1 The minimum ℓ_1-norm interpolator

$\{(x_i, y_i)\}_{i=1}^n$ $x_i \in \mathbb{R}^d$, $y_i = w_*^T x_i \in \mathbb{R}$

Consider $\min \|w\|_1$ s.t. $w^T x_i = y_i$, $i=1, \cdots n$ solution $\hat{w}_n$

$$\hat{R}_n(w) = \frac{1}{n}\sum_{i=1}^n |w^T x_i - y_i|^2, \quad R(w) = \mathbb{E}_{x,y}\left[(w^T x - y)^2\right]$$

$\|x\|_\infty \le 1$

(a) $L_B = \{|w^T x - w_*^T x|^2 : \|w\|_1 \le B\}$, $H_B = \{w^T x : \|w\|_1 \le B\}$

Prove $\widehat{Rad}_n(L_B) \lesssim (B + \|w_*\|_1)\widehat{Rad}_n(H_B)$

<u>Proof</u>. $\widehat{Rad}_n(L_B) = \mathbb{E}_\xi \sup_{\|w\|_1 \le B} \frac{1}{n}\sum_{i=1}^n \xi_i\left[(w^T - w_*^T)x_i\right]^2$

$\le 2(B + \|w_*\|_1)\widehat{Rad}_n(H_B)$

As x^2 is 2β-Lipschitz continuous in $[0, \beta]$

and $|(w^T - w_*^T)x_i| \le B + \|w_*\|_1$ #

(b) $\forall \delta \in (0,1)$, with probability $\ge 1-\delta$ over sampling of $x_1, \cdots x_n$,

$$R(\hat{w}_n) \lesssim \|w_*\|_1^2\left(\sqrt{\frac{\log(2d)}{n}} + \sqrt{\frac{\log(4/\delta)}{n}}\right)$$

<u>Proof</u>. $\widehat{Rad}_n(L_B) \lesssim (B + \|w_*\|_1)\widehat{Rad}_n(H_B)$

$\lesssim (B + \|w_*\|_1) B\sqrt{\frac{\log(2d)}{n}}$

According to Generalization error based on Rademacher complexity,

$$\sup_w |R(w) - \hat{R}(w)| \le 2\widehat{Rad}_n(L_B) + 4B\sqrt{\frac{2\log(4/\delta)}{n}}$$

$$\lesssim \|w_*\|_1^2\left(\sqrt{\frac{\log(2d)}{n}} + \sqrt{\frac{\log(4/\delta)}{n}}\right)$$

Notice that $\hat{R}(\hat{w}_n) = 0$, we are done. #

2 The reproducing kernel property

$k: \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ be a SPD kernel. $\mathcal{H}_k$: associated RKHS.

(a) if $k(x,x) \le C$ for all $x \in \mathcal{X}$, then $|f(x)| \le \sqrt{C}$ for all f in unit ball of $\mathcal{H}_k$.

Proof. We know $k(x,x') = \langle K_x, K_{x'} \rangle_{\mathcal{H}_k}$. For any $\|f\|_{\mathcal{H}_k} \le 1$,

$$
\begin{aligned}
|f(x)| &= |\langle f, K_x \rangle| \\
&\le \|f\|_{\mathcal{H}_k} \|K_x\|_{\mathcal{H}_k} \quad \text{(Cauchy - Schwarz)}
\end{aligned}
$$

But $\|K_x\|_{\mathcal{H}_k}^2 = \langle K_x, K_x \rangle_{\mathcal{H}_k} = k(x,x) \le C$

So we have $|f(x)| \le \sqrt{C}$. #

(b) MMD P, Q pro. distribution over $\mathcal{X}$

$$\mathrm{MMD}(P,Q) = \sup_{\|f\|_{\mathcal{H}_k} \le 1} \mathbb{E}_{x \sim P}[f(x)] - \mathbb{E}_{x \sim Q}[f(x)]$$

Show $\mathrm{MMD}^2(P,Q) = \mathbb{E}_{x,x' \sim P}[k(x,x')] + \mathbb{E}_{z,z' \sim Q}[k(z,z')] - 2\,\mathbb{E}_{x \sim P,\, z \sim Q}[k(x,z)]$

