Learning theory

- Set up / Assumption
- Bias variance
- Approx Estim
- Empirical Risk minimizer
- uniform convergence
- VC dimensions

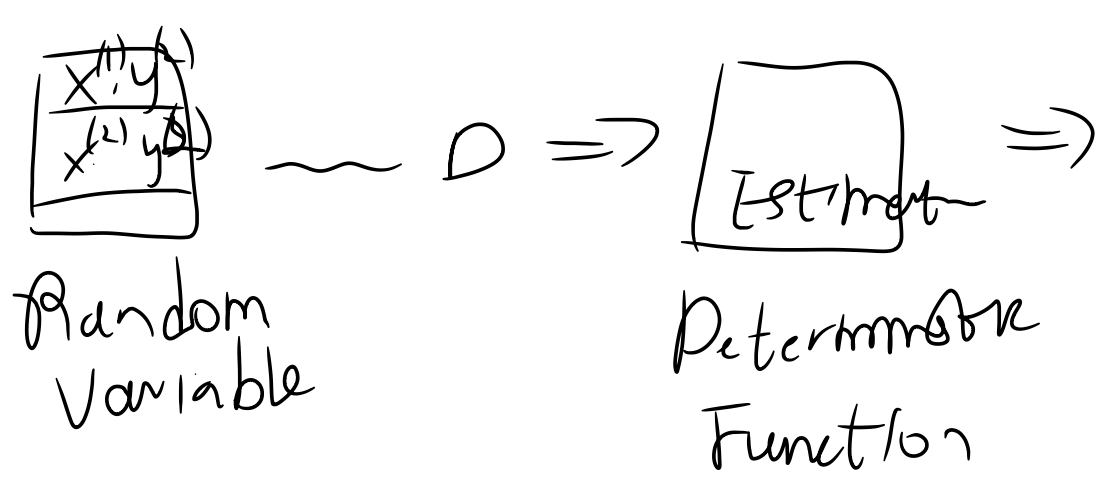
Assumptions

1 Data distribution D

$(X, Y) \sim D$ — Train — Test

2 Independent sampling

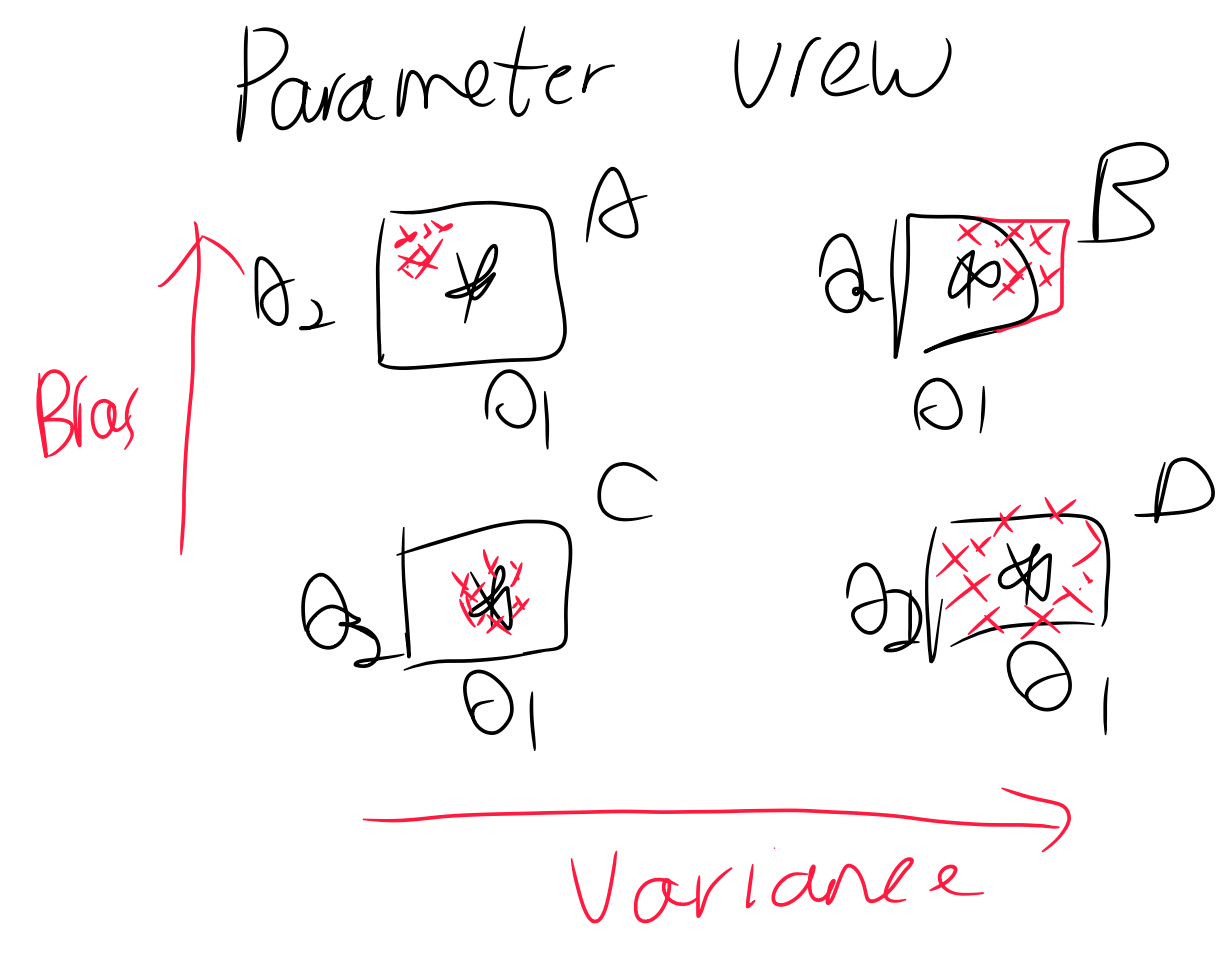
θ^* or h^* "True" Parameter



$\Rightarrow \hat{h}$ or $\hat{\theta}$ ~ sampling distribution

Random variable

Parameter view



$m \to \infty$

$Var[\hat{\theta}] \to 0$

"statistical efficiency" rate at which $Var[\hat{\theta}] \to 0$ as $m \to 0$

Algorithm consistent if $\hat{\theta} \to \theta^*$ as $m \to \infty$

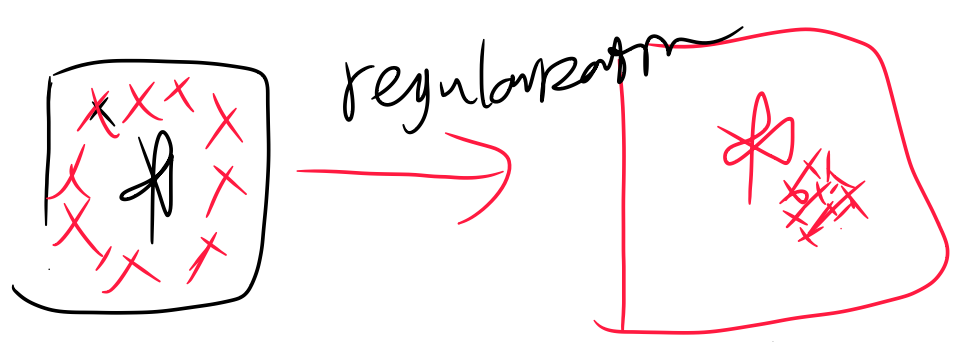
$E[\hat{\theta}] = \theta$ for all m

then good algorithm

