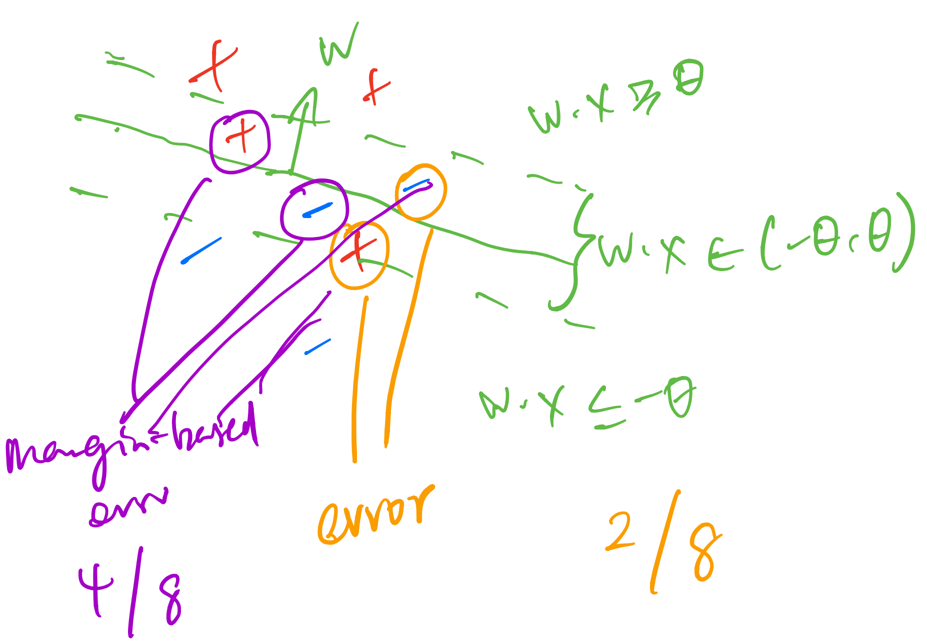
Thm (more abstract version) Suppose D supported on $\{x \in \mathbb{R}^d : \|x\|_\infty \le R_\infty\} \times \{\pm 1\}$. Fix margin value $\theta \in$ $\cancel{[0,1]}$ $(0, B_1 R_\infty]$. Then, w.p. $1-\delta$ over m samples S, for any predictor w such that $\|w\|_1 \le B_1$,

$\le B_1$ $\le R_\infty$ $\in [-B_1 R_\infty, +B_1 R_\infty]$

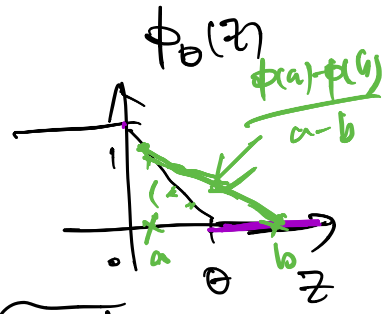
$$P_D(y\langle w,x\rangle \le 0) \le P_S(\underbrace{y\langle w,x\rangle}_{\text{margin of } w \text{ on } (x,y)} \le \theta) + O\left(\frac{B_1 R_\infty}{\theta}\sqrt{\frac{\ln d/\delta}{m}}\right).$$

$\sqrt{\frac{d}{m}}$



Define $\mathcal{F} = \{ l_{\theta,w} : \|w\|_1 \le B_1 \}$ (ramp loss)

$$l_{\theta,w}(x,y) = \phi_\theta(y\langle w, x\rangle)$$



Show $Rad_n(\mathcal{F}) \le O\left(\frac{B_1 R_\infty}{\theta}\sqrt{\frac{\ln d}{m}}\right)$.

Pf: **intuition**: $Rad_n(\mathcal{F})$

$$= \mathbb{E}_{S\sim D^m} Rad_S(\mathcal{F})$$

$$Rad_S(\mathcal{F}) = \mathbb{E}_{\sigma\sim U(\{\pm 1\})^m} \frac{1}{m} \sup_{f\in\mathcal{F}} \sum_{i=1}^m \sigma_i f(x_i, y_i)$$

$$= \frac{1}{m}\mathbb{E}_\sigma \sup_{w : \|w\|_1 \le B_1} \sum_{i=1}^m \sigma_i \cancel{\phi_\theta}(y_i\langle w, x_i\rangle)$$

step 1: use contraction inequality to "remove" ϕ_θ.

Lemma (contraction) suppose $S = (z_1 \dots z_m)$, suppose $\mathcal{G}$ is a fn class, and ϕ is a L-Lipschitz fn,

$\forall a, b \ |\phi(a) - \phi(b)| \le L|a-b|$.

