Agnostic Learnability of Finite Hypothesis Classes

Notations Recap

- X – Input
- Y- Output/Label
- *D* – Data distribution of X
- *f*(x) – target/labelling function
- Data generation model
  - x ~ *D*
  - *y = f(x)*
- *h(x)* ∈ *H* – Hypothesis

- Any domain
- {0, 1}
- Fixed but unknown
- X → Y

- Realizability assumption: *f*(x) ∈ *H*

Notations Recap

- Random Sample of size m
  - $S_m = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$
- $S_m \sim D^m$
- Assumption: *iid sample*

Quality of a Hypothesis

- Actual Risk of hypothesis *h* w.r.t *f,* for distribution *D*
  - $L_D(h) = \Pr_{x \sim D}[\, h(x) \neq f(x) \,]$
  - True measure of quality but unmeasurable since *D* is unknown
- Empirical Risk of hypothesis *h* w.r.t *f,* for sample *S*
  - $L_S(h) = 1/|S| \; \Sigma_S \, [\, h(x) \neq f(x) \,]$
  - Can be measured
  - Not a true measure of quality
  - Random variable depending on choice of *S*

Learner

- A mapping from sample space to hypothesis space
  - $S \rightarrow H$

- Takes as input a sample, and returns a hypothesis

Empirical Risk Minimization (ERM)

- A learning paradigm -
- $h_S = \text{argmin}_{h \in H} L_S(h)$
- Under the realizability assumption $f \in H$
  - $h_S = {}_{h \in H} L_S(h) = 0$

Example ERM

- If $x \in S$
  - $h(x) = f(x)$ (we know $f(x)$)
- Else
  - $h(x) = 0$

- Follows ERM under realizability assumption
- But “overfits” and do not generalize

Inductive Bias

- Introduce background knowledge about the hypothesis class
  - *Example: Tasty Guavas are in a Rectangle*

- $h_S = \text{argmin}_{h \in H} L_S(h)$
  - $h(x) \in H$ is the class of all rectangles

- Avoids overfitting

Theorem

- Finite hypothesis classes have small error with high probability under the realizability assumption

Proof

- Small error
  - error $\leq \varepsilon$
- Bad hypothesis
  - h_B if $L_D(h_B) > \varepsilon$ (true error)
- Set of all bad hypothesis
  - $H_B \subseteq H = \{ h_B \mid L_D(h_B) > \varepsilon \}$

Probability of Learning a Bad Hypothesis

- ERM looks at sample S and produces the hypothesis h_S
- Probability $h_S \in H_B$ should be small
- We want to upper-bound
  - $\Pr_D[\{h_S \mid L_D(h_S) > \varepsilon\}] \leq \delta$

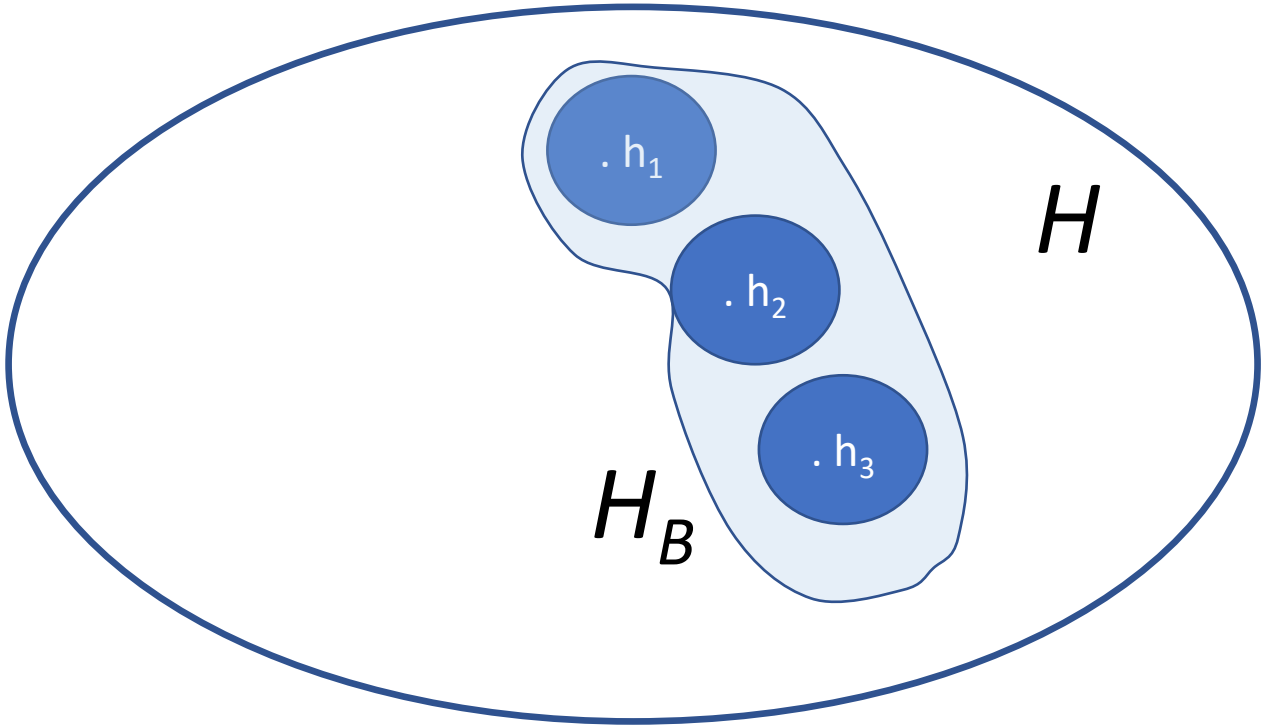
Quality of a Sample

- Misleading samples
  - M = $\{S \mid h_B \in H_B\,,\, L_S(h_B) = 0\}$
  - A sample is misleading if a bad hypothesis has zero error on S

