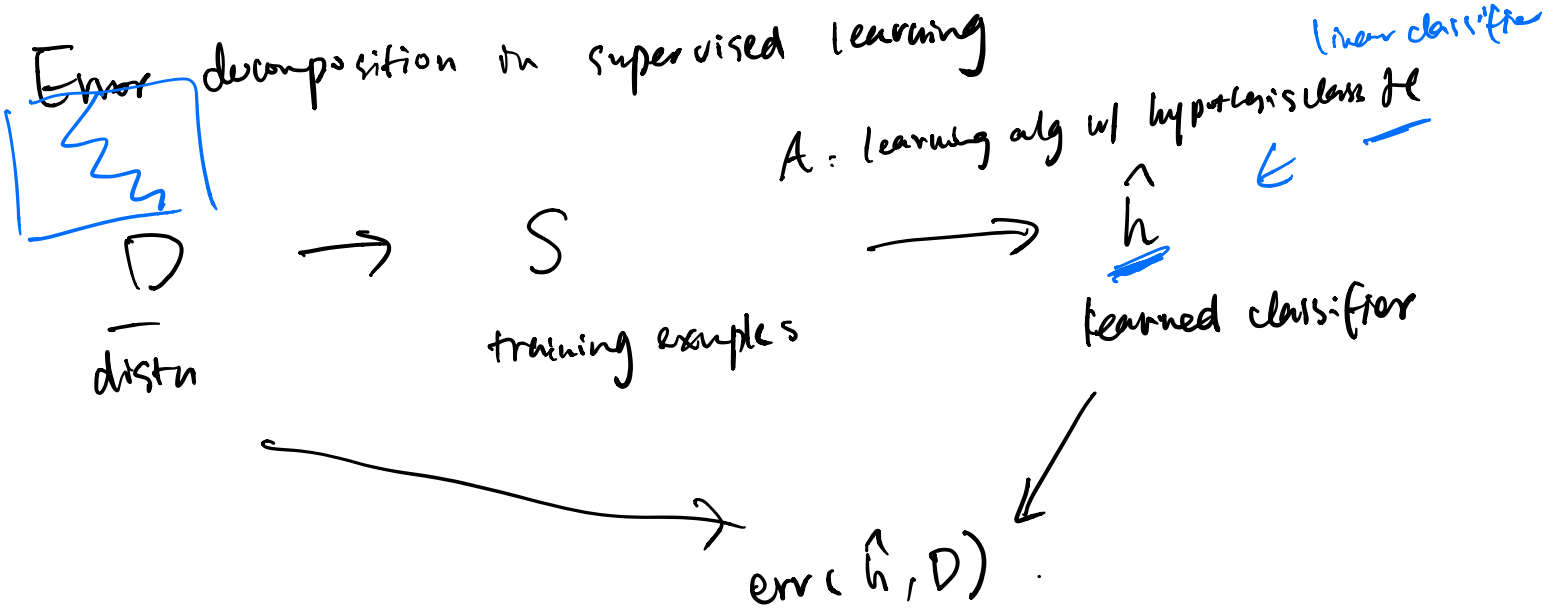
Model Selection

fixed hypothesis class $\mathcal{H}$

choose ---- $\mathcal{H}$ that are helpful for learning.

Error decomposition in supervised learning



Q: What are some important factors that contribute to the generalization error of $\hat{h}$?

- representativeness of training example
- complexity of $\hat{h}$, ($\mathcal{H}$)
- optimization accuracy of A.
- expressiveness of $\mathcal{H}$ relative to D.

