Support Vector Machine (SVM)

Oct / 26 / 2023

§1 Overview

(1) Classification Pb

Classification problem

$S \in (\mathcal{X} \times \mathcal{Y})^m$, $\mathcal{Y} = \{\pm 1\}$

SVM $h \in [\mathcal{X} \to \mathcal{Y}]$, $h(x) = \operatorname{sgn}(\langle w, x \rangle + b)$.

(2) Theoretical Analysis

① $\mathbb{P}[\forall h \in \mathcal{H}, \quad R(\mathcal{H}) \leq \underbrace{\quad}_{\text{independent of input dimension}}] \geq 1 - \delta$.

② $\mathcal{H}_{PL} = \{x \mapsto \operatorname{sgn}(\langle w, x \rangle + b) \mid w \in \mathbb{R}^n, b \in \mathbb{R}\}$, in comparison using VCDim,

$\operatorname{VCDim}(\mathcal{H}_{PL}) = n+1$

$\mathbb{P}[\forall h \in \mathcal{H}_{PL} \quad R(h) \leq \hat{R}_S(h) + \underbrace{\sqrt{\frac{2(n+1)\log(\frac{em}{n+1})}{m}} + \sqrt{\frac{\log(1/\delta)}{2m}}}_{\hookrightarrow n\text{-dependent.}}] \geq 1 - \delta$.

③ Dualisation via Lagrangian.

§2. Support Vector Machine

(1) **Linearly separable assumption**

$\mathcal{D} \in Pr(\mathcal{X} \times \mathcal{Y})$, $\mathcal{X} = \mathbb{R}^n$. s.t. $\exists w \in \mathbb{R}^n, b \in \mathbb{R}$ with

$\mathbb{P}_{(x,y) \sim \mathcal{D}}[(\langle w, x \rangle + b) y > 0] = 1$.

(2) **SVM**.

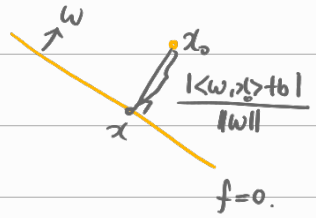
input $S = ((x_1, y_1), \cdots, (x_m, y_m))$

output $(w^*, b^*) \in \operatorname{argmin} \frac{1}{2}\|w\|^2$ subject to $y_i(\langle w, x_i \rangle + b) \geq 1 \quad \forall i \in [m]$

return $f^*(x) = \langle w^*, x \rangle + b^*$

(3) Maxgin Maximization

① $f(x) = \langle w, x \rangle + b$, $f(x) = 0$.



$x_0 \in \mathbb{R}^n$, $x_0 = x + \lambda w$

(i) $|f(x_0)| = |f(x) - f(x_0)| = |\langle w, x_0 - x \rangle|$
$= |\lambda| \|w\|^2$

$\Rightarrow \|x_0 - x\| = \|\lambda w\| = \frac{|f(x_0)|}{\|w\|}$

(ii) For any x with $f(x) = 0$,

$|f(x) - f(x_0)| = |\langle w, x_0 - x \rangle| \geq \|w\| \cdot \|x_0 - x\|$

$\Rightarrow \|x_0 - x\| \geq \frac{|f(x) - f(x_0)|}{\|w\|} = \frac{|f(x_0)|}{\|w\|}$

$\min_x \|x_0 - x\|^2$ s.t. $f(x) = 0$.

By Lagrangian multiplier method, $x - x_0 = \lambda \nabla f(x) = \lambda w$ for some λ

$\Rightarrow \|x - x_0\| = \lambda w$.

② $\sup_{w,b} \min_{i \in [m]} \rho_{w,b}(x_i) = \frac{|\langle w, x_i \rangle + b|}{\|w\|}$ subject to $y_i(\langle w, x_i \rangle + b) > 0$ $\forall i \in [m]$.

$= \frac{y_i(\langle w, x_i \rangle + b)}{\|w\|}$ //

$=$ // subject to $\min_{i \in [m]} y_i(\langle w, x_i \rangle + b) = 1$

∵) Objective is invariant to the scaling of (w, b).

$w' = \frac{w}{\eta}$, $b' = \frac{b}{\eta}$, $\eta = \min_{i \in [m]} y_i(\langle w, x_i \rangle + b) > 0$.

