CS861 Notes

Young Wu

April 13, 2019

1 Lecture 1

output, label space Y (classification: finite discrete) (regression $\mathbb{R}$)
input, item, instance, object, point space $X\left(e.g.\mathbb{R}^d\right)$
training set: $\left(x_i \in X, y_i \in Y\right)_{i=1:n} \overset{iid}{\sim} P_{X \times Y}$
test data $\overset{iid}{\sim} P_{X \times Y}$
finding "best" $h : X \to Y$
$h^\star(x) \in \arg\max_y P(y|x)$, unknown, cannot compute
loss function: $\ell : Y \times Y \to \mathbb{R}_{\geqslant 0}$
e.g. $0 - 1$ loss:

$$\ell(y_1, y_2) = \begin{cases} 1 & \text{if } y_1 \neq y_2 \\ 0 & \text{otherwise} \end{cases}$$

squared loss:

$$\ell(y_1, y_2) = \frac{1}{2}(y_1 - y_2)^2$$

Risk: $R(h) := \mathbb{E}_{(x,y)\sim P}\left[\ell(h(x), y)\right]$
Empirical Risk: $\hat{R}(h) := \frac{1}{n}\sum_{i=1}^{n} \ell(h(x_i), y_i)$
$h^\star = \arg\min_{\{h:X\to Y\}} R(h)$, cannot compute
Hypothesis space: $\mathcal{H} = \{h\}$
What ML does: Given $(x_i, y_i)_{i=1:n} := S$ or $D \overset{iid}{\sim} P_{XY}^n$

$$\hat{h} \in \arg\min_{h\in\mathcal{H}} \hat{R}(h)$$
$$\Rightarrow \hat{R}\left(\hat{h}\right) = 0$$

Assumption: Realizability

$$\exists h^\star \in \mathcal{H}, R(h^\star) = 0$$
$$\Rightarrow \hat{R}(h^\star) = 0$$

Probably Approximately Correct

$$R\left(\hat{h}\right) \leqslant \varepsilon$$

2 Lecture 2

1. Get dataset $S = (x_i, y_i)_{i=1:n}$
2. Run ML $\hat{h}_s = \mathrm{ML}(S)\,, \hat{R}_s\left(\hat{h}_s\right) = \frac{1}{n}\sum_{i=1}^{n}\ell\left(\hat{h}_s\left(x_j\right), y_j\right)$
3. Test set error $\hat{\hat{R}}\left(\hat{h}_s\right) = \frac{1}{m}\sum_{j=1}^{m}\ell\left(\hat{h}_s\left(x_j\right), y_j\right), e.g. 0.01, e.g. 0$

$$T = (x_j, y_j)_{j=1:m}$$
$$R\left(h_s\right) = \mathbb{E}_{(x,y)\sim P_{XY}}\left[\ell\left(\hat{h}_s\left(x\right), y\right)\right], \text{ head prob}$$

Given $\hat{\hat{R}} = 0$
Suppose $R > \varepsilon$
event prob
Prob (m trials all tails) $< (1-\varepsilon)^m$
Prob (T has $\hat{\hat{R}}\left(\hat{h}_s\right) = 0$) $\leqslant (1-\varepsilon)^m$

Suppose $\hat{R}\left(\hat{h}_s\right) = 0$
want: statement about $R\left(\hat{h}_s\right)$ being large (bad)

$$\hat{h}_s \in \arg\min_{h\in\mathcal{H}} \hat{R}\left(h\right)$$

$\in\leftarrow$ we picked ANY of them!
Empirical Risk Minimization
$R\left(\hat{h}_s\right)$: head prob of coin $\hat{h}_s$

$$\left\{S : \exists h \in \mathcal{H} : R\left(h\right) < \varepsilon \wedge \hat{R}_s\left(h\right) = 0\right\}$$
$$(X \times Y)^n \setminus \left\{S : R\left(\hat{h}_s\right) > \varepsilon\right\}$$
$$\mathbb{P}_{S\sim P^n}\left[\underbrace{\left\{S : R\left(\hat{h}_s\right) > \varepsilon\right\}}_{\mathcal{S}}\right] < \delta$$

fix $h \in \mathcal{H}_\varepsilon, \mathcal{H}_\varepsilon = \{h : R\left(h\right) > \varepsilon\} \subset \mathcal{H}, \mathcal{S}_h = \left\{S : \hat{R}_s\left(h\right) = 0\right\}$

$$\mathcal{S} \subseteq \{\mathcal{S}_h : h \in \mathcal{H}_\varepsilon\} := \bigcup_{h\in\mathcal{H}_\varepsilon} \mathcal{S}_h$$

$$
\begin{aligned}
\mathbb{P}(\mathcal{S}) &\leqslant \mathbb{P}\left(\bigcup_{h\in\mathcal{H}_\varepsilon}\mathcal{S}_h\right)\\
&\overset{UnionB}{\leqslant} \sum_{h\in\mathcal{H}_\varepsilon}\mathbb{P}(\mathcal{S}_h)\\
&= \sum_{h\in\mathcal{H}_\varepsilon}\mathbb{P}(n\text{ tails given }\mathbb{P}(\text{ head }) > \varepsilon)\\
&\overset{R(h)>\varepsilon}{\leqslant} \sum_{h\in\mathcal{H}_\varepsilon}(1-\varepsilon)^n\\
&\leqslant \sum_{h\in\boxed{\mathcal{H}}}(1-\varepsilon)^n\\
&\overset{\boxed{\mathcal{H}\text{ finite}}}{=} |\mathcal{H}|(1-\varepsilon)^n
\end{aligned}
$$

3 Lecture 3

- "task, world, environment, population" fixed unknown P_{XY}
- training data $S \overset{iid}{\sim} P^n$
- ERM: $\hat{h}_s = \arg\min_{h\in\mathcal{H}}\hat{R}(h) := \frac{1}{n}\sum_{(x,y)\in S}\ell(h(x),y)$
- Hypo space $\mathcal{H}$

"ideally" want:

$$h^\star \in \arg\min_{h\in\mathcal{H}} R(h) : = \mathbb{E}_{(x,y)\sim P}\ell(h(x),y)$$

A1: $\ell 0-1$ loss
A2: $R(h^\star) = 0$ "realizable case"
A3: $|\mathcal{H}| < \infty$

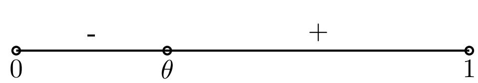
- "Wolf" $\mathcal{H}_\varepsilon := \{h\in\mathcal{H} : R(h) > \varepsilon\}$

A2: $\hat{R}\left(\hat{h}_s\right) = 0$ "looks like a sheep"

- "Sheep" $\mathcal{H}\backslash\mathcal{H}_\varepsilon$

Bad event:

- Want: $\mathbb{P}_{S\sim P^n}[\{S : S\text{ enables a wolf to look like the best sheep }\}] < \delta$



$$P_X = U[0,1]$$

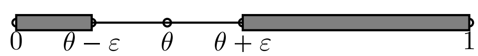
$$P\left(y=+|x\right)=\begin{cases}1 & \text{if } x\geqslant\theta\\ 0 & \text{otherwise}\end{cases}$$

$$y\in\{-,+\}$$

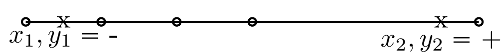
$$h_a=\begin{cases}+ & \text{if } x\geqslant a\\ - & \text{otherwise}\end{cases}$$

$$\mathcal{H}=\{h_a : a\in[0,1]\}$$
$$\ell : 0-1$$
$$R\left(h_a\right)=|a-\theta|$$

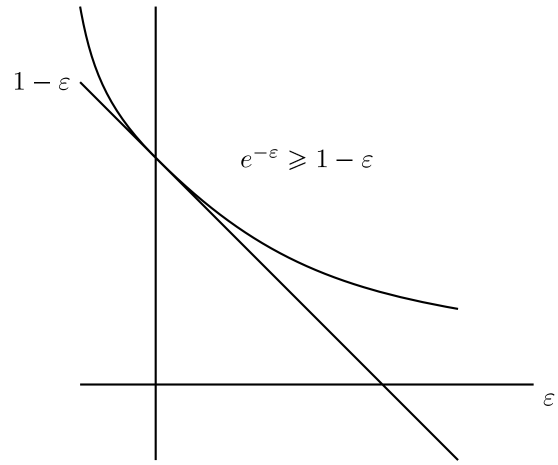
"wolves"



S



$$\begin{aligned}
&\mathbb{P}\left[\left\{S : \exists h\in\mathcal{H}_\varepsilon, \hat{R}\left(h\right)=0\right\}\right]\\
&=\mathbb{P}\left[\bigcup_{h\in\mathcal{H}_\varepsilon}\left\{S : \hat{R}\left(h\right)=0\right\}\right]\\
&\overset{Unionb.}{\leqslant}\sum_{h\in\mathcal{H}_\varepsilon}\mathbb{P}\left[\left\{S : \hat{R}\left(h\right)=0\wedge R\left(h\right)>\varepsilon\right\}\right]\\
&\overset{\mathcal{H}_\varepsilon}{\leqslant}\sum_{h\in\mathcal{H}_\varepsilon}\left(1-\varepsilon\right)^n\\
&\overset{\mathcal{H}_\varepsilon\subseteq\mathcal{H}}{\leqslant}\sum_{h\in\mathcal{H}}\left(1-\varepsilon\right)^n\\
&\overset{A3}{=}|\mathcal{H}|\left(1-\varepsilon\right)^n\\
&\overset{e^{-\varepsilon}\geqslant 1-\varepsilon}{\leqslant}|\mathcal{H}|e^{-\varepsilon n}:=\delta
\end{aligned}$$



$$
\begin{aligned}
-\varepsilon n &= \log\left(\frac{\delta}{|\mathcal{H}|}\right) \\
\varepsilon &= \frac{1}{n}\log\left(\frac{|\mathcal{H}|}{\delta}\right)
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{P}\left(\{\text{ Bad } S\}\right) &\leqslant \delta \\
\mathbb{P}\left((X \times Y)^n \setminus \{\text{ Bad } S\}\right) &> 1-\delta \\
\mathbb{P}\left[R\left(\hat{h}_s\right) \leqslant \varepsilon\right] &\geqslant 1-\delta
\end{aligned}
$$

With prob at least $1-\delta$,

$$
\begin{aligned}
R\left(\hat{h}_s\right) \leqslant \varepsilon &:= \frac{\log\left(|\mathcal{H}|\right) - \log\left(\delta\right)}{n} \\
n &:= \frac{\log\left(|\mathcal{H}|\right) - \log\left(\delta\right)}{\varepsilon}
\end{aligned}
$$

4 Lecture 4

A1: ℓ is $0-1$ loss

$$
\begin{aligned}
&\mathbb{P}_{S\sim P^n}\left(\left\{S : \exists h \in \mathcal{H}_\varepsilon, \quad \underbrace{\hat{R}_s\left(h\right) = 0}_{\text{A2: Realizable } \min_{h\in\mathcal{H}} R(h) = 0}\right\}\right) \\
\mathcal{H}_\varepsilon &= \left\{h \in \mathcal{H} : R\left(h\right) > \varepsilon\right\} \\
R\left(h\right) &= \mathbb{E}_{(x,y)\sim P}\ell\left(h\left(x\right), y\right) \\
\hat{R}_s\left(h\right) &= \frac{1}{n}\sum_{x,y\in S}\ell\left(h\left(x\right), y\right)
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{P}\left(\bigcup_{h\in\mathcal{H}_\varepsilon}\left\{S:\hat{R}_s\left(h\right)=0\right\}\right)\\
&\overset{\text{Union}}{\leqslant}\sum_{h\in\mathcal{H}_\varepsilon}\mathbb{P}\left(S:\hat{R}_s\left(h\right)=0\wedge R\left(h\right)>\varepsilon\right)\\
&\overset{\text{Realizability}}{\leqslant}\sum_{h\in\mathcal{H}_\varepsilon}\left(1-\varepsilon\right)^n\\
&\overset{A3:|\mathcal{H}|<\infty}{\leqslant}\left|\mathcal{H}\right|\left(1-\varepsilon\right)^n\\
&\leqslant|\mathcal{H}|e^{-\varepsilon n}:=\delta
\end{aligned}
$$

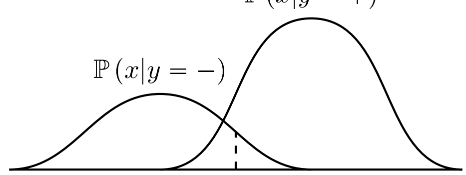
$$
\begin{aligned}
&n\geqslant\frac{\log|\mathcal{H}|-\log\delta}{\varepsilon}\\
&\varepsilon\leqslant O\left(\frac{1}{n}\right)
\end{aligned}
$$

Agnostic learning (wrt $\mathcal{H}$) , $R\left(h^\star\right)\geqslant 0$

$$
h^\star\in\arg\min_{h\in\mathcal{H}}R\left(h\right)
$$

e.g.

$$
\mathbb{P}\left(y=+\right)=\frac{1}{2}
$$



Want Uniform Convergence:

$$
\begin{aligned}
&\left\{S:\exists h\in\mathcal{H},\left|R\left(h\right)-\hat{R}_s\left(h\right)\right|>\varepsilon\right\}\\
&\forall\ h\in\mathcal{H},|R\left(h\right)-\hat{R}_s\left(h\right)|\leqslant\varepsilon
\end{aligned}
$$

" $R\left(\hat{h}_s\right)-R\left(h^\star\right)$ small"

$$
\begin{aligned}
&\mathbb{P}\left(\left\{S:\exists h\in\mathcal{H},\left|R\left(h\right)-\hat{R}_s\left(h\right)\right|>\varepsilon\right\}\right)\\
&\leqslant\sum_{h\in\mathcal{H}}\mathbb{P}\left(\left\{S:\left|R\left(h\right)-\hat{R}_s\left(h\right)\right|>\varepsilon\right\}\right)
\end{aligned}
$$

$$\overset{\text{Hoeffding's}}{\leqslant} \quad 2|\mathcal{H}|e^{-\dfrac{2n\varepsilon^2}{(b-a)^2}} := \delta$$

new A1: $\ell \in [a,b]$ or ${}_{(x,y)\sim P}\ell\left(h\left(x\right),y\right)$ is subGaussian
remove A2
retain A3
Hoeffding's Ineq. $\left|\mu - \dfrac{1}{n}\sum_{i}^{n}\theta_i\right|$

$$\mathbb{P}_{S\sim P^n}\left(\left\{S : \left|R\left(h\right) - \hat{R}_s\left(h\right)\right| > \varepsilon\right\}\right) \leqslant 2e^{-\dfrac{2n\varepsilon^2}{(b-a)^2}}$$

$$\boxed{\text{wp } \geqslant 1-\delta,\ \forall\ h \in \mathcal{H}, \left|R\left(h\right) - \hat{R}_s\left(h\right)\right| \leqslant \varepsilon}$$

$$\begin{aligned}
&R\left(\hat{h}_s\right) - R\left(h^\star\right)\\
&= R\left(\hat{h}_s\right) - \hat{R}_s\left(\hat{h}_s\right) + \hat{R}_s\left(\hat{h}_s\right) - \hat{R}_s\left(h^\star\right) + \hat{R}_s\left(h^\star\right) - R\left(h^\star\right)\\
&\leqslant \varepsilon + \underbrace{\left(\hat{R}_s\left(\hat{h}_s\right) - \hat{R}_s\left(h^\star\right)\right)}_{\text{ERM } \leqslant 0} + \varepsilon\\
&\leqslant 2\varepsilon
\end{aligned}$$

$$\begin{aligned}
\frac{2n\varepsilon^2}{(b-a)^2} &= \log\frac{2|\mathcal{H}|}{\delta}\\
n &= \frac{\log\left(2|\mathcal{H}|\right) - \log\delta}{2\varepsilon^2}(b-a)^2\\
\varepsilon &\leqslant O\left(\frac{1}{\sqrt{n}}\right)
\end{aligned}$$

"learning alg" $A\left(S\right) = h$

5 Lecture 5

$$\mathbb{P}\left(\left\{S \text{ bad }\right\}\right) \leqslant \delta = |\mathcal{H}|e^{-n\varepsilon}$$

i.e. $\exists h, \left|R\left(h\right) - \hat{R}_s\left(h\right)\right| > \varepsilon$
Today's goal: $\mathbb{P}\left(\left\{S \text{ bad }\right\}\right) \geqslant \delta$

Any learning Algorithm $A : \{S\} \to \mathcal{H}$, ERM is a A

$$S \overset{iid}{\sim} P^n\left(x,y\right), n \leqslant \frac{|X|}{2}$$

Theorem 1. $\forall\ A, \exists P$

1. $\exists h : X \to Y, R_P(h) = 0$
2. $\mathbb{P}\left(\left\{S : R_P(A(S)) \geqslant \frac{1}{8}\right\}\right) \geqslant \frac{1}{7}$

$$2n = \left|\{\text{ distinct } x's \text{ in } S\} \cup \{\text{ some other distinct } x \text{ from } X \text{ not in } S\}\right|$$

$$\overbrace{x_1, ..., x_n}^{S_n}, x_{n+1}, ..., x_{2n}$$

Construct a family of $P(x, y) = P(x) \cdot P(y|x)$

1. $P(x)$ uniform on the $2n$ items
2. $\mathcal{C} := 2^{2n}$ labelings over $2n$ items

$$\mathcal{C} \text{ rows } \Rightarrow \mathcal{C} \text{ joint } P_{XY} = \begin{cases} 0...0...0...00, \mathbb{P}(y = 1|x) = 0\ \forall\ x \\ 0...0...0...01, \mathbb{P}(y = 1|x_{2n}) = 1, \mathbb{P}(y = 1|x \neq x_{2n}) = 0 \\ 1...1...1...11, \mathbb{P}(y = 1|x) = 1\ \forall\ x \end{cases}$$

Key idea: $\max_{c \in [\mathcal{C}]} \mathbb{E}_{S \sim P_c^n} R_{P_c}(A(S))$

$$\boxed{z \text{ rv } \geqslant 0, \mathbb{E}(z) \geqslant a \Rightarrow \mathbb{P}(z \geqslant b) \geqslant c}$$

$$\boxed{z \in [0,1] \text{ rv }, \mathbb{E}(z) \geqslant u \Rightarrow \mathbb{P}(z \geqslant 1 - a) \geqslant \frac{u - (1 - a)}{a}}$$