$z, \quad \phi\circ g = \phi(g(z))$

$\mathcal{F} = \{ \phi \circ g : g \in \mathcal{G} \}$. then,

$$\mathrm{Rad}_S(\mathcal{F}) \leq L \cdot \mathrm{Rad}_S(\mathcal{G}) .$$

Apply contraction ineq w/

$$\mathcal{G} = \{ m_w : \|w\|_1 \leq B_1 \}$$

$$m_w(x, y) = y \langle w, x \rangle .$$

$$\phi = \phi_\theta .$$

$$\mathcal{F} = \{ \phi \circ g : g \in \mathcal{G} \} .$$

$$\Rightarrow \quad \mathrm{Rad}_S(\mathcal{F}) \leq L_{\phi_\theta} \cdot \mathrm{Rad}_S(\mathcal{G})$$

$\frac{1}{\theta}$

$(\sigma_1 \cdots \sigma_m) \overset{d}{=} (\sigma_1 y_1 \cdots \sigma_m y_m)$

Bounding $\mathrm{Rad}_S(\mathcal{G})$:

$$= \frac{1}{m} \mathop{\mathbb{E}}_{\sigma} \sup_{\|w\|_1 \leq B_1} \sum_{i=1}^{m} \boxed{\sigma_i \cdot y_i} \langle w, x_i \rangle$$

σ_i

$$= \frac{1}{m} \mathop{\mathbb{E}}_{\sigma} \sup_{\|w\|_1 \leq B_1} \langle w, \sum_{i=1}^{m} \sigma_i x_i \rangle \quad (*)$$

[**Fact**: given $\beta = (\beta_1 \cdots \beta_d)$

$$\max_{\alpha : \|\alpha\|_1 \le A} \underline{\langle \alpha, \beta \rangle} \overset{? \le}{=} A \cdot \|\beta\|_\infty \quad]$$

$\geq$ $\qquad$ $\geq$

$\frac{1}{p} + \frac{1}{q} = 1$

$p = 1$, $q = \infty$

pf of fact:

① $\le$: $\quad \forall \alpha.\ \sum_i |\alpha_i| \le A.$

$$\sum_i \alpha_i \beta_i \le \sum_i |\alpha_i| \cdot |\beta_i|$$

$\le \max_i |\beta_i| = \|\beta\|_\infty$

A

$$= \|\beta\|_\infty \cdot \underline{\underline{\|\alpha\|_1}}$$

$$\le A \cdot \|\beta\|_\infty.$$

② "=" $\quad$ pick $\alpha^* = \begin{cases} A \cdot e_{i^*}, & \beta_{i^*} > 0 \\ -A\, e_{i^*}, & \beta_{i^*} \le 0 \end{cases}$

$i^* = \arg\max_i |\beta_i| \qquad \|\alpha^*\|_1 \le A$

$\beta_{i^*} > 0. \qquad \langle \alpha^*, \beta \rangle = A|\beta_{i^*}| = A\|\beta\|_\infty.$

continuing (4):

$$\|v\|_\infty = \max(v_1, -v_1, v_2, -v_2, \ldots v_d, -v_d)$$

$$= \frac{B_1}{m} \mathbb{E}_\sigma \left\| \sum_{i=1}^{m} \sigma_i x_i \right\|_\infty$$

$$= \frac{B_1}{m} \mathbb{E}_\sigma \max\left(\max_{j=1}^{d} \sum_{i=1}^{m} \sigma_i x_{ij}, \max_{j=1}^{d} \sum_{i=1}^{m} \sigma_i (-x_{ij}) \right)$$

$\in [-R_\infty, +R_\infty]$

$\frac{(2 \cdot R_\infty)^2}{4} = R_\infty^2 - SG$

$m \cdot R_\infty^2 - SG$

Massart's Lemma:

(N σ^2-SG $X_1 \sim X_N$. $\mathbb{E} \max_i X_i \le \sigma \sqrt{2 \ln N}$)

Applying Massart:

$N = 2d$

$\sigma^2 = m \cdot R_\infty^2$

$$\le \frac{B_1}{m} \cdot \sqrt{m \cdot R_\infty^2 \cdot 2 \ln(2d)}$$

$$= B_1 R_\infty \cdot \sqrt{\frac{2 \ln(2d)}{m}} .$$

Pf of the contraction inequality:

$\mathcal{F}$. $\{ \overset{\text{Lip}}{\phi} \circ g : g \in \mathcal{G} \}$ $\mathcal{G}$.

$$m \, \mathrm{Rad}_S(\mathcal{F}) = \mathbb{E}_\sigma \sup_{g \in \mathcal{G}} \sum_{i=1}^{m} \sigma_i \phi(g(z_i)) \quad (\Delta)$$

$$\overset{①}{\leq} \mathbb{E}_{\sigma} \sup_{g \in \mathcal{G}} L \cdot \sigma_1 \cdot g(z_1) + \sum_{i=2}^{m} \sigma_i \phi(g(z_i))$$

$$\overset{②}{\leq} \mathbb{E}_{\sigma} \sup_{g} L \sigma_1 g(z_1) + L \sigma_2 g(z_2) + \sum_{i=3}^{m} \sigma_i \phi(g(z_i))$$

$$\vdots$$

Goal

$$\leq \mathbb{E}_{\sigma} \sup_{g \in \mathcal{G}} \sum_{i=1}^{m} \sigma_i \cdot L \cdot g(z_i)$$