$\Rightarrow \min_{i \in [m]} y_i(\langle w', x_i \rangle + b') = 1$

and objective value is the same.

$$= \sup_{w,b} \frac{1}{\|w\|} \text{ s.t. } \min_{i\in[m]} y_i(\langle w, x_i\rangle + b) = 1$$

$$= \sup_{w,b} \frac{1}{\|w\|} \text{ s.t. } \min_{i\in[m]} y_i(\langle w, x_i\rangle + b) \geq 1.$$

(∵) w, b with $\min_{i\in[m]} y_i(\langle w, x_i\rangle + b) = \mu \geq 1$,

$w' = \frac{w}{\mu}$, $b' = \frac{b}{\mu}$ has $\min_{i\in[m]} y_i(\langle w', x_i\rangle + b') = 1$

and $\frac{1}{\|w'\|} = \frac{\mu}{\|w\|} \geq \frac{1}{\|w\|}$.

Thus, sup. achieved in smaller domain.

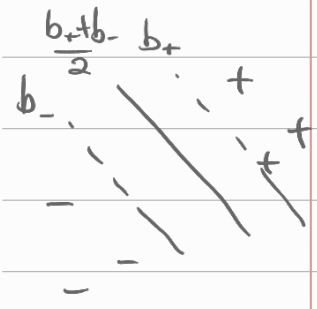
$$= \inf_{w,b} \frac{1}{2}\|w\|^2 \text{ s.t. } y_i(\langle w, x_i\rangle + b) \geq 1 \ \forall i \in [m].$$

(Question) Formulation without b?

$\min \frac{1}{2}\|w\|^2$ s.t. $\forall j \in J \ \forall i \in [m]\setminus J$

$$-1 - \langle w, x_i\rangle \geq 1 - \langle w, x_j\rangle,$$

where $J = \{j \mid y_j = +1\}$.



($\Rightarrow$) Take $b \in \left[\max_{j\in J} 1 - \langle w, x_j\rangle, \min_{i\in[m]\setminus J} -1 - \langle w, x_i\rangle\right]$.

Then, for $j \in J$, $b \geq 1 - \langle w, x_j\rangle \Rightarrow \langle w, x_j\rangle + b \geq 1$

$i \notin J$, $b \leq -1 - \langle w, x_i\rangle \Rightarrow \langle w, x_i\rangle + b \leq -1$.

($\Leftarrow$ $-(\langle w, x_i\rangle + b) \geq 1$)

($\Leftarrow$) $\langle w, x_j\rangle + b \geq 1 \ \forall j \in J$ $\langle w, x_i\rangle + b \leq -1 \ \forall i \in [m]\setminus J$

Thus, $\max_{j\in J} 1 - \langle w, x_j\rangle \leq b \leq \min_{i\in[m]\setminus J} -1 - \langle w, x_i\rangle$

§ SVM for Separable Cases

(1) $\min_{w,b} \frac{1}{2}\|w\|^2$ s.t. $y_i(\langle w, x_i\rangle + b) \geq 1 \quad \forall i \in [n]$

$\|w\|^2 = F(w,b)$

(2) Dualisation

① Dual variables $\alpha_i \geq 0$, $0 \geq 1 - y_i(\langle w, x_i\rangle + b) =: g_i(w,b)$

(Lagrangian Variables)

$\alpha = (\alpha_1, \cdots, \alpha_m) \in \mathbb{R}^m_{\geq 0}$

② $\mathcal{L}(w,b,\alpha) = \frac{1}{2}\|w\|^2 + \sum_{i=1}^m \alpha_i g_i(w,b)$

Fact

For $\forall \alpha \in \mathbb{R}^m$,

(1) $\inf_{w,b} \mathcal{L}(w,b,\alpha) \leq F(w',b') \quad \forall w',b'$ s.t. $0 \geq g_i(w',b') \ \forall i \in [m]$.

$\because$) $\mathcal{L}(w',b',\alpha) = F(w',b') + \sum_{i=1}^m \alpha_i g_i(w',b')$

$\leq F(w',b')$

Corollary

$\sup_{\alpha} \inf_{w,b} \mathcal{L}(w,b,\alpha) \leq \inf_{w',b' \text{ s.t. } 0 \geq g_i(w',b') \ \forall i \in [m]} F(w',b')$.

(2) (w^*,b^*,α^*) is a saddle of $\mathcal{L} \Leftrightarrow (w^*,b^*)$ is a soln. of the original problem.

$\mathcal{L}(w^*,b^*,\alpha) \leq \mathcal{L}(w^*,b^*,\alpha^*) \leq \mathcal{L}(w,b,\alpha^*) \quad \forall \alpha \geq 0, \ \forall w,b$.

$\because$) $\mathcal{L}(w^*,b^*,\alpha) \leq \mathcal{L}(w^*,b^*,\alpha^*)$.