Lemma B.1 Markov's ineq

$Q = (2n)^n$ distinct S sequences, $S_1, S_2, ..., S_Q$

6 Lecture 6

Take $2n$ distinct item from X

$$x_1 ... x_{2n}$$
$$0...00$$
$$0...01$$
$$...$$

$\mathcal{C} := 2^{2n}$ different labelings of the $2n$ items

$$P_1(y|x) \text{ "world" } P_1(x, y) = P(x) P_1(y|x)$$

$$
\begin{aligned}
&...\\
&P_{\mathcal{C}}\left(y|x\right)\\
P\left(x\right) = &\frac{1}{2n}\\
R_c\left(h\right) = &\mathbb{E}_{(x,y)\sim P_c(x,y)} \underbrace{\ell}_{0-1 \text{ loss}} \left(h\left(x\right), y\right), c \in [\mathcal{C}]
\end{aligned}
$$

$$
n \leqslant \frac{|X|}{2}
$$

No-Free lunch Thm

$$
\forall\ A, \exists P_c\left(c \in [\mathcal{C}]\right)
$$

1. $\exists h, R_c\left(h\right) = 0$
2. $\mathbb{P}_{S\sim P_c^n}\left[R_c\left(A\left(S\right)\right) \geqslant \frac{1}{8}\right] \geqslant \frac{1}{7}$

$$
\max_{c\sim[\mathcal{C}]} \mathbb{E}_{S\sim P^n} R_c\left(A\left(S\right)\right) \geqslant \frac{1}{4} \overset{\text{Markov}}{\Rightarrow} (2)
$$

$$
\begin{aligned}
S = &\left(x^1, x^2, ..., x^n\right), x^{i \leftarrow \text{ position in } S}\\
&S_1 : x_1, x_1, ..., x_1\\
&S_2 : x_1, ..., x_1, x_2\\
&...\\
&S_Q : x_{2n}, ..., x_{2n}\\
Q := &\left(2n\right)^n
\end{aligned}
$$

$$
\begin{aligned}
&\max_{c\in[\mathcal{C}]} \mathbb{E}_{S\sim P_c^n} R_c\left(A\left(S\right)\right)\\
&= \max_{c\in[\mathcal{C}]} \frac{1}{Q}\sum_{q=1}^{Q} R_c\left(A\left(S_q\right)\right)\\
&\geqslant \frac{1}{\mathcal{C}}\sum_{c=1}^{\mathcal{C}} \frac{1}{Q}\sum_{q=1}^{Q} R_c\left(A\left(S_q\right)\right)\\
&= \frac{1}{Q}\sum_{q=1}^{Q} \frac{1}{\mathcal{C}}\sum_{c=1}^{\mathcal{C}} R_c\left(A\left(S_q\right)\right)\\
&\overset{P_x \text{ Unif}}{=} \frac{1}{Q}\sum_{q=1}^{Q} \frac{1}{\mathcal{C}}\sum_{c=1}^{\mathcal{C}} \frac{1}{2n}\sum_{i=1}^{2n} \mathbb{1}_{[A(S_q)(x_i)\neq y_c(x_i)]}
\end{aligned}
$$

$$
t := \left|\left\{x_1, ..., x_{2n}\right\} \setminus \left\{S_q\right\}\right| \geqslant n
$$

Let $\{v_1, ..., v_t\}$ be the set

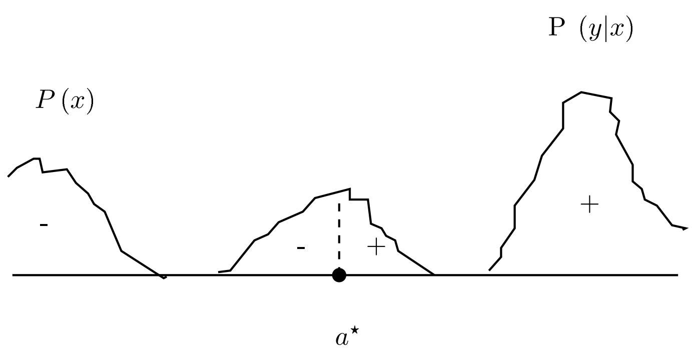
$$
\begin{aligned}
&\overset{\text{only consider } \{v\}}{\geqslant} \frac{1}{Q}\sum_{q=1}^{Q}\frac{1}{\mathcal{C}}\sum_{c=1}^{\mathcal{C}}\frac{1}{2n}\sum_{i=1}^{t}\mathbb{1}_{[A(S_q)(v_i)\neq y_c(v_i)]} \\
&= \frac{1}{Q}\sum_{q=1}^{Q}\frac{1}{2n}\sum_{i=1}^{t}\underbrace{\frac{1}{\mathcal{C}}\sum_{c=1}^{\mathcal{C}}\mathbb{1}_{[A(S_q)(v_i)\neq y_c(v_i)]}}_{\frac{1}{2}} \\
&= \frac{1}{Q}\sum_{q=1}^{Q}\frac{1}{2n}\sum_{i=1}^{t}\frac{1}{2} \\
&= \frac{1}{Q}\sum_{q=1}^{Q}\frac{1}{2n}\frac{1}{2}t \\
&\geqslant \frac{1}{2}\sum_{q=1}^{Q}\frac{1}{2n}\frac{1}{2}n \\
&= \frac{1}{4}
\end{aligned}
$$

$$h_c(x) = \begin{cases} 0 & \text{if } x \notin \mathbb{Z} \\ y_c(x) & \text{otherwise} \end{cases}$$

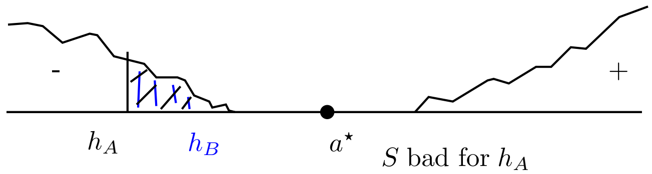
$$|\mathcal{H}| = \infty$$

ex 6.1

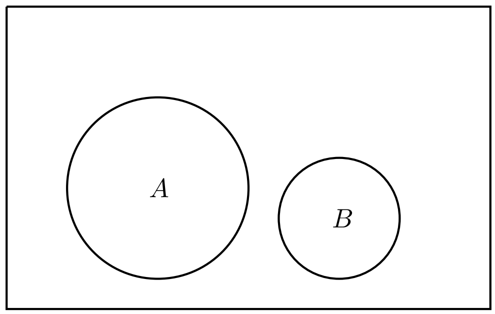
$$\mathcal{H} = \{h_A : a \in \mathbb{R}\}$$

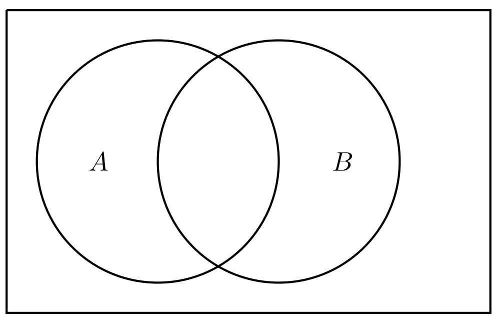


$$h_a(x) = \begin{cases} 1 & \text{if } x \geqslant a \\ 0 & \text{otherwise} \end{cases}$$



$$\mathbb{P}\left(\bigcup_{h\in\mathcal{H}_\varepsilon}\{S \text{ makes } h, 0 \text{ training risk }\}\right) \leqslant \sum_{h\in\mathcal{H}_\varepsilon}\mathbb{P}\left(\{S_{\text{ bad}}\}\right)$$





$$P\left(A \text{ or } B\right) \leqslant P\left(A\right) + P\left(B\right)$$

7 Lecture 7

Recall: finite $\mathcal{H}$

$$\mathbb{P}_{S\sim P^n}\left(\max_{h\in\mathcal{H}} R\left(h\right) - \hat{R}_s\left(h\right) \leqslant \varepsilon\right) \geqslant 1-\delta$$

$$\varepsilon = \sqrt{\frac{\log|\mathcal{H}| + \log\dfrac{1}{\delta}}{2n}}$$

Today: Any $\mathcal{H} = \{h\}, h : X \to Y$

ex. $\mathcal{H} = \left\{ \underbrace{h_a(x)}_{a \in \mathbb{R}} = \operatorname{sign}[\sin(ax)] \right\}$

$$VC = \infty$$

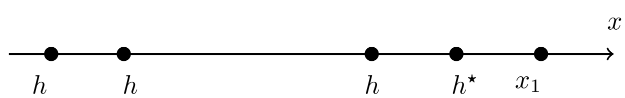
Growth number

$$G(n) := \sup_{x_1, \ldots, x_n \in X} \left| \left\{ \mathbb{1}_{[h(x1) \neq y_1]}, \ldots, \mathbb{1}_{[h(x_n) \neq y_n]} : h \in \mathcal{H} \right\} \right|$$

$$(y_i = h^\star(x_i))$$

$$d := VC(\mathcal{H}) = \arg\max_n G_{\mathcal{H}}(n) = 2^n$$

ex: $\mathcal{H} = \left\{ \underbrace{h_a(x)}_{X = \mathbb{R}} = \{1, x \geqslant a, 0 \text{ ow } \}, a \in \mathbb{R} \right\}$



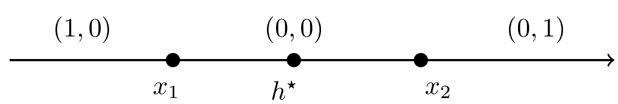
$$\left(\mathbb{1}_{[h(x_1) \neq 0]} \right)$$

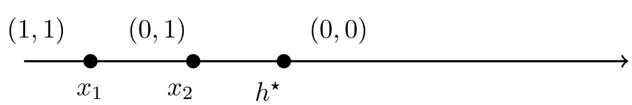
$$|\{(0), (1)\}| = 2$$

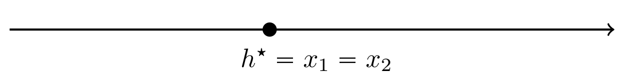
$$G(1) = 2, VC(\mathcal{H}) = 1$$

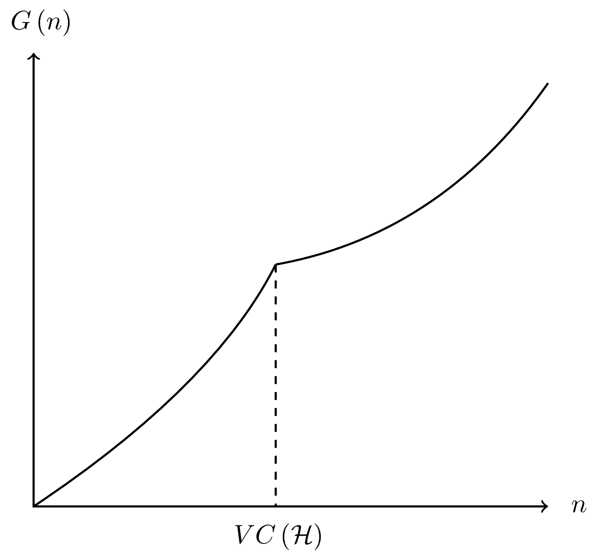
$$G(2) = 3$$

$$G(3)$$









Proof outline

- Introduce "ghost sample" $S' \sim P^n$

$$\hat{R}'_{s'}(h) = \frac{1}{n} \sum_{x', y' \in S'} \ell\left(h\left(x'\right), y'\right)$$

- Symmertrization Lemma

$$\boxed{\forall\ \varepsilon \geqslant \sqrt{\frac{2\log 2}{n}}}, \mathbb{P}_{s\sim P^n}\left(\sup_{h\in\mathcal{H}} R(h) - \hat{R}_s(h) > \varepsilon\right) \overset{\text{Sym lemma}}{\leqslant} 2\mathbb{P}_s\left(\sup_{h\in\mathcal{H}} \hat{R}'_{s'}(h) - \hat{R}_s(h) > \frac{\varepsilon}{2}\right)$$

define

$$\boxed{\left\{\left(\underbrace{\ell(h(x_1), y_1), ..., \ell(h(x_n), y_n)}_{S}, \underbrace{\ell\left(h\left(x'_1\right), y'_1\right), ...\ell\left(h\left(x'_n\right), y'_n\right)}_{S'}\right) : h \in \mathcal{H}\right\} := Vec(2n)}$$

$$\overset{\text{def } Vec(2n)}{=} 2\mathbb{P}_s\left(\max_{h\in Vec(2n)} \hat{R}'_{s'}(h) - \hat{R}_s(h) > \frac{\varepsilon}{2}\right)$$

$$\overset{\text{Union b}}{\leqslant} 2|Vec(2n)|\mathbb{P}\left(\hat{R}'_{s'}(h) - \hat{R}_s(h) > \frac{\varepsilon}{2}\right)$$

$$\overset{\text{Growth num}}{\leqslant} 2G_{\mathcal{H}}(2n)\mathbb{P}\left(\hat{R}'_{s'}(h) - \hat{R}_s(h) > \frac{\varepsilon}{2}\right)$$

$$\overset{\text{Hoeffding's (2 samples)}}{\leqslant} 2G_{\mathcal{H}}(2n)e^{-\frac{n\left(\frac{\varepsilon}{2}\right)^2}{2}}$$

$$\overset{\varepsilon\geqslant\sqrt{\frac{2\log 2}{n}}}{\Rightarrow} \text{wp } \geqslant 1-\delta, \sup_{h\in\mathcal{H}} R(h) - \hat{R}_s(h) \leqslant 2\sqrt{2\frac{\log G_{\mathcal{H}}(2n) + \log\frac{2}{\delta}}{n}}$$

"Tighter than VC-bound"

8 Lecture 8

Growth number $G_{\mathcal{H}}(n) := \sup_{x_1 \dots x_n \in X} |\{(h(x_1), ..., h(x_n)) : h \in \mathcal{H}\}|$
symmetrization lemma

$$\forall \underbrace{\boxed{\varepsilon \geqslant \sqrt{\frac{2\log 2}{n}}}}_{\text{Assumption A1}},$$

$$P\left(\sup_{h\in\mathcal{H}} R(h) - \hat{R}(h) > \varepsilon\right) \leqslant 2P\left(\sup_{h\in\mathcal{H}} \underbrace{\hat{R}'(h)}_{\text{ghost sample}} - \hat{R}(h) > \frac{\varepsilon}{2}\right)$$

$$\Rightarrow P\left(\sup_{h\in\mathcal{H}} R(h) - \hat{R}(h) > \varepsilon\right) \leqslant 2G_{\mathcal{H}}(2n)\, e^{-\frac{n\left(\frac{\varepsilon}{2}\right)^2}{2}} := \delta$$

$$\varepsilon := \sqrt{\frac{8\left(\log 2G_{\mathcal{H}}(2n) - \log\delta\right)}{n}}$$

VC-dim of $\mathcal{H} := d$

$$d = \max_n n : G_{\mathcal{H}}(n) = 2^n$$

Sawer's Lemma:
Assuming $d < \infty$. Then

$$G_n(n) \leqslant \sum_{i=0}^{d} \binom{n}{i}$$

$$\Rightarrow \text{ If } n \geqslant d, G_{\mathcal{H}}(n) \leqslant \left(\frac{en}{d}\right)^d$$

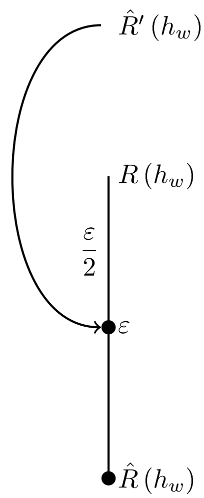
$$P\left(\sup_{h\in\mathcal{H}} R(h) - \hat{R}(h) > \varepsilon\right) \leqslant 2\left(\frac{2en}{d}\right)^d e^{-\frac{n\varepsilon^2}{8}} := \delta$$

$$\Rightarrow \varepsilon = \sqrt{8\frac{d\log n + d\log\frac{2e}{d} + \log\frac{2}{\delta}}{n}} \Rightarrow O\left(\sqrt{\frac{d}{n}}\right)$$

Proof of Sym Lemma. Let a "worst" hypo be $h_w \in \arg\sup_{h\in\mathcal{H}} R(h) - \hat{R}(h)$

$$\mathbb{1}_{\left[R(h_w)-\hat{R}(h_w)>\varepsilon\right]} \cdot \mathbb{1}_{\left[R(h_w)-\underbrace{\hat{R}'(h_w)}_{\text{ghost}}<\frac{\varepsilon}{2}\right]}$$

$$= \mathbb{1}_{\left[R(h_w)-\hat{R}(h_w)>\varepsilon \wedge \left(\hat{R}'(h_w)-R(h_w)\right)>-\frac{\varepsilon}{2}\right]}$$



$$\overset{\text{implication}}{\leqslant} \quad \mathbb{1}_{\left[\hat{R}'(h_w)-\hat{R}(h_w)>\frac{\varepsilon}{2}\right]}$$

$$\overset{\text{expectation over ghost sample } S'}{\Rightarrow} \quad \mathbb{1}_{\left[R(h_w)-\hat{R}(h_w)>\varepsilon\right]} \underbrace{P'}_{\text{wrt } S'\sim P_{XY}^n}\left(R\left(h_w\right)-\hat{R}'\left(h_w\right)<\frac{\varepsilon}{2}\right)$$

$$\leqslant P'\left(\hat{R}'\left(h_w\right)-\hat{R}\left(h_w\right)>\frac{\varepsilon}{2}\right) \rightarrow \boxed{1}$$

By Hoeffding's Ineq

$$P'\left(R\left(h_w\right)-\hat{R}'\left(h_w\right)<\frac{\varepsilon}{2}\right) \geqslant 1-e^{-2n\left(\frac{\varepsilon}{2}\right)^2} \overset{\text{A1}}{\geqslant} \frac{1}{2}$$

$$\boxed{1} \Rightarrow \mathbb{1}_{\left[R(h_w)-\hat{R}(h_w)>\varepsilon\right]} \leqslant 2P'\left(\hat{R}'\left(h_w\right)-\hat{R}\left(h_w\right)>\frac{\varepsilon}{2}\right)$$

$$\overset{h_w \text{ def}}{\Rightarrow} \quad \mathbb{1}_{\left[\sup_{h\in\mathcal{H}} R(h)-\hat{R}(h)>\varepsilon\right]} \leqslant \text{ RHS above } \overset{\text{implication}}{\leqslant} \quad 2P'\left(\sup_{h\in\mathcal{H}} \hat{R}'(h)-\hat{R}(h)>\frac{\varepsilon}{2}\right)$$

$$\overset{\text{sym } S \text{ and } S'}{=} 2P\left(\sup_{h\in\mathcal{H}} \hat{R}'(h)-\hat{R}(h)>\frac{\varepsilon}{2}\right)$$

$$\overset{\text{expectation over } S}{\Rightarrow} P\left(\sup_{h\in\mathcal{H}} R(h)-\hat{R}(h)>\varepsilon\right) \leqslant 2P\left(\sup_{h\in\mathcal{H}} \hat{R}'(h)-\hat{R}(h)>\frac{\varepsilon}{2}\right)$$

9 Lecture 9

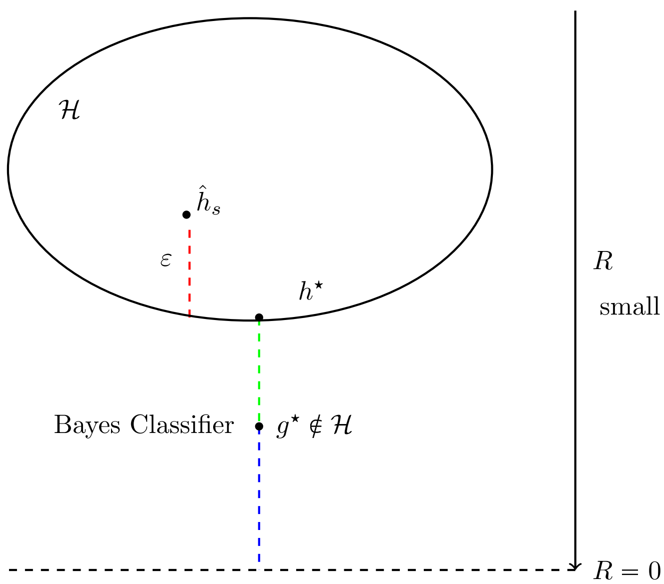
$$\mathbb{P}_{S\sim P^n}\left(\sup_{h\in\mathcal{H}} R(h)-\hat{R}_s(h) \leqslant \varepsilon\right) \geqslant 1-\delta$$

where $\varepsilon \sim O\left(\sqrt{\frac{VC(\mathcal{H}) + \log\frac{1}{\delta}}{n}}\right), \boxed{\varepsilon(n, \delta)}$

$$\mathbb{P}_S\left(R\left(\hat{h}_s^{\text{ERM}}\right) - R(h^\star) \leqslant \varepsilon\right) \geqslant 1 - \delta$$

