show Linear Regression is $\overset{\text{ERM}+}{}$ agnostically PAC learnable
- we can never achieve 0 loss on any D, so it's not realizable

L5 Rademacher complexity

distribution-specific complexity

covering number approach : depend on dimension d.
need bounded norm. B. scale sensetive
Rademacher complexity : do not depend on dimension d.

here we are going to bound on average instead of high probability.

how much can I overfit to sample S

$$\mathbb{E}_{S\sim D^m} \sup_{h\in\mathcal{H}} L_D(h)-L_S(h) = \mathbb{E}_{S\sim D^m}\left(\sup_{h\in\mathcal{H}} \underbrace{\mathbb{E}_{S'\sim D^m}[L_{S'}(h)}_{L_D(h)} - L_S(h)]\right)$$

① $\sup_y \mathbb{E}_x f_y(x) \le \mathbb{E}_x \sup_y f_y(x)$

Pf: for any y, we have $f_y(x) \le \sup_{y'} f_{y'}(x)$

then $\mathbb{E}_x f_y(x) \le \mathbb{E}_x \sup_{y'} f_{y'}(x)$

and then $\sup_y \mathbb{E}_x f_y(x) \le (\sup) \mathbb{E}_x \sup_y f_y(x)$

$$\overset{①}{\le} \mathbb{E}_{S,S'\sim D^m} \sup_{h\in\mathcal{H}} [L_{S'}(h) - L_S(h)]$$

$$= \mathbb{E}_{S,S'} \sup_h \frac{1}{m}\sum_{i=1}^m [\ell(h,z_i) - \ell(h,z_i')]$$

② let $\sigma_i \in [-1,1]$ for $i\in[m]$,

$$(u_i,u_i') = \begin{cases}(z_i, z_i') & \text{if } \sigma_i = 1\\ (z_i', z_i) & \text{if } \sigma_i = -1\end{cases}$$

for any $\vec{\sigma} = (\sigma_1,\dots,\sigma_m)$

then $\ell(h,z_i) - \ell(h,z_i') = \sigma_i(\ell(h,z_i) - \ell(h,z_i'))$

$\sigma_i \sim \text{Unif}$; $P(\sigma_i = 1) = P(\sigma_i = -1) = \frac{1}{2}$

$$\overset{②}{=} \mathbb{E}_{S,S',\vec\sigma}\, \mathbb{E}_{U,U'}\left[\sup_h \frac{1}{m}\sum_{i=1}^m \sigma_i(\ell(h,u_i) - \ell(h,u_i')) \,\middle|\, \underline{S,S',\vec\sigma}\right]$$

condition

not doing anything (deterministic $UU' \mid SS'\vec\sigma$)

switch order of expec

$P(SS')\cdot P(\sigma)\, P(UU' \mid SS'\sigma)$

$$= \mathbb{E}_{U,U',\vec\sigma}\, \mathbb{E}_{S,S'}\left[\sup_h \frac{1}{m}\sum_{i=1}^m \sigma_i(\ell(h,u_i) - \ell(h,u_i')) \,\middle|\, U,U',\vec\sigma\right]$$

$P(UU')\cdot P(\sigma)\, P(SS \mid UU'\sigma)$

deterministic

we can drop S

$U, U' \mid S, S', \vec{\sigma}$ is determinated
but $U, U' \sim D^m$, $U, U', \vec{\sigma}$ are independent
$S, S' \mid U, U', \vec{\sigma}$ is determinated

$$= \mathop{\mathbb{E}}_{U,U'} \mathop{\mathbb{E}}_{\sigma} \left[\sup_h \frac{1}{m} \sum_{i=1}^m \sigma_i \left(\ell(h, u_i) - \ell(h, u_i') \right) \right]$$

$U, U' \sim D^m$ rename $U \to S$

$$= \mathop{\mathbb{E}}_{S,S' \sim D^m} \mathop{\mathbb{E}}_{\sigma \sim Rad^m} \left[\sup_h \frac{1}{m} \sum_{i=1}^m \sigma_i \left(\ell(h, z_i) - \ell(h, z_i') \right) \right]$$

$\sup [\sigma_i (\ell(h, z_i)) + (-\sigma_i)\, \ell(h, z_i')]$

$$\le \mathop{\mathbb{E}}_{S} \mathop{\mathbb{E}}_{\sigma} \sup_h \frac{1}{m} \sum_{i=1}^m \sigma_i \, \ell(h, z_i) + \mathop{\mathbb{E}}_{S'} \mathop{\mathbb{E}}_{\sigma} \sup_h \frac{1}{m} \sum_{i=1}^m (-\sigma_i)\, \ell(h, z_i')$$

same

$$= 2 \mathop{\mathbb{E}}_{S,\sigma} \sup \frac{1}{m} \sum_{i=1}^m \sigma_i \, \ell(h, z_i)$$

③
$$= 2 \mathop{\mathbb{E}}_{S} Rad\left((\ell \circ \mathcal{H})|_S \right)$$

$$\mathop{\mathbb{E}}_{S \sim D^m} \sup_{h \in \mathcal{H}} L_D(h) - L_S(h) \le 2 \mathop{\mathbb{E}}_{S} Rad\left((\ell \circ \mathcal{H})|_S \right)$$

$$\mathop{\mathbb{E}}_{S \sim D^m} \sup_{h \in \mathcal{H}} L_S(h) - L_D(h) \le 2 \mathop{\mathbb{E}}_{S} Rad\left((\ell \circ \mathcal{H})|_S \right)$$

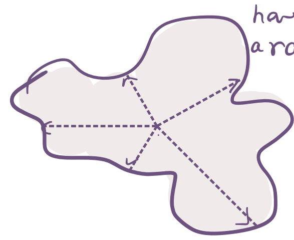
③ def

$\ell \circ \mathcal{H} = \{ z \mapsto \ell(h, z) : h \in \mathcal{H} \}$

$\mathcal{F}|_S = \{ (f(z_1), f(z_2), \dots, f(z_m)) : f \in \mathcal{F} \} \subseteq \mathbb{R}^m$ per-sample risk

$$Rad(V) = \mathop{\mathbb{E}}_{\vec{\sigma} \sim Rad^m} \sup_{v \in V} \frac{1}{m} \underbrace{\sum_{i=1}^m \sigma_i v_i}_{\vec{\sigma} \cdot v}$$

the biggest product one can get given a Radermacher vector



how far can get at a random direction

measure the size of the set

though Rad vector is axis direction

(i) suppose one-element hypothesis :

$$Rad(\{v\}) = \mathop{\mathbb{E}}_{\vec{\sigma}} \sup_{v \in \{v\}} \frac{\vec{\sigma} \cdot v}{m} = \mathop{\mathbb{E}}_{\vec{\sigma}} \frac{\vec{\sigma} \cdot v}{m} = \frac{(\mathbb{E} \vec{\sigma})\, v}{m} = 0$$

(ii) suppose $\mathcal{H}$ is bigger :

$$Rad(\{-1, 1\}^m) = \mathop{\mathbb{E}}_{\sigma} \sup_{v \in \{-1,1\}^m} \frac{\sigma \cdot v}{m} = 1$$

$$Rad([-1, 1]^m) = 1$$

$$Rad(\{v : \|v\| \le \sqrt{m}\}) = 1$$

$\mathcal{H}$ bigger, but Randemacher is equal

because Rad vector only looks at $\{\pm 1\}^m$

Gaussian complexity would look at any direction

property :

(i) $Rad(cV) = Rad(\{cv : v \in V\}) = \frac{1}{m} \mathop{\mathbb{E}}_{\sigma} \sup_{v \in V} \sigma \cdot (cv) = \frac{|c|}{m} \mathop{\mathbb{E}}_{\sigma} \sup_{v \in V} \mathrm{sign}(c)\, \vec{\sigma} \cdot v$

$c \in \mathbb{R}$ $= |c|\, Rad(V)$

(ii) $\mathrm{Rad}(V+W) = \mathrm{Rad}(\{v+w : v \in V, w \in W\}) = \frac{1}{m} \mathbb{E}_\sigma \sup_{v,w} \sigma(v+w) = \frac{1}{m} \mathbb{E}_\sigma \sup_v \sigma v + \sup_w \sigma w$

$= \mathrm{Rad}(V) + \mathrm{Rad}(W)$

if w is constant vector $\Rightarrow \mathrm{Rad}(w) = 0$

How to compute $\mathrm{Rad}(\ell \circ \mathcal{H}_{|S})$ practically?

- Talagrand's contraction lemma : 1-Lipschitz function

Let $\phi : \mathbb{R}^m \to \mathbb{R}^m$ be given by $\phi(t) = (\varphi_1(t_1), \cdots, \varphi_m(t_m))$, where each ϕ_i is ρ-Lip

then $\mathrm{Rad}(\phi \circ V) = \mathrm{Rad}(\{\phi(v) : v \in V\}) \leq \rho \, \mathrm{Rad}(V)$

Then for typical supervised learning losses

$(\ell \circ \mathcal{H})|_S = \{(\ell(h, z_1), \cdots, \ell(h, z_m)) : h \in \mathcal{H}\}$

$= \{(l_{y_1}(h(x_1)), \cdots, l_{y_m}(h(x_m))) : h \in \mathcal{H}\}$

$= (l_{S_y} \circ \mathcal{H})|_{S_x}$ might depend on S_y

if l_{y_i} (the loss func of a prediction for y_i) are all ρ-Lipschitz.

then Talagrand's lemma gives : $S_x = (x_1, \cdots, x_m)$

$\mathrm{Rad}(\ell \circ \mathcal{H}|_S) \leq \rho \, \mathrm{Rad}(\mathcal{H}|_{S_x})$

only holds for real-valued hypothesis ($\mathcal{H}$ don't need to be function) (can be param of Gaussian ..)

for linear regression, ℓ_Y is 1-Lipschitz $\mathcal{H}_B = \{x \mapsto w \cdot x : \|w\| \leq B\}$

$\mathrm{Rad}(\ell_y \circ \mathcal{H}|_S) = \mathrm{Rad}(\{\ell_{y_1}(h(x_1)), \ell_{y_2}(h(x_2)), \cdots, \ell_{y_m}(h(x_m))\})$

$\leq \mathrm{Rad}(\mathcal{H}|_S)$

$\mathrm{Rad}(\mathcal{H}_B) = \mathbb{E}_\sigma \sup_{\|w\| \leq B} \frac{1}{m} \sum_{i=1}^m \sigma_i \langle w, x_i \rangle$

$= \mathbb{E}_\sigma \sup_{\|w\| \leq B} \frac{1}{m} \langle w, \sum_{i=1}^m \sigma_i x_i \rangle$

$\leq \frac{1}{m} \mathbb{E}_\sigma B \| \sum_{i=1}^m \sigma_i x_i \|$

$\leq \frac{1}{m} B \sqrt{\mathbb{E}_\sigma \| \sum_i \sigma_i x_i \|^2}$ to get rid of σ

① $\|\sum_i a_i\|^2 = \langle \sum_i a_i, \sum_i a_i \rangle = \sum_{ij} \langle a_i, a_j \rangle$

$\overset{①}{=} \frac{B}{m} \sqrt{\mathbb{E}_\sigma \sum_{ij} \langle \sigma_i x_i, \sigma_j x_j \rangle} = \frac{B}{m} \sqrt{\mathbb{E}_\sigma \sum_i \sum_j \sigma_i \sigma_j x_i x_j}$

$= \frac{B}{m} \sqrt{\sum_i \mathbb{E}_{\sigma_i} \sigma_i^2 \|x_i\|^2 + \sum_{i \neq j} \mathbb{E} \sigma_i \sigma_j \langle x_i, x_j \rangle}$

$\sigma_i^2 = 1$; $\mathbb{E}\sigma_i\sigma_j = 0$: independent $= \mathbb{E}\sigma_i \mathbb{E}\sigma_j = 0$

$= \frac{B}{m} \sqrt{\sum_i \|x_i\|^2}$

$\mathbb{E}_S \mathrm{Rad}(\mathcal{H}_S) \leq \frac{B}{\sqrt{m}} \mathbb{E}_{S_x} \sqrt{\frac{1}{m} \sum_i \|x_i\|^2} \leq \frac{B}{\sqrt{m}} \sqrt{\mathbb{E}_{S_x} \|x\|^2}$

If $Pr(\|x\| \le C) = 1$, $Rad(\mathcal{H}_B|_{S_X}) \le \frac{BC}{\sqrt{m}}$ It is fine in some case to only look the average

hard bound

Proof for Talagrand's Lemma

① use Lemma 2 to prove Talagrand's contraction Lemma

Lemma2: if $\varphi: \mathbb{R} \to \mathbb{R}$ is 1-Lipschitz, $Rad(\{\varphi(V_1), V_2, \dots, V_m\}) \le Rad(V)$

Pf: $Rad(\{(\frac{1}{\rho}\varphi_1(V_1), V_2, \dots, V_m) : v \in V\}) \le Rad(V)$ $\swarrow \rho$-Lipschitz

$= Rad(\{(V_2, \frac{1}{\rho}\varphi_1(V_1), \dots, V_m) : v \in V\})$ Rotate doesn't change Rad

$= \vdots$

$= Rad(\{(V_2, V_3, \dots, \frac{1}{\rho}\varphi_1(V_1)) : v \in V\})$

the same for $\varphi_1, \varphi_2, \dots$

$Rad(\{\frac{1}{\rho}\varphi(V_1), \frac{1}{\rho}\varphi_2(V_2), \dots, V_m) : v \in V\}) \le Rad(V)$

$= Rad(\{\frac{1}{\rho}\varphi_2(V_2), \frac{1}{\rho}\varphi(V_1), \dots, V_m) : v \in V\})$

$\therefore Rad(\frac{1}{\rho}\phi \circ V) \le Rad(V)$

$\therefore Rad(\varphi \circ V) \le \rho\, Rad(V)$

② prove this Lemma 2

$m \cdot Rad(\{(\varphi(V_1), V_2, \dots, V_m) : v \in V\})$

$= \mathbb{E}_{\vec{\sigma} \sim Rad^m} \sup_{v \in V} \sigma_1 \varphi(V_1) + \underbrace{\sigma_2 V_2 + \dots + \sigma_m \cdot V_m}_{\vec{\sigma}_{2:} V_{2:}}$

$= \frac{1}{2} \mathbb{E}_{\vec{\sigma}_{2:} \sim Rad^m} \sup_{v \in V} [\varphi(V_1) + \vec{\sigma}_{2:} \cdot V_{2:}] + \frac{1}{2} \mathbb{E}_{\vec{\sigma}_{2:} \sim Rad^m} \sup_{v \in V} [-\varphi(V_1') + \vec{\sigma}_{2:} \cdot V_{2:}']$ two independent parts

switch v, v' won't change

$= \frac{1}{2} \mathbb{E}_{\vec{\sigma}_{2:}} \sup_{v, v' \in V} [|\varphi(V_1) - \varphi(V_1')| + \vec{\sigma}_{2:} (V_{2:} + V_{2:}')]$

switch v, v'

$\le \frac{1}{2} \mathbb{E}_{\vec{\sigma}_{2:}} \sup_{v, v' \in V} [|V_1 - V_1'| + \vec{\sigma}_{2:} (V_{2:} + V_{2:}')]$ 1-Lipschitz

$= \frac{1}{2} \mathbb{E}_{\vec{\sigma}_{2:}} \sup_{v, v' \in V} [V_1 - V_1' + \vec{\sigma}_{2:} (V_{2:} + V_{2:}')]$

$= \frac{1}{2} \mathbb{E}_{\vec{\sigma}_{2:}} \sup_{v \in V} [V_1 + \vec{\sigma}_{2:} \cdot V_{2:}] + \sup_{v' \in V} [V_1' + \vec{\sigma}_{2:} \cdot V_{2:}'] = \mathbb{E}_{\vec{\sigma}} \sup_{v} \vec{\sigma} \cdot v = m\, Rad(V)$

Lec McDiarmid's inequality

use Randermacher to get a high probability bound

$\Rightarrow$ McDiarmid's inequality (concentration inequality)