$\sum_{i=1}^m \alpha_i g_i(w^*,b^*) \leq \sum_{i=1}^m \alpha_i^* g_i(w^*,b^*)$

$\alpha_i^* g_i(w^*,b^*) = 0$

① (w^*, b^*) is feasible. i.e. $g_i(w^*, b^*) \le 0 \quad \forall i \in [m]$

Otherwise, we can choose $\alpha_i > 0$ large s.t. $\mathcal{L}(w^*, b^*, \alpha) > \mathcal{L}(w^*, b^*, \alpha^*)$

② $0 \le \sum_{i=1}^{m} \alpha_i^* g_i(w^*, b^*)$

As $\alpha_i^* g_i(w^*, b^*) \le 0 \quad \forall i \in [m]$, $\alpha_i^* g_i(w^*, b^*) = 0 \quad \forall i \in [m]$

③ $F(w^*, b^*) = \mathcal{L}(w^*, b^*, \alpha^*)$

$\le \mathcal{L}(w', b', \alpha^*)$

$\le F(w', b')$ (If (w', b') feasible).

Thus, (w^*, b^*) is the optimal solution of primal

~~$\inf_{w,b} \mathcal{L}(w, b, \alpha) = \mathcal{L}(w^*, b^*, \alpha^*) \ge \mathcal{L}(w^*,$~~

④ (w^*, b^*, α^*) is the optimal solution of dual:

$\sup_{\alpha \ge 0} \inf_{w,b} \mathcal{L}(w, b, \alpha) = \mathcal{L}(w^*, b^*, \alpha^*)$.

∵) $\mathcal{L}(w^*, b^*, \alpha^*) \le \mathcal{L}(w, b, \alpha^*) \Rightarrow \mathcal{L}(w^*, b^*, \alpha^*) \le \inf_{w,b} (w, b, \alpha^*)$

$\le \sup_{\alpha \ge 0} \inf_{w,b} (w, b, \alpha)$

(≤) by weak duality.

Fact3 (w^*, b^*) is a soln of the original problem if and only if $\exists \alpha^*$ s.t. (w^*, b^*, α^*) is a saddle of the associated Lagrangian.

$\sup_{\alpha} \inf_{w,b} \mathcal{L}(w, b, \alpha)$ s.t. $\nabla_{w,b} \mathcal{L}(w, b, \alpha) = 0$

$\alpha \ge 0$.

Note (Lagrangian of SVM).

$$\mathcal{L}(w,b,\alpha) = \frac{1}{2}\|w\|^2 + \sum_{i=1}^{m} \alpha_i (1 - y_i(\langle w, x_i\rangle + b))$$

$$\nabla_w \mathcal{L}(w,b,\alpha) = w + \left(\sum_{i=1}^{m} \alpha_i (-y_i x_i)\right) = 0 \Rightarrow w = \sum \alpha_i y_i x_i$$

$$\nabla_b \mathcal{L}(w,b,\alpha) = \sum_{i=1}^{m} \alpha_i (-y_i) = 0. \Rightarrow \sum \alpha_i y_i = 0.$$

$$\mathcal{L}(w,b,\alpha) = \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j x_i^T x_j + \sum_i \alpha_i - \sum_i \alpha_i y_i \langle \sum \alpha_j y_j x_j, x_i \rangle$$

$$= -\frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle + \sum_i \alpha_i.$$

$$= \mathcal{L}(\alpha).$$

Solve $\sup_{\alpha \geq 0} \mathcal{L}(\alpha)$

Then $w^* = \sum_i \alpha_i^* y_i x_i$.

If $\exists i_0$ s.t. $\alpha_{i_0} > 0$.

By KKT, $g_{i_0}(w^*, b^*) = 1 - y_{i_0}(\langle w^*, x_{i_0}\rangle + b^*) = 0$.

$$\Rightarrow b^* = y_{i_0} - \langle w^*, x_{i_0}\rangle$$

$$= y_{i_0} - \sum_{i=1}^{m} \alpha_i^* y_i \langle x_i, x_{i_0}\rangle$$

$$f(x) = \langle w^*, x\rangle + b^* \qquad \sum_{i=1}^{m} \alpha_i y_i = 0$$

$$= \left(\sum_i \alpha_i^* y_i \langle x_i, x - x_{i_0}\rangle\right) + y_{i_0}$$

$$\rho(w^*, b^*) = \frac{\min_{i \in [m]} y_i(\langle w^*, x_i\rangle + b)}{\|w^*\|} = \frac{1}{\|w^*\|}$$

$$\rho(w^*, b^*)^2 = \frac{1}{\|w^*\|^2} = \frac{1}{\sum_{i=1}^{m} \alpha_i^2}$$

$$\alpha_i^*(1 - y_i(\langle w_i^*, x_i\rangle + b)) = 0.$$

$$\alpha_i^* - \alpha_i^* y_i (\langle w^*, x_i\rangle) = \alpha_i^* y_i b$$

§1 Remainder/Motivation. Nov / 2 / 2023

(1) **SVM for Separable Case**

$$\underset{w,b}{\operatorname{argmin}} \frac{1}{2}\|w\|^2 \quad \text{s.t.} \quad y_i(\langle w, x_i\rangle + b) \geq 1 \quad \forall i \in [m].$$

(2) **Non-Separable Case**

① Relax the constrain by slack variable

② Additional opt objective that encourage minimum violation of constraint.