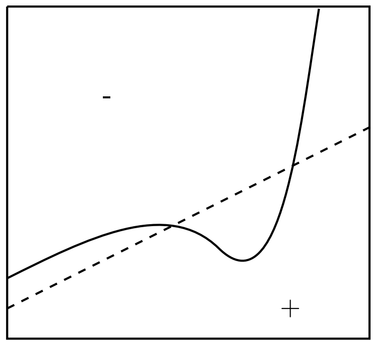
where $h^\star \in \arg\inf_{h \in \mathcal{H}} R(h)$

$$R(h^s) = \mathbb{E}_{(x,y)\sim P} \ell\left(\hat{h}_s(x), y\right)$$



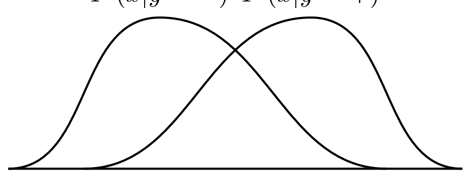
(red) Estimation error: $R\left(\hat{h}_s\right) - R(h^\star)$
(green) Apprimation error: $R(h^\star) - R\left(g^{\text{Bayes}}\right)$
(blue) Bayes error: $R\left(g^{\text{Bayes}}\right)$

$$P_X = \text{Unif } [0,1]^2$$
$$P_{XY}$$



$\mathcal{H}$ linear

$$
\begin{aligned}
&P_{Y|X} \\
&1 > P\left(y=+|x \in + \text{ region }\right) > \frac{1}{2} \\
&\frac{1}{2} > P\left(y=+|x \in - \text{ region }\right) > 0
\end{aligned}
$$



$$\mathcal{H} = \{\text{sign}\left[\sin\left(\alpha x\right)\right] : \alpha \in [1,2]\}$$

$$g^{\text{ Bayes }} \in \arg\max_{y \in Y} P\left(y|x\right)$$

Consider $\mathcal{H}_1, \mathcal{H}_2, ...$

$$VC\left(\mathcal{H}_i\right) < \infty,\ \forall\ i \in N$$

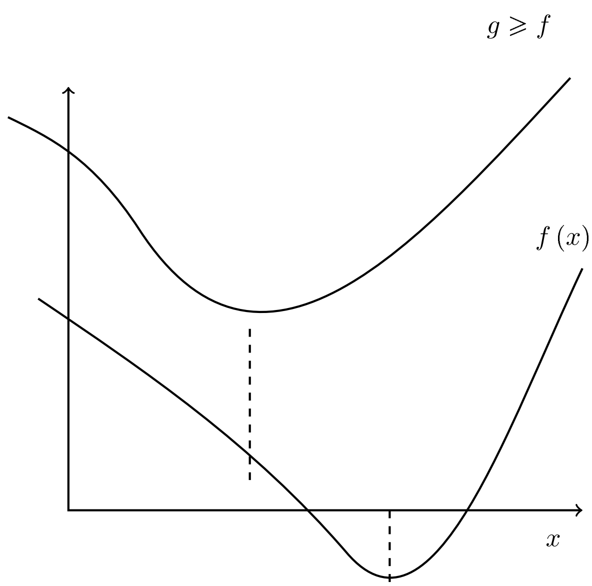
$$
\begin{aligned}
&\mathbb{P}\left(\sup_{h \in \mathcal{H}} R\left(h\right) - \hat{R}_s\left(h\right) > \varepsilon\left(n, \delta\right)\right) \leqslant \delta \\
&\Rightarrow \mathbb{P}\left(\sup_{h \in \mathcal{H}_i} R\left(h\right) - \hat{R}_s\left(h\right) > \varepsilon\left(n, w_i\delta\right)\right) \leqslant w_i\delta,\ \forall\ i \\
&\mathbb{P}\left(\ \forall\ i, \sup_{h \in \mathcal{H}_i} R\left(h\right) - \hat{R}_s\left(h\right) > \varepsilon\left(n, w_i\delta\right)\right) \leqslant \left(\sum_{i=1}^{\infty} w_i\right)\delta,\ \forall\ i
\end{aligned}
$$

where $\sum_{i=1}^{\infty} w_i \leqslant 1, w_i > 0$

$$R(h) - \hat{R}_s(h) \leqslant \varepsilon_i(n, w_i\delta)$$
$$R(h) \leqslant \hat{R}_s(h) + \varepsilon_i(n, w_i\delta)$$

"Alg 1"

$$\hat{h}_s \in \arg\inf_{h \in \cup_{i=1}^{\infty}\mathcal{H}_i} \hat{R}_s(h) + \min_{i: h \in \mathcal{H}_i} \varepsilon_i(n, w_i\delta)$$



"Alg 2" $\leftarrow$ Stuctural Risk Minimization

$$i^\star(h) = \left[\arg\min_{i=1,2,\ldots} h \in \mathcal{H}_i\right], \hat{h}_s \in \arg\inf_h \hat{R}_s(h) + \varepsilon_{i^\star(h)}\left(n, w_{i^\star(h)}\delta\right)$$

$$\min_\theta \sum_{i=1}^{n} \ell(\theta, x_i y_i) + \frac{\lambda}{2}\|\theta\|^2$$
$$\min_\theta \sum_{i=1}^{n} \ell(\theta, x_i y_i) + \frac{\lambda}{2}\|\theta - 86\|^2$$

10 Lecture 10

Structural Risk Minimization

Input: $\mathcal{H}_1, \mathcal{H}_2, \ldots$ each with finite VC

$$w_1, w_2, \ldots, \text{ such that } \sum_{i=1}^{\infty} w_i \leqslant 1, w_i \geqslant 0$$

$$\varepsilon_i\left(N, w_i \delta\right) \approx \sqrt{\frac{VC\left(\mathcal{H}_i\right) + \log \frac{1}{w_i \delta}}{N}}$$

$$\mathbb{P}_s\left(\sup_{h \in \mathcal{H}_i} R(h) - \hat{R}(h) > \varepsilon_i\left(N, w_i \delta\right)\right) < w_i \delta$$

Union bound over $i = 1, 2, \ldots$

$$\mathbb{P}_s\left(\exists i, \sup_{h \in \mathcal{H}_i} R(h) - \hat{R}(h) > \varepsilon_i\left(N, w_i \delta\right)\right) < \left(\sum_{i=1}^{\infty} w_i\right) \delta$$

$$\mathbb{P}_s\left(\forall\ i, \sup_{h \in \mathcal{H}_i} R(h) - \hat{R}(h) \leqslant \varepsilon_i\left(N, w_i \delta\right)\right) \geqslant 1 - \left(\sum_{i=1}^{\infty} w_i\right) \delta \overset{\left(\sum_i w_i \delta \leqslant 1\right)}{\geqslant} 1 - \delta$$

For each $\mathcal{H}_i$

$$\forall\ \varepsilon, \delta, P, \exists N,\ \forall\ n > N$$

$$\mathbb{P}_{s \sim P^n}\left(\sup_{h \in \mathcal{H}_i} R(h) - \hat{R}_s(h) > \varepsilon\right) < \delta$$

$$N \approx \frac{VC\left(\mathcal{H}_i\right) + \log \frac{1}{\delta}}{\varepsilon^2}$$

$$\varepsilon_i(w, \delta) \approx \sqrt{\frac{VC\left(\mathcal{H}_i\right) + \log \frac{1}{\delta}}{N}}$$

SRM:

$$\hat{h}^{\text{ SRM}} \in \arg \min_{i, h \in \mathcal{H}_i} \hat{R}(h) + \varepsilon_i\left(N, w_i \delta\right)$$

Special Case of SRM

Assumption: $\mathcal{H}$ countable

$$\begin{aligned}
\mathcal{H} &= \{h_1, h_2, \ldots\} \\
&= \{h_1\} \cup \{h_2\} \cup \ldots \\
\mathcal{H}_1 &= \{h_1\}, \mathcal{H}_2 = \{h_2\} \ldots
\end{aligned}$$

Need w_i for $\mathcal{H}_i$, equiv denote as $w_h, h \in \mathcal{H}$

$$\sum_{h \in \mathcal{H}} w_i \leqslant 1, w_h \geqslant 0 \; \forall \; h \in \mathcal{H}$$

SRM

$$\hat{h}^{\text{ SRM}} \in \arg\min_{h \in \mathcal{H}} \hat{R}(h) + \varepsilon_i(N, w_i \delta) = \arg\min_{h \in \mathcal{H}} \hat{R}(h) + \sqrt{\frac{\log \frac{1}{w_i} + \log \frac{2}{\delta}}{2N}}$$

$\mathcal{H}_i$ singleton $\Rightarrow$ Hoeffding's

$$\varepsilon_i = \sqrt{\frac{\log \frac{2}{w_i \delta}}{2N}}$$

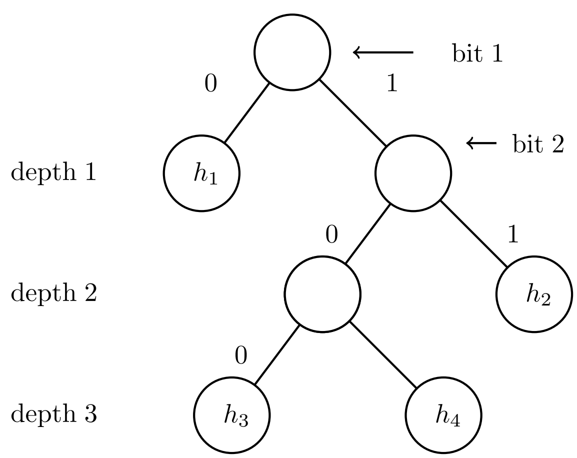
$\boxed{\text{Special}^{2}\text{ case}}$ prefix binary code

~~1 h_1~~

~~10 h_2~~

~~11 h_3~~

~~101 h_4~~



$$
\begin{array}{lll}
0 & h_1 & \\
11 & h_2 & \\
100 & h_3 & \\
101 & h_4 &
\end{array}
\Rightarrow
\begin{array}{l}
w_1 = \dfrac{1}{2} \\
w_2 = \dfrac{1}{4} \\
w_3 = \dfrac{1}{8} \\
w_4 = \dfrac{1}{8}
\end{array}
$$

$$w_h : = \frac{1}{2^{\text{ depth } (h)}}$$

Kreft's Thm

$$\sum_{h \in \mathcal{H}} w_h \leqslant 1$$

$$\arg\min \hat{R}(h) + \sqrt{\frac{\text{depth } (h) \log 2 + \log \frac{1}{\delta}}{2N}}$$

minimum description length (MDL)

$\Rightarrow$ Occam's Razor

PAC-Bayes bound

Set prior P over $\mathcal{H}$

For any $Q = A(S)$ over $\mathcal{H}$ that learner produces (i.e. equiv of $\hat{h}^{\text{ ERM}}$)

$$R(Q) : = \mathbb{E}_{(x,y) \sim P_{\text{dist}}} \mathbb{E}_{h \sim Q} \ell(h(x), y)$$

11 Lecture 11

PAC-Bayes bound

Prior P over $\mathcal{H}$, loss $\ell \in [0, 1]$

Risk $R(h) = \mathbb{E}_{P_{XY}} \ell(h(x), y), \hat{R}_s(h) = \dfrac{1}{|S| = n} \displaystyle\sum_{i=1}^{n} \ell(h(x_i), y_i)$

"Gibbs classifier" $h \sim P$

$$R(P) = \mathbb{E}_{h \sim P} R(h), \hat{R}_s(P) = \mathbb{E}_{h \sim P} \hat{R}_s(h)$$

Sample $S \sim P_{XY}^n, A(S) := Q$ distribution over $\mathcal{H}$

Theorem 2. *P given,* $\forall\, P_{XY}, \ell \in [0,1], \varepsilon, \delta,$

$$\mathbb{P}_{S\sim P_{XY}^n}\left(\forall\, Q, R(Q) \leqslant \hat{R}_s(Q) + \sqrt{\frac{KL(Q\|P) + \log\frac{n}{\delta}}{2(n-1)}}\right) \geqslant 1-\delta$$

$$KL(Q\|P) := \sum_{h\in\mathcal{H}} Q(h)\log\frac{Q(h)}{P(h)}$$

Let Q_1 *concentrated on* $\hat{h}_s^{ERM}, Q_1$ *minimizes* $\hat{R}_s(Q)$

$$KL(Q_1\|P) = Q\left(\hat{h}_s^{ERM}\right)\log\frac{Q\left(\hat{h}_s^{ERM}\right)}{P\left(\hat{h}_s^{ERM}\right)} = \log\frac{1}{P\left(\hat{h}_s^{ERM}\right)}$$

$$\lim_{x\to 0} x\log x = \lim_{x\to 0}\frac{(\log x)'}{\left(\frac{1}{x}\right)'} = \frac{\frac{1}{x}}{-\frac{1}{x^2}} = -x = 0$$

(read textbook for proof)

11.1 Rademacher Complexity

$$\mathcal{F} := \ell\circ\mathcal{H} = \{f(\cdot,\cdot) := \ell(h(\cdot),\cdot), h\in\mathcal{H}\}$$

$$R(f) = \mathbb{E}_{(x,y)\sim P_{XY}} f(x,y), \hat{R}_s(f) = \frac{1}{n}\sum_{(x,y)\in S} f(x,y)$$

Definition 1. Rademacher Complexity of $\mathcal{F}$

$$\mathbb{R}(\mathcal{F}\circ S) := \frac{1}{n}\mathbb{E}_{\vec{\sigma}}\sup_{f\in\mathcal{H}}\vec{\sigma}^T\vec{f}_s$$

Depends on S, (alos n)

$$[\sigma_1, ..., \sigma_n] := \vec{\sigma}, \sigma_i \in \{-1, 1\}, P(\sigma_i = 1) = \frac{1}{2}\ \forall\ i$$

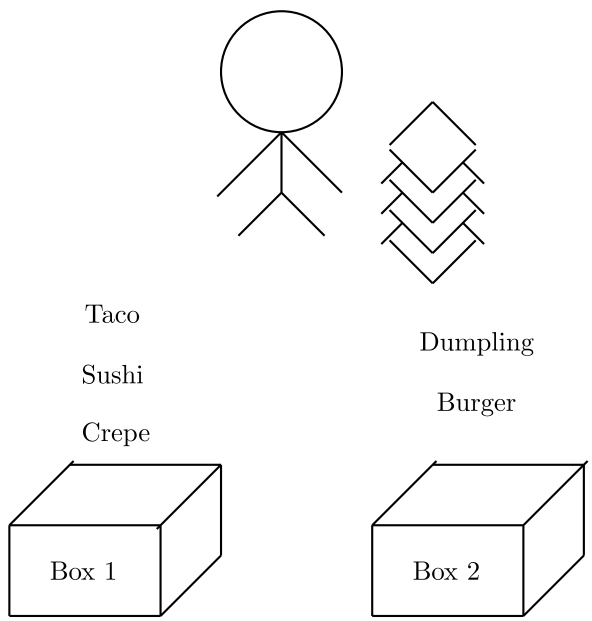
$$\begin{bmatrix} f(x_1,y_1) \\ ... \\ f(x_n,y_n)\end{bmatrix} (x_i,y_i)\in S, f\in\mathcal{F}$$

Theorem 3. 26.5 (3) *loss bound,* $|\ell(\cdot)| \leqslant c$

$$\forall\ h^\star \in \mathcal{H},\ \ wp\ \geqslant 1-\delta, R\left(\hat{h}_s^{ERM}\right) \leqslant R\left(h^\star\right) + 2\mathbb{R}\left(\mathcal{F}\circ S\right) + 5c\sqrt{\frac{2\log\frac{8}{\delta}}{n}}$$

data dependent bound

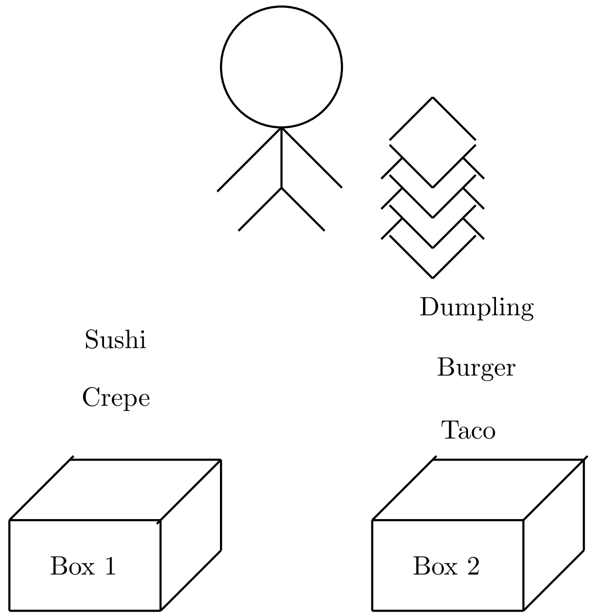
12 Lecture 12

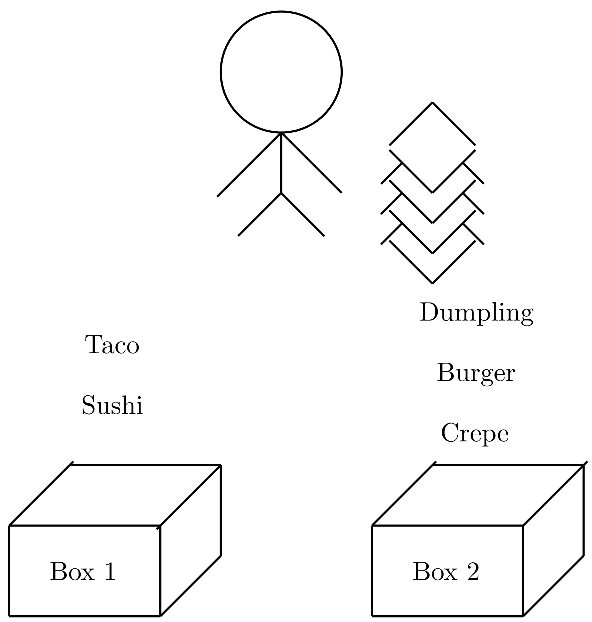


X = vocabulary

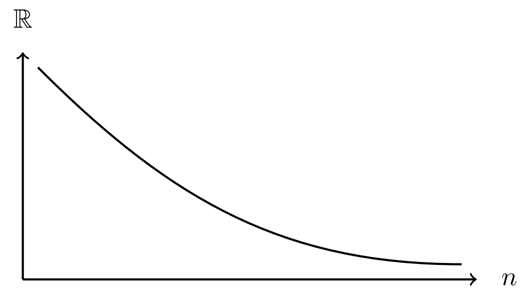
$$h \in \underbrace{\mathcal{H}}_{\text{all rules in your mind}}, X \to \{\text{ Box1 , Box2 }\}$$

$$\begin{aligned}\frac{1}{m} &= \sum_{j=1}^{m} \sup_{h\in\mathcal{H}} \sum_{i=1}^{5} \sigma_i^{(j)} h\left(x_i\right) \\ &\approx \mathbb{E}_{\sigma_1\ldots\sigma_5} \sup_{h\in\mathcal{H}} \vec{\sigma}^T \vec{h}\end{aligned}$$





$\mathbb{R}(\mathcal{H})$



$$A \subset \mathbb{R}^n$$
$$\mathbb{R}\left(A\right) = \frac{1}{n}\mathbb{E}_{\sigma_1 \ldots \sigma_n} \sup_{a \in A} \vec{\sigma}^T a$$

