1. Project proposal due on Mar 16.

2. HW 2 out (on class website)

(latest version on Tuesday Mar. 2)

due Mar 23

Weak PAC learning: $\mathcal{H}$. A. is a γ-weak PAC learner for $\mathcal{H}$. if $\exists$ a fn $f: (0,1) \to \mathbb{N}$, for any $\delta > 0$. for any D realizable by $\mathcal{H}$. w/ $m \geq f(\delta)$, (i.i.d samples from D)

A produces a classifier h. s.t. w.p. $1-\delta$:

$$\text{err}(h, D) \leq \frac{1}{2} - \gamma .$$

Adaboost

if weak learner A returns classifiers from base hypothesis class B,

for any $h^* \in H$, for any distrn D_X over X, (*)

$$\exists f \in B. \quad \mathop{\mathbb{E}}_{x \sim D_X} I(f(x) \neq h^*(x)) \leq \frac{1}{2} - \gamma.$$

Adaboost gives an algorithmic pf showing

for any $h^* \in H$, $\exists$ distribution D_B over B.

such that $h^*(x) = \text{sign}\left(\sum_{f \in B} D_B(f) \cdot f(x)\right)$.

A more direct pf:

rewrite (*):

$$\max_{D_X} \left(\min_{f \in B} \mathop{\mathbb{E}}_{x \sim D_X} \underbrace{I(f(x) \neq h^*(x))}_{\downarrow \quad 1} \right) \leq \frac{1}{2} - \gamma$$

$$\frac{1 - f(x)h^*(x)}{2}$$

$$\Downarrow$$

$$\max_{D_X} \max_{f \in B} \mathop{\mathbb{E}}_{x \sim D_X} f(x) h^*(x) \geq 2\gamma. \qquad A(f)$$

11

$$\min_{D_X} \max_{D_B \in \Delta(B)} \mathop{\mathbb{E}}_{f \sim D_B} \mathop{\mathbb{E}}_{x \sim D_X} f(x) h^y(x) \qquad (*)$$

D_X: p ; D_B: q set of all distns on B

Von Neumann's Minimax Theorem:

given $A \in \mathbb{R}^{m \times n}$. $\Delta_n^d = \{(v_1 \sim v_d) : \forall i : v_i \geq 0, \sum_i v_i = 1\}$

protocol 1 modified — protocol 2 modified.

$$\min_{p \in \Delta^m} \max_{q \in \Delta^n} p^T A q = \max_{q \in \Delta^n} \min_{p \in \Delta^m} p^T A q.$$

A:

Alice \ Bob	R	P	S
R	0	1	−1
P	−1	0	1
S	1	−1	0

$\min_i \max_j A_{ij} = 1$

protocol 1:

— Alice chooses ~~i~~ (row) $p \in \Delta^m$

— Bob chooses j after seeing $q \in \Delta^n$

— Alice suffer loss of ~~A_{ij}~~ $p^T A q = \mathbb{E}_{i \sim p} \mathbb{E}[A_{ij}]$ given i.

$\max_j A_{ij}$

If Alice & Bob behave optimally. what would be Alice's pay off ?

$\max_j \min_i A_{ij} = -1$

protocol 2:

- Bob chooses j (column)
- Alice chooses i after seeing j $\Big]$ $\min_i A_{ij}$
- Alice loss $-A_{ij}$

If Alice & Bob behave optimally. what would be Alice's pay off ?

Applying von Neumann's thm on (*).

$$\max_{D_B \in \Delta(B)} \min_{D_X} \mathop{\mathbb{E}}_{x \sim D_X} \mathop{\mathbb{E}}_{f \sim D_B} f(x)\, h^*(x) \geq 2\gamma.$$

$$\max_{D_B \in \Delta(B)} \min_{x} \left[\mathop{\mathbb{E}}_{f \sim D_B} f(x) \right] h^*(x) \geq 2\gamma.$$

$\exists D_B$, for all $x \in X$,

$$\sum_{f \in B} D_B(f)\, f(x) \cdot h^*(x) \geq 2\gamma.$$

$$> 0$$

if we define (say B is finite)

$$\tilde{x} = (f(x))_{f \in B} \in \{-1,+1\}^N \quad N = |B|$$

$(\tilde{x}, h^*(x))_{x \in \mathcal{X}}$ is linearly separable by

$$w = (D_B(f))_{f \in B} \in \mathbb{R}^N .$$

Generalization error bounds for boosting.

