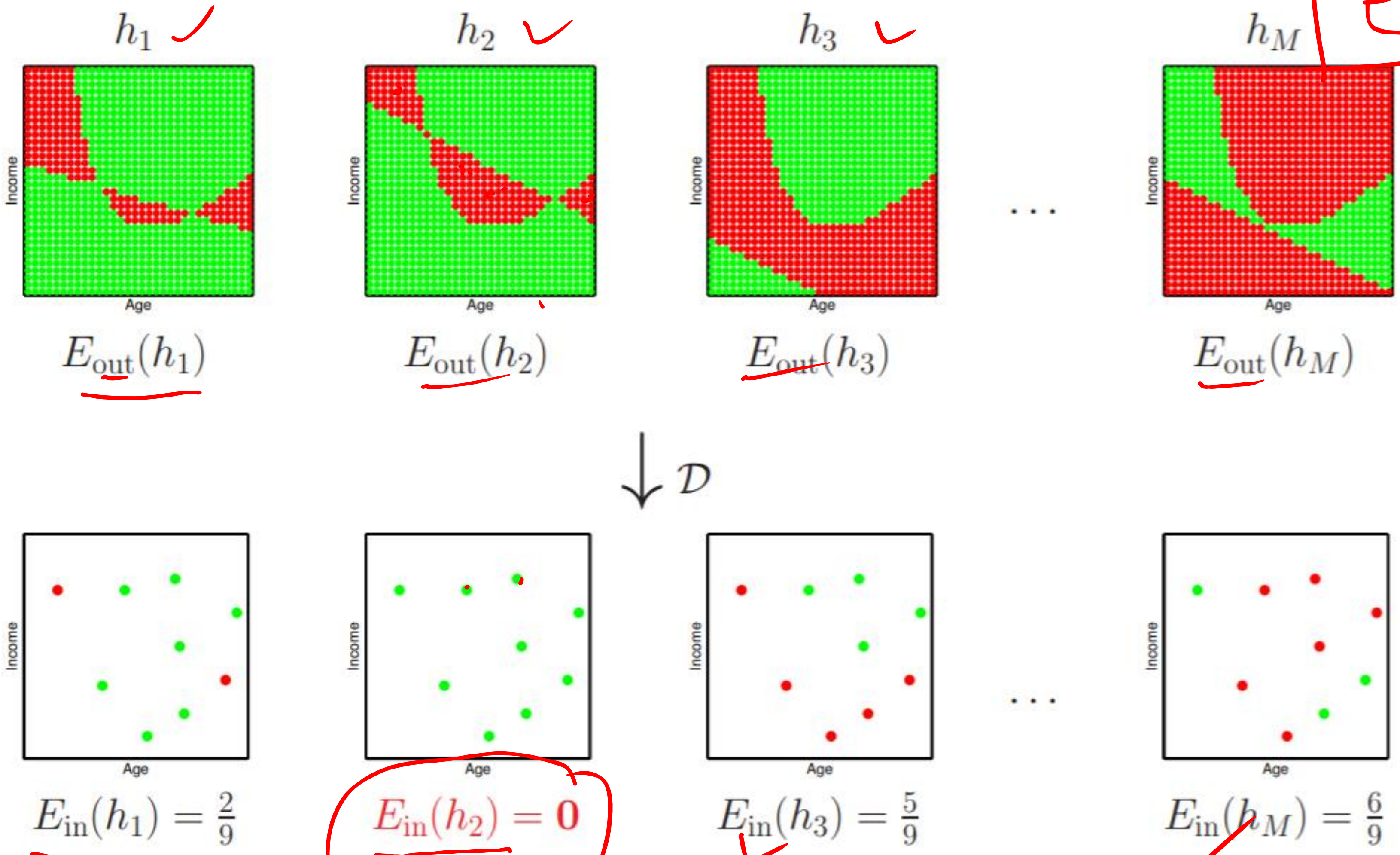
Machine Learning from Data

Lecture 4: Spring 2021

Today's Lecture

- Feasibility of Learning
- Two Step Solution to Learning
- Error and Noise

$E_{in} \simeq E_{out}$



$E_{in} \simeq 0$

Pick the hypothesis with minimum E_{in}; will E_{out} be small?