McDiarmid's inequality: Let $x_1, \cdots, x_m$ be independent, let $f(x_1, \cdots, x_m)$ be a real-valued function satisfying bounded differences

$$\forall i \in [m] \quad \sup_{x_1, \cdots, x_m, x_i'} |f(x_1, \cdots, x_m) - f(x_1, \cdots, x_{i-1}, x_i', x_{i+1}, \cdots, x_m)| \le c_i$$

change any variable at any one index doesn't change result too much

Then with probability at least $1-\delta$,

$$\begin{cases} f(x_1, \cdots, x_m) \le \mathbb{E} f(x_1, \cdots, x_m) + \sqrt{\frac{1}{2} (\sum_{i=1}^m c_i^2) \log \frac{1}{\delta}} \\ f(x_1, \cdots, x_m) \ge \mathbb{E} f(x_1, \cdots, x_m) - \sqrt{\frac{1}{2} (\sum_{i=1}^m c_i^2) \log \frac{1}{\delta}} \end{cases}$$

special case: $f(x) = \frac{1}{m} \sum_i x_i$, each $x_i \in [a, b]$

then $\forall i \in [m]$ $\frac{1}{m} \sup_{x_1 \cdots x_m x_k'} |(x_1 + x_2 + \cdots + x_m) - (x_1 + x_2 + \cdots + x_k' + \cdots x_m)| = \frac{1}{m} \sup_{x_k x_k'} |x_k - x_k'|$

$= \frac{1}{m}(b-a)$

so $c_i = \frac{b-a}{m}$, $\frac{1}{m} \sum_i x_i - \mathbb{E} x_i \le \sqrt{\frac{m}{2} (\frac{b-a}{m})^2 \log \frac{1}{\delta}} = (b-a) \sqrt{\frac{\log 1/\delta}{2m}}$

(McDiarmid's inequality implies Hoeffding)

in case of loss $\ell(h, z) \in [a, b]$ for all h, z, then with prob $\ge 1-\delta$

$$\sup_{h \in \mathcal{H}} L_D(h) - L_S(h) \le \mathbb{E} \sup_{h \in \mathcal{H}} L_D(h) - L_S(h) + (b-a) \sqrt{\frac{1}{2m} \log \frac{1}{\delta}} \quad (1)$$

if $\hat{h}_S$ is an ERM, with prob $\ge 1-\delta$ that

$$L_D(\hat{h}_S) - L_D(h^*) \le \mathbb{E} \sup_{h \in \mathcal{H}} [L_D(h) - L_S(h)] + (b-a) \sqrt{\frac{2}{m} \log \frac{2}{\delta}} \quad (2)$$

Pf. $L_D(\hat{h}_S) - L_D(h^*) \le \underbrace{L_D(\hat{h}_S) - L_S(\hat{h}_S)}_{①} + \underbrace{L_S(h^*) - L_D(h^*)}_{②}$

prove (1): Let $S^{(i)} = (z_1, \cdots, z_{i-1}, z_i', z_{i+1}, \cdots, z_m)$. we have

$$L_D(h) - L_S(h) = L_D(h) - L_{S^{(i)}}(h) + L_{S^{(i)}}(h) - L_S(h)$$

$$\sup_h L_D(h) - L_S(h) \le \sup_h L_D(h) - L_{S^{(i)}}(h) + \sup_h L_{S^{(i)}}(h) - L_S(h)$$

$$|\sup_h [L_D(h) - L_S(h)] - \sup_h [L_D(h) - L_{S^{(i)}}(h)]| \le |\sup_h L_{S^{(i)}}(h) - L_S(h)| \quad S, S^{(i)} \text{ changable}$$

$$|\sup_h [L_D(h) - L_S(h)] - \sup_h [L_D(h) - L_{S^{(i)}}(h)]| \le \sup_h |L_{S^{(i)}}(h) - L_S(h)| = \frac{b-a}{m}$$

apply McDiarmid's inequality ⌋

with prob at least of $1-\delta$:

$$\sup_h L_D(h) - L_S(h) \le \underbrace{\mathbb{E}_S \sup_h L_D(h) - L_S(h)}_{\mathbb{E}_S \mathrm{Rad}(\ell \circ \mathcal{H}|S)} + (b-a)\sqrt{\frac{1}{2m}\log\frac{1}{\delta}}$$

$\ne \sup_h \mathbb{E}(L_D(h) - L_S(h))$ $\mathbb{E}$ and max does not commute!

Pf for McDiarmid very general commonly used.

$X_{1:k-1} = (X_1, \cdots, X_{k-1})$. Fix some $k \in [m]$ freeze some arbitrary values $X_{1:k-1}$

$\mathbb{E}_{X_{k+1:m}} f(X_{1:k-1}, X_k, X_{k+1:m})$ is random depend on X_k

note that $\sup_{X_k} f(X_{1:m}) - \inf_{X_k} f(X_{1:m}) \le C_k$ by assumption

$$\Rightarrow \mathbb{E}_{X_{k+1:m}} \left(\sup_{X_k} f(X_{1:k-1}, X_k, X_{k+1:m}) + \sup_{X_k}\left(-f(X_{1:k-1}, X_k, X_{k+1:m})\right)\right) \le C_k \qquad \sup \mathbb{E} \le \mathbb{E}\sup$$

$$\sup_{X_k} \mathbb{E}_{X_{k+1:m}} f(X_{1:k-1}, X_k, X_{k+1:m}) + \sup_{X_k} \mathbb{E}_{X_{k+1:m}} (-f(X_{1:k-1}, X_k, X_{k+1:m})) \le C_k$$

$$\sup_{X_k} \mathbb{E}_{X_{k+1:m}} f(X_{1:k-1}, X_k, X_{k+1:m}) - \inf_{X_k} \mathbb{E}_{X_{k+1:m}} f(X_{1:k-1}, X_k, X_{k+1:m}) \le C_k$$

by Hoeffding's Lemma $\mathbb{E}_{X_{k+1:m}}$ is $SG(c_k/2)$

$X \sim SG(\sigma)$ $\mathbb{E}[e^{\lambda(X-\mu)}] \le e^{\frac{1}{2}\lambda^2\sigma^2}$

then $\mathbb{E}_{X_k} \exp(\lambda \mathbb{E}_{X_{k+1:m}} f(X_{1:k-1}, X_k, X_{k+1:m})) \le \exp\left(\lambda \mathbb{E}_{X_k} \mathbb{E}_{X_{k+1:m}} f(X_{1:k-1}, X_k, X_{k+1:m}) + \frac{1}{2}\lambda^2 (\frac{C_k}{2})^2\right)$

holds for any $X_{1:k-1}$, then average no $X_{1:k-1}$

$$\mathbb{E}_{X_{1:k}} \exp(\lambda \mathbb{E}_{X_{k+1:m}} f(X_{1:m})) \le \mathbb{E}_{X_{1:k-1}} \exp(\lambda \mathbb{E}_{X_{k:m}} f(X_{1:m}) + \frac{1}{8}\lambda^2 C_k^2)$$

take log

$$\underbrace{\log \mathbb{E}_{X_{1:k}} \exp(\lambda \mathbb{E}_{X_{k+1:m}} f(X_{1:m}))}_{a_k} \le \underbrace{\log \mathbb{E}_{X_{1:k-1}} \exp(\lambda \mathbb{E}_{X_{k:m}} f(X_{1:m}))}_{a_{k-1}} + \frac{1}{8}\lambda^2 C_k^2$$

sum over $k = 1, \dots, m$ $\quad a_m \le a_1 + \sum \frac{1}{8}\lambda^2 c_k^2$

$$\log \mathbb{E}_{X_{1:m}} \exp(\lambda f(X_{1:m})) \le \log \exp(\lambda \mathbb{E}_{X_{1:m}} f(X_{1:m})) + \sum_{k=1}^m \frac{1}{8}\lambda^2 C_k^2$$

$$\mathbb{E}_{X_{1:m}} \exp(\lambda f(X_{1:m})) \le \exp(\lambda \mathbb{E}_{X_{1:m}} f(X_{1:m})) \cdot \exp\left(\sum_{k=1}^m \frac{1}{8}\lambda^2 C_k^2\right)$$

$$\mathbb{E}_{X_{1:m}} \exp\left(\lambda f(X_{1:m}) - \lambda \mathbb{E}_{X_{1:m}} f(X_{1:m})\right) \le \exp\left(\frac{1}{2}\lambda^2 \sum_{k=1}^m (\frac{C_k}{2})^2\right)$$

$f(X_{1:m}) \in SG(\frac{1}{2}\sqrt{\sum_{k=1}^m C_k^2})$

with chernoff bound for sub-Gaussians

with prob at least $1-\delta$ $\quad f(X_{1:m}) \le \mathbb{E} f(X_{1:m}) + \frac{1}{2}\sqrt{\sum_{i=1}^m C_i^2} \cdot \sqrt{2\log\frac{1}{\delta}}$

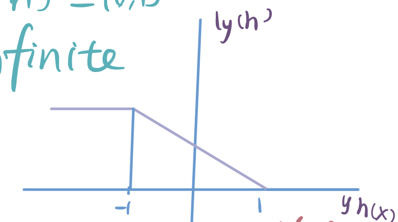
Rad is scale-sensitive : depend on the choice of function

Lec 6 Growth Function and VC Dimension

How to bound Rad for binary classifier?

For binary $\mathcal{H}|_{S_x} = \{(h(x_1), h(x_2), \cdots, h(x_m)) : h \in \mathcal{H}\} \subseteq \{0,1\}^m$

At most 2^m possible vectors even if $\mathcal{H}$ is infinite



- $L_{0\text{-}1}(h,(x,y)) = \mathbb{1}(h(x) \neq y)$

  $\mathcal{H}|_{S_x}$ finite

$$l_y(\hat{y}) = \begin{cases} 0 & \text{if } y\hat{y} \geq 1 \\ \frac{1}{2} - \frac{1}{2}y\hat{y}, & \text{if } y\hat{y} \in [-1,1) \\ 1, & \text{if } y\hat{y} \leq -1 \end{cases}$$

$\|L_y\|_{Lip} = \frac{1}{2}$

$Rad(\ell_{0\text{-}1} \circ \mathcal{H}_{\pm 1}|_S) \leq \frac{1}{2} Rad(\mathcal{H}_{\pm 1}|_S)$

- If $|V| < \infty$, $\|v\| \leq B$ for all $v \in V$

  Then $Rad(V) \leq \frac{B}{m}\sqrt{2\log|V|}$

special case: if $V = \mathcal{H}_{\pm 1}|_{S_x}$, $\|V\| = \sqrt{1+1+\cdots+1} = \sqrt{m}$

$Rad(\mathcal{H}_{\pm 1}|_{S_x}) \leq \sqrt{\frac{2}{m}\log|\mathcal{H}_{\pm 1}|_{S_x}|}$

Pf $Rad(V) = \mathbb{E}_{\sigma} \sup_{v \in V} \sum_{i=1}^{m} \frac{\sigma_i v_i}{m}$

$\sigma_i \in SG(1)$

$\frac{\sigma_i v_i}{m} \in SG\left(\frac{|v_i|}{m}\right) \Rightarrow \sum_i \frac{\sigma_i v_i}{m} \in SG\left(\sqrt{\sum_i \frac{v_i^2}{m^2}}\right)$ independent

$= SG\left(\frac{1}{m}\|v\|\right) \leq SG\left(\frac{B}{m}\right)$

- $T_i \in SG(\sigma)$, $\mathbb{E}T_i = 0$, T_i can be dependent

  then $\mathbb{E}\max_{i\in[m]} T_i \leq \sigma\sqrt{2\log m}$

  Pf. see in A2 Q(2)(4) Jensen's Inequality: $\exp(\mathbb{E}Y) \leq \mathbb{E}\exp(Y)$

when taking sup $\vec{\sigma}v$ can be dependent, so $\mathbb{E}_{\sigma} \sup_{v \in V} \underbrace{\sum_{i=1}^{m} \frac{\sigma_i v_i}{m}}_{T_i} \leq \frac{B}{m}\sqrt{2\log \underbrace{|V|}_{|\mathcal{H}|_{S_x}|}}$

$Rad(\mathcal{H}|_{S_x})$ depends on particular distribution

we can use $|\mathcal{H}|_{S_x}| < |\mathcal{H}|$ but it's very loose!

use growth function to drop dependence on particular S_x

- The growth function of $\mathcal{H}$ is

  $\Gamma_{\mathcal{H}}(m) = \sup_{x_1,\cdots,x_m \in X} |\mathcal{H}|_{(x_1,\cdots,x_m)}|$, $|\mathcal{H}_{S_x}| \leq \Gamma_{\mathcal{H}}(m)$

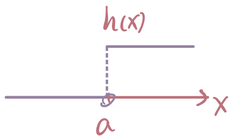
- (shatter): if $\mathcal{H}$ shatter a set $\underset{S_x \subseteq X}{S}$, means $\mathcal{H}$ can assign any possible label to the set i.e. $|\mathcal{H}|_{S_x}| = 2^m$

if the theory could explain any outcome then it's too general

VC is defined on the size of set you can shatter.

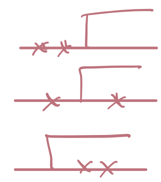
• (VC dimension) $\mathrm{VCdim}(\mathcal{H}) = \max\{m \geq 0 : \Gamma_{\mathcal{H}}(m) = 2^m\}$ worst case there is a set that can be shattered

example: threshold $h_a(x) = \mathbb{1}(x \geq a)$ VCdim = 1



can't shatter two points

0 0
0 1
1 0
1 1



• $\mathcal{H} = \{x \mapsto \mathrm{sgn}(w\cdot x) : w \in \mathbb{R}^d\}$ want to show $\mathrm{VCdim}(\mathcal{H}) = d$

$$\mathrm{sgn}(t) = \begin{cases} 1 & \text{if } t \geq 0 \\ -1 & \text{if } t < 0 \end{cases}$$

① $\mathcal{H}$ can shatter $\{e_1, e_2, \ldots, e_d\}$ to get $(y_1, \ldots, y_d)$

use $w = (y_1, \ldots, y_d)$, $w \cdot e_j = y_j$

(assume) ② Let $x_1, \ldots, x_{d+1}$ be shatterable by $\mathcal{H}$ ($x \in \mathbb{R}^d$, so $x_1, \ldots, x_{d+1}$ can't be linear dependent)

$\therefore \exists \alpha \in \mathbb{R}^{d+1} \setminus \{0\}$, s.t. $\sum_i^{d+1} \alpha_i x_i = 0$

Let $I_+ = \{i \in [d+1] : \alpha_i > 0\}$, $I_0 = \{i \in [d+1] : \alpha_i = 0\}$, $I_- = \{i \in [d+1] : \alpha_i < 0\}$

$\exists w$ s.t. $h_w(x_j) = \begin{cases} 1 & \text{if } j \in I_+ \\ -1 & \text{if } j \in I_- \end{cases}$

$$0 = w \cdot 0 = w \sum_{i=1}^{d+1} \alpha_i x_i = w \sum_{i \in I_+} \overbrace{\alpha_i x_i}^{>0} + w \sum_{i \in I_-} \overbrace{\alpha_i x_i}^{<0} + w \sum_{i \in I_0} \overbrace{\alpha_i x_i}^{=0}$$

$$= \sum_{i \in I_+} \underbrace{\alpha_i}_{>0} \underbrace{w x_i}_{\geq 0} + \sum_{i \in I_-} \underbrace{\alpha_i}_{<0} \underbrace{w x_i}_{<0} \geq 0$$ 取等 iff $\begin{cases} w x_i = 0 \text{ for } \forall i \in I_+ \\ I_- = \{\} \end{cases}$

we can also find some $\tilde{w}$ that $\tilde{w} \cdot x_i < 0$ for $\forall i \in I_+$

$$0 = \tilde{w} \cdot 0 = \tilde{w} \cdot \sum_{i \in I_+} \alpha_i x_i = \sum_{i \in I_+} \underbrace{\alpha_i}_{>0} \underbrace{\tilde{w} x_i}_{<0} < 0 \Rightarrow \text{contradiction!}$$

• $\mathrm{VCdim}(\{x \mapsto w\cdot x + b : w \in \mathbb{R}^d, b \in \mathbb{R}\}) = d+1$

• $\Gamma_{\mathcal{H}}(m) = O(m^{\mathrm{VCdim}(\mathcal{H})})$.