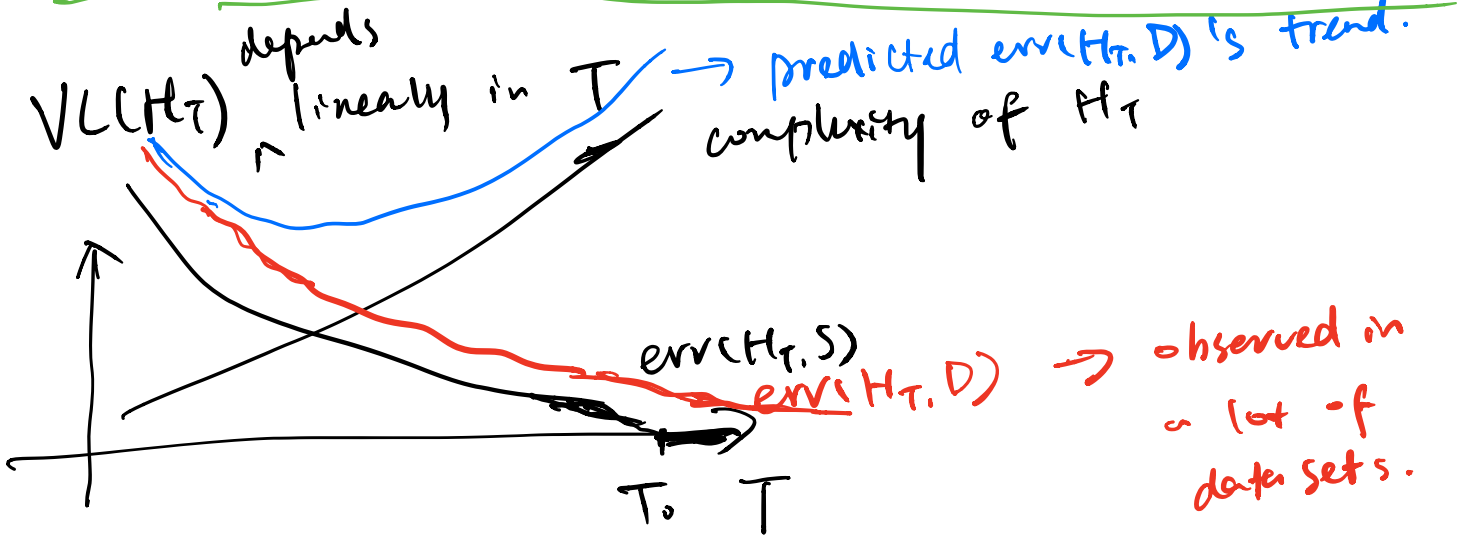
Choose T in AdaBoost:

$$\text{err}(H_T, S) \leq \exp(-2T\gamma^2)$$

$$\text{err}(H_T, D) \leq \text{err}(H_T, S) + \sqrt{\frac{VC(H_T)}{m} T}$$

$\leq N = |B|$

$$H_T \in \left\{ \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x) \right) : \alpha_1 \cdots \alpha_T \in \mathbb{R}, \ \forall t \ h_t \in B \right\}$$



Motivates the large margin theory for boosting.

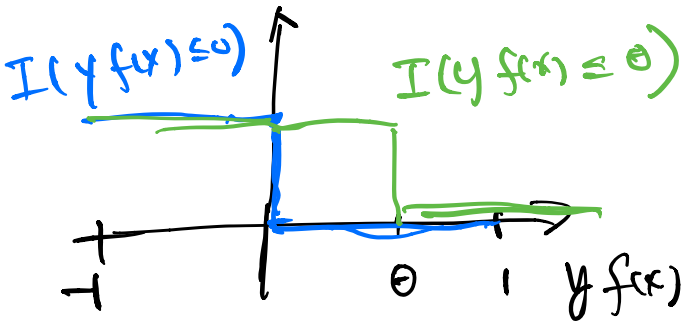
Thm: suppose B is finite. $C(B) = \left\{ \sum_{h \in B} \alpha_h h(x) : \sum_{h \in B} |\alpha_h| \leq 1 \right\}$ is the set of voting classifier over

B. Fix margin $\theta \in [0,1]$. Then, w.p. $1-\delta$: (in iid training examples $S \sim D$): for all $f \in C(B)$: ($\sum \alpha_h h(x) = \sum_{h \in \mathcal{H}} \alpha_h h(x)$)

$$P_D(yf(x) \le 0) \le P_S(yf(x) \le \theta) + O\left(\frac{1}{\theta}\cdot\sqrt{\frac{\ln\frac{|B|}{\delta}}{m}}\right)$$

margin of f on (x,y).

error of $\operatorname{sign}(f)$



Application to AdaBoost:

$$\overline{f}_T = \frac{\sum_{t=1}^{T} \alpha_t h_t(x)}{\sum_{t=1}^{T} \alpha_t} \in C(B).$$

Let $\theta = \frac{\gamma}{2}$.