§2 SVM for Non-separable Cases

(1) (Slak Variable)

$\xi_i \geq 0$ for each data point (x_i, y_i), $\xi = (\xi_1, \cdots, \xi_m)$.

(2) $p \geq 1$ $C \geq 0$,

$$(w^*, b^*, \xi^*) = \underset{w,b,\xi}{\operatorname{argmin}} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} \xi_i^p \quad \text{s.t.} \quad y_i(\langle w, x_i\rangle + b) \geq 1 - \xi_i, \quad \xi_i \geq 0 \quad \forall i.$$

return $x \mapsto \operatorname{sgn}(\langle w, x\rangle + b)$

(3) (Removing ξ ?)

Given (w, b), $\xi_i^* = \max(0, 1 - y_i(\underbrace{\langle w, x_i\rangle + b}_{f_{w,b}(x_i)}))$

$$(\xi_i^*)^p = \max(0, (1 - y_i f_{w,b}(x_i))^p) = \Phi_p(-y_i f_{w,b}(x_i)),$$

where $\Phi(a) = \max(0, (1+a)^p)$.

For $p \geq 0$, Φ is nondecreasing convex.

$$\underset{w,b}{\operatorname{argmin}} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} \Phi_p(-y_i f_{w,b}(x_i))$$

(4) Dualisation – $p=1$

$0 \geq 1-\xi_i - y_i(\langle w, x_i\rangle + b) \Rightarrow \alpha_i \geq 0$ $\quad \alpha = (\alpha_1, \dots, \alpha_m)$

$0 \geq -\xi_i \quad\quad \beta_i \geq 0 \quad\quad \beta = (\beta_1, \dots, \beta_m)$

$$L(w,b,\xi,\alpha,\beta) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^m \xi_i + \sum_{i=1}^m \alpha_i(1-\xi_i - y_i(\langle w, x_i\rangle + b)) + \sum_{i=1}^m \beta_i(-\xi_i)$$

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

$$\nabla_b L = \sum_{i=1}^m \alpha_i y_i = 0$$

$$\frac{\partial}{\partial \xi_i} L = C - \alpha_i - \beta_i = 0 \Rightarrow \alpha_i + \beta_i = C.$$

Thus, $w = \sum_{i=1}^m \alpha_i y_i x_i$, $\sum_{i=1}^m \alpha_i y_i = 0$, $\alpha_i + \beta_i = C$.

$$\inf_{w,b,\xi} L(w,b,\xi,\alpha,\beta) = \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j \langle x_i, x_j\rangle + \sum_{i=1}^m (C - \alpha_i - \beta_i)\xi_i + \sum_{i=1}^m \alpha_i - \sum_{i=1}^m \alpha_i y_i \Big\langle \sum_{j=1}^m \alpha_j y_j x_j, x_i \Big\rangle$$

$$= \sum_{i=1}^m \alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j \langle x_i, x_j\rangle$$

with $\sum \alpha_i y_i = 0$, $\quad 0 \leq \alpha_i = C - \beta_i \leq C$.

KKT condition:

$\alpha_i^*(1-\xi_i^* - y_i(\langle w^*, x_i\rangle + b^*)) = 0 \quad \forall i$

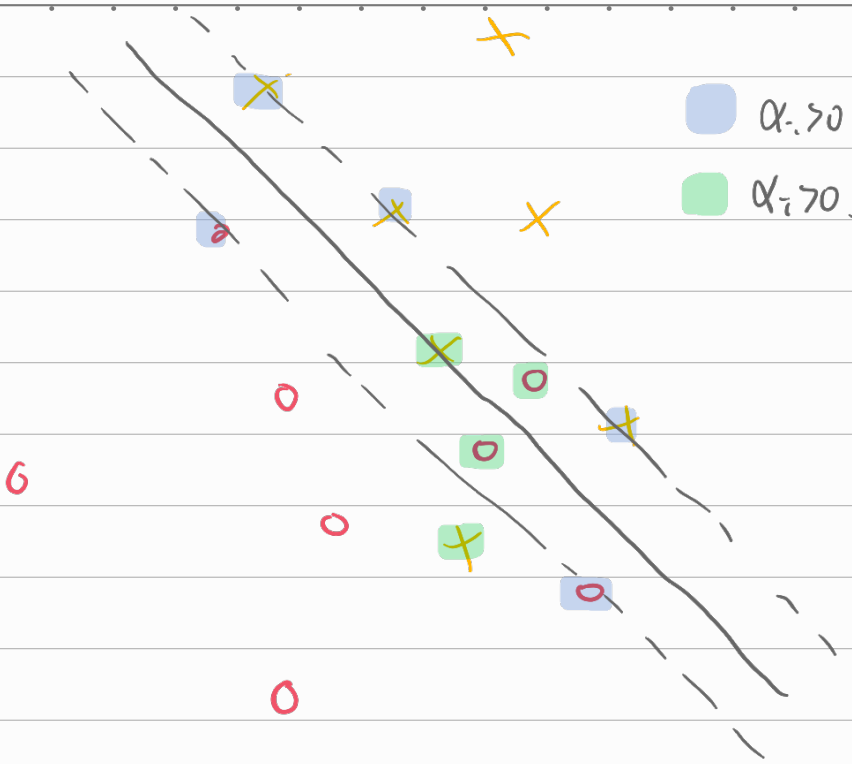
$\beta_i^*(-\xi_i^*) = 0 \quad \forall i$.

$\alpha_i^* \neq 0 \Rightarrow y_i\langle w^*, x_i\rangle + b^* = 1 - \xi_i^*$

① $\xi_i = 0 \Rightarrow y_i\langle w^*, x_i\rangle + b^* = 1$. (on marginal plane)

② $\xi_i \neq 0 \Rightarrow x_i$ is outlier

Soft Margin SVM



§3 Margin Theory

(1) Motivation

① $\mathcal{H}_{PL} = \{x \mapsto \mathrm{sgn}(\langle w, x\rangle + b) \mid w \in \mathbb{R}^n, b \in \mathbb{R}\}$

$\mathrm{VCDim}(\mathcal{H}_{PL}) = n+1$

$$\mathbb{P}_{S\sim D^m}\left[\forall h \in \mathcal{H}_{PL} \quad R(h) \le \hat{R}_S(h) + \sqrt{\frac{2\log(em/n+1)(n+1)}{m}} + \sqrt{\frac{\log 1/\delta}{2m}}\right] \ge 1-\delta.$$

Bound useless if $n \gg m$

(2) Goal

$r, r' > 0$ $D \in Pr(\mathcal{X} \times \mathcal{Y})$, $m \in \mathbb{N}$, $\delta > 0$

Assume $\mathbb{P}_{(x,y)\sim D}[\|x\| \le r] = 1$ and $\mathcal{H} = \{x \mapsto \mathrm{sgn}(\langle w, x\rangle) \mid w \in \mathbb{R}^n, \|w\| \le \Lambda\}$

Then, $\mathbb{P}_{S\sim D^m}\Big[\forall h_w \in \mathcal{H}, \forall \rho \in (0, r'], R(h_w) \le \frac{1}{m}\sum_{i=1}^{m} \max(0, 1 - \frac{y_i\langle w, x_i\rangle}{\rho}) + 4\sqrt{\frac{r^2\Lambda^2}{\rho^2 m}} + \sqrt{\frac{\log\log_2(2r'/\rho)}{m}} + \sqrt{\frac{\log(2/\delta)}{m}}\Big] \ge 1-\delta$

upper / /

Q. How can we manipulate w.r.t ρ to make bound small

$$w' = \frac{w}{\rho} \qquad \|w\| = \|w'\|\rho \le \Lambda \Rightarrow \rho \le \frac{\Lambda}{\|w'\|}.$$

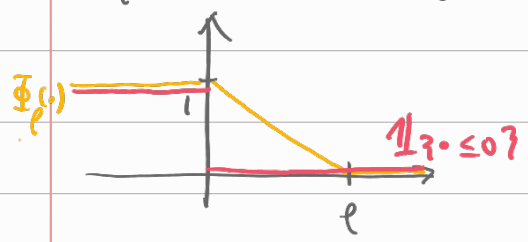
Theorem 5.8 $D \in Pr(X \times Y)$, $m \in \mathbb{N}$, $\delta > 0$, $\mathcal{F} \subseteq [X \to_m \mathbb{R}]$ s.t. $\forall x \in X$, $\exists M > 0$ with $H = \{\underbrace{\operatorname{sgn} \circ f}_{h_f} \mid f \in \mathcal{F}\}$, $\sup_{f \in \mathcal{F}} |f(x)| < M$

$$\Rightarrow \mathbb{P}_{S \sim D^m}\left[\forall f \in \mathcal{F} \;\; R(h_f) \le \hat{R}_{S,\rho}(f) + \frac{2}{\rho} R_m(\mathcal{F}) + \sqrt{\frac{\log 1/\delta}{2m}}\right] \ge 1-\delta.$$