Verification and Selection Bias

- If we pick the hypothesis with minimum in-sample error, it does not approximate out-of-sample error.
- Search Causes Selection Bias
- In Real Learning in-sample error cannot reach out to out-of-sample error.

$$E_{in}(h_2) = 0$$

$$P(\text{Bad}) \leq \text{Something} \cdot \text{small}$$

$$\boxed{\text{Prob}\left[\left|E_{out}(g) - E_{in}(g)\right| > \varepsilon\right] \leq 2e^{-2\varepsilon^2 N}}$$

Not Valid

Using Hoeffding's Inequality in Learning

- Definition - "Hoeffding's inequality provides an upper bound on the probability that the sum of bounded independent random variables deviates from its expected value by more than a certain amount."

1) For the validity of the bound we must fix our hypothesis before we see the data
2) To change h after looking at data

Violating Hoeffding's inequality.

Updating Hoeffding's Bound

Set $\mathcal{H}$ $\xrightarrow{LA}$ picks g based on D after generating data.

- ✗ $P\left[\,|E_{in}(h_m) - E_{out}(h_m)| > \varepsilon\,\right]$ is small for any fixed h_m

$$\underbrace{P\left[\,|E_{in}(g) - E_{out}(g)| > \varepsilon\,\right]}_{\text{Bound is updated}}$$ is small for the final hypothesis g.

$\mathcal{H}$, $h_m \in \mathcal{H}$

$$\underbrace{|E_{in}(g) - E_{out}(g)| > \varepsilon}\ \Longrightarrow\ |E_{in}(h_1) - E_{out}(h_1)| > \varepsilon$$

subset relation

or $|E_{in}(h_2) - E_{out}(h_2)| > \varepsilon$

$\vdots$

or $|E_{in}(h_M) - E_{out}(h_M)| > \varepsilon$ — (1)

A, B → events

* if $A \Rightarrow B$ or $A \subseteq B$, $P[A] \leq P[B]$

* A, B, C . . . . Z : $P(A \text{ or } B \text{ or } C \ldots \text{ or } Z) \leq P[A] + P[B] + \ldots P[Z]$

union bound.

Eq (1)

$$P[|E_{in}(g) - E_{out}(g)| > \epsilon] \le P[|E_{in}(h_1) - E_{out}(h_1)| > \epsilon]$$

$2e^{-2\epsilon^2 N}$

$$\text{or } P[|E_{in}(h_2) - E_{out}(h_2)| > \epsilon]$$

$$\vdots$$

$$P[|E_{in}(h_M) - E_{out}(h_M)| > \epsilon]$$

$$\le \sum_{m=1}^{M} P[|E_{in}(h_m) - E_{out}(h_m)| > \epsilon]$$

$$\boxed{P[|E_{in}(g) - E_{out}(g)| > \epsilon] \le 2M e^{-2\epsilon^2 N}}$$

(2)

Feasibility of Learning

- Two Questions to answer:
  - Can we make sure that $E_{out}(g)$ is close enough to $E_{in}(g)$ $E_{out}(g) \approx E_{in}(g)$
  - Can we make $E_{in}(g)$ small enough

D

Complexity of H and f → complex phenomena

$|H|$

Bigger choice set

Price

Complexity of H	Complexity f
$\lvert H \rvert \uparrow$	How hard is f?
→ Yes → Q1, keep M in check.	→ Yes → Q1 → No direct impact because not there in Hoeffding's inequality.
→ Yes → Q2, M ↑ due to the result of choice	→ D is a result of f
	f is complex
	$E_{in} \uparrow$
	$\lvert H \rvert \uparrow$ → bad for Q1