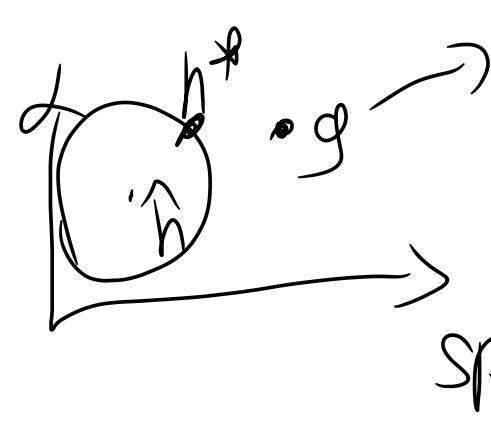
Fighting variance

1 by increasing m

2 regularization relax condition so avoid overfitting



Potentially introduce bias by lower variance



→ lowest error of all data distribution

space of hypothesis

g: best hypothesis h^* best in class hypothesis

$\hat{h}$: learnt from finite data

$\varepsilon(h)$ Risk / Generalization error

$= E_{(x,y)\sim D}[\mathbb{1}\{h(x) \neq y\}]$

$\hat{\varepsilon}_s(h)$ empirical error

$= \frac{1}{m} \sum \mathbb{1}\{h(x) \neq y\}$

$\varepsilon(g) =$ Bayes Error / Irreducible error

$\varepsilon(h^*) - \varepsilon(g) =$ Approximation error

$\varepsilon(\hat{h}) - \varepsilon(h^*) =$ Estimation error

$$\varepsilon(\hat{h}) = \text{Estimation error} + \text{Approximation error} + \text{Irreducible error}$$