Definition (Empirical ρ-margin Loss)

$$\hat{R}_{S,\rho}(f) := \frac{1}{m}\sum_{i=1}^{m} L_\rho(f(x_i), y_i),$$

$$L_\rho(f(x_i), y_i) := \Phi_\rho(y_i f(x_i)) = \min(1, \max(0, 1 - \tfrac{f(x_i)y_i}{\rho}))$$



(pf of thm 5.8)

$g \subseteq [Z \to [0,1]]$, $D \in Pr(Z)$

$$\mathbb{P}_{S \sim D^m}\left[\forall g \in g \;\; \mathbb{E}_{z \sim D}[g(z)] \le \frac{1}{m}\sum_{i=1}^{m} g(z_i) + 2R_m(g) + \sqrt{\frac{\log(1/\delta)}{2m}}\right] \ge 1-\delta.$$

$$g_\circ := \{(x,y) \in X \times Y \overset{\pi}{\mapsto} f(x)y \mid f \in \mathcal{F}\}.$$

$$g := \{(x,y) \mapsto \Phi_\rho(g_\circ(x,y)) \mid g_\circ \in g_\circ\}.$$

$$\mathbb{E}_{S \sim D^m}\left[\forall g \in g \;\; \mathbb{E}_{(x,y)\sim D}[g(x,y)] \le \frac{1}{m}\sum_{i=1}^{m} g(x,y) + 2R_m(g) + \sqrt{\frac{\log 1/\delta}{2m}}\right] \ge 1-\delta$$

$\forall f \in \mathcal{F}$ $\quad \mathbb{E}_{(x,y)\sim D}[\Phi_\rho(f(x)y)]$ $\quad \frac{1}{m}\sum_{i=1}^{m} \Phi_\rho(f(x)y)$

$$R(h_f) = \mathop{\mathbb{E}}_{(x,y)\sim D}\left[\mathbb{1}_{\{\operatorname{sgn} f(x) \neq y\}}\right] \le \mathop{\mathbb{E}}_{(x,y)\sim D}\left[\Phi_\rho(f(x)y)\right]$$

Note. Φ_ρ is $\frac{1}{\rho}$-Lipschitz.

By Telagrand's lemma, $\hat{R}_S(\mathcal{G}) \le \frac{1}{\rho}\hat{R}_S(\mathcal{G}_0) = \frac{1}{\rho}\hat{R}_S(\tilde{\mathcal{F}})$.

Exercise $\hat{R}_S(\mathcal{G}_0) = \hat{R}_S(\tilde{\mathcal{F}})$

(pf) $$\hat{R}_S(\tilde{\mathcal{F}}) = \frac{1}{m}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{f\in\mathcal{F}} \sum_{i=1}^m \sigma_i f(x_i)\right]$$

$$\hat{R}_S(\mathcal{G}) = \frac{1}{m}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{g\in\mathcal{G}_0} \sum_{i=1}^m \sigma_i g_0(x_i,y_i)\right]$$

$$= \frac{1}{m}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{f\in\mathcal{F}} \sum_{i=1}^m \sigma_i y_i f(x_i)\right]$$

$$= \frac{1}{m}\mathop{\mathbb{E}}_{y_i\sigma_i\sim \mathrm{Unif}\{\pm 1\}}\left[\sup_{f\in\mathcal{F}} \Sigma(\sigma_i y_i) f(x_i)\right]$$

$$= \frac{1}{m}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{f\in\mathcal{F}} \Sigma \sigma_i f(x_i)\right]$$

$$= \hat{R}_S(\mathcal{F})$$

Hence, $R_m(\mathcal{G}_0) = R_m(\tilde{\mathcal{F}})$, and $R_m(\mathcal{G}) \le \frac{1}{\rho} R_m(\mathcal{F})$.

Lemma (Talagrand)

$\mathcal{G}_0 \subseteq [Z \to \mathbb{R}]$ s.t. $\forall z \in Z$, $\exists M > 0$ with $\sup_{g_0\in\mathcal{G}_0} |g_0(z)| < M$

and $\Phi : \mathbb{R} \to \mathbb{R}$ is L-Lipschitz. and Bounded

Then, $\mathcal{G}_1 := \Phi(\mathcal{G}_0) = \{\Phi \circ g_0 \mid g_0 \in \mathcal{G}_0\}$ has

$\hat{R}_S(\mathcal{G}_1) \le L\hat{R}_S(\mathcal{G}_0)$

(pf) $$\hat{R}_S(\mathcal{G}_1) = \frac{1}{m}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{g_0\in\mathcal{G}_0} \sum_{i=1}^m \sigma_i \Phi(g_0(z_i))\right.$$