Interpreting the Hoeffding's Bound

$$\mathbb{P}\left[|E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon\right] \leq \boxed{2|\mathcal{H}|e^{-2\epsilon^2 N}}, \qquad \text{for any } \epsilon > 0.$$

$$\mathbb{P}\left[|E_{\text{in}}(g) - E_{\text{out}}(g)| \leq \epsilon\right] \geq 1 - \underbrace{2|\mathcal{H}|e^{-2\epsilon^2 N}}_{\delta}, \qquad \text{for any } \epsilon > 0.$$

$$\delta = 2|H|e^{-2\varepsilon^2 N}$$

Theorem: With Probability at least $1-\delta$

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{1}{2N} \log \frac{2|H|}{\delta}}$$

Proof: $P\left[\,|E_{in}(g)-E_{out}(g)| \le \varepsilon\,\right] \ge 1-\delta$

In other words: With probability at least $1-\delta$

$$|E_{in}(g)-E_{out}(g)| \le \varepsilon$$

$$\delta = 2|H|e^{-2\varepsilon^2 N}$$

$$\varepsilon = \underbrace{\sqrt{\frac{1}{2N}\log\frac{2|H|}{\delta}}}$$

$$\Rightarrow \quad E_{out}(g) \le E_{in}(g) + \varepsilon$$

$$\underbrace{E_{out}(g) \le E_{in}(g)} + \underbrace{\sqrt{\frac{1}{2N}\log\frac{2|H|}{\delta}}}_{\text{error bar}}$$

E_{in} *reaches out to* E_{out} when *H* is small

error bar ↑

$$\underline{E_{\text{out}}(g)} \leq \underbrace{E_{\text{in}}(g)} + \sqrt{\frac{1}{2N}\left(\log\frac{2|\mathcal{H}|}{\delta}\right)}.$$

sample

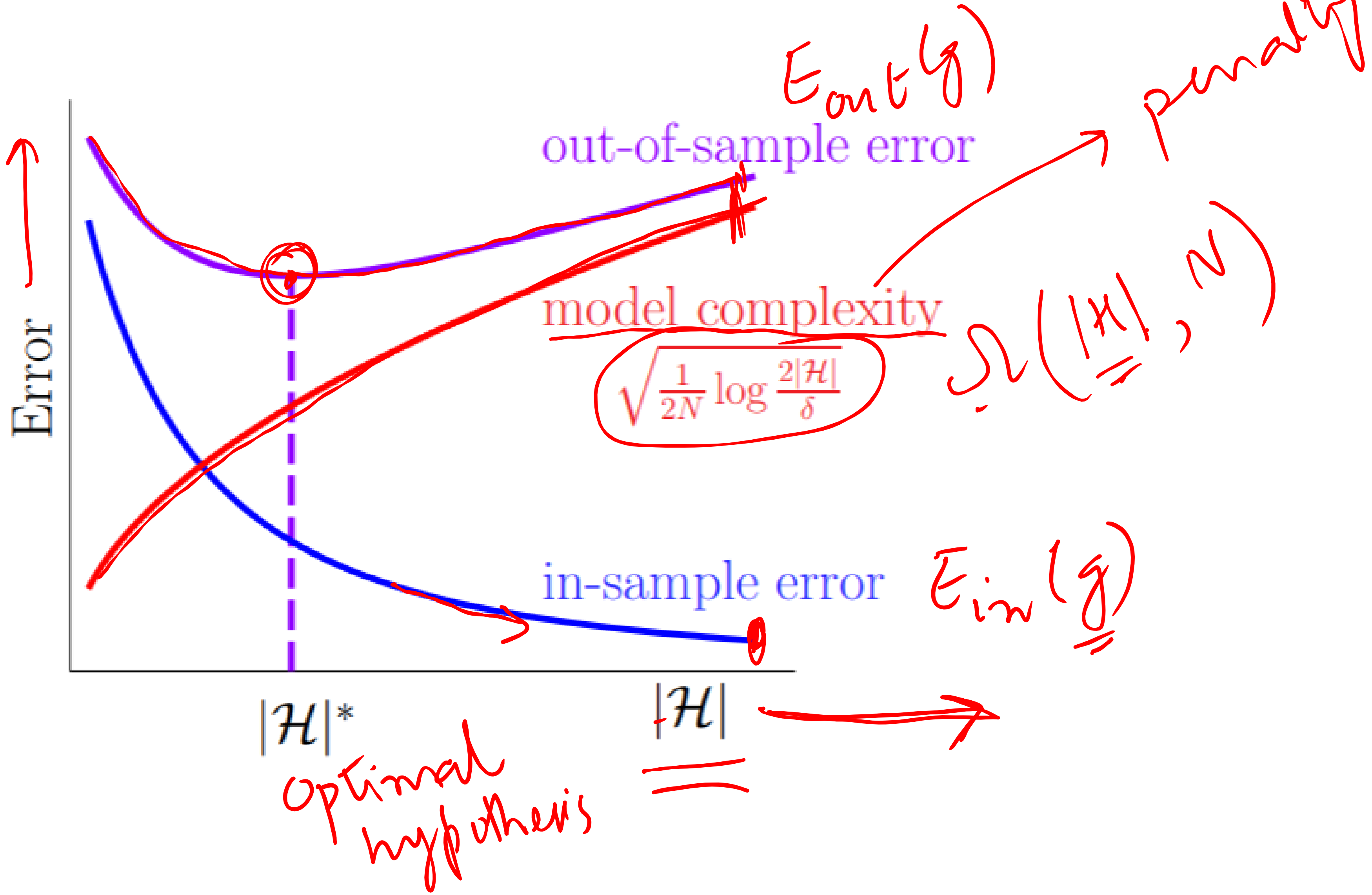
generalization bound.