Proof. Let $k(x,y) = \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}_k}$. (e.g. $\varphi = K_x$)

$$
\begin{aligned}
\mathrm{MMD}(P,Q) &= \sup_{\|f\|_{\mathcal{H}_k} \le 1} \mathbb{E}_{x \sim P}[f(x)] - \mathbb{E}_{Y \sim Q}[f(Y)] \\
&= \sup_{\|f\|_{\mathcal{H}_k} \le 1} \mathbb{E}_{x \sim P} \langle f, \varphi(X) \rangle_{\mathcal{H}_k} - \mathbb{E}_{Y \sim Q} \langle f, \varphi(Y) \rangle_{\mathcal{H}_k} \\
&= \sup_{\|f\|_{\mathcal{H}_k} \le 1} \langle f, \mathbb{E}_{x \sim P}[\varphi(X)] \rangle_{\mathcal{H}_k} - \langle f, \mathbb{E}_{Y \sim Q}[\varphi(Y)] \rangle_{\mathcal{H}_k} \\
&= \sup_{\|f\|_{\mathcal{H}_k} \le 1} \langle f, \mathbb{E}_{x \sim P}\varphi(X) - \mathbb{E}_{Y \sim Q}\varphi(Y) \rangle \\
&= \| \mathbb{E}_{x \sim P}\varphi(X) - \mathbb{E}_{Y \sim Q}\varphi(Y) \|_{\mathcal{H}_k}
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{MMD}^2(P,Q) = {} & \langle \mathbb{E}_{x \sim P}\varphi(X), \mathbb{E}_{x' \sim P}\varphi(X') \rangle_{\mathcal{H}_k} + \langle \mathbb{E}_{Y \sim Q}\varphi(Y), \mathbb{E}_{Y' \sim Q}\varphi(Y') \rangle_{\mathcal{H}_k} \\
& - 2 \langle \mathbb{E}_{X \sim P}\varphi(X), \mathbb{E}_{Y \sim Q}\varphi(Y) \rangle_{\mathcal{H}_k} \\
= {} & \mathbb{E}_{x,x' \sim P}[k(x,x')] + \mathbb{E}_{z,z' \sim Q}[k(z,z')] - 2\,\mathbb{E}_{x \sim P,\, z \sim Q}[k(x,z)] \quad \#
\end{aligned}
$$

3. L^∞ approximation of two-layer NN

$\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$ $f \in C(\mathbb{S}^{d-1})$ target function

$\exists$ prob. distribution ρ s.t. $f(x) = \mathbb{E}_{(a,b)\sim\pi}[a\sigma(b^T x)]$, $b \in \mathbb{S}^{d-1}$, $|a| \le 1$ a.s.

σ = ReLu.

(a) $h_x : [-1,1] \times \mathbb{S}^{d-1} \mapsto \mathbb{R}$, $h_x(a,b) = a\sigma(b^T x)$ $\mathcal{H} = \{h_x : \|x\|_2 \le 1\}$

Prove $\widehat{\mathrm{Rad}}_m(\mathcal{H}) \le \frac{2}{\sqrt{m}}$

Proof. $\widehat{\mathrm{Rad}}_m(\mathcal{H}) = \mathbb{E}_\xi \sup_{h_x \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^m \xi_i a_i \sigma(b_i^T x_i) = \frac{1}{m} \mathbb{E}_\xi \sup_{h_x \in \mathcal{H}} \sum_{i=1}^m \mathrm{sign}(a)(|a| b_i^T x)$

$= \frac{1}{m} \mathbb{E}_\xi \sup_{h_x \in \mathcal{H}} \sum_{i=1}^m \xi_i' \sigma(|a_i| b_i^T x) \le \beta\, \widehat{\mathrm{Rad}}_m(\mathcal{H})$ by concentration lemma.

where $\mathcal{H} = \{w^T x : \|w\|_2 \le 1\}$ and $\beta = 2$ (Lip Const for Relu for $-1 \le x \le 1$)