• Binary classifier with $\mathrm{VCdim}(\mathcal{H}) = d$, zero-one loss for any $m > d$:

$$\mathbb{E} \sup_{h \in \mathcal{H}} L_D(h) - L_S(h) \leq \mathbb{E}_{S_x} \sqrt{\frac{2}{m} \log|\mathcal{H}_{|S_x}|} \leq \sqrt{\frac{2}{m} \log \Gamma_{\mathcal{H}}(m)} \leq \sqrt{\frac{2d}{m}[\log m + 1 - \log d]}$$

Corollary. If $m \geq d := \mathrm{VCdim}(\mathcal{H})$, then $\Gamma_{\mathcal{H}}(m) \leq \left(\frac{em}{d}\right)^d$ $\quad \left(\log \Gamma_{\mathcal{H}}(m) \leq d\left(\log\frac{m}{d} + 1\right)\right)$

Sauer-Shelah Lemma ①: If $d = VCdim(\mathcal{H}) < \infty$, then $\tau_{\mathcal{H}}(m) \le \sum_{i=0}^{d} \binom{m}{i}$

(we do not prove the lemma)

use S-S lemma to prove corollary, want to show $\sum_{i=0}^{d} \binom{m}{i} \le \left(\frac{em}{d}\right)^d$ for $m \ge d$

pf for corollary

$$\sum_{i=0}^{d} \binom{m}{i} \le \sum_{i=0}^{d} \binom{m}{i} \underbrace{\left(\frac{m}{d}\right)^{d-i}}_{\ge 1}$$

$$\le \sum_{i=0}^{m} \binom{m}{i} \left(\frac{m}{d}\right)^{d-i} \quad \text{add non-negative}$$

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

$$= \left(\frac{m}{d}\right)^d \left(1+\frac{d}{m}\right)^m \qquad \left(1+\frac{x}{n}\right)^n \le e^x$$

$$\le \left(\frac{m}{d}\right)^d e^d$$

Lemma ②: For all finite $S_x \subseteq \mathcal{X}$, $|\mathcal{H}|_{S_x}| \le \left|\{T \subseteq S_x : T \text{ is shattered by } \mathcal{H}\}\right|$

Pf for ① use ② $|\mathcal{H}|_{S_x}|$ is upper bounded by the number of subsets of S of size at most d.

Pf for ② inductive

These are equivlant for binary class, 0-1 loss:

- $\Pr\left(\sup_{h \in \mathcal{H}} L_D(h) - L_S(h) \le \varepsilon\right) \ge 1-\delta$ for $m \ge m(\varepsilon, \delta)$
- ERM agnostically PAC-learn $\mathcal{H}$
- $\mathcal{H}$ is ag PAC-learnable
- ERM PAC-learns $\mathcal{H}$
- $\mathcal{H}$ is PAC-learnable
- $VCdim(\mathcal{H}) < \infty$ implies uniform convergence.

Theorem. $\mathcal{H}$ is a binary classifer on $\mathcal{X}$. $|\mathcal{X}| = 2m$ $d = VCdim(\mathcal{H})$

Assume $m \le \frac{d}{2}$. then $\inf_{A} \sup_{\substack{D \\ \text{realizable} \\ \text{by } \mathcal{H}}} \Pr_{S \sim D^m}\left(L_D(A(S)) \ge \frac{1}{8}\right) \ge \frac{1}{7}$

for a best-case learning algorithm there is a worst-case distribution

$$\mathbb{E}_{f \sim Unif(\hat{\mathcal{X}} \to \mathcal{Y})} \mathbb{E}_{S \sim D^m(f)} L_{D(f)}(\hat{h}_S) = \mathbb{E}_{f, S_x} \mathbb{E}_{x \sim D_x} \mathbb{1}(\hat{h}_S(x) \ne f(x))$$

not in the training set

$$\mathbb{E}_{f, S_x, x} \mathbb{1}(\hat{h}_S(x) \ne f(x)) = \mathbb{E}_{f, S_x} [\Pr(x \notin S_x) \mathbb{E}_{x \sim D_x} [\mathbb{1}(\hat{h}_S(x) \ne f(x)) \mid x \notin S_x]$$

$$+ [\Pr(x \in S_x) \mathbb{E}_{x \sim D_x} [\mathbb{1}(\hat{h}_S(x) \ne f(x)) \mid x \in S_x]$$

$$P(x \notin S_x) = \frac{|\hat{\mathcal{X}} \setminus S_x|}{|\hat{\mathcal{X}}|} \geq \frac{m}{2m} = \frac{1}{2}$$

$S_x = ((x_1, f(x_1)), \cdots, (x_m, f(x_m))$

$$\geq \frac{1}{2} \mathbb{E}_{f, S_x, x} [\mathbb{1}(\hat{h}_S(x) \neq h(x) \mid x \notin S_x]$$

$$= \frac{1}{4}$$

average over $f \geq \frac{1}{4}$, so $\exists g \in \mathcal{F}$. $\mathbb{E}_{S \sim D_g} L_{D_{(g)}}(\hat{h}_S) \geq \frac{1}{4}$

Want to show $\Pr_{S \sim D^m_{(g)}} (L_{D_{(g)}}(\hat{h}_S) < \frac{1}{8}) \leq \frac{6}{7}$

Markov : $\Pr(1 - L_{D(g)} \geq \frac{7}{8}) \leq \frac{\mathbb{E}(1 - L_{D(g)})}{\frac{7}{8}} \leq \frac{1 - \frac{1}{4}}{\frac{7}{8}} = \frac{6}{7}$

with no prior, all algorithms perform the same on average.

Let $\mathcal{H}$ be a set of binary classifer over $\mathcal{X}$. For any $m \geq 1$

$$\inf_{A} \sup_{\substack{D \text{ realizable} \\ \text{by } \mathcal{H}}} \Pr_{S \sim D^m} (\underset{\text{0-1 loss}}{L_D(A(S))} > \frac{\text{VCdim}(\mathcal{H}) - 1}{32m}) \geq \frac{1}{100}$$

Let $d = \text{VCdim}(\mathcal{H})$

① $d = 1$, holds almost trivally

② $d \geq 2m$, this is a corollary of the above theorem

③ $2 \leq d < 2m$

VC dimension doesn't say anything about distribution

If $\mathcal{H} = \{x \to w \cdot x : w \in \mathbb{R}^d, \|w\| \leq B\}$ $VC(\mathcal{H}) = d$ (scale doesn't matter)

What do we do for approximation error?

Lec Structural Risk Minimization not uniform over all H_k

- $\mathcal{H} = \overset{w_1}{\mathcal{H}_1} \cup \overset{w_2}{\mathcal{H}_2} \cup \cdots$ 

weight: how much do I like $H_1, H_2, \ldots$ preference on each hypothesis

$\sum w_k \le 1$, $w_k > 0$

$\forall k\ \forall D.\ \Pr_{S\sim D^m}\left(\sup_{h\in\mathcal{H}_k} L_D(h) - L_S(h) \le \varepsilon_k(m,\delta)\right) \ge 1-\delta$ with $\varepsilon_k(m,\delta) \xrightarrow{m\to\infty} 0$

$\therefore \Pr_{S\sim D^m}\left(\forall h\in\mathcal{H}.\ L_D(h) \le L_S(h) + \varepsilon_{k_h}(m, \delta w_{k_h})\right) \ge 1-\delta$

$k_h := \operatorname{argmin}_{k\ge 1,\ h\in\mathcal{H}_k} \varepsilon_k(m, \delta w_k)$ index that minimize ε_k corresponding to h

h lives in some of the hypo classes at least one H_k

① Def: $SRM_{\mathcal{H}}(S) \in \operatorname{argmin}_{h\in\mathcal{H}} [L_S(h) + \varepsilon_{k_h}(m, \delta w_{k_h})]$ commit to a δ beforehand

higher w_k for simpler hypothesis, tighter for complex hypo

- SRM algorithm

We can implement this minimization by a finite number of calls to an "ERM oracle", as long as our loss is lower-bounded by $a \le \ell(h,z)$ (typically $a = 0$):

```
function SRM_H(S)
    best ← ∞
    for k = 1, 2, ... do
        h_k ← ERM_{H_k}(S)
        cand_loss ← L_S(h_k) + ε_k(m, w_k δ)
        if cand < best then
            ĥ ← h_k
            best ← cand
        if min_{k'>k} a + ε_{k'}(m, w_{k'} δ) > best then
            break
    return ĥ
```

wired because depend on δ

Note that if we "decompose" as $\mathcal{H}_1 = \mathcal{H}$, then SRM becomes just $\mathrm{ERM}_{\mathcal{H}}$.

- general bound of SRM

non-uniform learnability because m function depend on h

$\hat{h}_S := SRM_{\mathcal{H}}(S) \in \operatorname{argmin}_{h\in\mathcal{H}} L_S(h) + \varepsilon_{k_h}(m, \delta w_{k_h})$

$L_D(\hat{h}_S) \le L_S(\hat{h}_S) + \varepsilon_{k_{\hat{h}_S}}(m, \delta w_{k_{\hat{h}_S}})$ (def. prob $\ge 1-\delta$)

$\le L_S(h^*) + \varepsilon_{k_{h^*}}(m, \delta w_{k_{h^*}})$

some h can be learned with less data

$= L_D(h^*) + L_S(h^*) - L_D(h^*) + \varepsilon_{k_{h^*}}(m, \delta w_{k_{h^*}})$

$\le L_D(h^*) + \underbrace{(b-a)\sqrt{\tfrac{1}{2m}\log\tfrac{1}{\delta}}}_{\text{prob}\ge 1-\delta} + \varepsilon_{k_{h^*}}(m, \delta w_{k_{h^*}})$

$\therefore \Pr\left(L_D(\hat{h}_S) - L_D(h^*) \le \varepsilon_{k_{h^*}}(m, \delta w_{k_{h^*}}) + (b-a)\sqrt{\tfrac{1}{2m}\log\tfrac{2}{\delta}}\right) \ge 1-\delta$

$h^*\in\mathcal{H}$ is any fixed hypothesis

- Specific bound use Rademacher

$R_k = \mathbb{E}_{S\sim D^m} \mathrm{Rad}((\ell \circ \mathcal{H}_k)|_S)$

$L_D(h) \le L_S(h) + \underline{2R_{k_h}} + \underline{(b-a)\sqrt{\frac{1}{2m}\log\frac{1}{w_{k_h}\delta}}}$ but depend on δ

$w_k = \frac{6}{\pi^2 k^2}$ $\quad \log\frac{1}{w_{k_h}\delta} = \log(k_h^2 \cdot \frac{\pi^2}{6} \cdot \frac{1}{\delta}) = 2\log k_h + \log\frac{\pi^2}{6} + \log\frac{1}{\delta}$

$< 2\log k_h + \frac{1}{2} + \log\frac{1}{\delta}$

$= 2\log k_h + \log\frac{\sqrt{e}}{\delta}$

$\sqrt{\log\frac{1}{w_{k_h}\delta}} < \sqrt{2\log k_h} + \sqrt{\log\frac{\sqrt{e}}{\delta}}$

$\Pr(\forall h\in\mathcal{H}.\ L_D(h) \le L_S(h) + 2R_{k_h} + (b-a)\sqrt{\frac{1}{m}\log k_h} + (b-a)\sqrt{\frac{1}{2m}\log\frac{\sqrt{e}}{\delta}}) \ge 1-\delta$

② SRM $(S) \in \arg\min L_S(h) + 2R_{k_h} + (b-a)\sqrt{\frac{1}{m}\log k_h}$ does not commit to a δ
$\hat{h}_S$

does not depend on δ, better in practice

hold for $\forall h$

$L_D(\hat{h}_S) \le L_S(\hat{h}_S) + 2R_{k_{\hat{h}_S}} + (b-a)\sqrt{\frac{1}{m}\log k_{\hat{h}_S}} + (b-a)\sqrt{\frac{1}{2m}\log\frac{\sqrt{e}}{\delta}}$ (SRM $\hat{h}_S$) with $p \ge 1-\delta$

then

$L_D(\hat{h}_S) \le L_S(\hat{h}_S) + 2R_{k_{\hat{h}_S}} + (b-a)\sqrt{\frac{1}{m}\log k_{\hat{h}_S}} + (b-a)\sqrt{\frac{1}{2m}\log\frac{3}{\delta}}$ $\quad \sqrt{e}<3$

(holds with probabilty $1-\frac{\sqrt{e}}{3}\delta$)

union $\le L_S(h^*) + 2R_{k_{h^*}} + (b-a)\sqrt{\frac{1}{m}\log k_{h^*}} + (b-a)\sqrt{\frac{1}{2m}\log\frac{3}{\delta}}$ $\quad (1-\frac{\sqrt{e}}{3}\delta)$

Hoeffding $\quad L_S(h^*) \le L_D(h^*) + (b-a)\sqrt{\frac{1}{2m}\log\frac{3}{\delta}}$ $\quad (1-\frac{\delta}{3})$

$\le L_D(h^*) + 2R_{k_{h^*}} + (b-a)\sqrt{\frac{1}{m}\log k_{h^*}} + (b-a)\sqrt{\frac{2}{m}\log\frac{3}{\delta}}$ $\quad (1-\frac{\sqrt{e}+1}{3}\delta \ge 1-\delta)$

Comparison: if we know the correct $\mathcal{H}_{k_h}$ from the start and run ERM

ERM: $L_D(\mathrm{ERM}_{\mathcal{H}_{k^*}}) \le L_D(h^*) + 2R_{k_{h^*}} + (b-a)\sqrt{\frac{2}{m}\log\frac{1}{\delta}}$

- varying bounded losses

e.g. logistic regression $\quad \mathcal{H}_k = \{x \mapsto w\cdot x : \|w\| \le B_k\}$ $\quad \Pr(\|x\| \le C) = 1$

$h^*(x) = w^*\cdot x$

if choose $B_k = 2^k$ then $2^{k-1} < \|w\| \le 2^k \Rightarrow k_h < \log_2(2\|w\|) \Rightarrow \sqrt{\log k_h} < \sqrt{\log\log_2(2\|w^*\|)}$

$(b-a)_{k_h} < (2C\|w\|+1)$, thus (not commit to a δ)

$\hat{h}_S = h_{\hat{w}_S}$; $\quad \hat{w}_S \in \arg\min_{w\in\mathbb{R}^d} L_S(h_w) + \frac{4C\|w\|}{\sqrt{m}} + (2C\|w\|+1)\sqrt{\frac{1}{m}\log\log_2(2\|w\|)}$

$L_D(\hat{h}_S) \le L_D(h^*) + \frac{4C\|w^*\|}{\sqrt{m}} + (2C\|w^*\|+1)\sqrt{\frac{1}{m}\log\log_2(2\|w\|)} + (2C\|w^*\| + 2C\|\hat{w}_S\| + 2)\sqrt{\frac{1}{2m}\log\frac{3}{\delta}}$

$\approx$

- relationship to regularization

$\vec{w}_S \in \underset{w \in \mathbb{R}^d}{\operatorname{argmin}} L_D(h^*) + \frac{\lambda}{\sqrt{m}} \|w^*\|$ SRM $\sim$ regularization

- non-uniform learnability =

we have shown a bound

$\Pr(L_D(SRM_{\mathcal{H},S}(S)) \le L_D(h^*) + \varepsilon_{k_{h^*}}(m, w_{k_{h^*}}\delta) + (b-a)\sqrt{\frac{1}{2m}\log\frac{1}{\delta}}) \ge 1-2\delta$

the bound does not show PAC learnability because ε depend on h^* !!

Def (compete) : $A(S)$ (ε, δ)-competes with $h \in \mathcal{H}$ on D with m samples, $\Pr_{S \sim D^m}(L_D(A(S)) \le L_D(h) + \varepsilon) \ge 1-\delta$

Def (non-uniform learning) : A nonuniformly learns $\mathcal{H}$ if $\exists$ finite $m(\varepsilon, \delta, h)$, s.t. $\forall \varepsilon, \delta \in (0,1)$, $\forall h \in \mathcal{H}$, $\forall D$, $\forall m \ge m(\varepsilon, \delta, h)$, $A(S)$ (ε, δ)-competes with h on D

this is looser than PAC because m depend on h

PAC: $\exists m \ge m(\varepsilon, \delta)$ $\Pr(L_D(A(S)) \le \inf_h (L_D(h) + \varepsilon)) \ge 1-\delta$ for $\forall \varepsilon, \delta \in (0,1)$

If $\mathcal{H} = \mathcal{H}_1 \cup \mathcal{H}_2 \cup \cdots$ with $VCdim(\mathcal{H}_k) < \infty$

then SRM non-uniformly learns $\mathcal{H}$ (in 0-1 binary classification)

Pf Let $\mathcal{H}_k = \{h \in \mathcal{H} : m(\frac{1}{8}, \frac{1}{7}, h) \le k\}$

$\therefore \mathcal{H} = \mathcal{H}_1 \cup \mathcal{H}_2 \cup \cdots$

consider D realizable by $\mathcal{H}_k$

then $\Pr_{S \sim D^k}(L_D(A(S)) \le \frac{1}{8}) \ge \frac{1}{7}$ using $h^* \in \mathcal{H}_k$ with $L_D(h^*) = 0$

countable

Assume $\mathcal{H} = \{h_1, h_2, \cdots\}$ ~ divide singleton classes

Use $\mathcal{H}_k = \{h_k\}$

Then $\varepsilon_k(m, w_k\delta) \le (b-a)\sqrt{\frac{1}{2m}\log\frac{1}{w_h\delta}}$

$\le (b-a)\sqrt{\frac{1}{2m}\log\frac{1}{w_h}} + (b-a)\sqrt{\frac{1}{2m}\log\frac{1}{\delta}}$

how to assign weights?

how to decide the order?

