HW1

Convexify the constraint

$$\rho = \frac{1}{\|\omega\|}$$

$$\Rightarrow \quad \|\omega\|^2 = \frac{1}{\rho^2}$$

$g(\omega) \leq 0$: $\lambda g(\omega)$ is added to Lagrangian where $\lambda \geq 0$.

$g(\omega) = 0$

↳ $g(\omega) \geq 0 \Rightarrow -g(\omega) \leq 0$ ($\lambda g(\omega)$ to Lagrangian where $\lambda \leq 0$)

and $g(\omega) \leq 0$

↓

$\lambda g(\omega)$ to Lagrangian where $\lambda \geq 0$

For an equality constraint, we add $\lambda g(\omega)$ to Lagrangian. $\lambda \in \mathbb{R}$.

Set up

Empirical risk minimization

Ingredients

Data: (x_i, y_i) $i \in [m]$ $x \in \mathcal{X}$ $y \in \mathcal{Y}$ $i = 1, 2, \ldots, m$.

$(x_i, y_i) \sim \mathcal{D}$ iid.

Loss: $\ell : (\mathcal{X} \times \mathcal{Y}) \times \mathcal{H} \to \mathbb{R}^+$

$\mathcal{H}$: hypothesis class or space of functions over which we want to learn to fit the data.

Mean-square loss:

squared loss $\ell((x, y), h) = \| h(x) - y \|^2$

ERM:

$$\frac{1}{m} \sum_{x,y \in S} \ell((x,y), h) = \frac{1}{m} \sum_{(x,y) \in S} \| h(x) - y \|^2$$

MSE (pytorch. func.)
error

ERM: solve optimization:

$$\min_{h \in \mathcal{H}} \frac{1}{m} \sum_{(x,y) \in S} \ell((x,y), h)$$

Sample set

Prob. dist of sample

$S \sim \mathcal{D}^m = \mathcal{D} \times \ldots \times \mathcal{D}$

joint prob. dist of $\{(x_1, y_1), (x_2, y_2) \ldots, (x_m, y_m)\}$

Goal: $\min_{h \in \mathcal{H}} \mathbb{E}_{(x,y) \sim \mathcal{D}} \ell((x,y), h)$ = generalization error over $\mathcal{H}$

ML: h that "generalizes".

$\mathcal{D}$: unknown!

- Replacing / approximating $\mathbb{E}_{(x,y) \sim \mathcal{D}} \ell((x,y), h)$ (R.V.) with sample average.
- conver. in SLLN. Emp risk of h $\xrightarrow{m \to \infty}$ gener. error of h.

Emp. risk: $\hat{R}_S(h) = \frac{1}{m} \sum_{i=1}^{m} \ell((x_i, y_i), h)$

$S = \{(x_i, y_i)\}_{i \in [m]}$ $= \frac{1}{m} \sum_{(x,y) \in S} \ell((x,y), h)$

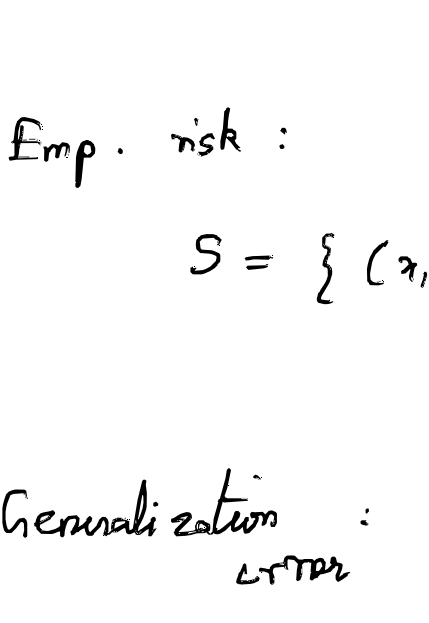
Generalization error: $R(h) = \mathbb{E}_{(x,y) \sim \mathcal{D}} \ell((x,y), h)$

Generalization gap:

$R(h) - \hat{R}_S(h)$

Overfitting: $\hat{R}_S(h) \approx 0$ but $R(h)$ is $\uparrow$.

Memorization: open questions! (generative models)



$y = \begin{cases} +1 & \text{when } x \in ▨ \\ -1 & \text{o.w.} \end{cases}$

$Ar(▨) = \frac{1}{2} Ar(\square)$

$(x, y) \sim Unif(\square \times \{-1, 1\})$.

My predictor:

$h(x) = \begin{cases} y_i, & x = x_i \\ 1 & \text{o.w.} \end{cases}$

Given $S = \{(x_i, y_i)\}_{i=1}^m$, $\hat{R}_S(h) = ? = 0$.