/ /

It suffices to prove

$$\mathbb{E}_{\sigma}\Big[\sup_{g_0}\Big(\sum_{i=1}^{k}\sigma_i\,\phi(g_0(z_i))+L\sum_{j=k+1}^{m}\sigma_j\,g_0(z_j)\Big)\Big]$$
$$\le \mathbb{E}_{\sigma}\Big[\sup_{g_0}\Big(\sum_{i=1}^{k-1}\sigma_i\,\phi(g_0(z_i))+L\sum_{j=k+1}^{m}\sigma_j\,g_0(z_j)\Big)\Big]$$

Which reduces to $\mathbb{E}_{\sigma_k}[\quad] \le \mathbb{E}_{\sigma_k}[\quad]$ (*) (**)

$$\sup_{g_0}\Big(\sum_{i=1}^{k-1}\sigma_i\phi(g_0(z_i)) + \phi(g_0(z_k)) + L\sum_{j=k+1}^{m}\sigma_j g_0(z_j)\Big) \quad \varphi(g_0)$$
$$\sup_{g_0}\Big(\text{〃} - \phi(g_0(z_k)) \text{〃}\Big) \quad \psi(g_0)$$

For $\varepsilon\in(0,1)$, $\exists\, g_1,g_2\in G_0$ st. $\varphi(g_1) > (1-\varepsilon)\sup_{g_0}\varphi(g_0)$
$\psi(g_2) > (1-\varepsilon)\sup_{g_0}\psi(g_0)$

Let us denote $\sum_{i=1}^{k-1}\sigma_i\,\phi(g_0(z_i)) + \sum_{j=k+1}^{m}\sigma_j\,g_0(z_j) =: u(g_0)$ $\Rightarrow \varphi(g_0) = u(g_0)+\phi(g_0(z_k))$
$\psi(g_0) =$ 〃 $-$ 〃

$$(1-\varepsilon)\times(*) = (1-\varepsilon)\Big[\tfrac{1}{2}\sup_{g_0}\big(u(g_0)+\phi(g_0(z_k))\big) + \tfrac{1}{2}\sup_{g_0}\big(u(g_0)-\phi(g_0(z_k))\big)\Big]$$
$$\le \tfrac{1}{2}\big(u(g_1)+\phi(g_1(z_k)) + u(g_2)-\phi(g_2(z_k))\big)$$
$$\le \tfrac{1}{2}\big(u(g_1)+u(g_2)+L|g_1(z_k)-g_2(z_k)|\big)$$
$$= \tfrac{1}{2}\big(u(g_1)+u(g_2)+Ls(g_1(z_k)-g_2(z_k))\big), \quad \text{with } s:=\operatorname{sgn}(g_1(z_k)-g_2(z_k))$$
$$= \tfrac{1}{2}\big(u(g_1)+Lsg_1(z_k)\big) + \tfrac{1}{2}\big(u(g_2)-Lsg_2(z_k)\big)$$
$$\le \tfrac{1}{2}\sup_{g}\big(u(g)+Lsg(z_k)\big) + \tfrac{1}{2}\sup_{g}\big(u(g)-Lsg(z_k)\big)$$
$$= \mathbb{E}_{\sigma_k}\Big[\sup_{g}\big(u(g)+L\sigma_k g(z_k)\big)\Big]$$
$$= (**)$$

As $\varepsilon>0$ is arbitrary, $(*)\le(**)$

/ /

Theorem 5.10 $S_x \in \{x \in \mathcal{X} \mid \|x\| \le r\}$. $\mathcal{F} = \{x \mapsto \langle w, x\rangle \mid \|w\| \le \Lambda\}$. Then,

$$\hat{R}_{S_x}(\mathcal{F}) \le \sqrt{\frac{r^2\Lambda^2}{m}}.$$

(pf)

$$\hat{R}_{S_x}(\mathcal{F}) = \frac{1}{m}\mathop{\mathbb{E}}_{\vec{\sigma}}\Big[\sup_{\substack{w\in\mathbb{R}^n \\ \|w\|\le\Lambda}} \sum_{i=1}^m \sigma_i \langle w, x_i\rangle\Big]$$

$$\le \frac{1}{m}\mathop{\mathbb{E}}_{\sigma}\Big[\sup \Big|\sum \sigma_i \langle w, x_i\rangle\Big|\Big]$$

$$\le \frac{1}{m}\mathop{\mathbb{E}}_{\sigma}\Big[\sup \|w\| \Big\|\sum \sigma_i x_i\Big\|\Big] \quad (C\text{-}S)$$