Notation: $h' = \underset{h \in \mathcal{H}}{\operatorname{argmin}}\ err(h, S)$

$h^* = \underset{h \in \mathcal{H}}{\operatorname{argmin}}\ err(h, D)$

$err(\hat{h}, D)$

Thm: w.p. $1-\delta$, $\hat{h} = h'$ (ERM)

$$\text{err}(\hat{h}, D) \le \underbrace{\Sigma_{gen}}_{\le \sqrt{\frac{\ln|\mathcal{H}|/\delta}{m}}} + \underbrace{\Sigma_{opt}}_{\le 0} + \underbrace{\text{err}(h^*, D)}_{\text{bias}} + \sqrt{\frac{\ln 1/\delta}{2m}}$$

$\text{err}(\hat{h}, D) - \text{err}(\hat{h}, S)$: generalization gap of $\hat{h}$

$\text{err}(\hat{h}, S) - \text{err}(h', S)$: optimization error of $\hat{h}$.

Pf:

$$\text{err}(\hat{h}, D) = \text{err}(\hat{h}, S) + \Sigma_{gen}$$

$$= \text{err}(h', S) + \Sigma_{opt} + \Sigma_{gen}$$

$$= \text{err}(h^*, S) + \Sigma_{opt} + \Sigma_{gen} + (\text{err}(h', S) - \text{err}(h^*, S))$$

w.p. $1-\delta$. by Hoeffding within $\sqrt{\frac{\ln|\mathcal{H}|/\delta}{2m}}$

$$\le \text{err}(h^*, D) + \sqrt{\frac{\ln 1/\delta}{2m}} + \Sigma_{opt} + \Sigma_{gen} + (\text{err}(h', S) - \text{err}(h^*, S))$$

≤ 0

$$= \text{err}(h^*, D) + \sqrt{\frac{\ln 1/\delta}{2m}} + \Sigma_{opt} + \Sigma_{gen}$$

can be quite loose

Remark: 1. $\text{err}(h^*, D)$ is called the bias of $\mathcal{H}$ on D. ($h' = \min_{h \in \mathcal{H}} \text{err}(h, D)$)

2. When m is large, $\sqrt{\frac{\ln|\mathcal{H}|/\delta}{2m}}$ can be ignored.

3. tightness of the above bound

$err(h^*, S) - err(h', S)$ can be quite large. In this case,

$$err(\hat{h}, D) \le \varepsilon_{gen} + \varepsilon_{opt} + \sqrt{\frac{1}{m}} + (err(h', S) - err(h^*, S)) + err(h^*, D)$$

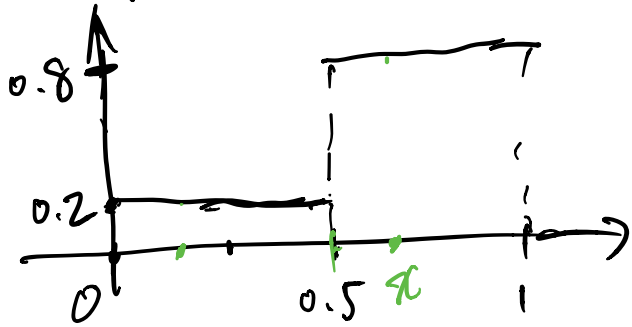
$\le err(h^*, D)$

$$err(h^*, S) - err(h^{*\prime}, S) \le \varepsilon_{gen} + \varepsilon_{opt} + \sqrt{\frac{1}{m}}$$

large — small

$\Rightarrow$ at least one of ε_{opt} and ε_{gen} would be large. $p(Y=1|X)$

Ex:



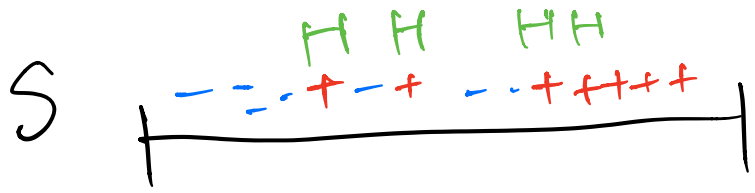
$$P_X(x) = \begin{cases} 1 & x \in [0,1] \\ 0 & \text{o.w.} \end{cases}$$

optimal classification decision:

$$\begin{cases} x \ge 0.5 & +1 \\ x < 0.5 & -1 \end{cases}$$

$$\mathcal{H} = \left\{ 2I\left(x \in \bigcup_{i=1}^{k} [a_i, b_i] \right) - 1 : k \in \mathbb{N},\ a_i, b_i \in \mathbb{R}\ \forall i \right\}$$