If $N >> \ln|H|$

$$\underbrace{E_{out}(g)} \simeq \underbrace{E_{in}(g)}$$

$E_{out}(g)$?

Trade-off



$E_{out}(g)$
out-of-sample error
penalty
model complexity
$\sqrt{\frac{1}{2N}\log\frac{2|\mathcal{H}|}{\delta}}$
$\Omega(|\mathcal{H}|, N)$
in-sample error
$E_{in}(g)$
Error
$|\mathcal{H}|^*$
$|\mathcal{H}|$
optimal hypothesis

2 Step Approach

-

1) Ensure $\underline{E_{out}(g) \simeq E_{in}(g)}$
2) Get $\underline{E_{in}(g) \simeq 0}$

Sample $\rightarrow$ Universe

$E_{in}(g) \simeq E_{out}(g)$ ✓

$g \simeq f$

$\underbrace{E_{out}(g) \simeq 0}$?

Implies that $\underline{E_{out}(g) \simeq 0}$

Summarize

- Is Learning Feasible? (Finite $\mathcal{H}$)

A) Yes, provide we accept a 2-step approach

$\rightarrow\ \rightarrow E_{out}(g) \simeq E_{in}(g)$ { Theoretically & with high probability

$\rightarrow\ E_{in}(g) \simeq 0$

Conditions: 1) Fix $\mathcal{H}$ ahead of time $\Rightarrow g \in \mathcal{H}$

2) Data must be IID $(P(x)) \leftarrow P(x)$ { Same Bin

3) $E_{out}(g) \rightarrow x \rightarrow$ comes from $P(x)$

B) $\mathcal{H}$ has a possibility to fail

A + B $\longrightarrow$ Summarize the fact that Learning is feasible.

Our Learning Approach is General

- 1) Applies to any target function f.
  2) Applies to any $P(x)$
  3) Applies to any hypothesis set $\mathcal{H}$ (PLA, NN, SVM)
  4) Applies to any LA.