Lemma 26.2, $F = \ell \circ \mathcal{H}$

$$\mathbb{E}_s\left[\sup_{f \in \mathcal{F}} R\left(f\right) - \hat{R}_s\left(f\right)\right] \leqslant 2\mathbb{E}_s \mathbb{R}\left(F \circ S\right)'$$

Proof:

$$\begin{aligned}\sup_{f \in F} R\left(f\right) - \hat{R}_s\left(f\right) &\overset{\text{ghost } s'}{=} \sup_{f \in F} \mathbb{E}_{s'}\left[\hat{R}_{s'}\left(f\right) - \hat{R}_s\left(f\right)\right] \\ &\overset{\text{Jensen's}}{\leqslant} \mathbb{E}_{s'} \sup_{f \in F}\left[\hat{R}_{s'}\left(f\right) - \hat{R}_s\left(f\right)\right]\end{aligned}$$

$$\mathbb{E}_s\left[\sup Rf - \hat{R}_s f\right] \leqslant \mathbb{E}_{s,s'} \sup_f \left[\hat{R}_{s'} f - \hat{R}_s f\right] \tag{1}$$

$$\begin{aligned}S &= \left\{z_1, ..., z_n\right\} \\ S' &= \left\{z_1', ..., z_n'\right\} \\ \text{(1) RHS } &= \frac{1}{n}\mathbb{E}_{s',s} \sup_f \sum_{i=1}^n \left(f\left(z_i'\right) - f\left(z_i\right)\right) \overset{\text{introduce } \sigma = \left\{\sigma_1, \ldots, \sigma_n\right\}}{=} \frac{1}{n}\mathbb{E}_{s',s,\sigma} \sup_f \sum_{i=1}^n \sigma_i \left(f\left(z_i'\right) - f\left(z_i\right)\right) \\ &\leqslant \frac{1}{n}\mathbb{E}_{s',s,\sigma}\left\{\left[\sup_{f \in F}\sum_i^n \sigma_i f\left(z_i'\right)\right] + \left[\sup_{g \in F}\sum_i^n \sigma_i\left(-g\left(z_i\right)\right)\right]\right\} \\ &= \mathbb{E}_s \mathbb{R}\left(F \circ S\right) + \mathbb{E}_s \mathbb{R}\left(F \circ S\right) = 2\mathbb{E}_s \mathbb{R}\left(F \circ S\right) \\ &\mathbb{E}_{s',s} \sup_f f\left(z_1'\right) - f\left(z_1\right) + \sum_{i=2}^n \left(f\left(z_i'\right) - f\left(z_i\right)\right) \\ &\overset{s,s' \text{ iid}}{=} \mathbb{E}_{s',s} \sup_f f\left(z_1\right) - f\left(z_1'\right) + \sum_{i=2}^n \left(f\left(z_i'\right) - f\left(z_i\right)\right) \\ &= \mathbb{E}_{s',s,\sigma_1} \sup_f \sigma_1\left(f\left(z_1'\right) - f\left(z_1\right)\right) + \sum_{i=2}^n \left(f\left(z_i'\right) - f\left(z_i\right)\right)\end{aligned}$$

$$\mathbb{P}_s\left(\overbrace{\sup_{f\in F} R(f)-\hat{R}_s(f)}^{X}\geqslant\varepsilon\right)\leqslant\frac{\mathbb{E}[\sup]}{\varepsilon}\leqslant\frac{2\mathbb{E}_s\mathbb{R}(F\circ S)}{\varepsilon}:=\delta$$

$$\Rightarrow\varepsilon=\frac{2\mathbb{E}_s\mathbb{R}(F\circ S)}{\delta}$$

Markov ineq r.v. $\boxed{X\geqslant 0}$ $\forall\ a>0$

$$P(X>a)\leqslant\frac{\mathbb{E}[X]}{a}$$

$$\boxed{\text{VC style } \varepsilon,\varepsilon=\sqrt{\frac{VC(\mathcal{H})+\log\frac{1}{\delta}}{n}}}$$

McDiarmid's Ineq, $f:V^n\to\mathbb{R},V\subseteq\mathbb{R}$
foral $i\in[n]$, $\forall\ x_i,x_i'\in V,|f(x_1,x_2,...,x_i,...,x_n)-f(x_1,x_2,...,x_i',...,x_n)|\leqslant c$
Lecture $X_1,...,X_n$ be r.v. in V

$$\mathbb{P}\left(|f(X_1,...,X_n)-\mathbb{E}f(X_1,...,X_n)|\leqslant c\sqrt{\left(\log\frac{2}{\delta}\right)\frac{n}{2}}\right)\geqslant 1-\delta$$

13 Lecture 13

$$\mathbb{P}_s\left(\sup_{f\in F} R(f)-\hat{R}_s(f)>\varepsilon\right)$$

Lemma 26.2

$$\mathbb{E}_s\left[\sup_{f\in F} R(f)-\hat{R}_s(f)\right]\leqslant 2\mathbb{E}_s\mathbb{R}(F\circ S)$$

Assumption: $|\ell|\leqslant c$
McDiarmid's Ineq: For f which satisfies $|f(x_1...x_i...x_n)-f(x_1...x_i'...x_n)|\leqslant c,\ \forall\ x,i,x_i'$

$$\text{wp }\geqslant 1-\delta,|f(X_1...X_n)-\mathbb{E}f(X_1...X_n)|\leqslant C_0\sqrt{\frac{n\log\frac{2}{\delta}}{2}}$$

$$\sup_{f\in F} R(f)-\hat{R}_s(f)=\sup_{h\in\mathcal{H}}\mathbb{E}_{(x,y)\sim P_{XY}}\ell(h(x),y)-\frac{1}{n}\sum_{i=1}^{n}\ell(h(x_i),y_i)$$
$$s=(x_1,y_1),...,(x_n,y_n)$$
$$\Rightarrow C_0=\frac{2C}{n}$$

$$\overset{\text{McDiarmid}}{\Rightarrow} \quad \text{wp} \geqslant 1-\delta, \left| \left[\sup_{f \in F} R(f) - \hat{R}_s(f) \right] - \mathbb{E}_{s'} \left[\sup_{f \in F} R(f) - \hat{R}_{s'}(f) \right] \right| \leqslant \frac{2c}{n} \sqrt{\frac{n \log \frac{2}{\delta}}{2}} = c\sqrt{\frac{2 \log \frac{2}{\delta}}{n}}$$

$$\forall\ x, i, x_i' \rightarrow \ \text{wp} \geqslant 1-\delta, \sup_{f \in F} R(f) - \hat{R}_s(f) \leqslant \mathbb{E}_{s'} \left[\sup_{f \in F} R(f) - \hat{R}_{s'}(f) \right] + c\sqrt{\frac{2 \log \frac{2}{\delta}}{n}}$$

$$\overset{\text{Lemma 26.2}}{\leqslant} \quad 2\mathbb{E}_{s'} \mathbb{R}\left(F \circ S'\right) + c\sqrt{\frac{2 \log \frac{2}{\delta}}{n}}$$

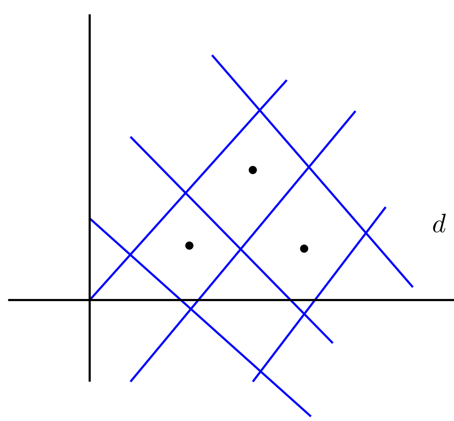
Thm 26.5 (1)

$$\boxed{\text{ERM } \arg\min_{h \in \mathcal{H}} \hat{R}_s(h)}$$

$$\mathcal{H} = \left\{ h_{w,b} : X \rightarrow Y, w \in \mathbb{R}^d, b \in \mathbb{R}, h_{w,b}(x) = \text{sign}\left(w^T x + b\right) \right\}$$
$$X \subseteq \mathbb{R}^d, Y = \{-1, 1\}$$

$$\text{sign}(z) = \begin{cases} 1, & \text{if } z \geqslant 0 \\ -1, & \text{if } z < 0 \end{cases}$$

$$\left\{ x : w^T x + b = 0 \right\}$$
$$VC(\mathcal{H}) = d + 1$$



$d = 2$

$\ell = 0 - 1$ loss

Assume: $\exists h^\star \in \mathcal{H},\ \forall\ (x, y) \in S, y = h^\star(x)$

(Batch) Perceptron Alg

Init, $t = 0, w_0 = \vec{0} \in \mathbb{R}^d$

pick $(x, y) \in S$

if sign $\left(w_t^T x\right) \neq y$

$$w_t = w_t + yx$$

Terminates?!
How many mistakes were made? $\neq \hat{R}_s\left(h\right)$

$$C := \min_{i \in |S|} y_i \left(w^\star\right)^T x_i$$
$$C > 0$$
$$w^{\star\star} := \frac{w^\star}{c}$$
$$\forall\ i, y_i \left(w^{\star\star}\right)^T x_i \geqslant 1, \boxed{1}$$
$$\boxed{B := \min \|w\|, w \in \left\{w :\ \forall\ i \in S : y_i w^T x_i \geqslant 1\right\}}$$
$$\boxed{w^\star \in \arg\min}$$

$$\frac{w_t^T w^\star}{\|w_t\| \, \|w^\star\|} \overset{\cos}{\leqslant} 1$$

$$w_{t+1}^T w^\star - w_t^T w^\star = \left(w_{t+1} - w_t\right)^T w \cdot = \left(y_t x_t\right)^T w^\star \overset{\boxed{1}}{\geqslant} 1$$
$$\sum_{t=0}^{T} \left(w_{t+1}^T w^\star - w_t^T w^\star\right) = w_{T+1}^T w^\star - \underbrace{w_0^T w^\star}_{0} \geqslant T, \boxed{2}$$

14 Lecture 14

Batch Perceptron (Ch. 9)
Give $\left(x_i, y_i\right)_{i=1:n}, Y = \{-1, 1\}$
Assume realizability $\exists w \in \mathbb{R}^d\ \forall\ i \in [n], y_i w^T x_i > 0$

$$\Rightarrow \exists w\ \forall\ - i, y_i w^T x_i \geqslant 1$$
$$\bar{W} := \left\{w :\ \forall\ i, y_i w^T x_i \geqslant 1\right\}$$
$$w^\star \in \arg\min_{w \in \bar{W}} \|w\|$$
$$B := \|w^\star\|$$

Alg: $w_0 = \varnothing$
if $y_i w_t^T x_t \leqslant 0$ (misclassification)
$w_{t+1} = w_t + y_t x_t$ (repeat)

Claim: Alg makes bounded number of mistakes on any sequence x_t, y_t (in training set)

Proof:

$$\cos\left(w_{t+1}, w^{\star}\right) = \frac{w_{t+1}^{T} w^{\star}}{\|w_{t+1}\| \cdot \|w^{\star}\|} \leqslant 1$$

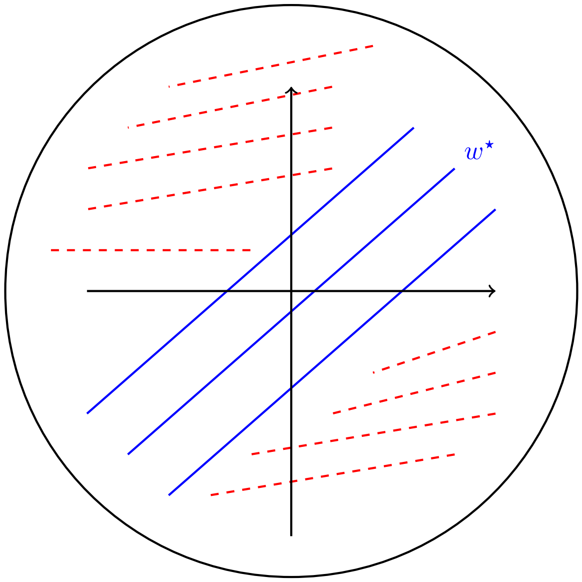
step 1 : $w_{t+1}^{T} w^{\star}$ grows $O\left(t\right)$

$$\begin{aligned}
w_{t+1}^{T} w^{\star} &= \left(w_t + y_t x_t\right)^{T} w^{\star} = w_t^{T} w^{\star} + y_t \underbrace{x_t^{T} w^{\star}}_{\geqslant 1\left(def w^{\star} \in \bar{W}\right)} \\
&\Rightarrow w_{t+1}^{T} w^{\star} - w_t^{T} w^{\star} \geqslant 1 \\
&\Rightarrow \sum_{t=1}^{T-1} \left(w_{t+1}^{T} w^{\star} - w_t^{T} w^{\star}\right) \geqslant T \\
&\Rightarrow w_T^{T} w^{\star} - w_0^{T} w^{\star} \geqslant T \\
&\underbrace{\Rightarrow}_{w_0 = \varnothing} w_T^{T} w^{\star} \geqslant T
\end{aligned}$$

step 2 : $\|w_{t+1}\| \sim o\left(t\right)$

$$\begin{aligned}
\|w_{t+1}\|^2 &= \|w_t + y_t x_t\|^2 = \left(w_t + y_t x_t\right)^{T} \left(w_t + y_t x_t\right) \\
&= w_t^{T} w_t + 2 y_t w_t^{T} x_t + y_t^2 x_t^{T} x_t \\
&\underbrace{=}_{Y \in \{-1, 1\}} \|w_t\|^2 + 2 \underbrace{y_t w_t^{T} x_t}_{\text{misclassification } \leqslant 0} + \|x_t\|^2 \\
&\leqslant \|w_t\|^2 + \|x_t\|^2 \\
&\Rightarrow \|w_{t+1}\|^2 - \|w_t\|^2 \leqslant \|x_t\|^2 \leqslant R^2 \left(\text{ Assumption 2, } \forall\, i \in [n], R \geqslant \|x_t\|\right) \\
&\Rightarrow \|w_T\|^2 \leqslant T R^2 \\
&\Rightarrow \|w_T\| \leqslant \sqrt{T} R
\end{aligned}$$

$$\begin{aligned}
1 \geqslant \frac{w_{t+1}^{T} w^{\star}}{\|w_{t+1}\| \cdot \|w^{\star}\|} &\underbrace{\geqslant}_{\text{steps } 1,2} \frac{T}{\sqrt{T} \cdot R \cdot B} = \frac{\sqrt{T}}{RB} \\
\sqrt{T} &\leqslant RB \\
T &\leqslant R^2 B^2
\end{aligned}$$



$$S = \{(x_1, y_1), ..., (x_n, y_n)\}$$
$$s \in \bigcup_{T=1}^{\infty} S^T$$

$$X, Y \in \{-1, 1\}$$
$$s \in \bigcup_{T=1}^{\infty} (X \times Y)^T$$

Interaction protocol:

$\boxed{0}$ world picks s

At time t :

$\boxed{1}$ wrold shows x_t in S

$\boxed{2}$ learner predict $\hat{y}_t \in Y$

$\boxed{3}$ world reveals y_t in S (learner "learns, updates")

ADD Realizability assumption: $\mathcal{H}$

$$\exists h^\star \in \mathcal{H} : \ \forall \ s_x \in \bigcup_{T=1}^{\infty} X^T, \ \forall \ x_t \in s_x, y_t = h^\star(x_t)$$

Special case: $|\mathcal{H}| < \infty$

Version space

$$VS := \{h \in \mathcal{H} : h \text{ is consistent with all data so far } \}$$

15 Lecture 15

Online learning:
→ realizable assumption: mistake bound
→ randomization: regret

Given $\mathcal{H}, \exists h^\star \in \mathcal{H}$ in each iteration t: world chooses $x_t \in X$ (not necessarily iid), learner predicts $\hat{y}_t \in Y$, world rewards $y_t := h^\star(x_t)$, learner incurs error $\mathbb{1}_{(\hat{y}_t \neq y_t)}$, "learns"

Given sequence $S = x_1, ..., x_n$, alg A, define the number of mistakes A makes on S by $M_A(S)$
If $\exists A : \max_{S \in \bigcup_{n=0}^{\infty} X^n} M_A(S)$ is finite (mistake bound), then $\mathcal{H}$ is online-learnable.
Also, what's $\min_A \max_{S \in \bigcup_{n=0}^{\infty} X^n} M_A(S)$?

"Baby steps" Assume $|\mathcal{H}| < \infty$
$\exists A_{\text{consistent}}$ or $A_{\text{version space}}$, maintains a version space
init: $VS_0 = \mathcal{H}$
At time t, pick any $h \in VS_t$, predict $\hat{y}_t = h(x_t)$
Update $VS_{t+1} = \{h' \in VS_t : h'(x_t) = y_t\}$

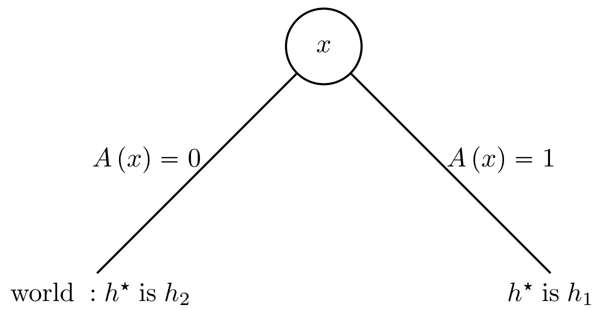
$$\begin{aligned}
X &= \{x^1, ..., x^m\} \\
h_1 &= 1, 0, ..., 0 \\
h_2 &= 0, 1, 0, ..., 0 \\
&... \\
h_m &= 0, ..., 0, 1 \\
h_{m+1} &= h^\star = 0, ..., 0 \\
M &\leqslant |\mathcal{H}| - 1
\end{aligned}$$

$$\begin{aligned}
&A_{\text{halving}} \\
&\hat{y}_t = \arg\max_{y \in \{0,1\}} |\{h \in VS : h(x_t) = y\}| \\
&1 \overset{h^\star \in VS}{\leqslant} |VS| \leqslant \frac{|\mathcal{H}|}{2^M} \Rightarrow 2^M \leqslant |\mathcal{H}|, M \leqslant \log_2 |\mathcal{H}|
\end{aligned}$$

$d := \max_{S \in \bigcup_{n=0}^{\infty} X^n} M_A(S)$ (no longer assume $|\mathcal{H}| < \infty$) (still $\exists h^\star \in \mathcal{H}$)
Fix any A. Assume $\exists x \in X$ such that $\exists h_1, h_2 \in \mathcal{H}$ such that $h_1(x) \neq h_2(x)$