$R(h) = \frac{1}{2}$.

Realizability assumption: $\exists\ h^* \in \mathcal{H}$ s.t.

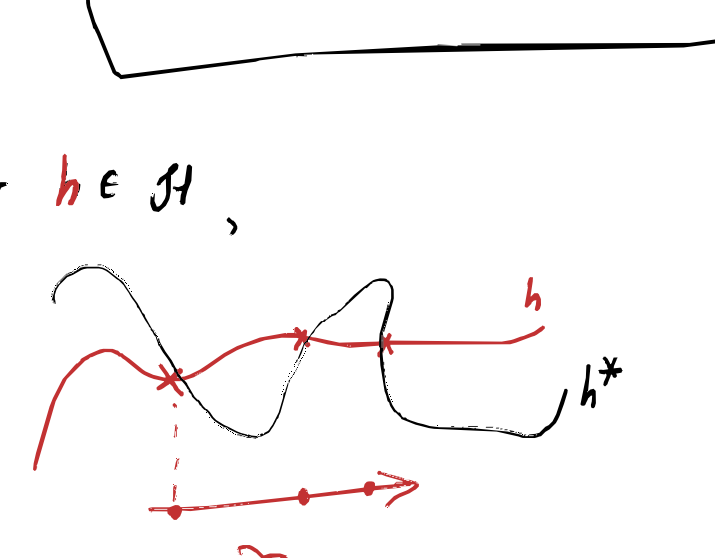
$R(h^*) = 0$.

$\Rightarrow$

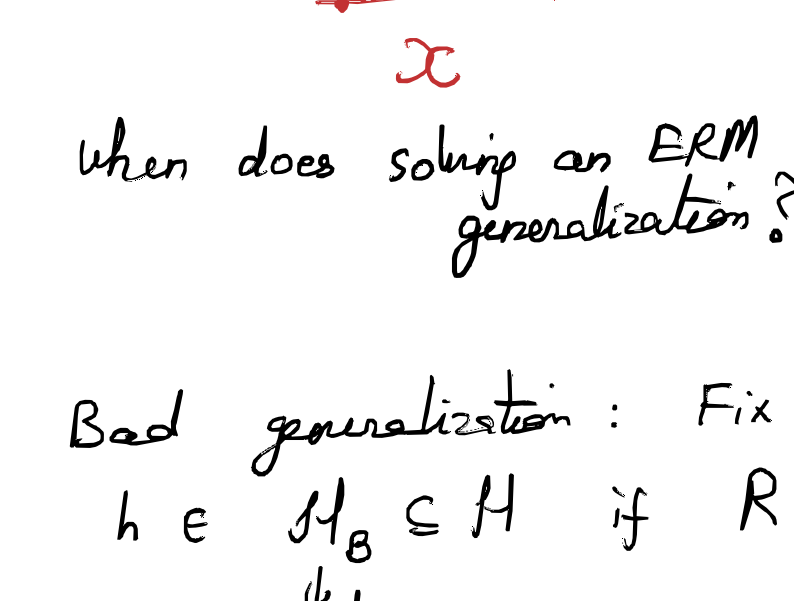
For for almost every S, $\hat{R}_S(h_S) = 0$ if h_S is ERM solution for S

ie. $h_S = \arg\min_{h \in \mathcal{H}} \frac{1}{m} \sum_{(x,y) \in S} \ell((x,y), h)$

$\min_{h \in \mathcal{H}} \frac{1}{m} \sum_{(x,y) \in S} \ell((x,y), h) = 0$.



For $h \in \mathcal{H}$,



When does solving an ERM lead to generalization?

Bad generalization: Fix $\varepsilon > 0$.

$h \in \mathcal{H}_B \subseteq \mathcal{H}$ if $R(h) > \varepsilon$.

bad hypothesis class

$\mathbb{P}_r(R(h) > \varepsilon) = ?$

$\mathcal{D}^m(\text{bad samples}) = \mathcal{D}^m(\{S \in (\mathcal{X} \times \mathcal{Y})^m : R(h_S) > \varepsilon\}) = ?$