· Minimum Description Length

choose a weight according to a binary language for determining $\mathcal{H}$ (prefix-free)

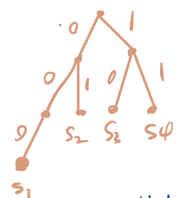
Kraft's inequality:

If $S \subseteq \{0,1\}^*$ (a set of binary strings) is prefix-free

(if '00' is a valid s, we don't have anything else that starts with '00')

(or read the data before '0' in C language)

then $\sum_{s\in S} 2^{-|s|} \le 1$ (probability distribution)



Then, we can choose a representation for $\mathcal{H}$ and assign $w_h = 2^{-|h|}$

$MDL_S \in \underset{h\in\mathcal{H}}{\operatorname{argmin}}\ L_S(h) + (b-a)\sqrt{\frac{1}{2m}\log\frac{1}{\delta 2^{-|h|}}}$

$\sqrt{\frac{1}{2m}\left(\log(\frac{1}{\delta}) + |h|\log 2\right)}$

or $MDL_S \in \underset{h\in\mathcal{H}}{\operatorname{argmin}}\ L_S(h) + \sqrt{\frac{\log 2}{2}\frac{|h|}{m}}$

$\therefore\ L_D(MDL(S)) \le L_D(h^*) + (b-a)\left(\sqrt{\frac{\log 2}{2}\cdot\frac{|h^*|}{m}} + \sqrt{\frac{2}{m}\log\frac{2}{\delta}}\right)$

Occam's razor: if there are multiple explanations of the data ($L_S(h_1) = L_S(h_2)$)

prefer the simpler one (shortest explanation)

If we choose $|h|$ to be the length of the shortest possible implementation of h in some programming language, is known Kolmogoru Complexity

MAP inferance with a Kolmogorov prior

Margins Theory

we do ERM with 0-1 loss

VC theory

$\mathcal{H} = \{ x \mapsto \mathrm{sgn}(w \cdot x) : x \in \mathbb{R}^d \}$ $\quad$ $VCdim(\mathcal{H}) = d$ $\quad$ known from previous lec

$$\sup_{h \in \mathcal{H}} L_D(h) - L_S(h) \le \sqrt{\frac{2d}{m}(\log m + 1 - \log d)} + \sqrt{\frac{1}{2m} \log \frac{1}{\delta}} \quad (\text{prob} \ge 1-\delta)$$

$$L_D(\hat{h}_S) - \inf_{h \in \mathcal{H}} L_D(h) \le \sqrt{\frac{2d}{m}(\log m + 1 - \log d)} + \sqrt{\frac{2}{m} \log \frac{2}{\delta}} \quad (\text{prob} \ge 1-\delta)$$

two problems:

① ERM with 0-1 loss is NP hard if D is not realizable by $\mathcal{H}$

② if d is really big, the bound doesn't tell anything until $\frac{m}{\log m} > 2d$

like in kernel methods, d is sometimes infite!

we like a better bound when in high dimension (big d) (hope it is not d-depedent)

Randemacher

If $\mathcal{H}_B = \{ x \mapsto w \cdot x : \|w\| \le B \}$ then $\mathbb{E}_S \, Rad(\mathcal{H}_B|_S) \le \frac{B\sqrt{\mathbb{E}\|x\|^2}}{\sqrt{m}}$

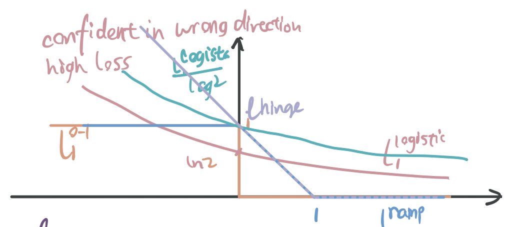
but ↑ is continuous, logistic loss

$$Rad(\ell_{logistic} \circ \mathcal{H}_B|_S) \le 1 \cdot Rad(\mathcal{H}_B|_S)$$

$$\therefore L_D^{logistic}(\operatorname*{argmin}_{h \in \mathcal{H}_B} L_D^{logistic}(h)) \le \frac{2B\sqrt{\mathbb{E}\|x\|^2}}{\sqrt{m}} + (B\sqrt{\mathbb{E}\|x\|^2} + 1)\sqrt{\frac{2}{m}\log\frac{2}{\delta}})$$

what if we want to bound on accuracy

$Rad(l_{0-1} \circ \mathrm{sgn} \circ \mathcal{H}_B)$ $\left\{ \begin{array}{l} l_{0-1} \text{ not Lipschitz} \\ VCdim(\mathcal{H}_B) = d \end{array} \right.$



if $y = 1$

0-1 loss	$1\{h \le 0\}$
logistic loss (LR)	$\log(1 + \exp(-h))$
hinge loss (SVM)	$[1-h]_+$
ramp loss (s)	$[1-h]_+ - [s-h]_+$ $(s<0)$

we can use a surrogate loss

$$\forall h, z \quad \ell_{surr}(h, z) \ge l_{0-1}(h, z)$$

$$L_D^{surr}(h) \ge L_D^{0-1}(h)$$

$$L_D^{0-1}(\mathrm{sgn} \circ h) \le L_D^{surr}(h) \le L_S^{surr}(h) + 2Rad(\ell_{surr} \circ \mathcal{H}|_S) + (b-a)_{surr}\sqrt{\frac{1}{2m}\log\frac{1}{\delta}}$$

① use logistic loss

let $\ell_{surr}(h, z) = \frac{1}{\log 2} \ell_{logistic}(h, z)$

but logistic loss gives a loose bounds ($\ell_{logistic} \sim \ell_{0-1}$ not close enough)

② use ramp loss $\quad \ell^{ramp}(h,(x,y)) = \ell_y^{ramp}(h(x)) = \begin{cases} 1 & yh(x) \le 0 \\ 1-yh(x) & 0 \le yh(x) \le 1 \\ 0 & 1 \le yh(x) \end{cases}$

1-Lipschitz, and bounded in $[0,1]$

$\forall h \quad L_D^{0-1}(\mathrm{sgn} \circ h) \le L_D^{ramp}(h) \le L_S^{ramp}(h) + 2\frac{BC}{\sqrt{m}} + \sqrt{\frac{1}{2m}\log\frac{1}{\delta}} \quad (P \ge 1-\delta)$

better than logistic loss

the larger the norm of w
the smaller the margin

so ... what about L_S^{ramp} term? when will $L_S^{ramp} \to 0$?

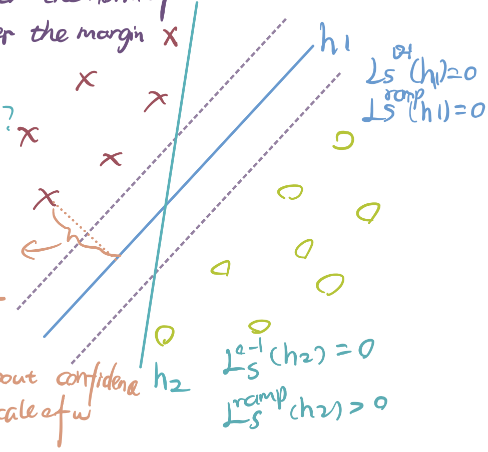
margin $= \min_i w \cdot x_i y_i$

least confident sample

is sensitive to scale of w

geometrical margin

$y_i \frac{w \cdot x_i}{\|w\|}$



doesn't care about confidence
invariant to scale of w

Assume $\exists h^* \in \mathcal{H}$, s.t. $L_D^{ramp}(h^*) = 0$. And $\mathbb{E}\|x\|^2 \le C^2$

D is linearly separable with margin

$\therefore L_S^{ramp}(h^*) = 0 \qquad w^* = \arg\min \|w^*\| \ \text{s.t.}\ L_D^{ramp}(h^*) = 0$

For ERM: $\hat{w} = \arg\min \|w\| \quad \text{s.t.}\ L_S^{ramp}(h_w) = 0 \quad \Rightarrow \|\hat{w}\| \le \|w^*\|$

$\therefore L_D^{0-1}(\hat{h}_S) \le \frac{2C\|w^*\|}{\sqrt{m}} + \sqrt{\frac{1}{2m}\log\frac{1}{\delta}} \quad = \frac{2C}{\rho\sqrt{m}} + \sqrt{\frac{1}{2m}\log\frac{1}{\delta}}$

$\min_i \frac{w^* \cdot x_i}{\|w^*\|} y_i \ge \frac{1}{\|w^*\|} \qquad$ Let $\rho = \frac{1}{\|w^*\|}$ (scale $w_\rho = \rho w^*$)

then $\min_i \frac{w^* \cdot x_i y_i}{\|w^*\|} = \min_i \rho w^* x_i y_i$

this bound depends on $\|w^*\|$, which we don't really know

so use a SRM-like argument to get bound only on $\|\hat{w}\|$

$$L_D^{0-1}(\hat{h}_S) \le L_S^{ramp}(h_w) + \frac{1}{\sqrt{m}}\left(\sqrt{\frac{1}{2}\log\frac{3}{\delta}} + \begin{cases} 4Cr & \text{if } \|w\| \le r \\ 4C\|w\| + \sqrt{\log\log_2\frac{2\|w\|}{r}} & \text{if } \|w\| > r \end{cases}\right.$$

<u>Pf</u> $\quad$ Define $B_k = r2^k \qquad \delta_k = \frac{6\delta}{\pi^2 k^2}$ for all $k \ge 1 \quad \sum_{k=1}^{\infty} \delta_k = \delta$

$$\forall k, \Pr_{S \sim D^m}\Big(\forall h \in \mathcal{H}_{B_k},\ L_D^{0-1}(\mathrm{sgn} \circ h) \le L_D^{ramp}(h) \le L_S^{ramp}(h) + \underbrace{2\mathbb{E}_{S \sim D^m} Rad(\mathcal{H}_{B_k} | S_k)}_{\le \frac{2B_k C}{\sqrt{m}}} + \underbrace{\sqrt{\frac{1}{2m}\log\frac{1}{\delta_k}}}_{\le \frac{1}{\sqrt{2m}}\left(\sqrt{\log\frac{3}{\delta}} + \sqrt{2\log k}\right)}\Big) \ge 1-\delta_k$$

To get a union bound, look a particular w and ask which k

If $\|w\| \le 2r$, $h \in \mathcal{H}_B \quad \therefore k=1 \quad \therefore \log k = 0$, $B_k = 2r$

$\quad \forall w: \|w\| \le 2r, \ L_D^{0-1}(\mathrm{sgn} \circ h) \le L_S^{ramp}(h) + \frac{4Cr}{\sqrt{m}} + \sqrt{\frac{1}{2m}\log\frac{3}{\delta}}$

If $\|w\| > 2r$, $L_D^{0-1}(\mathrm{sgn} \circ h) \le L_S^{ramp}(h) + \frac{4C\|w\|}{\sqrt{m}} + \sqrt{\frac{1}{2m}\log\frac{3}{\delta}} + \sqrt{\frac{1}{m}\log_2\log_2\frac{2\|w\|}{r}}$

$\Uparrow$ If $\|w\| > 2r$, $\|w\| \le r2^k \Rightarrow k_w = \lceil \log_2 \frac{\|w\|}{r} \rceil$ $\therefore h \in H_{B_{k_w}}$

$B_{k_w} = r \cdot 2^{\lceil \log_2 \frac{\|w\|}{r} \rceil} < r \cdot 2^{\log_2 \frac{\|w\|}{r} + 1} = 2\|w\|$

$\log k_w < \log (\log_2 \frac{\|w\|}{r} + 1) = \log_2 \log_2 \frac{2\|w\|}{r}$ can choose $r = \|w\|$

this bound doesn't imply non-uniform learning because its dependence on data distribution

Lec SVM

D is separable with a margin if $\exists h^*$ s.t. $L_D^{ramp}(h^*) = 0$ "h_{w^*}"

Expand L_S^{ramp}

$\hat{h} = h_{\hat{w}}$; $\hat{w} \in \underset{w}{\operatorname{argmin}} \|w\|^2$ s.t. $\forall i \in [m]$, $y_i w \cdot x_i \ge 1$ (Hard SVM)

convex quadratic program

$\hat{w} = \underset{w}{\operatorname{argmin}} \|w\|^2$ s.t. $\forall i \in [m]$, $y_i w \cdot x_i \ge 1$

$= \underset{w}{\operatorname{argmax}} \frac{1}{\|w\|} [\underbrace{\min_i y_i w \cdot x_i}_{\text{at least } 1}]$, s.t. $\forall i \in [m]$, $y_i w \cdot x_i \ge 1$

, also no big than 1 because could scale down w

$\ge \underset{w}{\operatorname{argmax}} \min_i \frac{y_i w \cdot x_i}{\|w\|}$ s.t. $\forall i \in [m]$, $y_i w \cdot x_i > 0$

maximize the worst case geometric margin

invariant to scale of w

so we could relax it to positive by scaling

$L_S^{0-1}(h_w) = 0$

Duality :

$\min_w \frac{1}{2}\|w\|^2$ s.t. $\forall i \in [m]$, $y_i w \cdot x_i \ge 1$

$= \min_w \max_{\alpha_i \ge 0} \frac{1}{2}\|w\|^2 + \sum_{i=1}^m \underbrace{\alpha_i}_{\text{dual variable}} (1 - y_i w \cdot x_i)$

$= \max_{\alpha_i \ge 0} \min_w \frac{1}{2}\|w\|^2 + \sum_{i=1}^m \alpha_i (1 - y_i w \cdot x_i)$

$\min_x \max_y f(x,y) \ge \max_y \min_x f(x,y)$

(weak duality)

$\min_x \max_y f(x,y) = \max_y \min_x f(x,y)$

(strong duality)

we have strong duality here by Slater's condition

$\nabla_w \frac{1}{2}\|w\|^2 + \sum_{i=1}^m \alpha_i (1 - y_i w \cdot x_i)$

$= w - \sum_{i=1}^m \alpha_i y_i x_i = 0$

$\Rightarrow w = \sum_{i=1}^m \alpha_i y_i x_i$ $\quad \|w\|^2 = \sum_{i=1}^m \sum_{j=1}^m \alpha_i y_i x_i \cdot x_j y_j \alpha_j = \alpha^T \operatorname{diag}(y) X X \operatorname{diag}(y) \alpha$

$X \in \mathbb{R}^{m \times d}$

$w \cdot x_j = \sum_{i=1}^m \alpha_i y_i x_i \cdot x_j$

$= \max_{\alpha \ge 0} 1^T \alpha - \frac{1}{2} \alpha^T \operatorname{diag}(y) \underline{X X^T} \operatorname{diag}(y) \alpha$

$X X^T \in \mathbb{R}^{m \times m}$: Gram matrix $\quad (X X^T)_{ij} = x_i \cdot x_j$

which problem to solve (min or max ?) depends on m, d which is smaller

in $\mathbb{R}^d$, a lot of α_i will be zero

what if the data is not linealy separable ?

Soft SVM :

recall the SRM-like bound, two problems :

1). the "r" is annoying

we choose a small enough "r" that $r \le \|w\|$ for all reasonable w

but not so small that $\sqrt{\log_2 \log_2 \frac{1}{r}}$ is relevant to anything.