Pf of ①:

$$(\Delta) = \mathbb{E}_{\sigma_{2:n}} \Big[\frac{1}{2} \sup_{g \in \mathcal{G}} \Big(\phi(g(z_1)) + \sum_{i=2}^{m} \sigma_i \phi(g(z_i)) \Big) + \frac{1}{2} \sup_{g' \in \mathcal{G}} \Big(-\phi(g'(z_1)) + \sum_{i=2}^{m} \sigma_i \phi(g'(z_i)) \Big) \Big]$$

$$= \mathbb{E}_{\sigma_{2:n}} \frac{1}{2} \sup_{g, g' \in \mathcal{G}} \Big[\underbrace{\phi(g(z_1)) - \phi(g'(z_1))}_{\leq L \cdot |g(z_1) - g'(z_1)|} + \sum_{i=2}^{m} \sigma_i \phi(g(z_i)) + \sum_{i=2}^{m} \sigma_i \phi(g'(z_i)) \Big]$$

symmetrize wrt g, g'.

$$\sup_{g, g' \in \mathcal{G} : g(z_1) \geq g'(z_1)}$$

$$\le \sup_{g, g' \in \mathcal{G} : g(z_1) \ge g'(z_1)}$$

$$\le \mathbb{E}_{\sigma_{2:n}} \frac{1}{2} \sup_{g, g' \in \mathcal{G}} L\left(g(z_1) - g'(z_1)\right) + \sum_{i=2}^{m} \sigma_i \phi\left(g(z_i)\right) + \sum_{i=2}^{m} \sigma_i \phi\left(g'(z_i)\right)$$

split

$$= \mathbb{E}_{\sigma_{2:n}} \frac{1}{2} \sup_{g \in \mathcal{G}} \overset{\sigma_1 = 1}{L g(z_1)} + \sum_{i=2}^{m} \sigma_i \phi\left(g(z_i)\right)$$

$$+ \frac{1}{2} \sup_{g' \in \mathcal{G}} \overset{\sigma_1 = -1}{- L g'(z_1)} + \sum_{i=2}^{m} \sigma_i \phi\left(g'(z_i)\right)$$

$$= \mathbb{E}_{\sigma_{2:n}} \mathbb{E}_{\sigma_1} \sup_{g \in \mathcal{G}} L \cdot \sigma_1 g(z_1) + \sum_{i=2}^{m} \sigma_i \phi\left(g(z_i)\right).$$

$\boxtimes$.

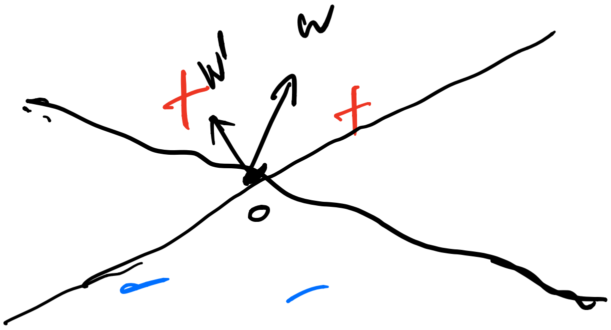
Algorithm inspired by Margin-based generalization bound:

fix $\theta = 1$.

we would like to find a w. s.t.

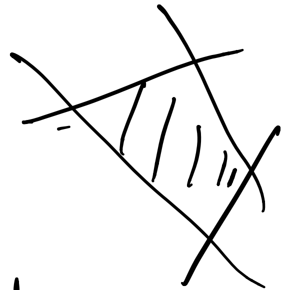
① $P_S\left(y\langle w, x\rangle \le 1 \right) = 0$.

② $\|w\|_1$ as small as possible.



W direction is better because we don't need a large scaling factor to ensure all examples to have margin ≥ 1.

$\|w\|_1$

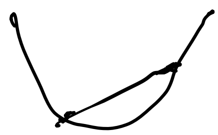


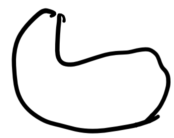
$\min \quad \|w\|_1$

s.t. for every $i \in \{1..m\}$. $\quad y_i\langle w, x_i\rangle \geq 1$.

ℓ_1 - support vector machine (SVM).

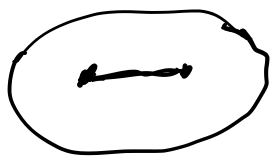
convex optimization problem:





$$\min_{x} \ f(x) \leftarrow \text{convex}$$

$$\text{s.t. } x \in K \rightarrow \text{convex set}$$



ℓ_2 - SVM.

$$\min_{w} \ \|w\|_2$$

s.t. $\forall i$. $\quad y_i\langle w, x_i\rangle \geq 1$.

Next class: margin-based generalization error bounds for ℓ_2-SVMs.