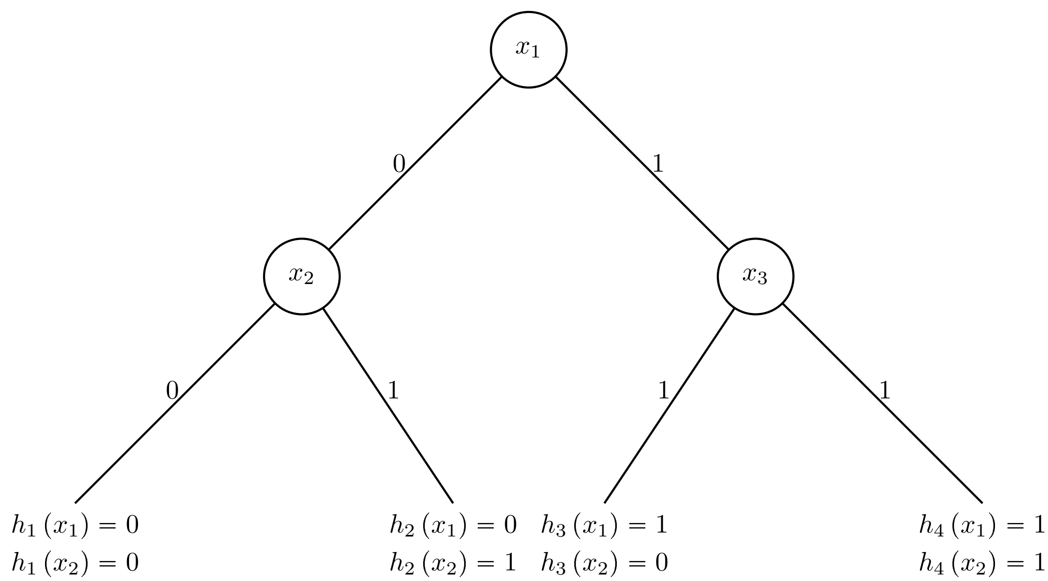
$$h_1(x) = 0$$

$$h_2(x) = 1$$



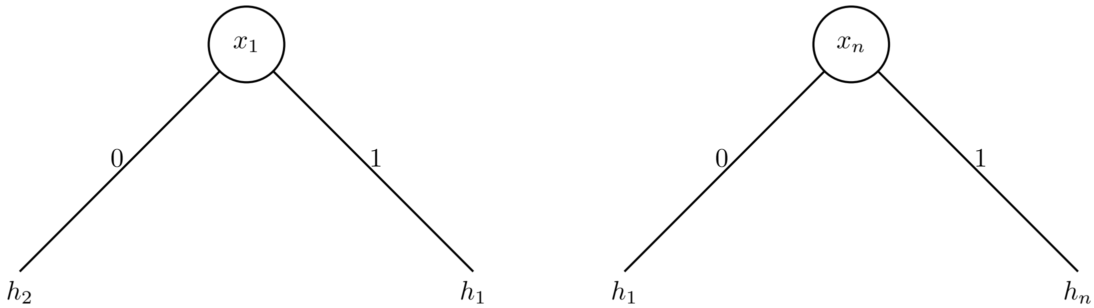
World wants

$\mathcal{H}$ such that $\exists$

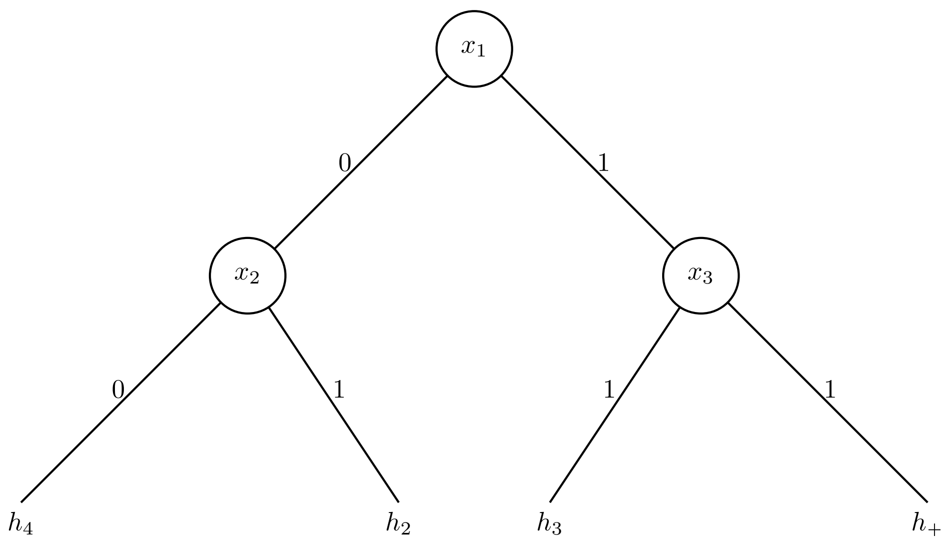


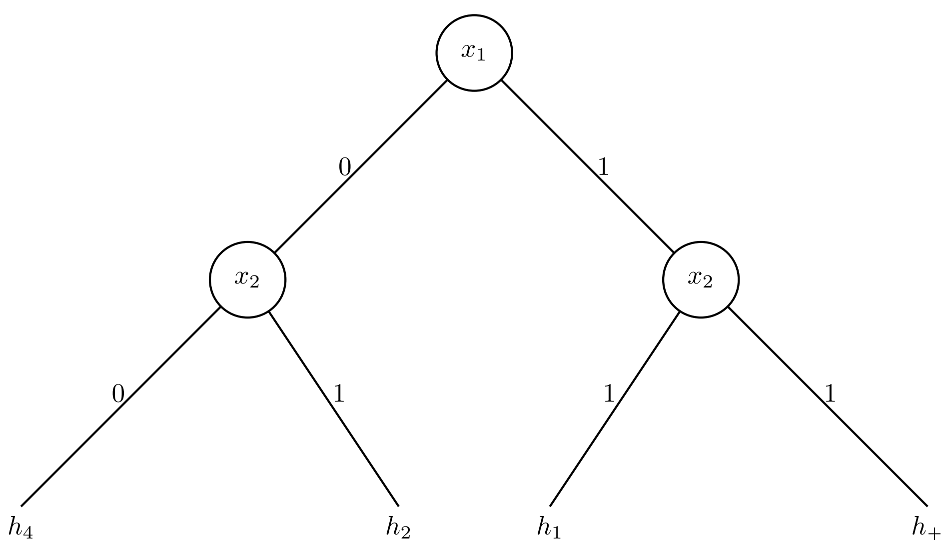
World asks $\mathcal{H}$ "is there a tree with h at the leaves?"

.	x_1	x_2	x_3	...
h_1	0	0	$\star$	
h_2	0	1	$\star$	
h_3	1	$\star$	0	
h_4	1	$\star$	1	
...				



$$\mathcal{H} = \{h_i, i \in [n]\}$$
$$h_i^\star = \mathbb{1}_{(x=x_i)}\ \forall\, i \in [n]$$
$$\mathcal{H}^2 = \mathcal{H} \cup \{h_+ : h_+(x) = 1,\ \forall\, x\}$$





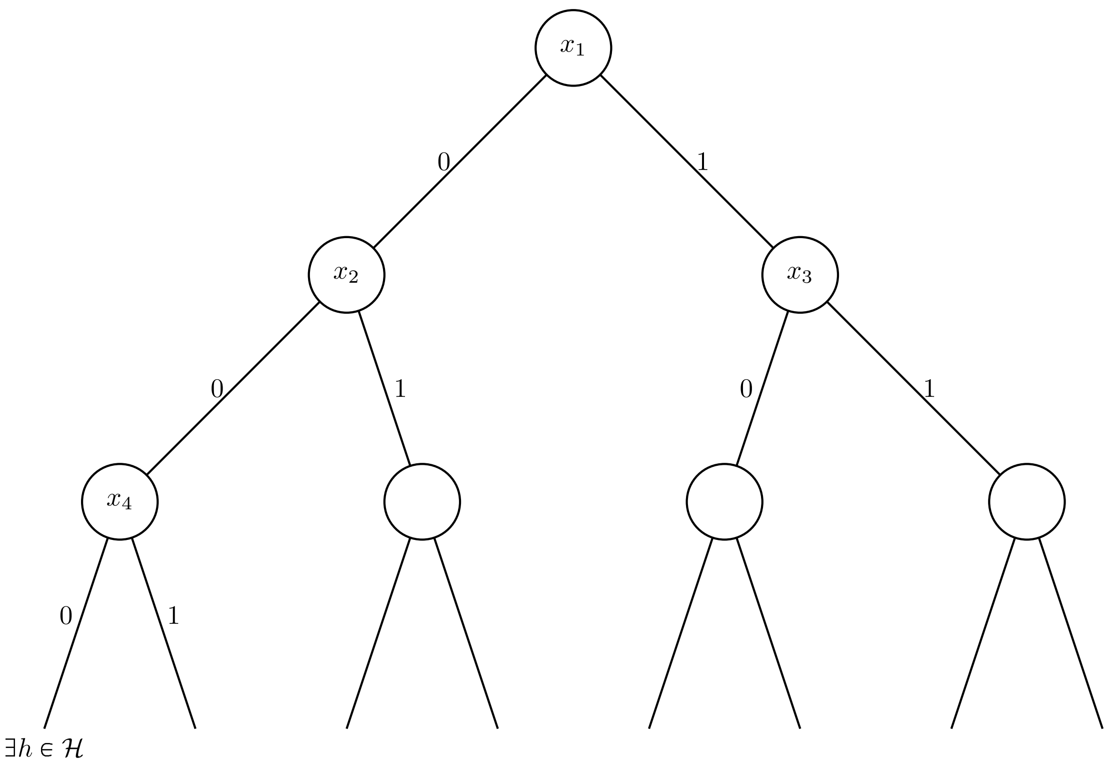
What is the deepest complete tree with h at the leaves?
depth = Littlestone dimension of $\mathcal{H}$.

16 Lecture 16

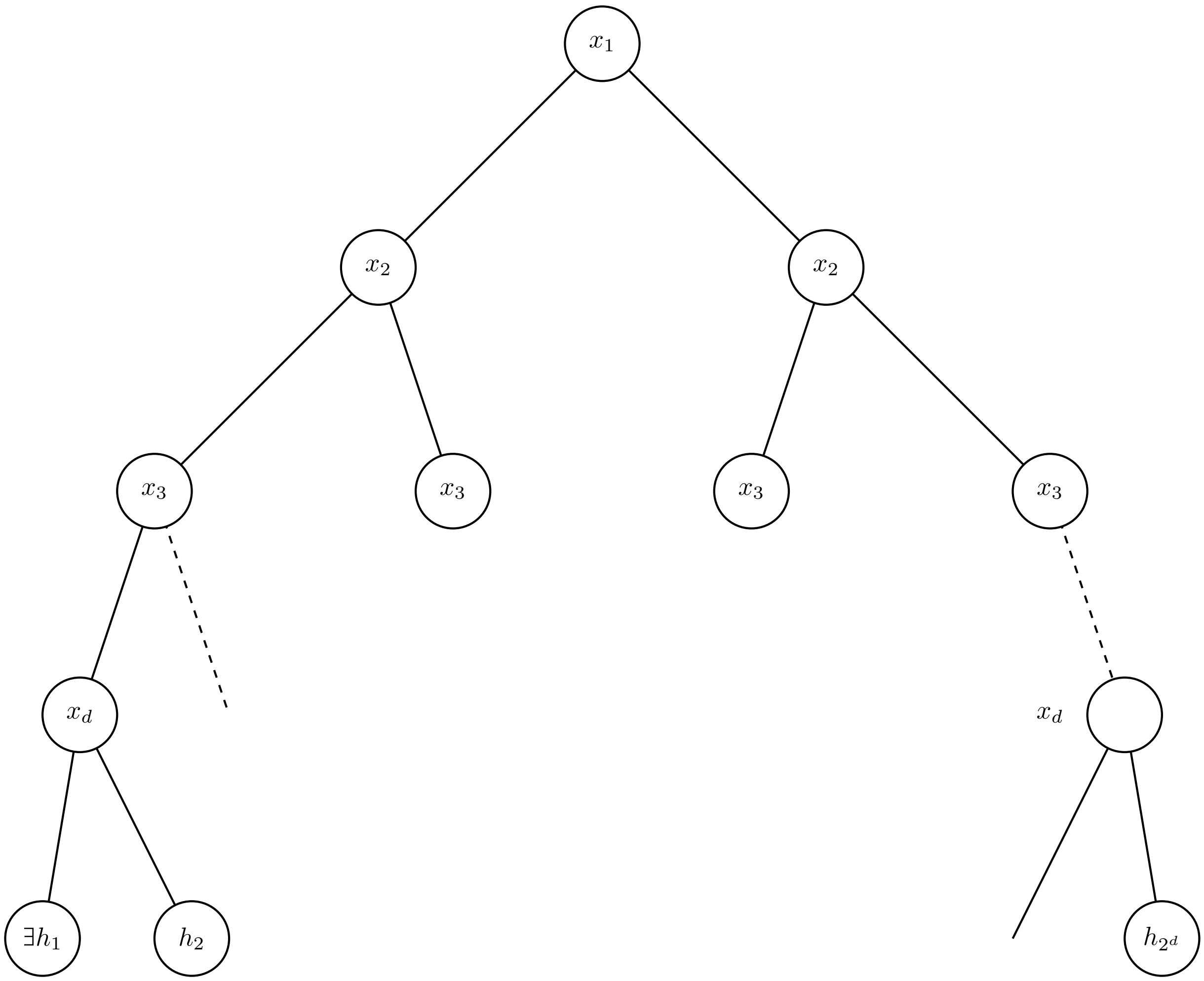
Realizable online learning

$$\max_A \max_{h^\star \in \mathcal{H}} \max_{s \in \bigcup_{n=0}^{\infty} X^n} M_A(S) \geqslant \text{ depth of "that tree" } := Ldim(\mathcal{H})$$

Given $X, \mathcal{H}$, find the deepest complete tree of the form



Ex. Suppose $VC(\mathcal{H}) = d$.
What about $Ldim(HsC)$?
$\exists x_1 ... x_d$ shattered by $\mathcal{H}$.

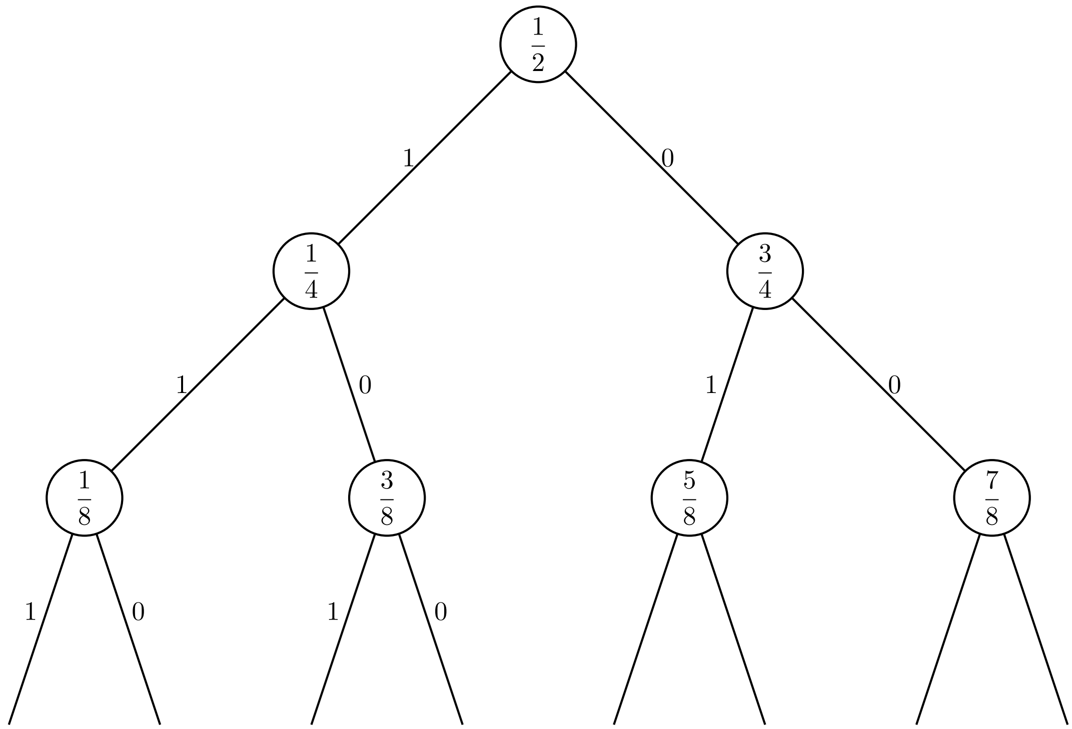


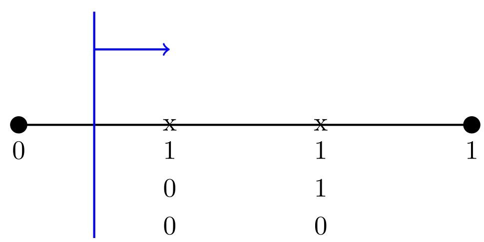
Theorem 4. $Ldim\left(\mathcal{H}\right) \geqslant VC\left(\mathcal{H}\right)$

Ex: "One-hot" $\{h_x : x \in C\} := \mathcal{H}$
$Ldim\left(\mathcal{H}\right) = 1$ possibly $|\mathcal{H}| = \infty$
Ex: $\mathcal{H} = \left\{h_a, a \in [0,1] : h_a\left(x\right) = \mathbb{1}_{(x \geqslant a)}\right\}$,

$$VC\left(\mathcal{H}\right) = 1$$

$$Ldim\left(\mathcal{H}\right) = \infty$$





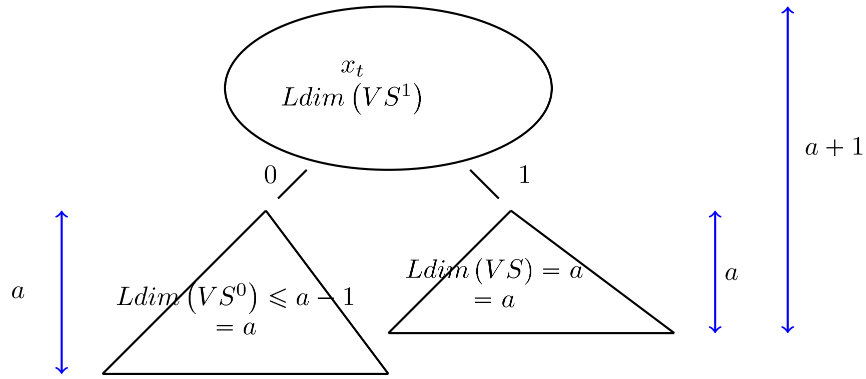
Standard Optimal Alg: A_{SOA}

receive x_t

$$VS^0 \cup VS^1, VS^p = \{h \in VS : h(x_t) = p\}$$
$$\hat{y}_t = \arg\max_{p \in Y} Ldim(VS^p)$$

receive y_t

$$VS \leftarrow VS^{y_t}$$



World: x_t
Alg: predicts: $\hat{y}_t$
World: give $y_t \in Y$

$$\text{regret } : \sum_{t=1}^{T} \ell\left(\hat{y}_t\left(x_t\right), y_t\right) - \min_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell\left(h\left(x_t\right), y_t\right) \geqslant_{\text{Show this!}} \frac{T}{2}$$

17 Lecture 17

No longer assume $\exists h^{\star} \in \mathcal{H}$ such that $y_t = h^{\star}\left(x_t\right)$
(expected) regret wrt $\mathcal{H}$ (want: $o\left(T\right)$, in fact $\sqrt{T}$)

$$\sup_{(x,y)_{1:T}} \mathbb{E} \sum_{t=1}^{T} \mathbb{1}\left(\underbrace{\hat{y}_t\left(x_t\right)}_{\text{made by } A} \neq y_t\right) - \min_{h \in \mathcal{H}} \sum_{t=1}^{T} \mathbb{1}_{\left(h\left(x_t\right) \neq y_t\right)}$$

Ex: $\mathcal{H} = \left\{h_0, h_1\right\}, A$ deterministic, regret $\geqslant \dfrac{T}{2}$
A randomized

17.1 Weighted Majority

(learn from expert advise)
Given d experts, horizon T, stepsize $\eta > 0$

1. init $\tilde{w}^{(1)} = \underbrace{(1, ..., 1)}_{d}$ (unnormalized weights)
2. For $t = 1, 2, ..., T$
3. $w^{(t)} = \dfrac{\tilde{w}^{(t)}}{Z_t}$ where $Z_t = \sum_{i=1}^{d} \tilde{w}_i^{(t)}$
4. Choose expert $i \sim w^{(t)}$