$$\le \frac{\Lambda}{m}\mathop{\mathbb{E}}_{\sigma}\Big[\Big\|\sum \sigma_i x_i\Big\|\Big] \quad (\|w\| \le \Lambda)$$

$$= \frac{\Lambda}{m}\mathop{\mathbb{E}}_{\sigma}\Big[\Big\|\sum \sigma_i x_i\Big\|\Big]$$

$$\le \frac{\Lambda}{m}\left(\mathbb{E}\Big[\Big\|\sum \sigma_i x_i\Big\|^2\Big]\right)^{\frac{1}{2}} \quad (\text{Jensen})$$

$$= \frac{\Lambda}{m}\left(\mathbb{E}\Big\langle \sum_i \sigma_i x_i, \sum_j \sigma_j x_j\Big\rangle\right)^{\frac{1}{2}}$$

$$= \frac{\Lambda}{m}\left(\mathbb{E}\sum_i \sigma_i^2 \|x_i\|^2 + \sum_{i\ne j} \sigma_i\sigma_j \langle x_i, x_j\rangle\right)^{\frac{1}{2}}$$

$$\le \frac{\Lambda}{m}\left(mr^2\right)^{\frac{1}{2}} = \sqrt{\frac{r^2\Lambda^2}{m}}. \quad (\|x_i\| \le r).$$

Theorem 5.9 $\mathcal{F}, D, \delta, m$ same as thm 5.8. Then for $r > 0$,

$$\mathop{\mathbb{P}}_{S\sim D^m}\Big[\forall f \in \mathcal{F}\ \forall \rho \in (0, r']\ R(h_f) \le \hat{R}_{S,\rho}(f) + \frac{4}{\rho}R_m(f) + \sqrt{\frac{\log\log(2r'/\rho)}{m}} + \sqrt{\frac{\log(2/\delta)}{2m}}\Big] \ge 1-\delta.$$

(proof) Let $\varepsilon = \sqrt{\frac{\log(2/\delta)}{2m}}$, so that $\delta = 2e^{-2m\varepsilon^2}$.

Let $\varepsilon_k = \varepsilon + \sqrt{\frac{\log k}{m}}$, $\rho_k = r'/2^k$.

By thm 5.8, $\mathbb{P}[\exists f \in \mathcal{F} \text{ s.t. } R(h_f) - \hat{R}_{S,\rho}(f) - \frac{2}{\rho_k}R_m(\mathcal{F}) - \varepsilon_k > 0] \le e^{-2m\varepsilon_k^2}$

$$\mathbb{P}\Big[\sup_{\substack{f\in\mathcal{F}\\ k\ge 1}} \text{''} \Big] \le \sum_k e^{-2m\varepsilon_k^2}$$

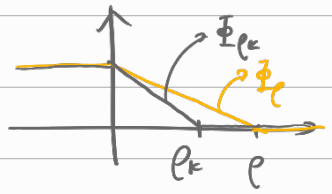
$$\le \sum_k e^{-2m(\varepsilon^2 + \frac{\log k}{m})} = \sum_k \frac{1}{k^2} e^{-2m\varepsilon^2}$$

$$= \frac{\pi^2}{6} e^{-2m\varepsilon^2} \le 2e^{-2m\varepsilon^2} = \delta.$$

For $\rho \in (0, r']$, $\rho \in (\rho_k, \rho_{k-1}]$ for some $k \geq 1$.

$\frac{1}{\rho_k} = \frac{2}{\rho_{k-1}} \leq \frac{2}{\rho}$ $\quad \sqrt{\log k} = \sqrt{\log \log_2 r'/\rho_k} \leq \sqrt{\log \log_2 2r'/\rho}$

$\hat{R}_{S,\rho_k}(f) \leq \hat{R}_{S,\rho}(f)$. :)



$$① = \sup_{\substack{f \in \mathcal{F} \\ \rho \in [0, r')}} R(h_f) - \hat{R}_{S,\rho}(f) - \frac{4}{\rho} R_m(\mathcal{F}) - \sqrt{\frac{\log \log_2 (2r'/\rho)}{m}} - \varepsilon$$

$$\leq \sup_{\substack{f \in \mathcal{F} \\ k \geq 1}} R(h_f) - \hat{R}_{S,\rho_k}(f) - \frac{2}{\rho_k} R_m(\mathcal{F}) - \sqrt{\frac{\log k}{m}} - \varepsilon \quad = ②$$

Thus, $\mathbb{P}[① > 0] \leq \mathbb{P}[② > 0] \leq 2e^{-2m\varepsilon^2} = \delta$

$\Rightarrow \mathbb{P}[① \leq 0] \geq 1 - \delta$.