Ignore $\sqrt{\log_2 \log_2 \frac{2\|w\|}{r}}$ term to pick

$$\min_w L_S^{ramp}(h_w) + \frac{4C}{\sqrt{m}}\|w\| \quad \text{still NP hard}$$

2) the problem is still NP hard. not a pratical algorithm.

we can take $\ell_{hinge} \ge \ell_{ramp} \ge \ell_{0-1}$, $\ell^{hinge}(h,(x,y)) = \begin{cases} 1-yh(x) & \text{if } yh(x) \le 1 \\ 0 & \text{if } yh(x) > 1 \end{cases}$

① 1-Lipschitz ② not bounded ③ convex

get more loss for a more-confident wrong answer

Algorithm :

hinge version $\approx \min_w L_S^{hinge}(h_w) + \frac{4C}{\sqrt{m}}\|w\|$

(a) $\approx \min_w L_S^{hinge}(h_w) + \lambda \|w\|^2$ square norm is much easier to optimize

c does not necessarily provide tight bound, λ should scale with $\frac{1}{\sqrt{m}}$, usually

(b) $= \frac{1}{2\lambda} \min_{0 \le \alpha_i \le \frac{1}{2\lambda m}} \mathbf{1}^T\alpha - \frac{1}{2}\alpha^T \operatorname{diag}(y) XX^T \operatorname{diag}(y)\alpha$

not equal in value but in min operation

$(XX^T)_{ij} = x_i \cdot x_j$

(b) predict with $h(x) = \sum_{i=1}^m \alpha_i y_i x_i \cdot x$

hinge ERM + Bounded weights $\mathcal{H}_B = \{x \mapsto w\cdot x : w \in \mathbb{R}^d, \|w\| \le B\}$

if $\hat{h}_B = \operatorname{argmin}_{h \in \mathcal{H}_B} L_S^{hinge}(h_w)$ since $\ell^{hinge} \ge \ell^{ramp}$

$$L_D^{0-1}(\operatorname{sgn} \circ \hat{h}_B) \le L_S^{hinge}(\hat{h}_B) + \frac{2BC}{\sqrt{m}} + \sqrt{\frac{1}{2m}\log\frac{1}{\delta}} \quad (P_r \ge 1-\delta)$$

ℓ^{hinge} is unbounded, but $h(x)$ is $\sup_{h,x}|h(x)| \le H \sup_{h,x}\|w\|\|x\| = HBC$

$$L_D^{0-1}(\operatorname{sgn} \circ \hat{h}_B) \le \inf_{h \in \mathcal{H}_B} L_D^{hinge}(h) + \frac{2BC}{\sqrt{m}} + (2+BC)\sqrt{\frac{2}{m}\log\frac{1}{\delta}} \quad (P_r \ge 1-\delta)$$

Lec. kernel (real-valued)

$h(x) = w_0 + w_1 x + w_2 x^2 + w_3 x^3$ $c \in \mathbb{R}_+$

$$= \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ w_3 \end{bmatrix} \cdot \underbrace{\begin{bmatrix} 1 \\ x \\ x^2 \\ x^3 \end{bmatrix}}_{\phi'(x)} = \begin{bmatrix} 1/\sqrt{c^3}\, w_0 \\ 1/\sqrt{3c^2}\, w_1 \\ 1/\sqrt{3c}\, w_2 \\ w_3 \end{bmatrix} \cdot \underbrace{\begin{bmatrix} \sqrt{c^3} \\ \sqrt{3c^2}\, x \\ \sqrt{3c}\, x^2 \\ x^3 \end{bmatrix}}_{\phi_c(x)}$$ scale them with constant c

$\phi'(x) \cdot \phi'(x') = 1 + xx' + (xx')^2 + (xx')^3$

$\underbrace{\phi_c(x) \cdot \phi_c(x')}_{K(x,x')} = c^3 + 3c^2(xx') + 3c(xx')^2 + (xx')^3 = (xx'+c)^3$ ← easier to compute

$K(x,x')$ kernel function

$\mathcal{F}$ is a vector space of functions

$f, f' \in \mathcal{F}$ $\quad f + f' \in \mathcal{F}$, $af \in \mathcal{F}$ $\forall a \in \mathbb{R}$, $a(bf) = (ab)f$ ..

$\langle f, f' \rangle_{L^2} = \int f(x) f'(x)\, dx$

so $\mathcal{F}_c = \{x \mapsto w \cdot \phi_c(x) : w \in \mathbb{R}^d\}$

$\quad \shortparallel$

$\mathcal{F}' = \{x \mapsto w \cdot \phi(x) : w \in \mathbb{R}^d\}$

Def: f has weights w, f' has weights w', f and $f' \in \mathcal{F}$

$\Rightarrow$
1. $f + f'$ have weights $w + w'$
2. αf have weights αw

} vector space

3. $\langle f, f' \rangle_{\mathcal{F}} = w \cdot w'$
4. $\|f\|_{\mathcal{F}} = \sqrt{\langle f, f \rangle_{\mathcal{F}}} = \|w\|$

} (real) Hilbert space
1. symmetric $\langle y, x \rangle = \langle x, y \rangle$
2. linear $\langle ax_1 + bx_2, y \rangle = a\langle x_1, y \rangle + b\langle x_2, y \rangle$
3. complete $\sum_{i=1}^{\infty} \|x_i\| < \infty$

property: $\forall x,\ \alpha f(x) + f'(x) = [\alpha f + f'](x)$

$\langle 0, 0 \rangle = 0$

$\langle f, f \rangle > 0$ if $f \neq 0$

$\langle \alpha f + g, h \rangle = \langle \alpha f, h \rangle + \langle g, h \rangle$

$\langle f, g \rangle = \langle g, f \rangle$

although $\mathcal{F}_c$ and $\mathcal{F}'$ are the same set, $\|f\|_{\mathcal{F}_c} \neq \|f\|_{\mathcal{F}'}$ for any f

if $\mathcal{F} = \{x \mapsto w \cdot \phi(x) : w \in \mathbb{R}^d\}$ $\underset{\text{identity func}}{\phi(x) = x}$ then $\|f_w\|_{\mathcal{F}} = \|w\|$

Soft SVM: $\min_w L_S^{hinge}(h_w) + \lambda \|w\|^2$

$\Rightarrow \min_{h \in \mathcal{F}} L_S^{hinge}(h_w) + \lambda \|h\|_{\mathcal{F}}^2$

what kind of function can be a kernel?

kernel is the inner product of feature maps over Hilbert space

if and only if

An example of kernel

let $\phi(x) = (\sqrt{c^3}, \sqrt{3c^2}\,x, \sqrt{3c}\,x^2, x^3) \in \mathbb{R}^4$, $\mathcal{F} = \{x \mapsto w \cdot \phi(x) : w \in \mathbb{R}^4\}$

consider $\phi(x)$ as a weight vector for an element in $\mathcal{F}$

$\Rightarrow f(x): x' \to \sqrt{c^3}\sqrt{c^3} + \sqrt{3c^2}x\sqrt{3c^2}x' + \sqrt{3c}x^2\sqrt{3c}x'^2 + x^3x'^3$

$$= (xx'+c)^3 = k(x, x') = \langle \phi(x), \phi(x') \rangle$$

define a kernel

- know the feature function $\phi(x) \in \mathcal{F}$, then $k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{F}}$
- don't know $\phi(x)$, make sure the kernel matrix is positive semi-definite

Def. A function $k: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive definte kernel if and only if there exist some Hilbert space $\mathcal{F}$ and feature map $\phi(x): \mathcal{X} \mapsto \mathcal{F}$ such that $k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{F}}$

positive semi-definite matrix is equivalently characterized as

① For all $\alpha \in \mathbb{R}^m$, $\alpha^T K \alpha \geq 0$

② All eigenvalues are non-negative

③ $K = LL^T$ for some $L \in \mathbb{R}^{m \times m}$

Thm. A function $k: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive definte kernel if and only if for all $m \geq 1$ and $x_1, \dots, x_m \in \mathcal{X}$, the kernel matrix K is positive semi-definite

$$K = \begin{bmatrix} k(x_1, x_1) & \cdots & k(x_1, x_m) \\ \vdots & \ddots & \vdots \\ k(x_m, x_1) & \cdots & k(x_m, x_m) \end{bmatrix} \in \mathbb{R}^{m \times m}.$$

Pf: ① k is a kernel $\Rightarrow$ $\alpha^T K \alpha \geq 0$

if $k(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{F}}$, then

$$\alpha^T K \alpha = \sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i \langle \varphi(x_i), \varphi(x_j) \rangle \alpha_j = \left\| \sum_{i=1}^{m} \alpha_i \varphi(x_i) \right\|_{\mathcal{F}}^2 \geq 0$$

② $\alpha^T K \alpha \geq 0 \Rightarrow k$ is a kernel $\begin{cases} \exists \text{ Hilbert space } \mathcal{H} \\ \exists\ \phi: \mathcal{X} \to \mathcal{H} \end{cases}$ s.t. $k(x,x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$

Reproducing kernels

define $\varphi(x) = [x' \mapsto k(x,x')]$ for all x

let $\mathcal{F}_0$ be set of all linear combinations of these functions $\sum_{i=1}^{m} \alpha_i \varphi(x_i)$

for all $x_1, \dots, x_m \in \mathcal{X}$, $a_1, \dots, a_m \in \mathbb{R}$. $Range(\varphi) = \{\varphi(x) : x \in \mathcal{X}\} \subseteq \mathcal{F}_0$ linearity

example (1) $k(x,x') = x \cdot x'$ $x \in \mathbb{R}^d$, $\varphi(x) = [x' \mapsto x \cdot x']$ is linear on x'

the set of all $\varphi(x)$ $\{\varphi(x) : x \in \mathcal{X}\} = \{x' \mapsto w \cdot x' : w \in \mathbb{R}^d\}$

so $\sum a_i \varphi(x_i)$ also in ↑ don't need to worry about linearity

$\|\varphi(x)\|_{\mathcal{F}} = \sqrt{\langle \varphi(x), \varphi(x) \rangle_{\mathcal{F}}} = \sqrt{k(x,x)} = \sqrt{\|x\|^2} = \|x\|$

(2) $k(x,x') = (x \cdot x' + c)^3$, $\phi(x) = (\sqrt{c^3}, \sqrt{c^2}x, \sqrt{c}x^2, x^3) \in \mathbb{R}^4$ or $(1, x, x^2, x^3)$ ...

$\varphi(x) = [x' \mapsto (xx'+c)^3] \in \mathcal{F}$ feature map $\nexists w$ s.t. $f = \varphi(w)$

↑ it cannot represent functions like $f(x) = -10 + x^3$, but this could be represented through linear combination of $\varphi(x)$, so $f(x) \in \mathcal{F}$

(3) $k(x,x') = \exp\left(\frac{1}{2\sigma^2}\|x - x'\|^2\right)$ functions in RKHS

Gaussian kernel linear combination of Gaussian

to make $\mathcal{F}_0$ Hilbert

① linear symmetric define $\left\langle \sum_{i=1}^{m} \alpha_i \varphi(x_i), \sum_{j=1}^{n} \beta_j \varphi(x_j') \right\rangle_{\mathcal{F}_0} = \sum_{i=1}^{m} \sum_{j=1}^{n} \alpha_i \beta_j k(x_i, x_j')$

then we have $\langle \varphi(x), \varphi(x') \rangle_{\mathcal{F}_0} = k(x,x') =$

reproducing property → $\langle f, \varphi(x) \rangle_{\mathcal{F}_0} = f(x)$ $f \in \mathcal{F}_0$

$\left\langle \sum_{i=1}^{m} \alpha_i \varphi(x_i), f' \right\rangle_{\mathcal{F}_0} = \sum_{i=1}^{m} \alpha_i f'(x_i)$

$|f(x)| = |\langle f, \varphi(x) \rangle_{\mathcal{F}}| \leq \|f\|_{\mathcal{F}} \|\varphi(x)\|_{\mathcal{F}} = \sqrt{k(x,x)}\,\|f\|_{\mathcal{F}}$

$|f(x) - f(x')| \leq |\langle f, \varphi(x) - \varphi(x') \rangle_{\mathcal{F}}| \leq \|f\|_{\mathcal{F}} \sqrt{k(x,x) + k(x',x') - 2k(x,x')}$

② complete add the limits of all Cauchy sequences and define the inner product to construct RKHS $\mathcal{F}$ based on $\mathcal{F}_0$

not all $f \in \mathcal{F}$ can be written as $\sum_{i=1}^{n} \alpha_i \varphi(x_i)$, but can get close

Optimizing in the RKHS

$\underset{f: \|f\|_{\mathcal{F}} \leq B}{\arg\min}\ L_S(f)$ for Gaussian kernel f is in infinite dimension space, $\|f\|_{\mathcal{F}}$ is a problem

$\underset{f \in \mathcal{F}}{\arg\min}\ L_S(f) + \lambda \|f\|_{\mathcal{F}}^2$ $f = \sum_{i=1}^{m} \alpha_i y_i \varphi(x_i)$ (soft SVM) general form: $\underset{f\in\mathcal{F}}{\operatorname{argmin}}\ A(f(x_1), \dots, f(x_m)) + R(\|f\|_{\mathcal{F}})$

non-decreasing, like 0-inf indicator func

the solution will be some linear combination of $\varphi(x_i)$

Theorm (Representer theorem). If $\mathcal{F}$ is an RKHS with feature map φ, then for any function $A: \mathbb{R}^m \to \mathbb{R}$ and nondecreasing function $R: \mathbb{R} \to \mathbb{R}\cup\{\infty\}$

$$\underset{f\in\mathcal{F}}{\operatorname{argmin}}\ A(f(x_1), \dots, f(x_m)) + R(\|f\|)$$

turn into finite dimension

contains a solution $f = \sum_{i=1}^{m} \alpha_i \varphi(x_i)$, if R is strictly increasing then every solution is this form.

so can solve $\underset{\alpha\in\mathbb{R}^m}{\operatorname{argmin}}\ A(\dots) + R(\sqrt{\alpha^T K \alpha})$

$f(x_j) = \sum_{i=1}^{m} \alpha_i k(x_i, x_j)$

and then $\hat{f} = \sum_i \hat{\alpha}_i \varphi(x_i)$

Pf.

$$A(f(x_1), \dots, f(x_m)) = A(\langle f, \varphi(x_1)\rangle_{\mathcal{F}}, \dots, \langle f, \varphi(x_m)\rangle_{\mathcal{F}})$$

Let $\mathcal{F}_{\parallel} = \operatorname{span}(\{\varphi(x_j) : j \in [m]\}) = \{\sum_i \alpha_i \varphi(x_i) : \alpha \in \mathbb{R}\}$ Parrall to data

Let $\mathcal{F}_{\perp}$ = orthogonal subspace of $\mathcal{F}_{\parallel}$ in $\mathcal{F}$

$f_{\parallel} \in \mathcal{F}_{\parallel}$, $f_{\perp} \in \mathcal{F}_{\perp}$, then $\langle f_{\parallel}, f_{\perp}\rangle_{\mathcal{F}} = 0$

$\forall f \in \mathcal{F}$, $f = f_{\parallel} + f_{\perp}$ $\quad f(x_i) = \langle f_{\parallel} + f_{\perp}, \varphi(x_i)\rangle_{\mathcal{F}} = \langle f_{\parallel}, \varphi(x_i)\rangle_{\mathcal{F}}$

$\therefore A(f(x_1), \dots, f(x_m)) = A(\underline{f_{\parallel}(x_1)}, \dots, \underline{f_{\parallel}(x_m)})$ having nonzero $f_{\perp}$ doesn't change A

$\|f\|_{\mathcal{F}}^2 = \|f_{\parallel} + f_{\perp}\|_{\mathcal{F}}^2 = \underline{\|f_{\parallel}\|_{\mathcal{F}}^2} + \underline{\|f_{\perp}\|_{\mathcal{F}}^2} + \underbrace{2\langle f_{\parallel}, f_{\perp}\rangle_{\mathcal{F}}}_{0}$ having nonzero $f_{\perp}$ does not help R

$\therefore$ always a solution in $\mathcal{F}_{\parallel}$

Eg kernel ridge regression

$\underset{f\in\mathcal{F}}{\operatorname{argmin}}\ L_S^{sq}(f) + \lambda \|f\|_{\mathcal{F}}^2 \qquad L_S^{sq}(f) = \frac{1}{m}\sum_{i=1}^{m}(f(x_i) - y_i)^2$

$\to \underset{\alpha\in\mathbb{R}^m}{\operatorname{argmin}}\ \frac{1}{m}\sum_{i=1}^{m}\left(\sum_{j=1}^{m}\alpha_j k(x_i, x_j) - y_i\right)^2 + \lambda\alpha^T K\alpha$

$= \underset{\alpha\in\mathbb{R}^m}{\operatorname{argmin}}\ \frac{1}{m}\underbrace{\|K\alpha - y\|^2}_{\alpha^T K^2 \alpha - 2y^T K\alpha + y^T y} + \lambda\alpha^T K\alpha$

$$= \underset{\alpha\in\mathbb{R}^m}{\operatorname{argmin}} \ \alpha^T K(K\alpha + m\lambda I)\alpha - 2y^T K\alpha$$

let gradient=0 $\quad 2K(\overset{\text{p.s.d.}}{K+m\lambda I})\alpha = 2Ky$

$$\alpha = (K+m\lambda I)^{-1}y$$

what kind of things kernel can / can't do ?