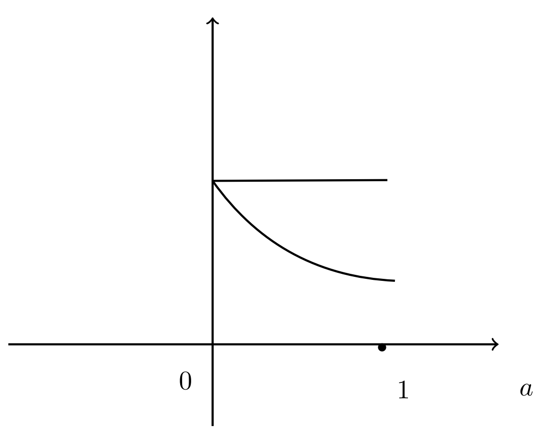
5. Observe loss vector $V_t = \left(v_{t_1}, ..., \overbrace{\boxed{v_{t_i}}}^{MAB}, ..., t_{t_d} \right) \in [0,1]^d$ "bounded", pay expected loss $w^{(t)^T} V_t$

6. update $\tilde{w}_j^{(t+1)} = \tilde{w}_j^{(t)} e^{-\eta v_{t_j}}\ \forall\ j \in [d]$

$$\left(\sum_{t=1}^{T} w^{(t)^T} V_t \right) - \min_{j \in [d]} \sum_{t=1}^{T} V_{t_j} \leqslant \sqrt{2 (\log d) T}$$

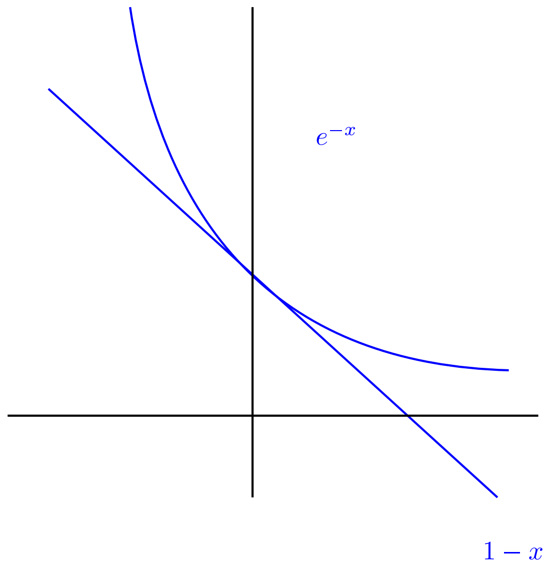
Proof:

$$\begin{aligned}
\log \frac{Z_{t+1}}{Z_t} &= \log \frac{\sum_{j=1}^{d} \tilde{w}_j^{(t)} e^{-\eta V_{t_j}}}{Z_t} \\
&= \log \sum_{j}^{d} w_j^{(t)} e^{-\eta V_{t_j}} \\
&\leqslant \log \sum_{j} w_j^{(t)} \left[1 - \eta V_{t_j} + \frac{\eta^2 V_{t_j}^2}{2} \right] \\
&= \log \left[1 - \underbrace{\sum_{j} w_j^{(t)} \left(\eta V_{t_j} - \frac{\eta^2 V_{t_j}^2}{2} \right)}_{:= b \in (0,1)} \right] \\
&\leqslant \log e^{-b} \\
&= -\eta \left(w^{(t)^T} V_t \right) + \eta^2 \sum_{j} w_j^{(t)} \frac{V_{t_j}^2}{2} \\
&\leqslant -\eta w^{(t)^T} V_t + \frac{\eta^2}{2}
\end{aligned}$$



$a \in (0,1)$

$$e^{-a} \leqslant 1 - a + \frac{a^2}{2}$$



$$1 - x \leqslant e^{-x}$$

18 Lecture 18

18.1 Online learning

(no assumption on $\exists h^\star \in \mathcal{H}, y_t = h^\star(x_t)$)

- subroutine: wighted Majority

Input: $d = \#$ experts, T rounds
init $\tilde{w}^{(1)} = (1, ..., 1)_d$
for $t = 1, 2, ...$

$$Z_t = \tilde{w}^{(t)^T} 1, w^{(t)} = \frac{\tilde{w}^{(t)}}{Z_t}$$

pick $i \sim w^{(t)}$ to predict
suffer expected loss $w^{(t)^T} V^{(t)}$

$$\tilde{w}^{(t+1)} = \tilde{w}^{(t)} e^{-\eta V_i^{(t)}} \ \forall \ i = 1 ... d$$

$$V_i^{(t)} := \ell\left(\overbrace{\text{expert}_i(x_t)}^{h_i(x_t)}, y_t\right)$$

$$V^{(t)} := \begin{bmatrix} V_1^{(t)} \\ \dots \\ V_d^{(t)} \end{bmatrix} \in [0,1]^d$$

Theorem 5. $\left(\sum_{t=1}^{T} w^{(t)^T} V^{(t)}\right) - \left(\underbrace{\min_{i\in[d]}}_{\text{"best expert"}} \sum_{t=1}^{T} V_i^{(t)}\right)$

$$\begin{aligned}
\log \frac{Z_{t+1}}{Z_t} &= \log \sum_{i}^{d} w_i^{(t)} e^{-\eta V_i^{(t)}} \\
&\overset{\left[e^{-x}\leqslant 1-x+\frac{x^2}{2}\right]}{\leqslant} \log\left[\sum_{i}^{d} w_i^{(t)}\left[1-\eta V_i^{(t)}+\frac{\eta^2 V_i^{(t)^2}}{2}\right]\right] \\
&= \log\left[1-\sum_i w_i^{(t)}\left(\eta V_i^{(t)}-\frac{\eta^2 V_i^{(t)^2}}{2}\right)\right] \\
&\overset{1-x\leqslant e^{-x}}{\leqslant} \log e^{-\sum_i w_i^{(t)}\left(\eta V_i^{(t)}-\frac{\eta^2 V_i^{(t)^2}}{2}\right)} \\
&= -\eta w^{(t)^T} V^{(t)} + \sum_i w_i^{(t)} \frac{\eta^2 V_i^{(t)^2}}{2} \\
&\overset{V\in[0,1]}{\leqslant} -\eta w^{(t)^T} V^{(t)} + \frac{\eta^2}{2} \leftarrow \boxed{1}
\end{aligned}$$

Telescope $\boxed{1}$ over t :

$$\begin{aligned}
\log Z_{T+1} - \log d &= \log Z_{T+1} - \log Z_1 \\
&= \sum_{t=1}^{T} \log \frac{Z_{t+1}}{Z_t} \\
&\overset{\boxed{1}}{\leqslant} -\eta \sum_{t}^{T} w^{(t)^T} V^{(t)} + \frac{\eta^2 T}{2} \leftarrow \boxed{2}
\end{aligned}$$

$$\begin{aligned}
Z_{T+1} &= \sum_i \tilde{w}_i^{(T+1)} \\
&= \sum_i 1 \cdot e^{-\eta \sum_{t}^{T} V_i^{(t)}},
\end{aligned}$$

$$\log Z_{T+1} = \log \sum_{i}^{d} e^{-\eta \sum_{t}^{T} V_i^{(t)}}$$

$$\geqslant \max_i \log e^{-\eta \sum_t^T V_i^{(t)}}$$

$$= \max_i -\eta \sum_t^T V_i^{(t)}$$

$$= -\eta \left(\min_i \sum_t^T V_i^{(t)} \right) \leftarrow \boxed{3}$$

"best expert"

$$\boxed{2}, \boxed{3} \Rightarrow$$

$$-\eta \left(\min_i \sum_t^T V_i^{(t)} \right) - \log d \overset{\boxed{3}}{\leqslant} \log Z_{T+1} - \log d$$

$$\leqslant -\eta \sum_t^T w^{(t)^T} V^{(t)} + \frac{\eta^2 T}{2}$$

$$\sum_t^T w^{(t)^T} V_i^{(t)} - \min_{i \in [d]} \sum_t^T V_i^{(t)} \leqslant \frac{\log d}{\eta} + \frac{\eta T}{2} \rightarrow \text{ sublinear}$$

$$\frac{\partial \text{ RHS}}{\partial \eta} = 0$$

$$\Rightarrow -\frac{\log d}{\eta^2} + \frac{T}{2} = 0$$

$$\Rightarrow \frac{1}{\eta} = \sqrt{\frac{T}{2 \log d}}$$

$$\Rightarrow \text{ RHS } = \sqrt{2T \log d}$$

If $|\mathcal{H}| < \infty$

$$\text{regret } \leqslant \sqrt{2T \log |\mathcal{H}|}$$

when $|\mathcal{H}| = \infty$?

- subroutine 2

SOA: $VS = \mathcal{H}$
for $t = 1, 2, ...$
receive x_t

$$VS^0 \cup VS^1$$

$$\hat{y}_t = \arg\max_y Ldim\,(VS^y)$$

$$VS \leftarrow VS^{y_t}$$

assuming $\exists h^{\star} \in \mathcal{H}$
s.t. $y_t = h^{\star}(x_t)\ \forall\ t$
then SOA mistakes $\leqslant Ldim(\mathcal{H})$

expert $(i_1, i_2, ..., i_L)$

$$1 \leqslant i_1 < i_2 < ...i_L \leqslant T, L \leqslant Ldim(\mathcal{H}) < \infty$$

init $VS = \mathcal{H}$
for $t = 1, 2, ..., T$
receive x_t
$VS^0 = \{h \in VS, h(x_t) = 0\}$, same for VS^1
if $t \in \{i_1, ..., i_L\}$

$$\hat{y}_t = \arg\min_y Ldim(VS^y) \leftarrow \text{ anti-SOA}$$

else

$$\hat{y}_t = \arg\max_y Ldim(VS^y) \leftarrow \text{ SOA}$$
$$VS \leftarrow VS^{\hat{y}_t}$$

Given $x_1...x_T$

$$\forall\ h \in \mathcal{H}$$
$$\Rightarrow h(x_1)...h(x_T)$$

want: $\exists L, i_1...i_L s.$t. expert $(i_1...i_L)$ produces the same predictions

Run SOA on input

$$x_1, h(x_1)$$
$$...$$
$$x_T, h(x_T)$$

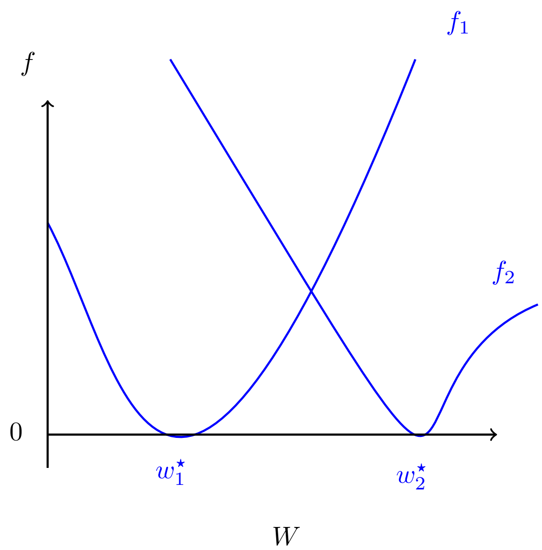
Lemma 21.13

19 Lecture 19

19.1 Online Convex Optimization

for $t = 1, 2, ...$
learner chooses $w^{(t)} \in$ Convex set W
environment chooses loss function $f_t : W \to \mathbb{R}$ convex (subgradient)
learner suffers $f_t\left(w^{(t)}\right)$

$$\text{regret } := \sum_{t=1}^{T} f_t\left(w^{(t)}\right) - \inf_{w \in W} \sum_{t=1}^{T} f_t(w)$$



Online gradient descent

$$w^{(1)} = 0$$

for $t = 1, 2, ...$
predict $w^{(t)}$

receive $f_t(\cdot)$, suffer $f_t\left(w^{(t)}\right)$

$$w^{(t+1)} = \text{Proj}_W \left[\underbrace{w^{(t)} - \underbrace{\eta}_{\text{stepsize}} \cdot \underbrace{\partial}_{\text{subgradient}} f_t\left(w^{(t)}\right)}_{W^{(t+1/2)}}\right]$$

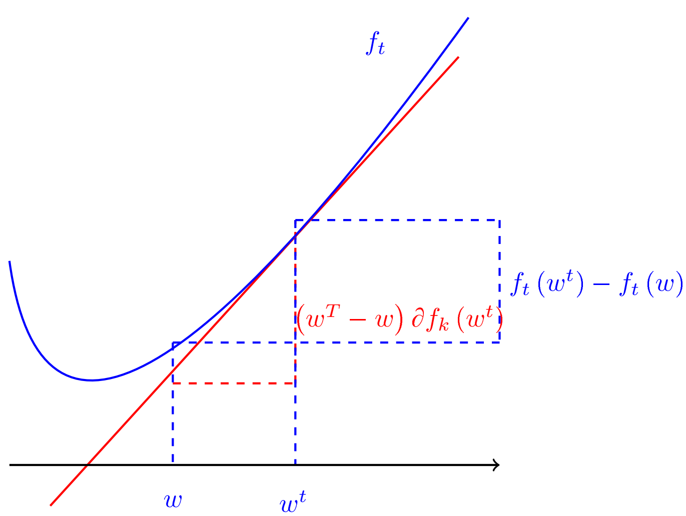
Theorem 6. *(Thm 21.15)*

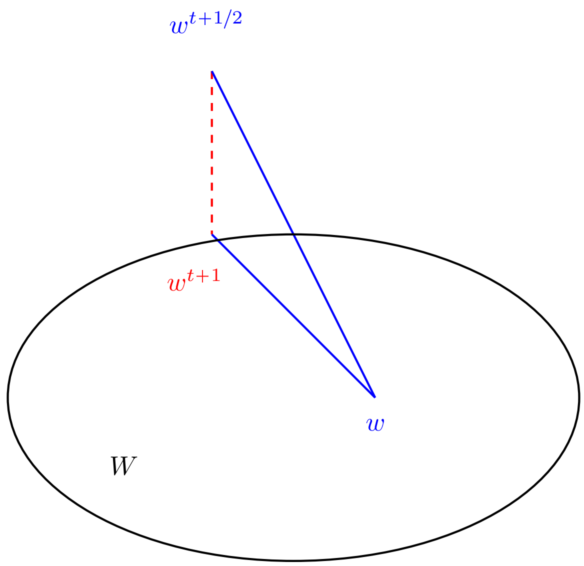
$$\begin{aligned} \textit{Regret} &\leqslant \inf_{w \in W} \frac{\|w\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\left\|\partial f_t\left(w^{(t)}\right)\right\|^2 \\ \textit{OR } \forall\, w \in W,\ \textit{Regret}\ (w) &\leqslant \frac{\|w\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\left\|\partial f_t\left(w^{(t)}\right)\right\|^2 \end{aligned}$$

Proof. Fix $w \in W$

$$\begin{aligned} \left\|w^{(t+1)} - w\right\|^2 - \left\|w^{(t)} - w\right\|^2 &= \left(\left\|w^{t+1} - w\right\|^2 - \left\|w^{t+1/2} - w\right\|^2\right) + \left(\left\|w^{t+1/2} - w\right\|^2 - \left\|w^t - w\right\|^2\right) \\ &\leqslant 0 + \left(\left\|w^t - \eta\partial f_t\left(w^t\right) - w\right\|^2 - \left\|w^t - w\right\|^2\right) \\ &= \left(-2\left(w^t - w\right)^T \eta\partial f_t\left(w^t\right) + \eta^2\left\|\partial f_t\left(w^t\right)\right\|^2\right) \\ &\leqslant 2\eta\left(f_t\left(w^t\right) - f_t(w)\right) + \eta^2\left\|\partial f_t\left(w^t\right)\right\|^2 \leftarrow \boxed{1} \end{aligned}$$

□





$$
\begin{aligned}
\sum_{t=1}^{T} \text{LHS } \boxed{1} &= \left\| w^{(T+1)} - w \right\|^2 - \left\| w^{(1)} - w \right\|^2 \\
&\leqslant \sum_{t=1}^{T} \text{RHS } \boxed{1} - = 2\eta \sum_{t=1}^{T} \left(f_t\left(w^t\right) - f_t\left(w\right) \right) + \eta^2 \sum_{t}^{T} \left\| \partial f_t\left(w^t\right) \right\|^2 \\
\sum_{t}^{T} f_t\left(w^t\right) - \sum_{t}^{T} f_t\left(w\right) &\leqslant \frac{-\left\| w^{T+1} - w \right\|^2 + \|w\|^2}{2\eta} + \frac{\eta}{2} \sum_{t}^{T} \left\| \partial f_t\left(w^t\right) \right\|^2 \\
&\leqslant \frac{\|w\|^2}{2\eta} + \frac{\eta}{2} \sum_{t}^{T} \left\| \partial f_t\left(w^t\right) \right\|^2 \leftarrow \boxed{2}
\end{aligned}
$$

Further assumptions:

1. W is norm bounded: $\forall\, w \in W, \|w\| \leqslant B$
2. $f_t,\ \forall\, t$ is ρ Lipschitz $\|\partial f_t\left(w\right)\| \leqslant \rho\ \forall\, w \in W$

$$
\begin{aligned}
\boxed{2} &\overset{\text{Assump 1.2}}{\Rightarrow} \sum_{t}^{T} f_t\left(w^t\right) - \sum_{t}^{T} f_t\left(w\right) \leqslant \frac{B^2}{2\eta} + \frac{\eta}{2} \sum_{t}^{T} \rho^2 \\
&= \underbrace{\frac{B^2}{2\eta} + \frac{\eta T \rho^2}{2}}_{\boxed{B\rho\sqrt{T}}} \\
&- \frac{B^2}{2\eta^2} + \frac{T\rho^2}{2} = 0 \\
&\eta = \sqrt{\frac{B^2}{T\rho^2}} = \frac{B}{\sqrt{T}\rho}
\end{aligned}
$$

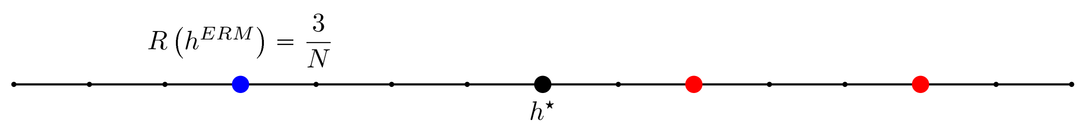
doubling trick

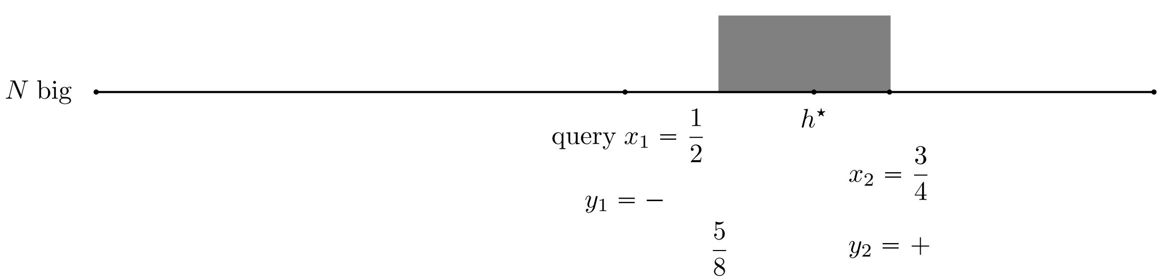
run OGD on $t = 1$
run OGD on $t = 2, 3$
run OGD on $t = 4, 5, 6, 7$