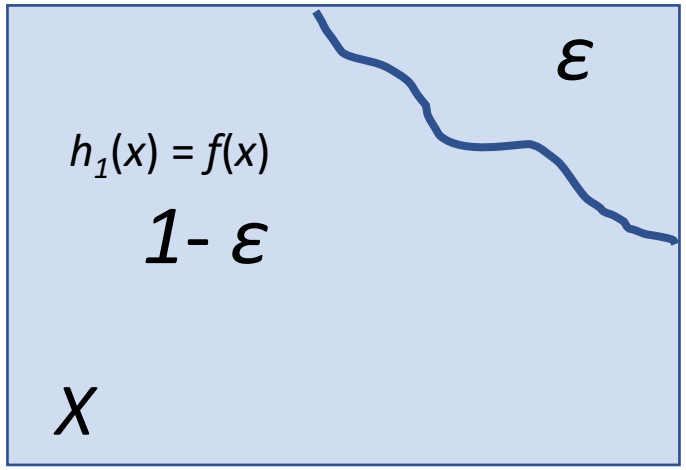
- Bad samples
  - $EB = \{S \mid L_D(h_S) > \varepsilon\}$
  - Set of samples using which ERM produces a bad hypothesis

- $EB \subseteq M$ (under the realizability assumption)

Bad Hypothesis Set

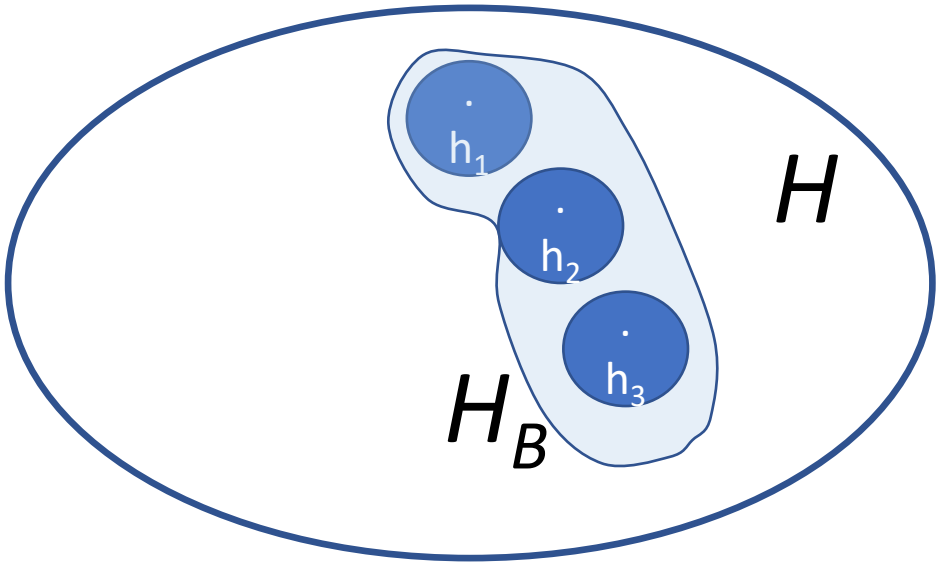


Probability of having a misleading sample *S*



- Pickup a Bad Hypothesis h_1
- Probability h_1 has zero error on S
- $\Pr_{x \sim D}[\, h_1(x) = f(x) \,] \leq 1 - \varepsilon$ (probability it is correct on x)
- $\Pr_{S \sim D}{}^{m}[\, h_1(x) = f(x), x \in S \,] \leq (1 - \varepsilon)^m$ (prob. It is correct on m iid x's)

Probability of having a misleading sample



- Probability the sample is misleading for any bad hypothesis
- $\Pr_{S\sim D}{}^{m}\,[\, h_B(x) = f(x),\, x \in S\,] \leq |H_B|(1-\varepsilon)^m$ (union inequality)
- $= \Pr_{S\sim D}{}^{m}[M]$

Proof

- $EB \subseteq M$

- $\Pr_D[\{h_S \mid L_D(h_S) > \varepsilon\}] = \Pr_D[\text{EB}] \leq \Pr_D[\text{M}] \leq |H_B|(1- \varepsilon)^m \leq |H|(1- \varepsilon)^m$
- $\leq \delta$

Sample Complexity

- $m(\varepsilon, \delta)$ – a mapping from two real numbers to an integer

- $m \geq m(\varepsilon, \delta) = 1/\ \varepsilon\ log(|H|/\ \delta)$

PAC Learnability

- Finite sample complexity
- Finite Hypothesis Classes are PAC Learnable under the Realizability Assumption