Rademacher complexity

$f\in\mathcal{F}$, RKHS for kernel. $K(x,x') = \langle \varphi(x), \varphi(x')\rangle_{\mathcal{F}}$

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

$\mathcal{H}_B = \{f\in\mathcal{F} : \|f\|_{\mathcal{F}} \le B\}$

$|f(x)| = |\langle f, \varphi(x)\rangle_{\mathcal{F}}| \le \|f\|_{\mathcal{F}}\,\|\varphi(x)\|_{\mathcal{F}}$

$\|\varphi(x)\|_{\mathcal{F}}^2 = \langle\varphi(x),\varphi(x)\rangle_{\mathcal{F}} = K(x,x)$

$$
\begin{aligned}
m\cdot \mathrm{Rad}(\mathcal{H}_B|S_x) &= \mathbb{E}_\sigma \sup_{\|f\|\le B} \sum_i \sigma_i f(x_i) \\
&= \mathbb{E}_\sigma \sup_{\|f\|\le B} \langle f, \sum_i \sigma_i \varphi(x_i)\rangle_{\mathcal{F}} \\
&\le B\,\mathbb{E}_\sigma \|\sum_i \sigma_i \varphi(x_i)\|_{\mathcal{F}} \\
&\le B\sqrt{\mathbb{E}_\sigma \|\sum_i \sigma_i \varphi(x_i)\|_{\mathcal{F}}^2} \\
&= B\sqrt{\sum_i \|\varphi(x_i)\|_{\mathcal{F}}^2} = B\sqrt{\sum_i \underbrace{K(x_i,x_i)}_{=1 \text{ for gaussian kernel}}}
\end{aligned}
$$

$$\mathrm{Rad}(\mathcal{H}_B|S_x) \le \frac{B}{\sqrt{m}}\sqrt{\frac{1}{m}\sum_i K(x_i,x_i)}$$

Lec Universal approximation

$\mathcal{X}$ is a compact (includes all limiting values of points) metric space e.g. a closed bounded subset of $\mathbb{R}^d$ (Euclidean space); $[0,1]^d$

$C(\mathcal{X})$ is the space of continuous functions $\mathcal{X}\to\mathbb{R}$

with $\|f\|_\infty = \sup_{x\in\mathcal{X}} |f(x)|$

Def (universal kernel) (universal approximation). A kernel $k: \mathcal{X}\times\mathcal{X}\to\mathbb{R}$ is universal if its RKHS $\mathcal{F}$ is dense in $C(\mathcal{X})$:

$\forall g\in C(\mathcal{X}),\ \forall\varepsilon>0,\ \exists f\in\mathcal{F}$ s.t. $\|g-f\|_\infty = \sup_{x\in\mathcal{X}} |g(x)-f(x)| \le \varepsilon$

every finite dimension kernel will approach 0 if $x \to \infty$

linear kernel is not universal

Prop. For any universal RKHS, $VCdim(\mathcal{F}) = \infty$

Let $V, W \in X$ be compact, disjoint sets. Let K be universal

Then $\exists f \in \mathcal{F}_K$ s.t. $\forall x \in V$, $f(x) > 0$ shattering

$\forall x \in W$, $f(x) < 0$

$\Rightarrow$ which means $VC\dim(\mathcal{F}) = \infty$

Pf. Let $dist_V(X) = \min_{v \in V} \|x - v\|$

Let $g(x) = \dfrac{dist_W(x) - dist_V(x)}{dist_W(x) + dist_V(x)}$ $\begin{cases} \forall x \in V \ \ g(x) = 1 & (dist_V(x) = 0) \\ \forall x \in W \ \ g(x) = -1 & (dist_W(x) = 0) \end{cases}$

then $\exists f \in \mathcal{F}$ s.t. $\|f - g\|_\infty < 1$ separate compact set

Prop. For any universal RKHS, $Rad(\mathcal{F}|_{S_x}) = \infty$

for any σ_i, can find some value that are positively correlated.

so Rad at least positive. If multiply f by a (a really large)

then the whole thing is scaled by a. thus $Rad(\mathcal{F}|_{S_x})$ is infinite.

universal approximation of neural networks

$f(x) = f^{(D)}(x)$ D layers

$f^k(x) = \sigma_k(W_k f^{k-1}(x) + b_k)$ $f^0(x) = x$

$ReLU(z) = [\max(z_i, 0)]_i$

$LTU(z) = [\mathbb{1}(z_i > 0)]_i$

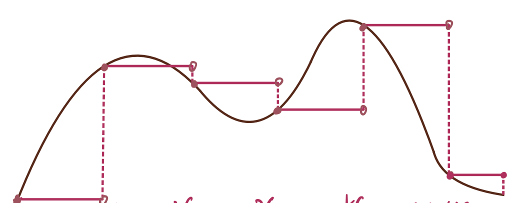
Thm. Let $g: [0,1]^d \to \mathbb{R}$ be ρ-Lipschitz. For any $\varepsilon > 0$, $\exists$ a two-layer net f with one hidden layer with $N = \lceil \frac{\rho}{\varepsilon} \rceil$ LTU units and a linear output unit

ceiling func

s.t. $\|f - g\|_\infty \le \varepsilon$

Pf For $i \in \{0, \dots, N-1\}$, let $b_i = \frac{i\varepsilon}{\rho}$

$b_0 = 0$, $b_1 = \frac{\varepsilon}{\rho}$, ..., $b_{N-1} = (\lceil \frac{\rho}{\varepsilon} \rceil - 1)\frac{\varepsilon}{\rho} < 1$



$$f(x) = \begin{cases} g(0) & 0 \le x < b_1 \\ g(b_1) & b_1 \le x < b_2 \\ \vdots & \vdots \\ g(b_{N-1}) & b_{N-1} \le x \le 1 \end{cases}$$

piece-wise constant approximation

$f(x) = \sum_{k=0}^{N-1} a_k \mathbb{1}(x \ge b_k)$ $a_0 = g(0)$, $a_1 = g(b_1) - a_0$, ..., $\sum_{i=0}^{k} a_i = g(b_k)$, $a_i = g(b_i) - g(b_{i-1})$

For any input x, let $k = \max\{i : b_i \le x\}$, then

$$|g(x) - f(x)| \le |g(x) - g(b_k)| + \underbrace{|g(b_k) - f(b_k)|}_{0} + \underbrace{|f(b_k) - f(x)|}_{0}$$
$$\le \rho |x - b_k| \le \rho \frac{\varepsilon}{\rho} = \varepsilon$$

another way to show universal

Thm. (Stone-Weierstrass) if $\mathcal{F}$ of functions $\mathcal{X} \to \mathbb{R}$ satisfies:

1). $\mathcal{F} \subseteq C(X)$ continuous

2). $\forall x \in \mathcal{X}$, $\exists f \in \mathcal{F}$ with $f(x) \neq 0$

3). $\mathcal{F}$ is an algebra: $\forall f, g \in \mathcal{F}$, $\forall \alpha \in \mathbb{R}$ $\begin{cases} \alpha f + g \in \mathcal{F} \\ fg = (x \mapsto f(x) g(x)) \in \mathcal{F} \end{cases}$

4). $\mathcal{F}$ seperates points: $\forall x, x' \in \mathcal{X}$ with $x \neq x'$, $\exists f \in \mathcal{F}$ with $f(x) \neq f(x')$

Then $\mathcal{F}$ is dense in $C(X)$

$$\mathcal{F}_{exp} = \{x \mapsto \sum_{i=1}^{N} \alpha_i \exp(w_i \cdot x) : N \ge 1, w_1, \cdots, w_N \in \mathbb{R}^d, \alpha_1, \cdots, \alpha_N \in \mathbb{R}\}$$

$\mathcal{F}_{exp}$ is dense in $C(X)$

✓ continuous ✓ $\exists f(x) \neq 0$ $(x \mapsto 1) \in \mathcal{F}$ ✓ $\alpha f + g \in \mathcal{F}$

✓ $fg = \sum_{i=1}^{N} \sum_{j=1}^{M} \alpha_i \beta_j \exp((w_i + v_j) x) \in \mathcal{F}$

x_1 vs x_2: use $f(x) = \exp((x_1 - x_2) \cdot x)$

$$\frac{f(x_1)}{f(x_2)} = \frac{\exp(\|x_1\|^2 - x_2 \cdot x_1)}{\exp(x_1 \cdot x_2 - \|x_2\|^2)} = \exp(\|x_1\|^2 + \|x_2\|^2 - 2x_1 x_2) = \exp(\|x_1 - x_2\|^2) \neq 1$$

MLP approximation: $\mathcal{F}_\sigma = \{x \mapsto \sum_{i=1}^{m} a_i \sigma(w_i \cdot x) : m \ge 1; w_1, \cdots, w_m \in \mathbb{R}^d, a_1, \cdots, a_m \in \mathbb{R}\}$

$\sigma : \mathbb{R} \to \mathbb{R}$ is continuous with $\lim_{z \to -\infty} \sigma(z) = 0$, $\lim_{z \to +\infty} \sigma(z) = 1$

① use $f_0 \in \mathcal{F}_{exp}$ to approximate g s.t. $\|f_0 - g\|_\infty \le \frac{\varepsilon}{2}$

② use linear combination of σ s.t. $\exp(z) \approx \sum_j c_j \sigma(t_j z)$

then replace each $\exp(w_i x)$ in f_0 by $\sum_j c_j \sigma(t_j w_i \cdot x)$ to find $f \in \mathcal{F}_\sigma$ s.t. $\|f - f_0\| \le \frac{\varepsilon}{2}$

this also holds if σ is anything but polynomial

because linear combination of polynomials of degree d is also degree d.

can't approximate any function (needs to be arbitrarily high to do this!)

Circuit complexity

Two-layer network, threshold activations, can represent all $g: \{\pm 1\}^d \to \pm 1$

- might take exponential width to do it
- if g can be computed in time T, $\exists$ a network of size $O(T^2)$ that implements g

there exists doesn't mean we can find it

Is ERM enough ?

- ERM is NP-hard to compute for NN
  - $f(x) = ReLU(w \cdot x)$ with square loss is NP-hard
- Uniform convergence ERM bounds might not be good enough

VC dimension param counting bound

P: params D: depth ReLU (piecewise-linear) net

$VCdim = O(PD(\log P)), \Omega(PD \log_{D} \frac{P}{D})$

big O: less than equal

big Ω: greater than equal

big Θ: equal

in practice, more params improve generalization, but worse bounds

DNN has lots of params !

norm-based bounds

Radermacher complexity, covering number ... tend to be vacuous

- ERM might not generalize well

Lec Stability, Regularization, Convex problems

Regularized loss minimization:

$\underset{h\in\mathcal{H}}{\operatorname{argmin}}\ L_S(h)+\lambda R(h)$

Lagrange dual

$\underset{h:\,R(h)\le\frac{1}{2}B^2}{\operatorname{argmin}}\ L_S(h)$

e.g

kernel ridge regression

$\underset{h\in\mathcal{F}}{\operatorname{argmin}}\ L_S^{sq}(h)+\frac{1}{2}\lambda\|h\|_{\mathcal{F}}^2$ $\qquad$ $\underset{h:\,\|h\|_{\mathcal{F}}\le B}{\operatorname{argmin}}\ L_S^{sq}(h)$

soft SVM

$\underset{h\in\mathcal{F}}{\operatorname{argmin}}\ L_S^{hinge}(h)+\frac{1}{2}\lambda\|h\|_{\mathcal{F}}^2$ $\qquad$ $\underset{h:\,\|h\|_{\mathcal{F}}\le B}{\operatorname{argmin}}\ L_S^{hinge}(h)$

equivalent in the sense that for any λ there is some B s.t. they are the same (vice versa)

easier to find some λ

motivation: the failure of bounding ERM with uniform convergence

we want a "stable" algorithm: If $S\approx S'$, $A(S)\approx A(S')$

the algorithm is not very sensitive to the small change in samples

implies a small $\mathcal{H}$!

in RLM, $R(h)$ could be seen as a stabilizer

$S=(z_1,\cdots,z_m)$ $\qquad$ $\underset{\text{change } z_i}{S^{(i\leftarrow z')}}=(z,\cdots,z_{i-1},z',z_{i+1},\cdots,z_m)$

Theorem $\underset{S\sim\mathcal{D}^m,A}{\mathbb{E}}[L_\mathcal{D}(A(S))-L_S(A(S))]=\underset{\substack{S\sim\mathcal{D}^m\\ z'\sim\mathcal{D}\\ i\sim\text{Uniform}([m])\\ A}}{\mathbb{E}}[\underbrace{\ell(A(S^{i\leftarrow z'}),z_i)}_{\text{generalization error}}-\underbrace{\ell(A(S),z_i)}_{\text{training error}}]$

A could be random but independent of data e.g. SGD

Def (average stable) A is $\varepsilon(m)$-on-average-replace-one stable if for all $\mathcal{D}$.

$$\underset{\substack{S\sim\mathcal{D}^m,\,z'\sim\mathcal{D}\\ i\sim\text{unif}([m]),\,A}}{\mathbb{E}}[\ell(A(S^{i\leftarrow z'}),z_i)-\ell(A(S),z_i)]\le\varepsilon(m)$$

average-case generalization gap

Def (uniformly stable) A is $\beta(m)$-uniformly stable if for all $m \geq 1$

stronger notion

$$\sup_{\substack{S \in D^m \\ z, z' \in D}} \left| \mathbb{E}_A \ell(A(S^{i \leftarrow z'}), z) - \mathbb{E}_A \ell(A(S), z) \right| \leq \beta(m)$$

Thm. (Uniform-stability) $\ell \in [a, b]$, A is $\beta(m)$-uniformly stable, with Prob $\geq 1-\delta$ over $S \sim D^m$

$$\mathbb{E}_A [L_D(A(S)) - L_S(A(S))] \leq \beta(m) + (2m\beta(m) + b - a)\sqrt{\frac{1}{2m} \log \frac{1}{\delta}}$$

if A is $\beta(m)$-uniformly stable, it's average-replace-one stable

let $f(S) = \mathbb{E}_A [L_D(A(S)) - L_S(A(S))] \leftarrow \mathbb{E}_{z' \sim D, i \sim \mathrm{Unif}[m], A} [\ell(A(S^{i \leftarrow z'}), z_i) - \ell(A(S), z_i)]$

If A is $\beta(m)$-uniformly stable. $\mathbb{E}_S f(S) \leq \beta(m)$

Then we can use McDiarmid's inequality if $f(S)$ has bounded difference

$$|f(S) - f(S^{i \leftarrow z'})| = |\mathbb{E}_A L_D(\hat{h}) - L_S(\hat{h}) - L_D(\hat{h}^i) + L_S(\hat{h}^i)| \leq |\mathbb{E}_A L_D(\hat{h}) - L_D(\hat{h}^i)| + |\mathbb{E}_A L_S(\hat{h}) - L_S(\hat{h}^i)|$$

$$\left| \sum_{j \neq i} \ell(\hat{h}, z_j) - \ell(\hat{h}^i, z_j) + \ell(\hat{h}, z_i) - \ell(\hat{h}^i, z') \right|$$

Assume $\ell \in [a, b]$, let $\hat{h} = A(S)$, $\hat{h}^i = A(S^{i \leftarrow z'})$

① $$\left|\mathbb{E}_A L_D(\hat{h}^i) - \mathbb{E}_A L_D(\hat{h})\right| = \left|\mathbb{E}_{A, z} [\ell(\hat{h}^i, z) - \ell(\hat{h}, z)]\right|$$
$$\leq \mathbb{E}_{z \sim D} \left|\mathbb{E}_A \ell(\hat{h}^i, z) - \mathbb{E}_A \ell(\hat{h}, z)\right|$$
$$\leq \beta(m)$$