We also have $\widehat{\mathrm{Rad}}_m(\mathcal{H}) \le \sqrt{\frac{1}{m}}$ (Linear class)

So $\widehat{\mathrm{Rad}}_m(\mathcal{H}) \le \frac{2}{\sqrt{m}}$. #

(b) Let $(a_i, b_i) \overset{iid}{\sim} \pi$ $\forall \delta \in (0,1)$, with probability $1-\delta$,

$$\sup_{x \in \mathbb{S}^{d-1}} \left| \frac{1}{m} \sum_{i=1}^m a_i \sigma(b_i^T x) - f(x) \right| \lesssim \frac{1}{\sqrt{m}} + \sqrt{\frac{\log(4/\delta)}{m}}$$

Proof. $0 \le f \le 1$ a.s.

From Generalization error based on Rademacher complexity,

$$\sup_{x \in \mathbb{S}^{d-1}} \left| \frac{1}{m} \sum_{i=1}^m a_i \sigma(b_i^T x) - f(x) \right|$$

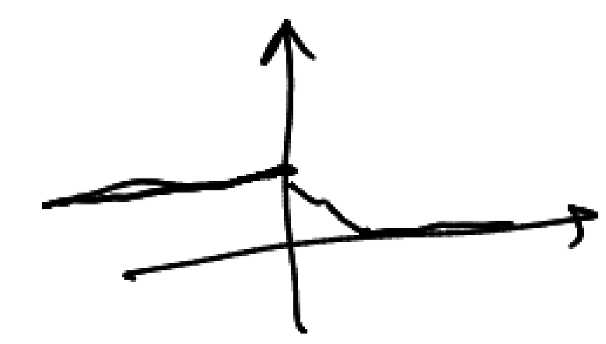
$$\le 2\widehat{\mathrm{Rad}}_m(\mathcal{H}) + 4\sqrt{\frac{2\log(4/\delta)}{m}}$$

$$\lesssim \frac{1}{\sqrt{m}} + \sqrt{\frac{\log(4/\delta)}{m}}$$ #

4 Margin-based bounds for classification

$S=\{(x_i,y_i)\}$, $y_i^2=1$ $\quad \hat{n}_\gamma(f)=|\{i\in[n]: f(x_i)y_i<\gamma\}|$

$l_{0-1}(t)=\begin{cases}1 & \text{if } t<0\\ 0 & \text{else}\end{cases} \qquad l_\gamma(t)=\begin{cases}1 & \text{if } t<0\\ 1-t/\gamma & \text{if } 0\le t<\gamma\\ 0 & \text{if } t\ge\gamma\end{cases}$



$R_{0-1}(f)=\mathbb{E}_{x,y}[l_{0-1}(f(x)y)]$

(a) $R_\gamma(f)=\mathbb{E}_{x,y}[l_\gamma(f(x)y)]$ $\quad \hat{R}_\gamma(f)=\frac{1}{n}\sum_{i=1}^{n} l_\gamma(f(x_i)y_i)$

$\forall f\in\mathcal{F},\ \gamma>0,\ R_{0-1}(f)\le R_\gamma(f),\ \hat{R}_\gamma(f)\le\frac{\hat{n}_\gamma(f)}{n}$

Proof. As $l_{0-1}(t)\le l_\gamma(t)$ for all $t\in\mathbb{R}$

$\mathbb{E}_{x,y}\, l_\gamma(f(x)y)\ge\mathbb{E}_{x,y}\, l_{0-1}(f(x)y)$, $R_{0-1}(f)\le R_\gamma(f)$

$l_\gamma(t)=0$ for all $t\ge\gamma$ and $l_\gamma(t)\le 1$ for all $t\in\mathbb{R}$.