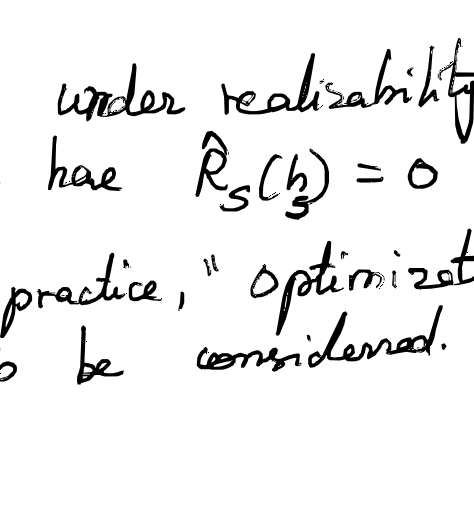
$\{S : R(h_S) > \varepsilon\} \subseteq \{S : \hat{R}_S(h) = 0 \text{ and } h \in \mathcal{H}_B\}$

$\mathcal{D}^m(\text{bad samples}) \leq \mathcal{D}^m(\{S : \hat{R}_S(h) = 0 \ \& \ h \in \mathcal{H}_B\})$

Caveat: under realizability assumption, we have $\hat{R}_S(h_S) = 0$ a.s.

But in practice, "optimization error" needs to be considered.

$\mathcal{D}^m(\text{bad samples}) \leq \bigcup_{h \in \mathcal{H}_B} \mathcal{D}^m(\{S : \hat{R}_S(h) = 0\})$



Fix $h \in \mathcal{H}_B$. ($\mathcal{H}_B = \{h \in \mathcal{H} : R(h) > \varepsilon\}$)

$\mathcal{D}^m(\{S : \hat{R}_S(h) = 0\})$

$= \mathcal{D}^m(\{S = \{(x_i, y_i)\}_{i \in [m]} : h(x_i) = y_i \ \forall i \in [m]\})$

$= (\mathcal{D}(\{(x, y) : h(x) = y\}))^m$

(Read as Prob. of finding $(x, y) \sim \mathcal{D}$ s.t. $y = h(x)$)

$\mathcal{D}$: prob. measure

$\mathcal{D}(\{\ \}) =$ Pr of "event" represented by set (Borel set)

$= \mathbb{E}_{(x,y) \sim \mathcal{D}} \mathbb{1}_{\{\ \}}((x, y))$

$h \in \mathcal{H}_B$ $\mathcal{X} \times \mathcal{Y}$

$\mathcal{D}(\{(x,y) : h(x) = y\})$ $\mathcal{D}$: data dist.

$= \mathbb{E}_{(x,y) \sim \mathcal{D}} \mathbb{1}_{\{(x,y) : h(x) = y\}}((x, y))$

(Indicator function $\mathbb{1}_A((x, y)) = \begin{cases} 1 & (x,y) \in A \\ 0 & \text{o.w.} \end{cases}$)

$\mathcal{D}(\{(x,y) : h(x) \neq y\}) = 1 - \mathcal{D}(\{(x,y) : h(x) = y\})$

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

$= \mathbb{E}_{(x,y) \sim \mathcal{D}} \ell((x,y), h)$ (zero-one loss)

$= R(h) > \varepsilon$

$\Rightarrow (\mathcal{D}(\{(x,y) : h(x) = y\}))^m \leq (1 - \varepsilon)^m$

$\Rightarrow \bigcup_{h \in \mathcal{H}_B} (\mathcal{D}(\{(x,y) : h(x) = y\}))^m \leq |\mathcal{H}_B| (1 - \varepsilon)^m$

(Recall Unionbound) Cardinality of set $\mathcal{H}_B$

We have shown,

$\mathcal{D}^m(\{\text{bad samples}\}) \leq |\mathcal{H}_B| (1 - \varepsilon)^m$

$\leq |\mathcal{H}| e^{-\varepsilon m}$

$((1 - \varepsilon)^m \leq e^{-m\varepsilon})$.

In other words,

$1 - \mathcal{D}^m(\{\text{bad samples}\}) \geq 1 - |\mathcal{H}| e^{-\varepsilon m}$

For every $\varepsilon > 0$, with Prob $\geq 1 - |\mathcal{H}| e^{-\varepsilon m}$, the generalization error $R(h_S) < \varepsilon$ for the ERM h_S of S

Set $\delta = |\mathcal{H}| e^{-\varepsilon m}$

$\log \delta = \log |\mathcal{H}| - \varepsilon m$

$\varepsilon m = \log \frac{|\mathcal{H}|}{\delta}$

$\varepsilon = \frac{1}{m} \log \frac{|\mathcal{H}|}{\delta}$

Alt: For any $\delta \in (0, 1)$, with Prob $\geq 1 - \delta$ over training samples $S \sim \mathcal{D}^m$, $R(h_S) < \frac{1}{m} \log \frac{|\mathcal{H}|}{\delta}$ for an ERM h_S over the training samples S.

Probably approximately correct bounds (PAC)