② $$\left|\mathbb{E}_A L_S(\hat{h}) - \mathbb{E}_A L_{S^{i \leftarrow z'}}(\hat{h}^i)\right| \leq \frac{1}{m} \sum_{j \neq i} \left|\mathbb{E}_A \ell(\hat{h}, z_j) - \mathbb{E}_A \ell(\hat{h}^i, z_j)\right| + \frac{1}{m} \left|\mathbb{E}_A \ell(\hat{h}, z_i) - \mathbb{E}_A \ell(\hat{h}^i, z')\right|$$
$$\overset{\frac{m-1}{m} \leq 1}{\leq} \beta(m) + \frac{b-a}{m} \quad \ell \text{ is bounded}$$

Combine ① + ②

f has bounded diffs with rate $c_i = 2\beta(m) + \frac{b-a}{m}$

$\Rightarrow$ with prob at least $1-\delta$ over $S \sim D^m$

$$f(S) \leq \mathbb{E} f(S) + c\sqrt{\frac{m}{2} \log \frac{1}{\delta}} \leq \beta(m) + (2m\beta(m) + b - a)\sqrt{\frac{1}{2m} \log \frac{1}{\delta}}$$

If $\beta(m)$ decays with $\frac{1}{m}$, then $f(S)$ decays $\frac{1}{\sqrt{m}}$

Convex Function

Def (convex set) set $C \subseteq X$ is convex if for all $x_0, x_1 \in C$ and $\alpha \in [0, 1]$

$$x_\alpha = (1-\alpha) x_0 + \alpha x_1 \in C$$

instead define a restricted domain, we define $f(x)=\infty$ for x out of domain. $\operatorname{dom} f = \{x \in \mathcal{X} : f(x) < \infty\}$

Def (convex function). A function $f: \mathcal{X} \to \mathbb{R} \cup \{\infty\}$ is called

- convex iff $\forall x_0, x_1 \in \mathcal{X}$, $\alpha \in (0,1)$ $f((1-\alpha)x_0 + \alpha x_1) \le (1-\alpha) f(x_0) + \alpha f(x_1)$
- m-strongly convex, ... $f((1-\alpha)x_0 + \alpha x_1) \le (1-\alpha) f(x_0) + \alpha f(x_1) - \frac{1}{2} m \alpha (1-\alpha) \|x_1 - x_0\|^2$
- strictly convex, ... $f((1-\alpha)x_0 + \alpha x_1) < (1-\alpha) f(x_0) + \alpha f(x_1)$

If f is differentiable, f is

- convex, iff $\forall x, x'$, $f(x') \ge f(x) + \langle \nabla f(x), x' - x \rangle$
- m-strongly convex, $f(x') \ge f(x) + \langle \nabla f(x), x' - x \rangle + \frac{1}{2} m \|x' - x\|^2$

If f is continuously differentiable, f is

- convex, iff $\forall x, x'$, $\langle \nabla f(x) - \nabla f(x'), x - x' \rangle \ge 0$
- m-strongly convex, $\langle \nabla f(x) - \nabla f(x'), x - x' \rangle \ge m \|x - x'\|^2$

If f is continuously twice differentiable, f is

- convex, iff $\forall x, x'$, $\nabla^2 f(x) \succeq 0$
- m-strongly convex, $\nabla^2 f(x) \succeq mI$ all eigs $\ge m$

$A \succeq 0 \Leftrightarrow A$ p.s.d

$A \succeq B \Leftrightarrow A - B \succeq 0$

RLM is often uniformly stable under some conditions

$$f_S(h) = \underbrace{L_S(h)}_{\text{convex}} + \lambda \underbrace{R(h)}_{\text{strongly convex}}$$

(set to 1-strongly convex here)

$f_S(h) - f_{S'}(h) + f_{S'}(h) - f_{S'}(g) + f_{S'}(g) - f_S(g)$ $\qquad S' = S^{i \leftarrow z'}$

$$\begin{aligned} f_S(h) - f_S(g) &= L_S(h) + \lambda R(h) - L_S(g) - \lambda R(g) \\ &= f_{S'}(h) - f_{S'}(g) + \frac{1}{m}\left(\ell(h, z_i) - \ell(g, z_i)\right) + \frac{1}{m}\left(\ell(g, z') - \ell(h, z')\right) \end{aligned}$$

let $\hat{h}^i$ minimize $f_{S^{i \leftarrow z'}}$ and $\hat{h}$ minimize f_S, then $f_{S'}(\hat{h}^i) - f_{S'}(\hat{h}) \le 0$

$\hat{h}^i = A(S^{i \leftarrow z'}) \in \operatorname{argmin}_h f_{S^{i \leftarrow z'}}(h)$ $\qquad \hat{h} = A(S)$

① $f_S(\hat{h}^i) - f_S(\hat{h}) \le \frac{1}{m}\left(\ell(\hat{h}^i, z_i) - \ell(\hat{h}, z_i)\right) + \frac{1}{m}\left(\ell(\hat{h}, z') - \ell(\hat{h}^i, z')\right) \left(+ \underbrace{f_{S'}(\hat{h}^i) - f_{S'}(\hat{h})}_{\le 0} \right)$

$\nabla f_S(\hat{h}) = 0$ as f_S is convex 0

② $f_S(\hat{h}^i) - f_S(\hat{h}) \ge \langle \nabla f_S(\hat{h}), \hat{h}^i - \hat{h} \rangle + \frac{1}{2}\lambda \|\hat{h}^i - \hat{h}\|^2$

①+② gives

$$\frac{1}{2}\lambda \|\hat{h}^i - \hat{h}\|^2 \le \frac{1}{m}\left(\ell(\hat{h}^i, z_i) - \ell(\hat{h}, z_i)\right) + \frac{1}{m}\left(\ell(\hat{h}, z') - \ell(\hat{h}^i, z')\right)$$

Assume $\forall z,\ h \mapsto \ell(h, z)$ is ρ-Lipschitz, then $\frac{1}{2}\lambda \|\hat{h}^i - \hat{h}\|^2 \le \frac{2\rho}{m} \|\hat{h} - \hat{h}^i\|$

$\Rightarrow \quad \|\hat{h} - \hat{h}^i\| \le \frac{4\rho}{m\lambda}$ use Lips twice

$\Rightarrow \quad |\ell(\hat{h}, z) - \ell(\hat{h}^i, z)| \le \frac{4\rho^2}{m\lambda}$

convex + Lipschitz RLM is $\frac{4\rho^2}{m\lambda}$-uniformly stable

Assume R is non-negative

$$\mathbb{E}_S L_D(A(S)) \le \mathbb{E}_S L_S(A(S)) + \frac{4\rho^2}{\lambda m} \le L_D(h^*) + \lambda R(h^*) + \frac{4\rho^2}{\lambda m}$$

$L_S(A(S)) \le L_S(A(S)) + \lambda R(A(S)) \le L_S(h^*) + \lambda R(h^*) \quad \forall h^* \in \mathcal{H}$

$\mathbb{E}_S L_S(A(S)) \le L_D(h^*) + \lambda R(h^*)$

Assume $R(h^*) \le \frac{1}{2}B^2$ how to tradeoff λ to get good model?

$$\mathbb{E}_S L_D(A(S)) \le \inf_{h: R(h) \le \frac{1}{2}B^2} L_D(h) + \frac{1}{2}\lambda B^2 + \frac{4\rho^2}{\lambda m}$$

$ax + \frac{b}{x} \ge 2\sqrt{ab}$ $(x>0)$

minimize the bound by opt λ

if $\lambda = \sqrt{\frac{2 \times 4\rho^2}{m B^2}} = \frac{\rho}{B}\sqrt{\frac{8}{m}} \iff B = \frac{\rho}{\lambda}\sqrt{\frac{8}{m}}$

$$\mathbb{E}_S L_D(A(S)) \le \inf_{h: R(h) \le \frac{1}{2}B^2} L_D(h) + B\rho\sqrt{\frac{8}{m}}$$

PAC learning

How to choose λ in practice?

$$\mathbb{E}_S L_D(A(S)) \le \inf_{h: R(h) \le \frac{1}{2}\left(\frac{\rho}{\lambda}\sqrt{\frac{8}{m}}\right)^2} L_D(h) + \frac{8\rho^2}{\lambda m}$$

constant λ can converge but B shrinks, can't compare every thing in $\mathcal{H}$.

λ can be $\frac{1}{\sqrt{m}}$, or we let $\begin{cases} \frac{1}{\lambda\sqrt{m}} \to \infty \\ \frac{1}{\lambda m} \to 0 \end{cases}$ $\Rightarrow \lambda \propto \frac{1}{m^\gamma} \quad \gamma \in (\frac{1}{2}, 1)$

to make B constant

Lec Gradient Descent

GD tries to find $\min_w f(w)$. start at w_1, then update

$w_{t+1} = w_t - \eta_t \nabla f(w_t)$ $f: \mathcal{W} \to \mathbb{R}$, $\mathcal{W} \subseteq \mathbb{R}^d$

repeat T steps then return

last iterate = w_T

average iterate : $\frac{1}{T} \sum_{i=1}^{T} w_T$

best iterate : $w_{\hat{t}}$: $\hat{t} \in \underset{t \in [T]}{\operatorname{argmin}} f(w_t)$, $w \in \{w_1, \cdots, w_T\}$

Assume η_t independent of data

what if w is optimized in constrained space?

Projected Gradient Descent

define $\operatorname{proj}_{\mathcal{W}}(w) \in \operatorname{argmin}_{\hat{w} \in \mathcal{W}} \|w - \hat{w}\|$ find the closest point in the set

Thm. If $\mathcal{W}$ is closed and convex , $\forall v \in \mathcal{W}$, $\|\operatorname{proj}_{\mathcal{W}}(w) - v\| \le \|w - v\|$

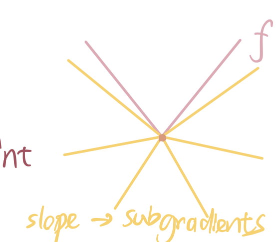
what if the function isn't differentiable ?

subgradient descent

subgradient : slope of planes that lower bound f

convex function $\Leftrightarrow$ has subgradient over every point

convexity : $f(w') \ge f(w) + \langle \nabla f(w), w' - w \rangle$



Def (subgradient). g is a subgradient of f at w if

$\forall w'$, $f(w') \ge f(w) + \langle g, w' - w \rangle$

The subdifferential of f at w , $\partial f(w)$, is the set of all subgradients of f at w

Prop 1. If f is convex and differentiable at w $\partial f(w) = \{\nabla f(w)\}$

Prop 2. f is convex iff $\forall w$, $\partial f(w)$ is non-empty

Prop 3. If f is ρ-Lipschitz, $\forall w$, $\forall g \in \partial f(w)$, $\|g\| \le \rho$

Pf. $\forall w$ in the interior of domain of f, Let $g \in \partial f(w)$

$$\rho \varepsilon \ge f(w + \varepsilon \frac{g}{\|g\|}) - f(w) \ge \langle g, \varepsilon \frac{g}{\|g\|} \rangle = \frac{\varepsilon \langle g, g \rangle}{\|g\|} = \varepsilon \|g\|$$

prop 4. Let $f(w) = \max_{i \in \mathcal{X}} f_i(w)$, $|\mathcal{X}| < \infty$, each f_i is convex.

$\forall w$, if $j \in \arg\max_i f_i(w)$, then $\partial f_j(w) \subseteq \partial f(w)$

Pf. $g \in \partial f_j(w)$, $f(w') \ge f_j(w') \ge f_j(w) + \langle g, w' - w \rangle$

$$= f(w) + \langle g, w' - w \rangle$$

Stochastic (projected) (sub)gradient descent

Lemma. Let $v_1, \dots, v_T$ be an arbitrary sequence, $w_{t+1} = \text{proj}_W (w_t - \eta v_t)$

Then $\sum_{t=1}^T \langle w_t - w^*, v_t \rangle \le \frac{1}{2\eta} \|w_1 - w^*\| + \frac{\eta}{2} \sum_{t=1}^T \|v_t\|^2$

Pf. $\langle w_t - w^*, v_t \rangle = \frac{1}{\eta} \langle w_t - w^*, \eta v_t \rangle$ $\qquad 2\langle a, b \rangle = \|a\|^2 + \|b\|^2 - \|a-b\|^2$

$$= \frac{1}{2\eta} (-\underbrace{\|w_t - \eta v_t - w^*\|^2}_{proj} + \|w_t - w^*\|^2 + \eta^2 \|v_t\|^2)$$

$$\le \frac{1}{2\eta} (-\underbrace{\|w_{t+1} - w^*\|^2}_{d_{t+1}} + \underbrace{\|w_t - w^*\|^2}_{d_t} + \eta^2 \|v_t\|^2)$$

$$\sum_{t=1}^T \langle w_t - w^*, v_t \rangle \le \frac{1}{2\eta} (\underbrace{-d_{T+1}}_{<0} + d_1) + \frac{\eta}{2} \sum_{t=1}^T \|v_t\|^2$$

$$\le \frac{1}{2\eta} \|w_1 - w^*\|^2 + \frac{\eta}{2} \sum_{t=1}^T \|v_t\|^2$$

projected GD bound

To bound $f(\bar{w}) - f(w^*)$ any $w^* \in W$

$w^* \in W$; f is convex and ρ-Lipschitz

$$f(\bar{w}) - f(w^*) = f(\frac{1}{T} \Sigma_{t=1}^T w_t) - f(w^*)$$

$$\le \frac{1}{T} \Sigma_{t=1}^T (f(w_t) - f(w^*))$$

Jensen's inequality:

convex $f \rightarrow f(\frac{1}{n} \sum_i x_i) \le \frac{1}{n} \sum_i f(x_i)$

$g_t \in \partial f(w_t)$, $f(w_t) - f(w^*) \le \langle g_t, w_t - w^* \rangle$

$$\le \frac{1}{T} \Sigma_{t=1}^T \langle g_t, w_t - w^* \rangle$$

use Lemma

if $\|w_1 - w^*\| \le B$ $\qquad \le \frac{1}{2\eta T} \|w_1 - w^*\|^2 + \frac{\eta}{2T} \sum_{t=1}^T \|g_t\|^2$

$$\le \frac{B^2}{2\eta T} + \frac{\eta \rho^2}{2}$$

If $\eta = \sqrt{\frac{B^2}{2T} \cdot \frac{2}{\rho^2}} = \frac{B}{\rho\sqrt{T}}$, $f(\bar{w}) - f(w^*) \le \frac{B\rho}{\sqrt{T}}$ tradeoff learning rate η

SGD bound

get an average direction of gradients

$f(w) = \frac{1}{m}\sum_{i=1}^{m} f_i(w) + \lambda R(w)$ $\quad f_i(w) = \ell(h_w, z_i)$

$f(w) = \mathbb{E}_{z\sim D}\, \ell(h_w, z) = L_D(h_w)$ but $\hat{g}_t$ may be depedent as $\hat{g}_1 \to w_2 \to \hat{g}_2$

Assume $\hat{g}_t | w_t$ are independent of each other use data only one pass

$w_{t+1} = \text{proj}_W(w_t - \eta \hat{g}_t)$ for $\mathbb{E}[\hat{g}_t | w_t] \in \partial f(w_t)$

$\mathbb{E}_z \nabla f_t(w) = \mathbb{E}_z \nabla \ell(h_w, z_t) = \nabla \underbrace{\mathbb{E}_{z\sim D}\ell(h_w, z)}_{L_D}$ unbiased gradient estimator

unseen z

$\hat{g}_{t:\tau} = (\hat{g}_t, \hat{g}_{t+1}, \cdots, \hat{g}_\tau)$

$\mathbb{E}_{\hat{g}_{1:T}}[f(\bar{w}) - f(w^*)] \le \mathbb{E}_{g_{1:T}} \frac{1}{T}\sum_{i=1}^{T} f(w_t) - f(w^*) = \frac{1}{T}\sum_{i=1}^{T} \mathbb{E}_{g_{1:T}} f(w_t) - f(w^*)$

randomness in selected data

$\mathbb{E}_{g_{1:T}} f(w_t) - f(w^*) = \mathbb{E}_{g_{1:t-1}} f(w_t) - f(w^*)$

$\mathbb{E}_{\hat{g}_{1:t-1}}[f(w_t) - f(w^*)] \le \mathbb{E}_{\hat{g}_{1:t-1}}[\langle w_t - w^*, \mathbb{E}_{\hat{g}_t}[\hat{g}_t | g_{1:t-1}]\rangle]$ $\quad g_t = \mathbb{E}[\hat{g}_t | w_t] = \mathbb{E}_{\hat{g}_t}[\hat{g}_t | \hat{g}_{t-1}]$