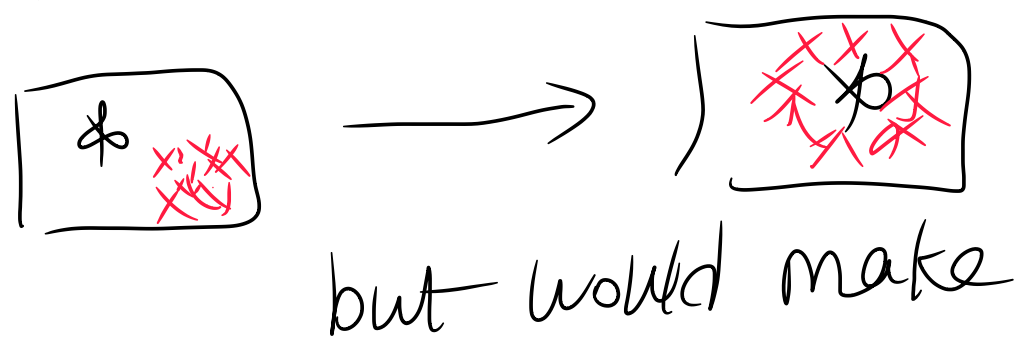
Estimation error = Estimation variance + estimation bias

$\underbrace{\text{Estimation variance}}_{\text{Variance}}$ $\underbrace{\text{estimation bias} + \text{Approximation error}}_{\text{bias}}$ $\underbrace{\text{Irreducible error}}_{\text{irreducible}}$

$\varepsilon(\hat{h}) =$ bias + variance + irreducible

Fighting bias

1 make $\mathcal{H}$ bigger by making model more general



but would make variance larger

Empirical Risk minimization

$\Rightarrow \hat{h}_{ERM}$

$$\hat{h}_{ERM} = \arg\min_{h \in \mathcal{H}} \frac{1}{m} \sum^{m} \mathbb{1}\{h(x^{(i)}) \neq y^{(i)}\}$$

How minimize?

Uniform convergence

① $\hat{\varepsilon}(h)$ vs $\varepsilon(h)$???

