Recap

Want to show: for every $h \in \mathcal{H}$, for every $\mathcal{D}$, for every $\delta > 0$, Fix margin $\rho > 0$.

$$R(h) \leq \hat{R}_{S,\rho}(h) + \frac{2}{\rho} Rad_S(\mathcal{H}) + 3\sqrt{\frac{\log(2/\delta)}{2m}},$$

with probability $> 1-\delta$.

Take $\Phi(S) = \sup_{h \in \mathcal{H}} \left(R(h) - \hat{R}_S(h)\right)$

Proof:

Recall:

$$R(h) = \mathbb{E}_{\underset{(x,y)}{z \sim \mathcal{D}}} \mathbb{1}_{\{y \neq h(x)\}}$$

$$\hat{R}_S(h) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{\{y_i \cdot h(x_i) < 0\}}$$

$$S = \{(x_i, y_i)\}_{i=1}^{m}$$

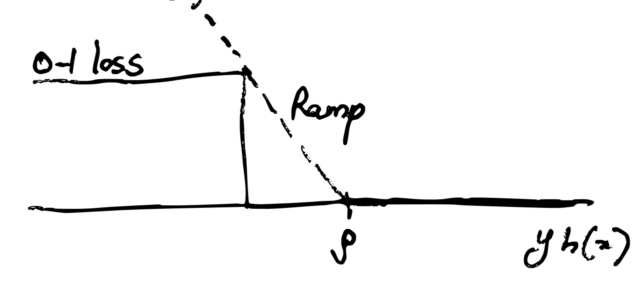
$$\mathbb{E}_{S \sim \mathcal{D}^m} \hat{R}_S(h) = R(h)$$

Suppose we proved

$$R(h) < \hat{R}_S(h) + \frac{2}{\rho} Rad_S(\mathcal{H}) + 3\sqrt{\frac{\log 2/\delta}{2m}} \rightarrow \text{A}$$

Ramp/
Margin loss

$$\hat{R}_{S,\rho}(h) = \min\left\{1, \max\left(0, 1 - \frac{y h(x)}{\rho}\right)\right\}$$



If A, we are done

$$R(h) < \hat{R}_S(h) + \cdots$$
$$< \hat{R}_{S,\rho}(h) + \cdots$$

(In practice, we solve ERM using ramp/hinge loss)

Proof sketch

$$\boxed{\text{Want: } R(h) < \hat{R}_S(h) + 2Rad_S(\mathcal{H}) + 3\sqrt{\frac{\log 2/\delta}{2m}}}$$

Today

and use

$$Rad_S(\mathcal{H}) < \frac{r\Lambda}{\sqrt{m}}$$

$$r = \sup_x \|x\| \qquad \Lambda = \sup_w \|w\|$$

$$\ell(z, h) = \mathbb{1}_{\{h(x) \neq y\}}$$

$$0 \leq \ell(z,h) \leq 1$$

$$0 \leq \hat{R}_S(h) \leq 1$$

$$0 \leq R_S(h) \leq 1 \qquad R_S(h) = \mathbb{E}_{z \sim \mathcal{D}} \ell(z,h)$$

In more generality, we will prove a Rademacher Complexity-based generalization bound when $\ell \in [0,1]$.

(Chapter 3 of Mohri + Chapter 5 of Mohri).

(Also for algorithmic-stability - different idea, depending on learning algorithm for generalization - useful inequality Mcdiarmid's inequality)

1) $\Phi(S) = \sup_{h \in \mathcal{H}} \hat{R}_S(h) - R_S(h)$

2) Apply M. inequality to $\Phi(S)$

Mcdiarmid's inequality

$S = z_1, \ldots, z_m$

If S' differs from S in one element such that

$$|\Phi(S) - \Phi(S')| \leq c \text{ holds}$$

for every $S, S' \sim D^m$,

Then, for any $t > 0$,

$$\Pr\left(\Phi(S) - \mathbb{E}_S \Phi(S) > t \right) \leq e^{-\frac{2t^2}{mc^2}}$$

ie with probability $> 1 - \delta$,

$$\Phi(S) - \mathbb{E}_S \Phi(S) < \sqrt{\frac{mc^2}{2} \log 1/\delta}$$

$$\Phi(S) = \sup_{h \in \mathcal{H}} \hat{R}_S(h) - R_S(h)$$

$$|\Phi(S) - \Phi(S')|$$

$$= \left| \sup_{h \in \mathcal{H}} \left(\hat{R}_S(h) - R_S(h)\right) - \sup_{h \in \mathcal{H}} \left(\hat{R}_{S'}(h) - R_{S'}(h)\right) \right|$$

$$\leq \left| \sup_{h \in \mathcal{H}} \left(\hat{R}_S(h) - \hat{R}_{S'}(h)\right) \right|$$

($\sup A - \sup B \leq \sup(A-B)$)

$$= \left| \frac{1}{m} \sup_{h \in \mathcal{H}} \sum_{i=1}^{m} \left(\ell(z_i, h) - \ell(z_i', h)\right) \right|$$

$$< \frac{1}{m}$$

(Φ satisfies B.D. property from Mc. ineq).