① $P_S(yf(x) \le \frac{\gamma}{2}) \le \exp(-T\gamma^2)$.

② complexity term $O\left(\frac{1}{\gamma}\cdot\sqrt{\frac{\ln\frac{|B|}{\delta}}{m}}\right)$

independent of T.

Thm (more abstract version) Suppose D supported on $\{x \in \mathbb{R}^d : \|x\|_\infty \le R_\infty\} \times \{\pm 1\}$. Fix margin value $\theta \in$

$[0,1]$. Then. w.p. $1-\delta$ over m samples S.

for any predictor w such that $\|w\|_1 \le B_1$.

$$P_D(y\langle w,x\rangle \le 0) \le P_S(y\langle w,x\rangle \le \theta) + \underbrace{\text{margin of } w \text{ on } (x,y)}$$

$$O\left(\frac{B_1 R_\infty}{\theta}\sqrt{\frac{\ln d/\delta}{m}}\right).$$

$B = \{h_1 \cdots h_N\}$ $\quad h_i(x) \in \{\pm 1\}$

$\tilde{x} = (h_1(x) \cdots h_N(x)) \quad \& \quad \|\tilde{x}\|_\infty \le 1 := R_\infty.$

$(\tilde{x}, y) \sim \tilde{D}.$

$$\sum_{i=1}^{N} \alpha_i h_i(x) \longleftrightarrow \langle \alpha, \tilde{x}\rangle$$

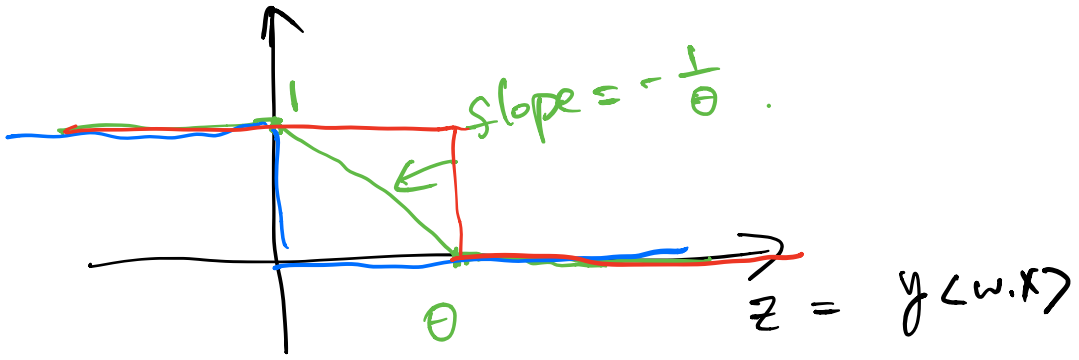
$\|\alpha\|_1 \le 1 \longrightarrow$ setting $B_1 = 1.$

choose $d = N.$

pf of thm:

define ramp loss $\ell_{\theta,w}(x,y) = \phi_\theta(y\langle w,x\rangle)$

where $$\phi_\theta(z) = \begin{cases} 1 & z \le 0 \\ 1 - \frac{z}{\theta} & z \in (0,\theta) \\ 0 & z \ge \theta \end{cases}$$



we will show : w.p. $1-\delta$.

① for every w, $\|w\|_1 \le B_1$.

$$\mathbb{E}_D \, \ell_{\theta,w}(x,y) \le \mathbb{E}_S \, \ell_{\theta,w}(x,y) + \sqrt{\frac{\ln 2/\delta}{2m}} + 2 \cdot \mathrm{Rad}_n(F)$$

($f(z)$)

(appear in HW2. Problem 3)

where $F = \{ \ell_{\theta,w} : \|w\|_1 \le B_1 \}$. ($f$)

② $$\mathbb{E}_D \, \ell_{\theta,w}(x,y) \ge P_D(y\langle w, x\rangle \le 0)$$

$$\mathbb{E}_S \, \ell_{\theta,w}(x,y) \le P_S(y\langle w, x\rangle \le \theta).$$

Next class:

focus on bounding

$$\mathrm{Rad}_n(F) \le O\left(\frac{B_1 R_\theta}{\theta}\sqrt{\frac{\ln d/\delta}{m}}\right)$$

contraction inequality of Rademacher complexity