All our conditions must hold.

Target Function

- 1) Complexity of target function → Need more data
  2) Noisy f. (stochastic) → Practice
  3) Error

Complex f → hard to learn, $E_{in}(g) \simeq 0$

$|H| \uparrow$ $\qquad \Omega(|H|, N) \uparrow$ $\qquad \sqrt{\frac{|H|}{N}}$ $|H| \to$ bigger, $N \to$ bigger

$\boxed{\text{Big } N}$ $\qquad E_{in} \simeq E_{out}$

Noisy target

- Same x $\begin{cases} \nearrow +1 \\ \searrow -1 \end{cases}$

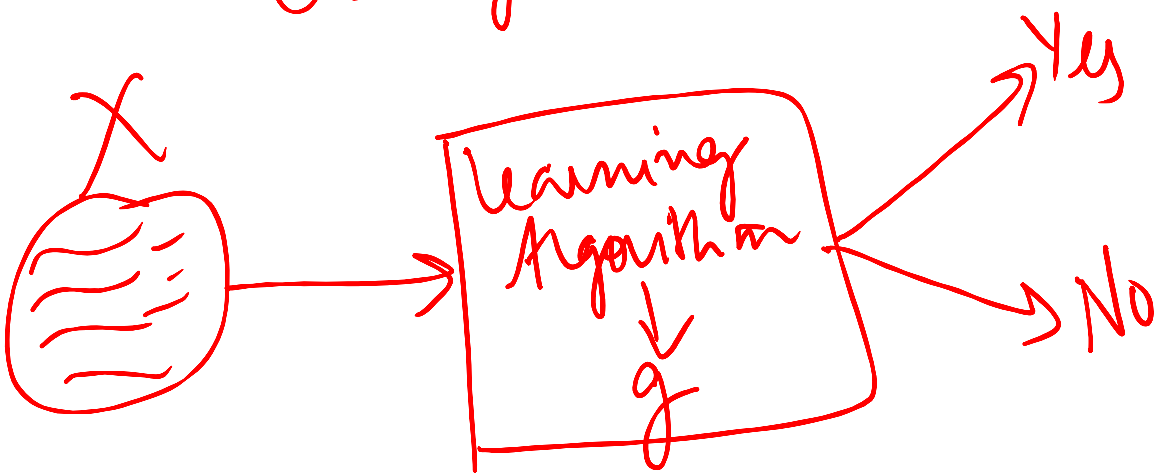
$f = P[y|x]$

$E_{in} \neq 0$

$f \rightarrow$ noisy

Error (User defined)

Classification problem

- 

Supermarket (SM)

1) Say No to the right person costly for the SM (False -ve)*
2) Say Yes to the wrong person not costly → okay (False +ve)

CIA access control

1) Say Yes to the wrong person, extremely costly for CIA (False +ve)*
2) Say No to the right person, okay (False -ve)

Interpretation of Error (Not good)

Risk matrix (SM)

g

Reality	Yes	No
Yes	0	10
No	1	0

Risk Matrix (CIA)

g

Reality	Yes	No
Yes	0	1
No	1000	0

Pointwise Errors

$$\boxed{e(h(x), y)}$$

- 1) Binary Errors (Classification problems)

$$e(h(x), y) = [\![h(x) \neq y]\!]$$

indicator function

2) Squared Errors (Regression problems)