So $\hat{R}_\gamma(f)=\frac{1}{n}\sum_{i=1}^{n} l_\gamma(f(x_i)y_i)\le\frac{1}{n}\hat{n}_\gamma(f)$. #

(b) $\mathcal{G}=\{(x,y)\mapsto f(x)y: f\in\mathcal{F}\}$ $\quad \mathcal{L}_\gamma=\{(x,y)\mapsto l_\gamma(f(x)y): f\in\mathcal{F}\}$

Show that $\widehat{\mathrm{Rad}}_n(\mathcal{L}_\gamma)\le\frac{1}{\gamma}\widehat{\mathrm{Rad}}_n(\mathcal{F})$

Proof. $\mathcal{L}_\gamma=l_\gamma\circ\mathcal{G}$

By contraction lemma,

$$\widehat{\mathrm{Rad}}_n(\mathcal{L}_\gamma)\le \mathrm{Lip}(l_\gamma)\,\widehat{\mathrm{Rad}}_n(\mathcal{G})$$

$$=\frac{1}{\gamma}\mathbb{E}_\xi\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\xi_i f(x_i)y_i$$

$$=\frac{1}{\gamma}\mathbb{E}_{\xi'}\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\xi_i' f(x_i)$$

$$=\frac{1}{\gamma}\widehat{\mathrm{Rad}}_n(\mathcal{F})$$

#

where $\xi,\xi'\sim B(\frac{1}{2},n)$.

(c) Fix $\gamma > 0$, $\delta \in (0,1)$. With prob. $\geq 1-\delta$ over sampling of S, $\forall f \in \mathcal{F}$

We have $R_{0\text{-}1}(f) \lesssim \frac{\hat{n}_\gamma(f)}{n} + \frac{1}{\gamma}\widehat{\text{Rad}}(\mathcal{F}_1) + \sqrt{\frac{\log(2/\delta)}{n}}$

Proof. We have

$$\begin{cases} R_{0\text{-}1}(f) \leq R_\gamma(f) \\ \sup_{f \in \mathcal{F}} |R_\gamma(f) - \hat{R}_\gamma(f)| \leq 2\,\text{Rad}_n(\mathcal{L}_\gamma) + \sqrt{\frac{2\log(2/\delta)}{n}} \\ 0 \leq \hat{R}_\gamma(f) \leq \frac{\hat{n}_\gamma(f)}{n} \\ \text{Rad}_n(\mathcal{L}_\gamma) = \mathbb{E}_{x,y}\widehat{\text{Rad}}_n(\mathcal{L}_\gamma) \leq \frac{1}{\gamma}\widehat{\text{Rad}}_n(\mathcal{F}_1) \end{cases}$$

Thus $R_{0\text{-}1}(f) \leq R_\gamma(f) \leq \hat{R}_\gamma(f) + 2\,\text{Rad}_n(\mathcal{L}_\gamma) + \sqrt{\frac{2\log(2/\delta)}{n}}$

$$\leq \frac{\hat{n}_\gamma(f)}{n} + \frac{1}{\gamma}\widehat{\text{Rad}}_n(\mathcal{F}_1) + \sqrt{\frac{2\log(2/\delta)}{n}} \quad . \#$$

(d) $\exists f^* \in \mathcal{F}_1$ s.t. $\mathbb{P}_{x,y}\{f^*(x)y \geq \gamma^*\} = 1$

$$\hat{f} = \arg\max_{f \in \mathcal{F}_1} \min_{i \in [n]} f(x_i)y_i$$

Show that $R_{0\text{-}1}(\hat{f}) \leq \frac{1}{\gamma^*}\widehat{\text{Rad}}_n(\mathcal{F}_1) + \sqrt{\frac{\log(2/\delta)}{n}}$

Proof.

Because we have $\mathbb{P}_{x,y}(f^*(x)y \geq \gamma^*) = 1$

We have $\min_{i \in [n]} \hat{f}(x_i)y_i \geq \min_{i \in [n]} f^*(x_i)y_i \geq \gamma^*$ a.s.

This implies $n_\gamma(\hat{f}) = 0$ a.s.

Using results in (c), we have

$$R_{0\text{-}1}(\hat{f}) \lesssim \frac{1}{\gamma^*}\widehat{\text{Rad}}_n(\mathcal{F}_1) + \sqrt{\frac{\log(2/\delta)}{n}} \quad \#$$