② $\varepsilon(\hat{h})$ vs $\varepsilon(h^*)$

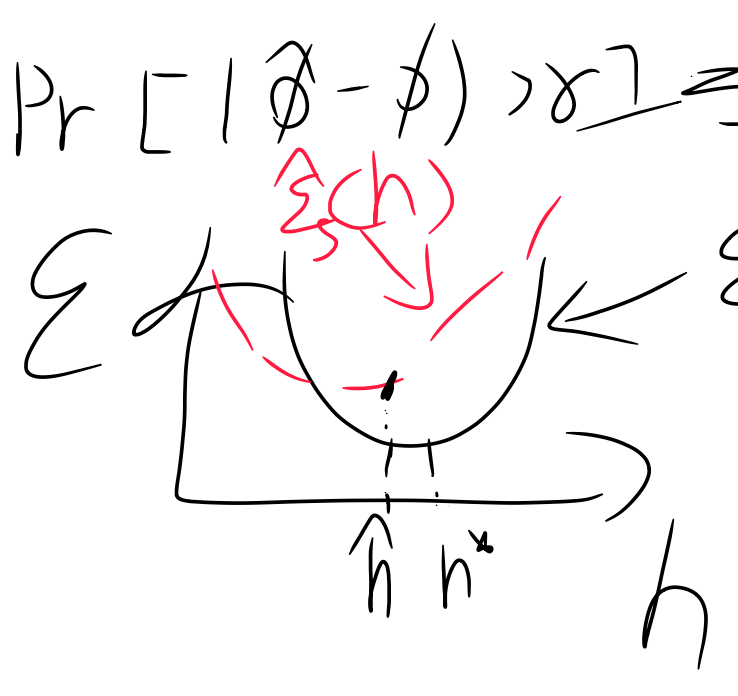
TOOLS (I) union bound (II) Hoeffding's Inequality

Let $z_1, \ldots, z_m \sim Bern(\phi)$ — these are samples

$\hat{\phi} = \frac{1}{m} \sum z_i$

Let $\gamma > 0$ [margin]

$$\Pr[|\hat{\phi} - \phi| > \gamma] \leq 2e^{(-2\gamma^2 m)}$$



$$E[\hat{\varepsilon}(h_i)] = \varepsilon(h_i)$$

$$\text{Prob}[|\hat{\varepsilon}(h_i) - \varepsilon(h_i)| > \gamma] \leq 2e^{-2\gamma^2 m}$$

What about the bound for all h_i's?

For finite hypothesis class $\mathcal{H}$

with $|\mathcal{H}| = k$

$$\text{Prob}[\exists h \in \mathcal{H} \; |\hat{\varepsilon}(h) - \varepsilon(h)| > \gamma] \leq k 2 e^{-2\gamma^2 m}$$

$$\Rightarrow \text{Prob}[\forall h \in \mathcal{H}, |\hat{\varepsilon}(h) - \varepsilon(h)| < \gamma] \geq 1 - \underbrace{2ke^{-2\gamma^2 m}}_{\delta}$$

δ — Prob of error

eg Fix $\gamma, \delta > 0$, $m \geq \frac{1}{2\gamma^2} \log \frac{2k}{\delta}$

sample complexity

$$\begin{aligned} \varepsilon(\hat{h}) &\leq \hat{\varepsilon}(\hat{h}) + \gamma \\ &\leq \hat{\varepsilon}(h^*) + \gamma \quad ??? \\ &\leq \varepsilon(h^*) + 2\gamma \end{aligned}$$

So with prob, $1 - \delta$, $\varepsilon(\hat{h}) \leq \varepsilon(h^*) + 2\sqrt{\frac{1}{2m} + \log \frac{2k}{\delta}}$

VC dimension

$VC(\mathcal{H}) =$ ~

"size" of inf dim hypothesis

$$\varepsilon(\hat{h}) \leq \varepsilon(h^*) + O\left(\sqrt{\frac{VC(\mathcal{H})}{m} \log \frac{m}{VC(\mathcal{H})} + \frac{1}{m} \log \frac{1}{\delta}}\right)$$