20 Lecture 20

passive learning: $(x, y) \sim P_{XY}$ "unit cost" PAC w.p. $\geqslant 1 - \delta, n = O\left(\frac{|\mathcal{H}|}{\varepsilon}\right)$
active learning: $x \sim P_X$ free, query x, "oracle" gets $y \sim P_{Y|X=x}$, realizable $(y = h^\star(x))$, unit cost, alg can choose x! adaptively
e.x. $X = 0 : \frac{1}{N} : 1, \mathcal{H} = \left\{\mathbb{1}_{(x \geqslant a)}, a \in X\right\}, P_X$ unif $(X) \overset{iid}{\sim} n$ training items (passive)
ERM, $h^{ERM} \in$ Version Space, $V = \{h \in \mathcal{H} : h(x_i) = y_i, i = 1 : n\}$

$$R\left(h^{ERM}\right) = O\left(\frac{1}{n}\right)$$

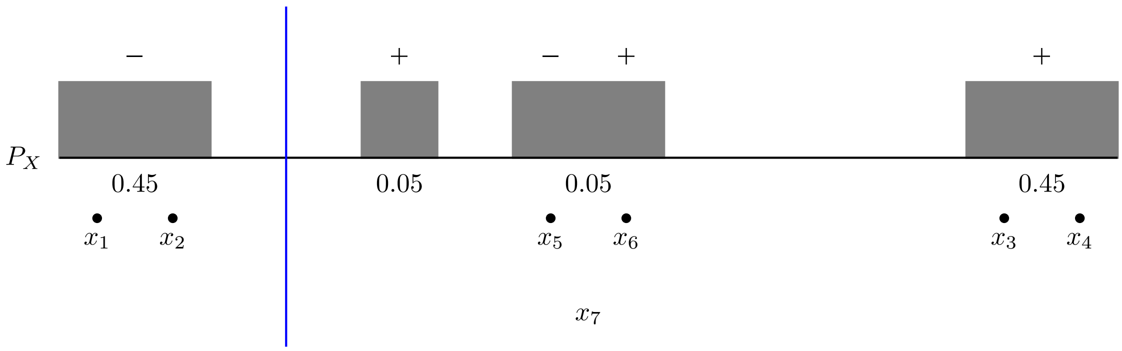




n queries

$$n = O\left(\log\left(\frac{1}{\varepsilon}\right)\right) \leqslant R(h \in VS) = O\left(\frac{1}{2^n}\right)$$

Unvertainty-based Active Learning



$R = 0.025$ $\quad$ $R = 0.05$ as $\#q \to \infty$

CAL

init $V_1 = \mathcal{H}$ version space

for epoch $r = 1, 2, ...$

$$x \sim P_X$$

if V_r disagrees on x (ie $\exists h, h' \in V_r, h(x) \neq h'(x)$)

query x 's label, oracle gives y

$$V_{r+1} = \{h \in V_r, h(x) = y\}$$

init $V_1 = \mathcal{H}$ version space

for epoch $r = 1, 2, ..., R$

(make sure this happens k times)

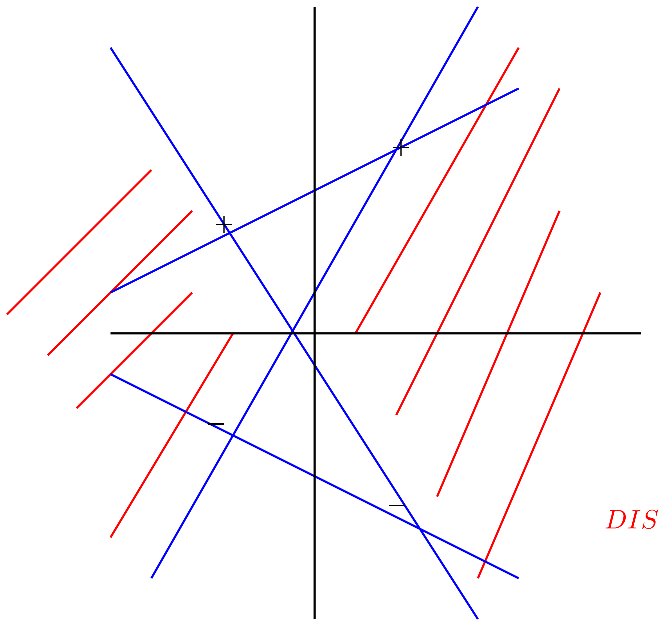
- $x \sim P_X$
- if V_r disagrees on x (ie $\exists h, h' \in V_r, h(x) \neq h'(x)$)
- query x 's label, oracle gives y

$$V_{r+1} = \{h \in V_r, h(x_i) = y_i, i = 1...k\}$$

Output any $h \in V_{R+1}$

version space V_r

Disagreement region $DIS(V_r) := \{x \in X : \exists h, h' \in V_r : h(x) \neq h'(x)\}$

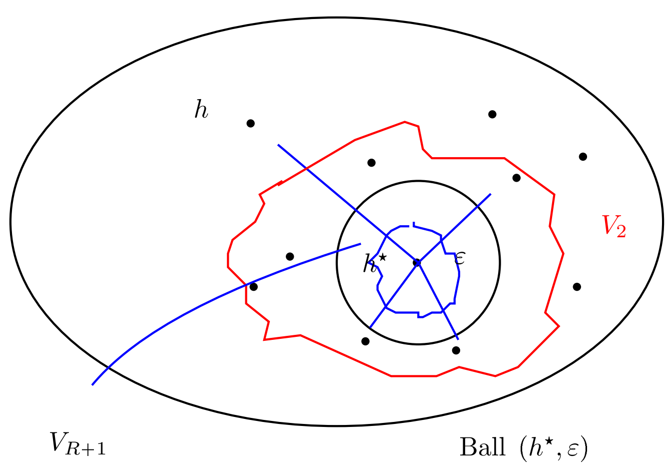


$$\Delta\left(V_r\right) := P_X\left(DIS\left(V_r\right)\right)$$
$$R\left(h^{CAL}\right)$$

pseudometric $d\left(h,h'\right), h,h' \in \mathcal{H} = \mathbb{E}_{x\sim P_X}\mathbb{1}_{(h(x)\neq h'(x))}$

$$\Rightarrow R(h) = d\left(h,h^\star\right)$$

$$\mathcal{H} = V_1$$



21 Lecture 21

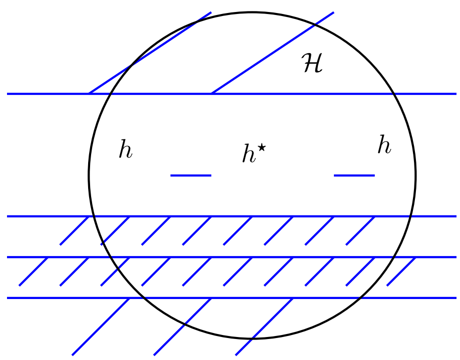
21.1 CAL (mini-batch version)

1. init $V_1 = \mathcal{H}$ (assume $|\mathcal{H}| < \infty$)
2. FOR epoches $i = 1...n$
3. Collect k items $x_1, .., x_k \overset{iid}{\sim} P_X$ such that they $\in DIS(V_i) = \{x \in X : \exists h, h' \in V_i, h(x) \neq h'(x)\}$
4. query them. Oracle gives their labels $y_1 ... y_k$
5. $V_{i+1} = \{h \in V_i : h(x_i) = y_i,\ \forall\, i \in [k]\}$
6. return any $h \in V_{n+1}$

Want: query complexity

$$O\left(\log \frac{1}{\varepsilon}\right)$$

w.p. $\leqslant 1 - \delta$



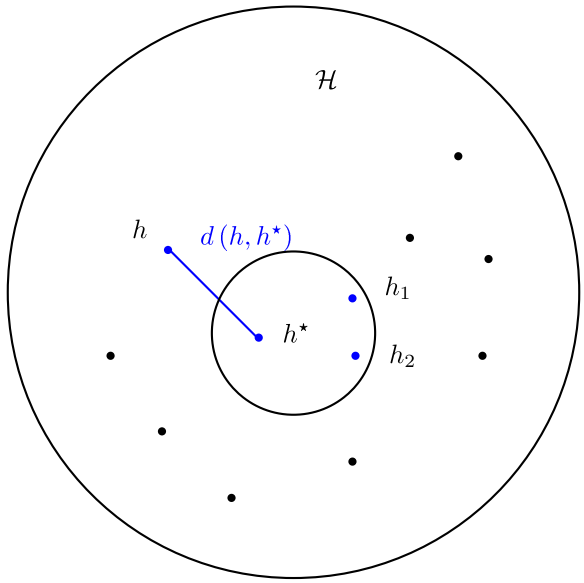
$$P_X(DIS(V_i)) \geqslant R(h),\ \forall\, h \in V_i$$

Define (pseudo) metric

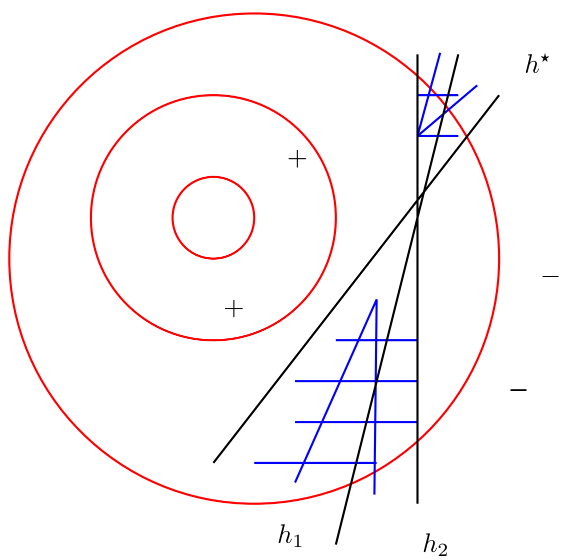
$$d(h, h') = P_x(DIS(\{h, h'\}))$$
$$\Rightarrow d(h, h^\star) = \mathbb{E}_{x \sim P_X} \mathbb{1}_{(h(x) \neq h^\star(x))} = R(h)$$

Want:

$$\boxed{P_X(DIS(V_{i+1})) \leqslant \frac{1}{2} P_x(DIS(V_i))\ \ \forall\, i \in [n]}$$
$$\Rightarrow R(h) \in V_{n+1} \leqslant P_X(DIS(V_{n+1})) \leqslant \frac{1}{2^n} := \varepsilon \Rightarrow R(h) < \varepsilon \text{ if } n = \log\left(\frac{1}{\varepsilon}\right)$$



$$B(h^\star, r) := \{h \in \mathcal{H}, d(h, h^\star) \leqslant r\}$$
$$P_X(DIS(B(h^\star, r)))$$



$$V_i = \{h^\star, h_1, h_2\}$$

Step 3 = draw iid $P_{X|V_i}$

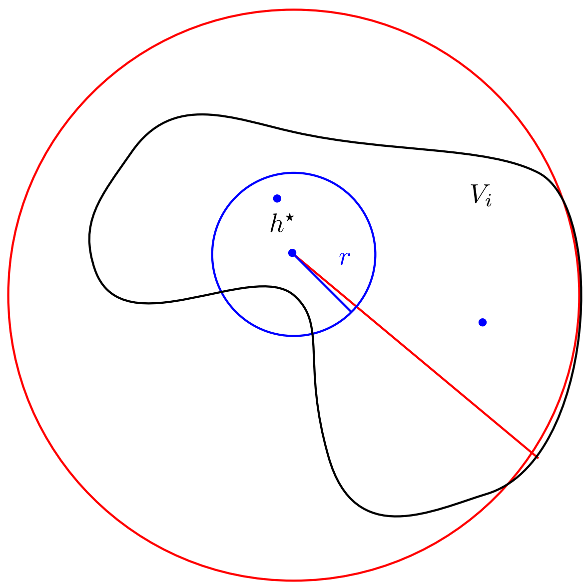
$$\boxed{P_{X|V_i} = \frac{P(x)}{\int_{x' \in DIS(V_i)} P(x')\, dx'} = \frac{P(x)}{P_X(DIS(V_i))}}$$

Define $Q_{XY} := P_{X|V_i} \cdot P(Y|X)$

$$R_Q(h) = \frac{R_P(h)}{P_X(DIS(V_i))}$$

Want:

$$P_X(DIS(V_{i+1})) \leqslant f\left(r_{V_{i+1}}\right), r_V := \max_{h \in V} d(h, h^\star)$$



$$d(h, h^\star) = R_P(h) = R_X(DIS(V_i)) \cdot R_Q(h),\ \forall\ h \in V_i$$

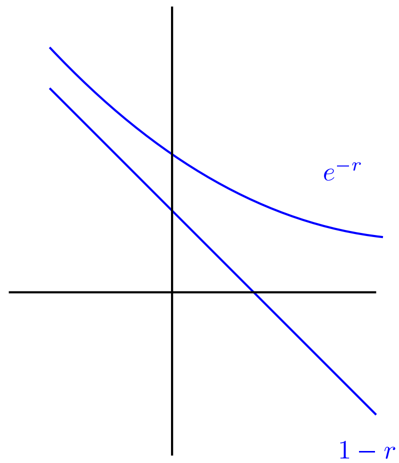
Suppose $h \in V_i$ survives k iid Q

$$\Rightarrow R_Q(h) \text{ small } := r$$

$$|V_i| \cdot \left[(1-r)^k\right] \leqslant |V_i|\, e^{-rk} := \frac{\delta}{n}$$

$$-rk = \log \frac{\delta}{n|V_i|}$$

$$\Rightarrow k = \frac{\log \frac{n|V_i|}{\delta}}{r}$$



22 Lecture 22

22.1 Missed Lecture

$$
\begin{aligned}
&V_i \\
Q &:= P_{X|DIS(V_i)} \cdot P_{Y|X} \\
R_Q(h) &:= \frac{R_p(h)}{P_X(DIS(V_i))}
\end{aligned}
$$

k labeled items $\overset{iid}{\sim} Q$

$$
\begin{aligned}
wp \geqslant 1 - \frac{\delta}{n}, \quad &\underbrace{\forall\, h \in V_{i+1}}_{\text{agrees on those } k \text{ items}}, R_Q(h) \leqslant r \text{ if } k \geqslant \frac{\log \frac{n|V_i|}{\delta}}{r} \\
&\Rightarrow \underbrace{R_p(h)}_{=d(h,h^\star)} \leqslant P_X(DIS(V_i)) \cdot r \\
&\Rightarrow V_{i+1} \subseteq B(h^\star, P_X(DIS(V_i)) \cdot r) \\
d(h,h^\star) &= \mathbb{E}_{x \sim P_X} \mathbb{1}_{(h(x) \neq h^\star(x))} \\
&= P_X(DIS(\{h, h^\star\}))
\end{aligned}
$$

want:

$$
\begin{aligned}
&P_X(DIS(V_{i+1})) \leqslant \frac{1}{2} P_X(DIS(V_i)) \\
&\Rightarrow \underbrace{P_X(DIS(V_{n+1}))}_{d(h,h^\star)\ \forall\ h \in V_{n+1}} \leqslant \frac{1}{2^n} P_X(DIS(V_1 = \mathcal{H})) \leqslant \frac{1}{2^n} = \varepsilon
\end{aligned}
$$

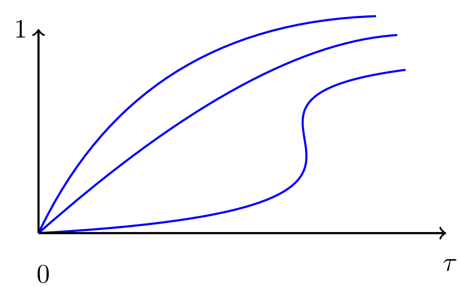
set $n = \log \frac{1}{\varepsilon}$

$$P_X \left(DIS \left(\overbrace{B\left(h^\star, \tau\right)}^{\{h \in \mathcal{H} : d(h, h^\star) \leqslant \tau\}} \right) \right)$$

$$\theta := \sup_{\tau \in (0,1)} \frac{P_X \left(DIS \left(B \left(h^\star, \tau \right) \right) \right)}{\tau} \text{ (Assume } \theta < \infty) \leftarrow \text{(disagreement coefficient)}$$

$$\Rightarrow \ \forall \tau \in (0,1), P_X \left(DIS \left(B \left(h^\star, \tau \right) \right) \right) \leqslant \theta \tau$$

$$P_X \left(DIS \left(V_{i+1} \right) \right) \leqslant P_X \left(DIS \left(B \left(h^\star, P_X \left(DIS \left(V_i \right) \right) r \right) \right) \right) \leqslant \theta P_X \left(DIS \left(V_i \right) \right) r$$



Choose $r = \frac{1}{2\theta}, k \geqslant \frac{\log \frac{n |V_i|}{\delta}}{\frac{1}{2\theta}}$

Set $k = \frac{\log \frac{n|V_i|}{\delta}}{\frac{1}{2\theta}} = 2\theta \left[\log \log \frac{1}{\varepsilon} + \log \frac{|\mathcal{H}|}{\delta} \right]$

Total number queries by CAL

$$k \cdot n = 2\theta \left[\log \log \frac{1}{\varepsilon} + \log \frac{|\mathcal{H}|}{\delta} \right] \log \frac{1}{\varepsilon}$$

22.2 End Missed Lecture

22.3 Stochastic Bandits

Arms $1 ... K$

Unknown but fixed reward distributions $V_1 ... V_k$ with means $U_1 ... U_k \in \mathbb{R}$

n total rounds

(pseudo) regret

$$U^\star = \max_{i \in [k]} U_i$$

Let $I_t \in [k]$ be the arm you pull at round $t, t \in [n]$

Let X_t be the reward you see at round t

$$nU^\star - \sum_{t=1}^{n} \mathbb{E} U_{I_t}$$

exploration then exploitation

1. pull each arm m times, estimate $\hat{U}_i, i \in [k]$
2. for $n - mk$ rounds, pull $I_t := \arg\max \hat{U}_i$

23 Lecture 23

23.1 Subgraussian tail bounds

Let $X_{i=1...n} - u$ be independent σ subgaussian random variables. Then,

$$\mathbb{P}(\hat{u} \geqslant u + \varepsilon) \leqslant e^{-\dfrac{n\varepsilon^2}{2\sigma^2}}$$

$$\text{wp } \geqslant 1 - \delta, u \leqslant \hat{u} + \sqrt{\frac{2\sigma^2 \log \dfrac{1}{\delta}}{n}}$$

23.2 K-arm bandit (stochastic)

"environment" k arms with reward
distributions $V_1 ... V_k$, 1-subgaussian
with mean $U_1 ... U_k$
def: $U^\star = \max\limits_{i \in [k]} U_i$
(assume $U_1 \geqslant U_2 \geqslant ... \geqslant U_k$)
def: $\Delta_i = U_1 - U_i, i \in [k]$

"learner, agent"
knows: k, 1-subgaussian, n time horizon, I_t = "policy" alg $(I_1, X_1, ..I_{t-1}, X_{t-1}), X_t \sim V_{I_t}, T_i(t)$, number of arm i pulls up to time t
Goal: minimize (psuedo) regret

$$Reg := nU^\star - \sum_{t=1}^{n} \mathbb{E} U_{I_t}$$

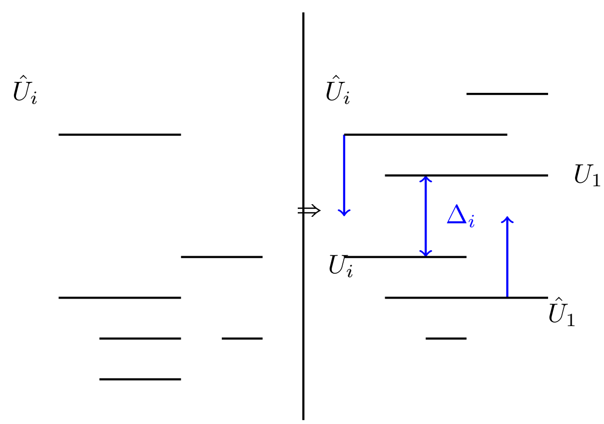
Alg: exploration-then-exploitation (m)