- h^*



$$err(h', S) = \min_{h \in \mathcal{H}} err(h, S) = 0.$$

$$err(h^*, D) = \min_{h \in \mathcal{H}} err(h, D) = 0.2$$

$$h^* = 2 \cdot I(x \in [0.5, 1]) - 1$$

Hoeffding

$$err(h^*, S) \geq 0.2 - \sqrt{\frac{1}{m}}$$

$\Rightarrow$ $err(h^*, S) - err(h', S)$ is large.

Bringing down ε_{opt}, ε_{gen}, $err(h^*, D)$?

ε_{opt} ↓ : change the ML optimization alg
make it simple to optimize

ε_{gen} ↓ : choose a less expressive $\mathcal{H}$.
collect more sample

$err(h^*, D)$ ↓ : choose a more expressive $\mathcal{H}$.

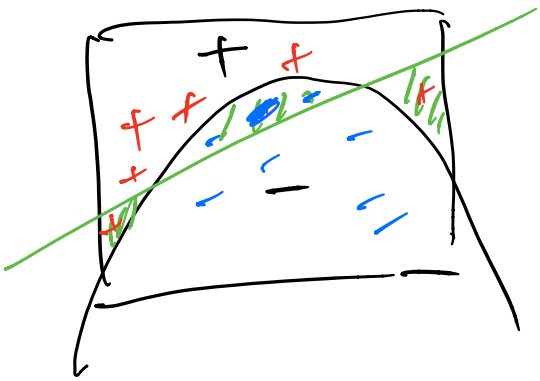
Important special case. $A = ERM(\mathcal{H})$.

Input → $VC(\mathcal{H})$

$$err(\hat{h}, D) \le \underbrace{err(h^*, D)}_{\text{bias}} + \underbrace{2\sqrt{\frac{\ln \frac{2|\mathcal{H}|}{\delta}}{2m}}}_{\text{complexity of } \mathcal{H}}.$$

bias-complexity tradeoff.

① underfitting: bias too large



$\mathcal{H} = \{$ linear classifiers $\}$

some fines can be caught by seeing

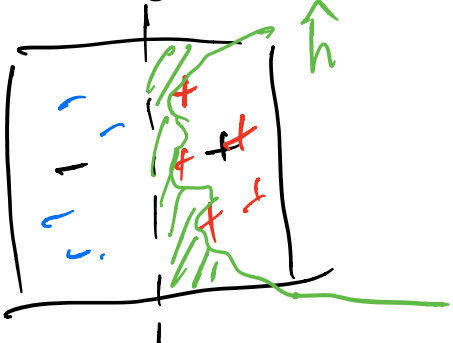
$err(\hat{h}, S)$ is too large $\Rightarrow$ $err(h^*, S)$ is also large

$\Downarrow$? $\Updownarrow$

$err(h^*, D)$ is large $err(h^*, D)$

② overfitting: $\mathcal{H}$ is too large so that complexity term is too large

$err(\hat{h}, D) - err(\hat{h}, S)$ is large.



$err(\hat{h}, S) \Rightarrow 0$

use fresh validation set V

$$err(\hat{h}, D) \approx err(\hat{h}, V) \quad \text{(Hoeffding)}$$

$$\varepsilon_{gen} \approx err(\hat{h}, V) - err(\hat{h}, S)$$

Model selection

- How can we choose a good learning alg in practice?
- only considers ERM over hypothesis classes.