Apply Mc. in. : with probability $1-\delta$

$$\underbrace{\sup_{h \in \mathcal{H}} R_S(h) - \hat{R}_S(h)}_{\Phi(S)} \leq \underline{\mathbb{E}[\Phi(S)]} + \sqrt{\frac{\log 1/\delta}{2m}}$$

$$\mathbb{E}_S[\Phi(S)] = \mathbb{E}_{S \sim \mathcal{D}^m} \sup_{h \in \mathcal{H}} \left(R_S(h) - \hat{R}_S(h)\right)$$

$$= \mathbb{E}_S \sup_{h \in \mathcal{H}} \left(\mathbb{E}_{S'} \hat{R}_{S'}(h) - \hat{R}_S(h)\right)$$

$$\leq \mathbb{E}_{S,S'} \sup_{h \in \mathcal{H}} \left(\hat{R}_{S'}(h) - \hat{R}_S(h)\right)$$

($\because \sup(A+B) \leq \sup A + \sup B$)

$$= \mathbb{E}_{S,S'} \sup_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \left(\ell(z_i', h) - \ell(z_i, h)\right)$$

$$= \mathbb{E}_{\sigma, S, S'} \sup_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i \left(\underline{\ell(z_i', h)} - \underline{\ell(z_i, h)}\right)$$

($\because \mathbb{E}_{S,S'}$, symmetric diff)

$$\leq 2 \mathbb{E}_{\sigma, S} \sup_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i \ell(z_i, h) \rightarrow B$$

($\sup(A+B) \leq \sup A + \sup B$)

Recall

$$Rad_S(\mathcal{H}) = \mathbb{E}_\sigma \sup_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i h(x_i)$$

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

$$\mathbb{E}_S \Phi(S) \leq 2 \mathbb{E}_{\sigma, S} \sup_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i \ell(z_i, h)$$

$$= 2 \mathbb{E}_{\sigma, S} \sup_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i \left(\frac{1 - y_i h(x_i)}{2}\right)$$

($\because$ 0-1 loss)

$$= \mathbb{E}_{\sigma, S} \sup_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i y_i h(x_i)$$

($y_i = \pm 1$)

$$\leq \mathbb{E}_{\sigma, S} \sup_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i h(x_i)$$

(h belongs to linear fns).

$$= Rad_m(\mathcal{H})$$

Implications:

Bias-complexity tradeoff

General form

$$R(h) \leq \hat{R}_S(h) + Rad(\text{sgn} \circ \mathcal{H})$$

with high probability

A class $\mathcal{H}$ is PAC-learnable with a "sample complexity" $m_{\mathcal{H}} : \mathbb{R}^{2+} \to \mathbb{N}$ if $m \geq m_{\mathcal{H}}(\varepsilon, \delta) \Rightarrow$

$$|R(h) - \hat{R}_S(h)| < \varepsilon$$

with probability $1-\delta$, for any $\mathcal{D}$.

For $h_S \in ERM(S, A)$

$$R(h_S) = \underbrace{R(h_S) - \min_{h \in \mathcal{H}} R(h)}_{\text{estimation error}} + \underbrace{\min_{h \in \mathcal{H}} R(h)}_{\text{approximation error}}$$

↓ depends on $\mathcal{H}$

↑ comp of H, ↓ approx error

Sample complexity: how representative is empirical error of $R(h)$

Computational complexity

Boosting

Freund Shapire 1995 , 1999

If you have a bunch of "rough" hypotheses that are "easy to Learn" (small computational complexity) can you combine ("boost") them to get better generalization/ training error

→ Approximation error .BAD
⇒ Estimation error good ✓

Practical algorithm: Adaboost
"adaptive"

→ Iterative algorithm I/P: $S = \{(x_i, y_i)\}_{i=1}^{m}$
algorithm for "weak" learning
→ how do you combine hypotheses?
"weak rules"

"Weak" learning returns $f_t(x) = \pm 1$

$$\underline{h_T(x)} = \text{sgn}\left(\sum_{t=1}^{T} w_t \underline{f_t}(x)\right)$$

(after T rounds of boosting)

O/P: h_T

There are no hyper parameters!
except for T