1. pull each arm m times $\Rightarrow \hat{U}_t, i \in [k]$
2. For remaining $n - mk$ pulls, $\arg\max\limits_{i \in [k]} \hat{U}_i$

$$m\Delta_1 + m\Delta_2 + ... + m\Delta_k = m \sum_{i=1}^{k} \Delta_i = \frac{n}{k} \sum \Delta_i$$

$$Reg = \sum_{i=1}^{k} \mathbb{E}(T_i(n)) \Delta_i$$

$$
\begin{aligned}
\mathbb{E}T_i(n) &= m + (n - mk)\,\mathbb{P}(I = i) \\
&= m + (n - mk)\,\mathbb{P}\left(\hat{U}_i \geqslant \max_{j\in[k], j\neq i} \hat{U}_j\right) \\
&\leqslant m + (n - mk)\,\mathbb{P}\left(\hat{U}_i \geqslant \hat{U}_1\right) \\
&= m + (n - mk) + \mathbb{P}\left(\hat{U}_i \leqslant \hat{U}_1 + U_i - U_i + \Delta_i\right) \\
&= m + (n - mk) + \mathbb{P}\left(\left(\hat{U}_i - U_i\right) - \left(\hat{U}_1 - U_1\right) \geqslant \Delta_i\right) \\
&\leqslant m + (n - mk)\, e^{-\frac{\Delta_i}{2\left(\frac{2}{m}\right)}}
\end{aligned}
$$

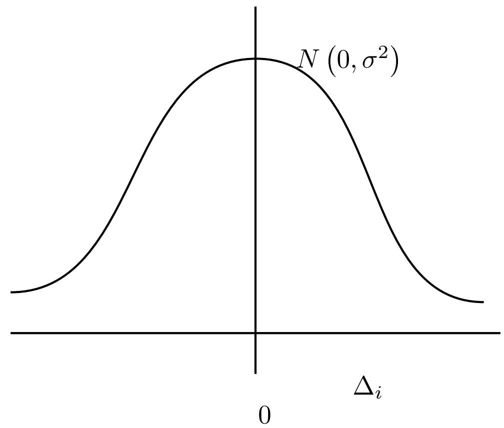


$$
\hat{U}_i = \frac{1}{m}\sum_{t=1, I_t = i}^{mk} X_t, X_t \overset{iid}{\sim} V_i, X_t - U_i, 1 - \text{subgaussian}
$$

$$
\Rightarrow \hat{U}_i - U_i, \frac{1}{\sqrt{m}} - \text{subgaussian } \forall\, i
$$

$$
\Rightarrow \left(\hat{U}_i - U_i\right) - \left(\hat{U}_1 - U_1\right), \sqrt{\frac{2}{m}} - \text{subgaussian}
$$

$$
\mathbb{P}(X \geqslant \Delta_i) \leqslant e^{-\frac{\Delta_i^2}{2\sigma^2}}
$$



$$
\begin{aligned}
Reg &= \sum_{i=1}^{k} \mathbb{E}T_i(n) \\
&\leqslant \sum_{i=1}^{k} \left(m + (n-mk)\, e^{-\dfrac{\Delta_i^2 m}{4}} \right) \Delta_i \\
&\leqslant \sum_{i}^{k} \left(m + n e^{-\dfrac{\Delta_i^2 m}{4}} \right) \Delta_i \\
&\leqslant \sum_{i}^{k} \left(m + n e^{-\dfrac{\Delta_2^2 m}{4}} \right) \Delta_2
\end{aligned}
$$

$$
\begin{aligned}
&\frac{\partial \text{ RHS } (m)}{\partial m} = 0 \\
&\Rightarrow \sum_{i=1}^{k} \Delta_i \left(1 - k e^{-\dfrac{\Delta_i^2 m}{4}} + (n-mk)\, e^{-\dfrac{\Delta_i^2 m}{4}} \left(-\frac{\Delta_i^2}{4} \right) \right) \\
&= \sum_{i} \Delta_i \left(1 - \left[k + (n-mk) \frac{\Delta_i^2}{4} \right] e^{-\dfrac{\Delta_i^2 m}{4}} \right) \\
&\approx \sum_{i} \Delta_i \left(1 - \left[k + n \frac{\Delta_i^2}{4} \right] e^{-\dfrac{\Delta_i^2 m}{4}} \right)
\end{aligned}
$$

...

$$
\frac{\partial \text{ RHS } (m)}{\partial m} = 0
$$

$$\Rightarrow \sum_i \Delta_i \left(1 - \frac{n\Delta_2^2}{4} e^{-\dfrac{\Delta_2^2 m}{4}}\right) = 0$$

$$\Rightarrow \sum \Delta_i = \left(\sum \frac{n\Delta_2^2 \Delta_i}{4}\right) e^{-\dfrac{\Delta_2^2 m}{4}}$$

$$\Rightarrow \frac{4}{\Delta_2^2} \log \frac{\sum \dfrac{n\Delta_2^2\Delta_i}{4}}{\sum \Delta_i} = m = \frac{4}{\Delta_2^2} \log \frac{n\Delta_2^2}{4}$$

24 Lecture 24

$$\begin{aligned}
U_1 &\geqslant \ldots \geqslant U_k \\
\Delta_i &= U_1 - U_i, i \in [k] \\
Reg &:= nU_1 - \sum_{t=1}^{n} \mathbb{E}\left[U_{I_t}\right] \\
&= \sum_{i=1}^{k} \Delta_i \mathbb{E}\left[T_i\left(n\right)\right] \\
&\leqslant \left(m + ne^{-\dfrac{\Delta_2^2 m}{4}}\right)\left(\sum \Delta_i\right) \leftarrow \boxed{1} \\
\text{optimal } m &= \frac{4}{\Delta_2^2} \log \frac{n\Delta_2^2}{4} \\
\boxed{1} &\overset{m}{=} \left(\frac{4}{\Delta_2^2} \log \frac{n\Delta_2^2}{4} + n\left[e^m\right]^{-\dfrac{\Delta_2^2}{4}}\right) \sum \Delta_i \\
&= \left(\frac{4}{\Delta_2^2} \log \frac{n\Delta_2^2}{4} + n\left[\frac{n\Delta_2^2}{4}\right]^{\dfrac{4}{\Delta_2^2}\left(-\dfrac{\Delta_2^2}{4}\right)}\right) \sum \Delta_i \\
&= \left(\frac{4}{\Delta_2^2} \log \frac{n\Delta_2^2}{4} + \frac{4}{\Delta_2^2}\right) \sum \Delta_i \\
&= \frac{4}{\Delta_2^2}\left(1 + \log \frac{n\Delta_2^2}{4}\right)\left(\sum \Delta_i\right) \\
&\overset{k=2, \Delta := \Delta_2}{=} \frac{4}{\Delta} \frac{1 + \log\left(n\Delta^2\right)}{4} \leftarrow \boxed{2}
\end{aligned}$$

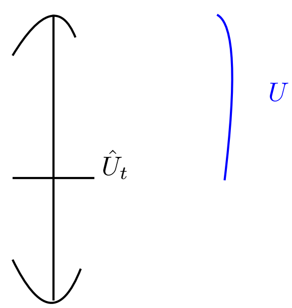
Alg needs Δ, k, n
Worst case Δ

$$\frac{\partial \boxed{2}}{\partial \Delta} = 0$$

$$
\begin{aligned}
-\frac{4}{\Delta^2}\left(1+\log\frac{n\Delta^2}{4}\right)+\frac{4}{\Delta}\frac{4}{n\Delta^2}\frac{2n\Delta}{4} &= 0\\
-\frac{4}{\Delta^2}-\frac{4}{\Delta^2}\log\frac{n\Delta^2}{4}+\frac{8}{\Delta^2} &= 0\\
\frac{4}{\Delta^2} &= \frac{4}{\Delta^2}\log\frac{n\Delta^2}{4}\\
\frac{n\Delta^2}{4} &= e\\
\Delta^\star &= \sqrt{\frac{4e}{n}}\\
\boxed{2} \overset{\Delta^\star}{=} \frac{4\sqrt{n}}{\sqrt{4e}}&\left(1+\log\frac{n\frac{4e}{n}}{4}\right)\\
&= \frac{8\sqrt{n}}{\sqrt{4e}} = \frac{4}{\sqrt{e}}\sqrt{n}
\end{aligned}
$$

24.1 Upper Confidence Bound (UCB)

$$
\begin{aligned}
x_1 \ldots x_t &\sim V\,(1-\text{subgaussian})\\
U := \frac{1}{t}\sum_{\tau=1}^{t} x_\tau &\leqslant \hat{U}_t + \sqrt{\frac{2\log\frac{1}{\delta}}{t}} \text{ wp } 1-\delta
\end{aligned}
$$



UCB(arm i, count t of arm i pulls) :=

$$
\begin{cases}
\hat{U}_{it} + \sqrt{\frac{2\log\frac{1}{\delta}}{t}} & \text{if } t > 0\\
\infty & \text{if } t = 0
\end{cases}
$$

Alg:
for $j = 1 \ldots n$
pull $I_j \in \arg\max_{i\in[t]} UCB\,(i, t_i)$

$$Reg := \sum_{i=1}^{k} \Delta_i \mathbb{E} T_i(n)$$

idea:

$$[\,\forall\, t : UCB(1,t) > U_1] := \text{ Event 1}$$

Fix $i \neq 1$, fix a magic number $\tau_i \in [n]$

$$[UCB(i,\tau_i) < U_1] := \text{ Event 2}$$

Define event $E_i = $ Event 1 $\wedge$ Event 2

$$\begin{aligned}
\mathbb{E}T_i(n) &= \mathbb{E}\mathbb{1}_{(E_i)} T_i(n) + \mathbb{E}\mathbb{1}_{(E_i^c)} T_i(n) \\
&\overset{\text{Claim 1}}{\leqslant} \tau_i + \mathbb{E}\mathbb{1}_{(E_i^c)} T_i(n) \\
&\overset{T_i(n)\leqslant n}{\leqslant} \tau_i + n\mathbb{P}(E_i^c)
\end{aligned}$$

Claim $1, E_i \Rightarrow T_i(n) \leqslant \tau_i$

$$\begin{aligned}
\mathbb{P}(E_i^c) &\overset{\text{Union}}{\leqslant} \mathbb{P}(\exists t, UCB(1,t) \leqslant U_1) + \mathbb{P}(UCB(i,\tau_i) > U_1) \\
&\leqslant n\mathbb{P}\left(\hat{U}_{it} + \sqrt{\frac{2\log\frac{1}{\delta}}{t}} \leqslant U_1\right) + \mathbb{P}(UCB(i,\tau_i) > U_1) \\
&\overset{\text{subgaussian}}{\leqslant} n\delta + \mathbb{P}\left(\hat{U}_{i,\tau_i} + \sqrt{\frac{2\log\frac{1}{\delta}}{\tau_i}} > U_i + \Delta_i\right) \\
&= n\delta + \mathbb{P}\left(U_{i\tau_i} - U_i > \underbrace{\Delta_i - \sqrt{\frac{2\log\frac{1}{\delta}}{\tau_i}}}_{\frac{1}{2}\Delta_i}\right) \leftarrow \boxed{3}
\end{aligned}$$

Choose τ_i such that

$$\Delta_i - \sqrt{\frac{2\log\frac{1}{\delta}}{\tau_i}} = \frac{1}{2}\Delta_i \leftarrow \boxed{4}$$

$$\boxed{3} \leqslant n\delta + e^{-\frac{\tau_i\left(\frac{1}{2}\Delta_i\right)^2}{2}} \leftarrow \boxed{5}$$

$$\boxed{4} \Rightarrow \frac{\Delta_i}{2} = \sqrt{\frac{2\log\frac{1}{\delta}}{\tau_i}}$$

$$\tau_i = \frac{4 \cdot 2\log\frac{1}{\delta}}{\Delta_i^2}$$

$$\boxed{5} = n\delta + e^{-\frac{8\log\frac{1}{\delta}}{\Delta_i^2}\frac{\Delta_i^2}{2\cdot 4}}$$

$$= (n+1)\,\delta$$

$$\mathbb{E}T_i(n) \leqslant \frac{8\log\frac{1}{\delta}}{\Delta_i^2} + n(n+1)\,\delta$$

25 Lecture 25

One version of UCB(δ)

$$UCB(i,t_i) := \hat{U}_{i,t_i} + \sqrt{\frac{2\log\frac{1}{\delta}}{t_i}}$$

repeat: pull $\arg\max_{i\in[k]} UCB(i,t_i)$
Last time

$$\begin{aligned}
\mathbb{E}T_i(n) &\leqslant \tau_i n(n+1)\,\delta \\
\tau_i &= \left\lceil \frac{8\log\frac{1}{\delta}}{\Delta_i^2} \right\rceil \\
Reg &= \sum_{i=1}^{k} \Delta_i \mathbb{E}T_i(n) \leqslant \sum_{i=1}^{k} \Delta_i (\tau_i + n(n+1)\,\delta) \\
&= \sum_i \Delta_i \left(\left\lceil \frac{8\log\frac{1}{\delta}}{\Delta_i^2} \right\rceil + n(n+1)\,\delta \right), \delta := \frac{1}{n^2} \\
&= \sum_i \Delta_i \left(\left\lceil \frac{16\log n}{\Delta_i^2} \right\rceil + 1 + \frac{1}{n} \right) \\
&\leqslant \sum_i \Delta_i \left(\frac{16\log n}{\Delta_i^2} + 3 \right) \\
&= \left(\sum_i \frac{16\log n}{\Delta_i} \right) + 3\sum_i \Delta_i
\end{aligned}$$

25.1 Contextual bandit

context vector $f_t := \begin{bmatrix} f_{\text{ user } (t)} \\ f_{\text{ arm } (i_t)} \end{bmatrix}$

$$\exists w^\star, \mathbb{E}\left[r_{\text{ reward } (t)}\right] = f_t^T w^\star$$

25.2 Optimal Teaching

Learner is ERM $= V(S)$
$\mathcal{H}$ realizable, S, ℓ is $0-1$ loss

$$\arg\min_{h \in \mathcal{H}} \frac{1}{|S|} \sum_{(x,y)\in S} \ell(h,x,y) = \text{ Version Space } V(S)$$

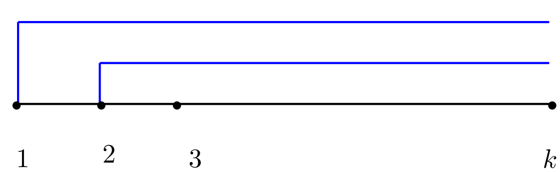
learner that only takes S
Learner is a function $\{S\} \to 2^{\mathcal{H}}$
$\mathcal{H}$ finite
learner $\left(S = \left(x^1, y^1\right) \dots (x^n, y^n)\right)$ creates a lookup table

$$T = \begin{bmatrix} h_1 & x_1 \\ \dots & \dots \\ h_{|S|} & x_{|S|} \end{bmatrix}$$

returns $T(x')$

$$S = (x^\star, y^\star) \dots$$

$X = [k]$



$$L(S) = \{h^\star\}$$
$$S = \{(2,-), (3,+)\}$$

$$\min_{S \in \mathbb{S}} |S| \text{ such that } L(S) = \{h^\star\}$$
$$\mathbb{S} = \bigcup_{n=0}^{\infty} (X \times Y)^n$$

Teaching dimension $(h^\star, \mathcal{H})$

$-$	x_1	x_2	...	x_k
h_1	1	0	...	0
...	0	1	...	0
...	...	...	...	...
h_k	0	...	0	1
h_{k+1}	0	...	...	0

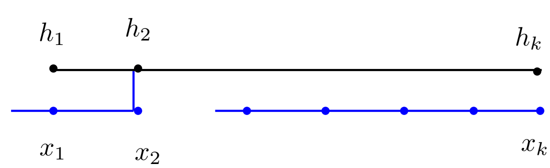
$$TD(\mathcal{H}) := \max_{h \in \mathcal{H}} TD(h, \mathcal{H})$$

$-$	x_1	x_2	...	x_k
h_1	1	0	...	0
...	0	1	...	0
...	...	...	...	...
h_k	0	...	0	1
h_{k+1}	1	...	...	1

$-$	x_1	x_2	...	x_k
h_1	1	0	...	0
...	0	1	...	0
...	...	...	...	...
h_k	0	...	0	1
h_{k+1}	1	...	...	1
h_{k+2}	0	...	...	0

26 Lecture 26

learner: $L(S) \in 2^{\mathcal{H}}$
Teacher: $T(h) = S$ such that $L(T(h)) = \{h\}$



$$T(h_2) = \{(x_1, -), (x_2, +), (x_3, +)\}$$

Teaching Dimension $(h, \mathcal{H}, L_{VS})$

$$= \min_{S \in 2^{X}} |S| \text{ such that } L(S) = \{h\}$$

$$TD(\mathcal{H}) = \max_{h \in \mathcal{H}} TD(h, \mathcal{H})$$

$-$	x_1	...	x_d	x_{d+1}	...	x_{d+2^d}
h_1	$-$	$-$	$-$	1	0	0
...	$-$	Whole truth Table	$-$	0	1	0
h_{2^d}	$-$	$-$	$-$	0	0	1

$$VC(\mathcal{H}) = d$$

$$TD(\mathcal{H}) = 1$$

$-$	x_1	...	x_d
h_1	1	0	0
...	0	1	0
h_d	0	0	1
h_{d+1}	0	0	0

$$VC(\mathcal{H}) = 1$$

$$TD(\mathcal{H}) = d$$

"Collusion"

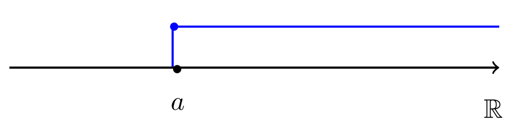
ONE definition of collusion-free teaching

If S teaches h, Then $\forall\ S' \supset S$ (labeled consistently by h) also teaches h

$-$	x_1	...	...	...	...	x_d
h_1	1	1	0	...	...	0
h_2	0	1	1	0	...	0
...	...	...	...	...	...	...
h_{d-1}	0	...	...	0	1	1
h_d	1	0	...	...	0	1

no-clash teaching

$T(h)$, if $\forall\ h, h', T(h)$ is INconsistent with h' OR $T(h')$ is INconsistent with h



$$\mathcal{H} = \left\{ h_a\left(x\right) = \mathbb{1}_{\left(x \geqslant a\right)} : a \in \mathbb{R} \right\}$$

Teaching Dimension $\left(h, \mathcal{H}, L_{VS}\right)$

$$= \min_{S \in 2^X} |S| \text{ such that } L\left(S\right) = \left\{\left(h - \varepsilon, h + \varepsilon\right)\right\}$$

Teaching Dimension $\left(h, \mathcal{H}, L_{SVM}\left(\text{ hard margin }\right)\right)$

$$= \min_{S \in 2^X} |S| \text{ such that } L\left(S\right) = \left\{h\right\}$$

Teaching Dimension $\left(h, \mathcal{H}, L_{SVM}\left(\text{ soft margin, logistic regression }\right)\right)$

$$= \min_{S \in 2^X} |S| \text{ such that } L\left(S\right) = \left\{h\right\}$$

26.1 Reinforcement Learning

$$Reg \sim O\left(\sqrt{n \log n}\right)$$