$$e(h(x), y) = (h(x) - y)^2$$

$$E_{in}(h) = \frac{1}{N} \sum_{i=1}^{N} e(h(x_i), y_i)$$

$$E_{out}(h) = E_x\left[e(h(x), f(x))\right]$$

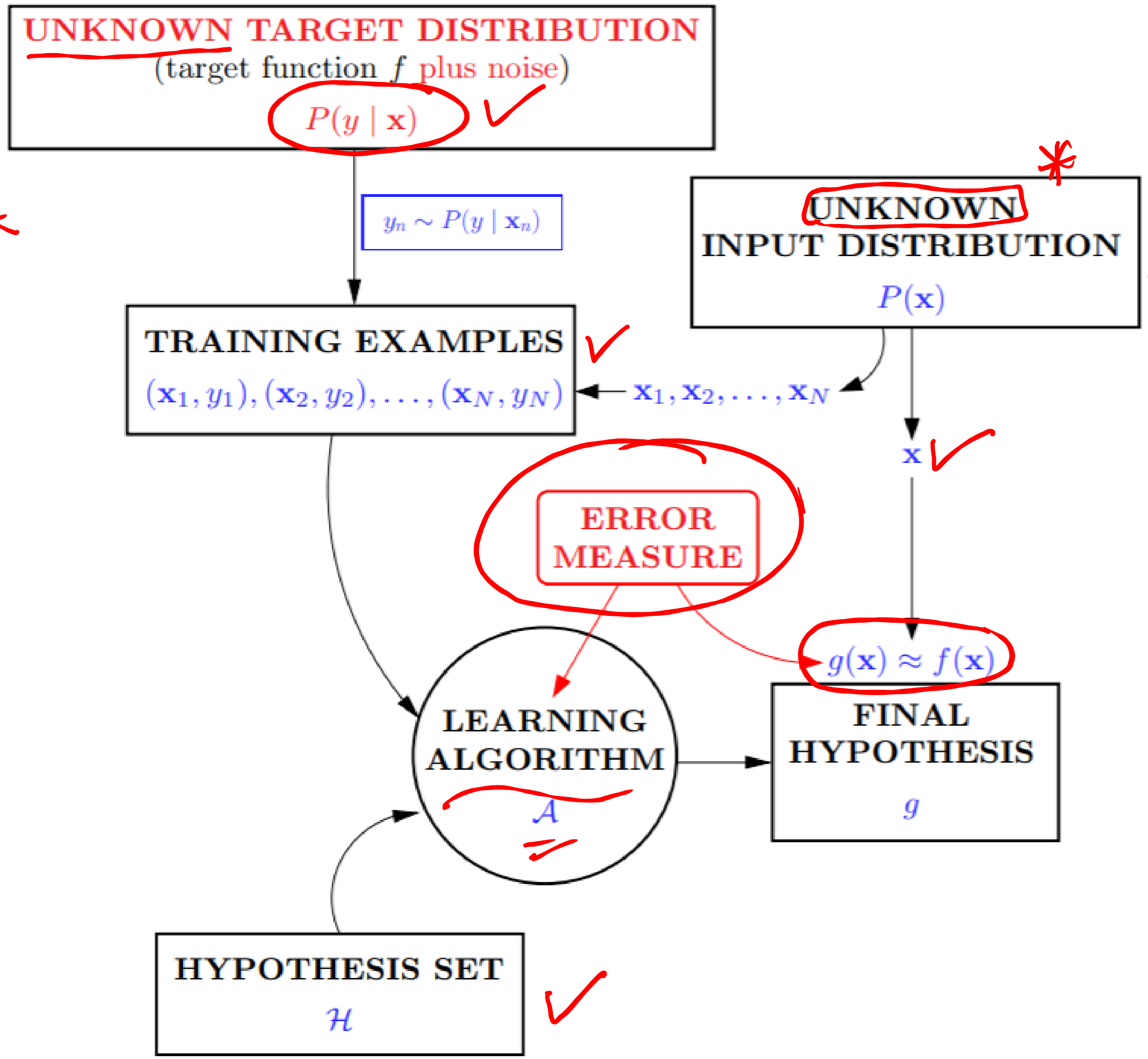
Learning Set-up

2-Step Process ✓

1) $E_{out}(g) \simeq E_{in}(g)$ *

2) $\boxed{E_{in} \simeq 0}$

can ~~will~~ be a failure



1D