Setup: $\mathcal{H}_1 \cdots \mathcal{H}_k$.
$= \{$ depth-1 decision trees$\}$ $= \{$ depth $\le k$ decision trees$\}$

$$h_i^* = \underset{h \in \mathcal{H}_i}{\operatorname{argmin}}\ err(h, D).$$

$\hat{h}_k$ may not be the best among $\{\hat{h}_1 \cdots \hat{h}_k\}$

$$\hat{h}_i = \underset{h \in \mathcal{H}_i}{\operatorname{argmin}}\ err(h, S).$$

How to use $\mathcal{H}_1 \cdots \mathcal{H}_k$ to find a good $\hat{h}$ w low error?

$$\hat{h} = \underset{h \in \bigcup_i \mathcal{H}_i}{\operatorname{argmin}}\ err(h, S) \quad ?$$

Ideal: validation:

$$\hat{\mathcal{H}} = \{\hat{h}_1 \cdots \hat{h}_k\}$$

$\hat{h} = \underset{h \in \hat{\mathcal{H}}}{\operatorname{argmin}} \ err(h, V)$ where V is a fresh validation sample.

Analysis:

① w.p. $1-\delta/2$: $\forall i$.

$$err(\hat{h}_i, D) \le err(h_i^*, D) + 2 \cdot \sqrt{\frac{\ln(|\mathcal{H}_i| \cdot 4 \cdot 2^i)/\delta}{2m}}$$

(standard ERM analysis + union bound over all i)

② w.p. $1-\delta/2$.

$$err(\hat{h}, D) \le \min_i err(\hat{h}_i, D) + 2 \cdot \sqrt{\frac{\ln 4/\delta}{|V|}}$$

①② ⇒ w.p. $1-\delta$:

$$err(\hat{h}, D) \le \min_i \left(err(h_i^*, D) + 2\sqrt{\frac{\ln(|\mathcal{H}_i| 4 \cdot 2^i)/\delta}{2m}} + 2\sqrt{\frac{\ln 4/\delta}{|V|}} \right)$$

$|V| = \Theta(m)$ $\quad g(x) = \Theta(f(x))$ $g(x) \in [c_1 f(x), c_2 f(x)]$ const c_1, c_2.

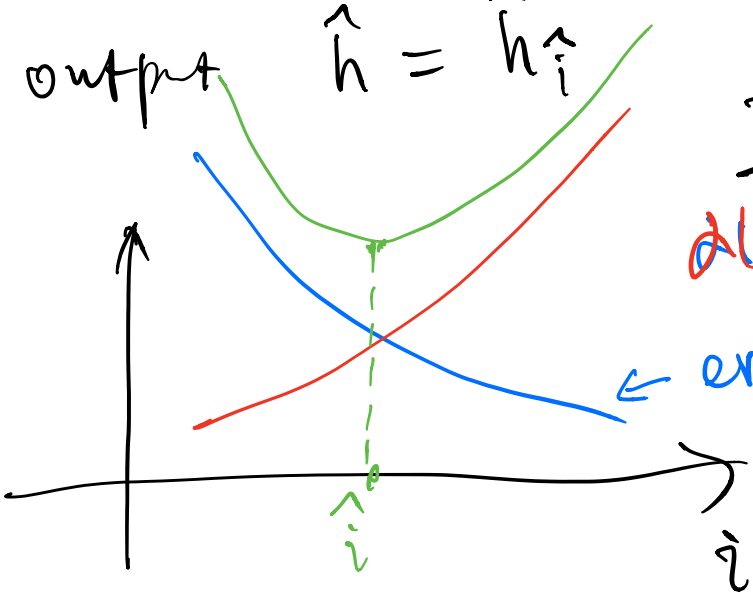
⇒ $\hat{h}$ has the best bias complexity tradeoff.

Idea 2: Structural risk minimization

(penalized ERM)

$$\hat{i} = \underset{i \in \{1,\dots,k\}}{\operatorname{argmin}} \; \operatorname{err}(\hat{h}_i, S) + \sqrt{\frac{d(H_i, m) \ln \frac{2k|H_i|}{\delta}}{2m}}$$

output $\hat{h} = \hat{h}_{\hat{i}}$



Ex: $H_1 \subseteq H_2 \subseteq \cdots \subseteq H_k$

$$\operatorname{err}(\hat{h}, D) \le \min_{i \in \{1,\dots,k\}} \operatorname{err}(h_i^*, D) + 4 \cdot d(H_i, m)$$