$= \mathbb{E}_{\hat{g}_{1:t-1}}[\mathbb{E}_{\hat{g}_t}[\langle w_t - w^*, \hat{g}_t\rangle | \hat{g}_{1:t-1}]]$

$= \mathbb{E}_{\hat{g}_{1:t}}[\langle w_t - w^*, \hat{g}_t\rangle]$

$\mathbb{E}_{\hat{g}_{1:T}}[f(\bar{w}) - f(w^*)] \le \mathbb{E}_{\hat{g}_{1:T}}[\frac{1}{T}\sum_{t=1}^{T}\langle w_t - w^*, \hat{g}_t\rangle]$

$\le \frac{1}{2\eta T}\|w_1 - w^*\| + \frac{\eta}{2T}\sum_{t=1}^{T}\mathbb{E}\|\hat{g}_t\|^2$

Lec Non-convex Optimization

2 layer NN:

$h_w(x) = w_2 \cdot \sigma(w_1 x)$ $\quad \sigma(t) = t$ linear model

$h_{-w}(x) = -w_2\, \sigma(-w_1 x) = h_w(x)$

$L_S(h_w) = \frac{1}{m}\sum_{i=1}^{m}(h_w(x_i) - y_i)^2$

If convex, $\ell(h_0, z) \le \ell(h_w, z)$ as $\ell(h_w, z) = \ell(h_{-w}, z)$ doesn't hold!
So deep linear model is not convex

Def (β-smooth). A function f is β-smooth if f is differentiable everywhere, and gradient ∇f is β-Lipschitz

prop. If f is twice-differentiable, f β-smooth iff $\nabla^2 f(x) \preceq \beta I$ for $\forall x$

prop. f is β-smooth. Then for any x and y in its domain

$$|f(y) - f(x) - \langle \nabla f(x), y-x \rangle| \le \frac{1}{2}\beta \|y-x\|^2$$

pf. Let $x_\alpha = (1-\alpha)x_0 + \alpha x_1$, $g(\alpha) = f(x_\alpha)$, $g'(\alpha) = \langle \nabla f(x_\alpha), x_1 - x_0 \rangle$

$$\begin{aligned} f(x_1) - f(x_0) &= g(1) - g(0) = \int_0^1 g'(\alpha)\, d\alpha = \int_0^1 \langle \nabla f(x_\alpha), x_1 - x_0 \rangle d\alpha \\ &= \int_0^1 \langle \nabla f(x_\alpha) - \nabla f(x_0) + \nabla f(x_0), x_1 - x_0 \rangle d\alpha \\ &= \langle \nabla f(x_0), x_1 - x_0 \rangle + \int_0^1 \langle \nabla f(x_\alpha) - \nabla f(x_0), x_1 - x_0 \rangle d\alpha \end{aligned}$$

$$\begin{aligned} |f(x_1) - f(x_0) - \langle \nabla f(x_0), x_1 - x_0 \rangle| &\le \int_0^1 |\langle \nabla f(x_\alpha) - \nabla f(x_0), x_1 - x_0 \rangle|\, d\alpha \\ &\le \int_0^1 \|\nabla f(x_\alpha) - \nabla f(x_0)\| \|x_1 - x_0\| d\alpha \\ &\le \int_0^1 \beta \|x_\alpha - x_0\| \|x_1 - x_0\| d\alpha \\ &= \int_0^1 \alpha\beta \|x_1 - x_0\|^2 d\alpha \\ &= \frac{1}{2}\beta \|x_1 - x_0\|^2 \end{aligned}$$

$\|x_\alpha - x_0\| = \|\alpha(x_1 - x_0)\|$

Descent Lemma: If ∇f is β-Lipschitz (or f β-smooth), $x_{t+1} = x_t - \eta \nabla f(x_t)$

$$\begin{aligned} f(x_{t+1}) &\le f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{1}{2}\beta \|x_{t+1} - x_t\|^2 \\ &= f(x_t) - \eta \langle \nabla f(x_t), \nabla f(x_t) \rangle + \frac{1}{2}\beta\eta^2 \|\nabla f(x_t)\|^2 \\ &= f(x_t) - \eta(1 - \frac{1}{2}\beta\eta) \|\nabla f(x_t)\|^2 \end{aligned}$$

need $1 - \frac{1}{2}\beta\eta > 0$ to minimize f, $\eta < \frac{2}{\beta}$

do GD on DNN will get to stationary point. ERM like

positive homogenuous:

ReLU σ $\quad W_3 \sigma(\underset{\times\frac{1}{2}}{W_2} \sigma(\underset{\times 2}{W_1} x))$

Lec Neural Tangent Kernel

two layer NN

$h(x;w) = \frac{1}{\sqrt{N}} \sum_{j=1}^{N} \underline{a_j}\, \sigma(w_j \cdot x)$ $\qquad w^{t=1} \sim \mathcal{N}(0, Id)$

fixed second layer

$\langle A, B\rangle_F = \sum_{ij} A_{ij} B_{ij}$

$h(x;w) \approx h_{\tilde{w}}(x;w) = h(x;\tilde{w}) + \langle \nabla_w h(x;\tilde{w}), w-\tilde{w}\rangle_F$ Talor approximation

good approximation if $w, \tilde{w}$ are near

$= \frac{1}{\sqrt{N}} \sum_{j=1}^{N} a_j [\sigma(\tilde{w}_j \cdot x) + \sigma'(\tilde{w}_j \cdot x) \cdot x \cdot (w_j - \tilde{w}_j)]$

$= \frac{1}{\sqrt{N}} \sum_{j=1}^{N} a_j [\underbrace{\sigma(\tilde{w}_j \cdot x) - \sigma'(\tilde{w}_j \cdot x) x \cdot \tilde{w}_j}_{\sigma(t) - \sigma'(t) \cdot t \text{ doesn't depend on } W \text{ constant term}} + \underbrace{\sigma'(\tilde{w}_j \cdot x) \cdot x \cdot w_j}_{\text{linear in } W}]$

For ReLU: $t>0 \quad = t - t = 0 \qquad t<0 \quad 0 - 0 = 0$

$h_{\tilde{w}}(x;W) = \langle \underbrace{\nabla_w h(x;\tilde{W})}_{\varphi_{\tilde{w}}(x)}, W\rangle : \{x \mapsto h_{\tilde{w}}(x;w) : W \in \mathbb{R}^{N \times d}\}$ is RKHS

$K_{\tilde{w}}(x, x') = \langle \nabla_w h(x;\tilde{w}), \nabla_w h(x';\tilde{w})\rangle$

For general activations

$h_{\tilde{w}}(x;w) - h(x;\tilde{w}) = \langle \nabla_w h(x;\tilde{w}), w - \tilde{w}\rangle_F$ is in RKHS

so how good is the linearization ?

$L_D(A(S)) - L^* = \underbrace{L_D(A(S)) - L_D(ERM_{\mathcal{H}}(S))}_{\text{optimization error}} + \underbrace{L_D(ERM_{\mathcal{H}}(S)) - \inf_{h^* \in \mathcal{H}} L_D(h^*)}_{\text{estimation error}} + \underbrace{\inf_{h^* \in \mathcal{H}} L_D(h^*) - L^*}_{\text{approximation error}}$

$|h(x;w) - h_{\tilde{w}}(x;w)| \le \frac{1}{\sqrt{N}} \sum_{j=1}^{N} |a_j|\, |\sigma(w_j \cdot x) - \sigma(\tilde{w}_j \cdot x) - \sigma'(\tilde{w}_j \cdot x)(w_j \cdot x - \tilde{w}_j \cdot x)|$

If σ is β-smooth

$|\sigma(t) - \sigma(\tilde{t}) - \sigma'(t)(t - \tilde{t})| \le \frac{1}{2}\beta \|t - \tilde{t}\|^2$

assume σ β-smooth

$\le \frac{1}{\sqrt{N}} \sum_{j=1}^{N} |a_j| \frac{1}{2} \beta \|w_j \cdot x - \tilde{w}_j \cdot x\|^2$

$\le \frac{\beta}{2\sqrt{N}} (\max_j |a_j|) \sum_j \|w_j - \tilde{w}_j\|^2 \|x\|^2$

$= \frac{\beta}{2\sqrt{N}} (\max_j |a_j|) \|x\|^2 \|W - \tilde{W}\|_F^2$ if N is large enough the linearization is good for

$\le B \quad \Rightarrow \quad \|W - \tilde{W}\|_F \le \sqrt{\frac{2B}{\beta}} \frac{1}{\sqrt{\max_j |a_j|}} \frac{1}{\|x\|} N^{\frac{1}{4}}$ large $\|w - \tilde{w}\|_F$

gradient flow

$$\frac{dw_t}{dt} = -\eta \nabla L_S(x \mapsto h(x; w_t))$$ easier to analyze than GD

square loss $= \frac{-2\eta}{m}\sum_{i=1}^{m} (h(x_i; w_t) - y_i)\nabla_w h(x_i; w_t))$

$$\frac{d}{dt} h(x_j; w_t) = \langle \nabla h(x_j; w_t), \frac{dw_t}{dt} \rangle$$

$$= \frac{-2\eta}{m}\sum_i (h(x_i; w_t) - y_i)\underbrace{\langle \nabla_w h(x_i; w_t), \nabla_w h(x_j; w_t)\rangle}_{K_{w_t}(x_i, x_j)}$$

kernel matrix $m \times m$

$$\frac{d}{dt} h|_{S_x}(t) = \frac{-2\eta}{m}(K_{w_t}|S_x)(h|_{S_x}(t) - S_y)$$

suppose K_{w_t} is constant in t then this dynamics will be the same as kernel gradient descent for kernel regression.

$$L_S(h) = \frac{1}{m}\sum_{i=1}^{m}(h(x_i) - y_i)^2$$

$$= \frac{1}{m}\sum_{i=1}^{m}(\langle h, \varphi(x_i)\rangle\langle\varphi(x_i), h\rangle - 2y_i\langle\varphi(x_i), h\rangle + y_i^2)$$ define $[a \otimes b]b' = \langle b, b\rangle a$

$$= \langle h, [\frac{1}{m}\sum_{i=1}^{m}\varphi(x_i)\otimes\varphi(x_i)]h\rangle - 2y_i\langle\frac{1}{m}\sum_{i=1}^{m}\varphi(x_i), h\rangle + \frac{1}{m}\sum_{i=1}^{m} y_i^2$$

$$\frac{dh_t}{dt} = -\eta\nabla_h L_S(h) = \frac{-2\eta}{m}\Sigma_{i=1}^{m}[\varphi(x_i)\otimes\varphi(x_i)]h + \frac{2\eta}{m}\Sigma_{i=1}^{m} y_i\varphi(x_i)$$

$$= \frac{-2\eta}{m}\Sigma_{i=1}^{m}(h(x_i) - y_i)\varphi(x_i)$$

$a_j = \{\pm 1\}$

$$\langle \nabla_w h(x; \tilde{w}), \nabla_w h(x'; \tilde{w})\rangle = \frac{1}{N}\Sigma_{j=1}^{N} a_j^2 \langle x\sigma'(\tilde{w}_j \cdot x), x'\sigma'(\tilde{w}_j \cdot x)\rangle$$

$$= x \cdot x' \frac{1}{N}\Sigma_{j=1}^{N}\sigma'(\tilde{w}_j \cdot x)\cdot\sigma'(\tilde{w}_j \cdot x')$$

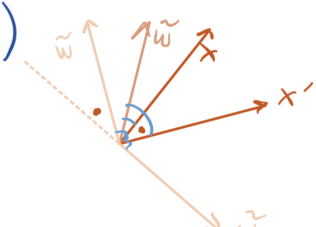
large number law ; converge surely to as $N \to \infty$

$$x \cdot x' \mathop{\mathbb{E}}_{\tilde{w} \sim N(0, I)}[\sigma'(\tilde{w}\cdot x)\sigma(w \cdot x)] = x \cdot x' \Pr(\frac{\tilde{w} \cdot x}{\|\tilde{w}\|\|x\|} > 0 \text{ and } \frac{\tilde{w} \cdot x'}{\|\tilde{w}\|\|x\|} > 0)$$ normalize

$\mathbb{E}(\mathbb{1}(a>0)) = \Pr(a > 0)$

$$= x \cdot x' \left(\frac{1}{2} - \frac{\arccos\frac{x \cdot x'}{\|x\|\|x'\|}}{2\pi}\right)$$

converge almost surely to some constant



empirical NTK

Looking at the prediction on x, based on a single SGD update on x_t

$$h(x; w_{t+1}) - h(x; w_t) = \eta B(x) K_w(x, x_t)(y_t - h(x_t; w_t)) + O(\eta^2)$$

Lec Implicit Regularization

$$f(w) = L_S^{sq}(x \mapsto w \cdot x) = \frac{1}{m}\|\overset{m\times d}{X} w - \overset{m\times 1}{y}\|^2$$

if $d > m$, have infinite many solutions

so what solution will we get when running GD?

kernel GD flow is not actually used because η is not $\to 0$

$$\nabla f(w) = \frac{2}{m}X^T(Xw - y)$$

$$w_{t+1} = w_t - \frac{2\eta_0}{m}X^TXw_t + \frac{2\eta_0}{m}X^Ty \qquad (\text{set } \eta = \frac{2\eta_0}{m})$$

$$= (I - \eta X^TX)w_t + \eta X^Ty$$

$$= (I - \eta X^TX)^2 w_{t-1} + (I - \eta X^TX)\eta X^Ty + \eta X^Ty$$

$$= (I - \eta X^TX)^t w_1 + \sum_{k=0}^{t-1}(I - \eta X^TX)^k \eta X^Ty$$

$$\overset{SVD}{=} (I - \eta V\Sigma^2V^T)^t w_1 + \sum_{k=0}^{t-1}(I - \eta V\Sigma^2V^T)^k \eta V\Sigma U^Ty$$

$$\overset{④}{=} (I - VV^T)^t w_1 + V(I - \eta\Sigma^2)^t V^Tw_1 + \eta\sum_{k=0}^{t-1}\left((I - VV^T) + V(I - \eta\Sigma^2)^k V^T\right)V\Sigma U^Ty$$

$$= (I - VV^T)^t w_1 + V\underline{(I - \eta\Sigma^2)^t}V^Tw_1 + \eta V\underline{\sum_{k=0}^{t-1}(I - \eta\Sigma^2)^k}\Sigma U^Ty$$

Assume $\eta < \frac{2}{\sigma_1^2}$ $\in$ descending order, σ_1: largest singular value of X

to ensure $|1 - \eta\sigma_i^2| < 1$, so $(I - \eta\Sigma^2)^{\infty} \to 0$ $\qquad \sigma^2(X) = \lambda(X^TX)$

Neumann series

$$w_\infty = (I - VV^T)^t w_1 + \underbrace{V\Sigma^{-1}U^Ty}_{X^+} = (I - VV^T)^t w_1 + X^+y$$

pseudoinverse of X

$(I - VV^T)^2 = I - 2VV^T + VV^T = I - VV^T$

there is a unique solution $X^+y = (X^TX)^{-1}X^Ty$

SVD singular value decomposition

If X is $m\times d$ of rank $r \le \min(m,d)$, then Σ is $r\times r$ diagonal with $\Sigma_{ii} > 0$

U is $m\times r$, V is $d\times r$ with $U^TU = I_r = V^TV$

$$\underset{m\times d}{X} = \underset{m\times r}{U}\ \underset{r\times r}{\Sigma}\ \underset{r\times d}{V^T}$$

① $X^TX = V\Sigma U^TU\Sigma V^T = V\Sigma^2V^T$ $\qquad (\Sigma^2)_{ii}$ is eigenvalue of X^TX / XX^T ($\sigma(X)$)

② $(VV^T)(VV^T) = V\underline{V^TV}V^T = VV^T$

③ $X(I - VV^T)y = U\Sigma V^Ty - U\Sigma\underline{V^TV}V^Ty = 0$

④ $(I - \eta X^TX)^k = (I - \eta V\Sigma^2V^T)^k$

$$= (I - VV^T + \underbrace{VV^T - \eta V\Sigma^2V^T}_{V(I - \eta\Sigma^2)V^T})^k$$

$$= (I - VV^T)^k + V(I - \eta\Sigma^2)^kV^T + \underbrace{V(I - \eta\Sigma^2)V^T - V(I - \eta\Sigma^2)V^T}_{0}$$