CSE 417T
Introduction to Machine Learning

Lecture 4

Instructor: Chien-Ju (CJ) Ho

Logistics: HW1

- Small updates for Problem 3
  - LFD Exercise 1.10 -> LFD Exercise 1.10 (a)-(d)

- Code submission
  - You only need to complete submit the **two Matlab files**
  - You need to write additional code for generating figures and conducting analysis but do not need to submit it

- You should be ready to answer Problem 1-5 now and problem 7 after today.
  - We'll cover the topic of Problem 6 before next Tuesday.

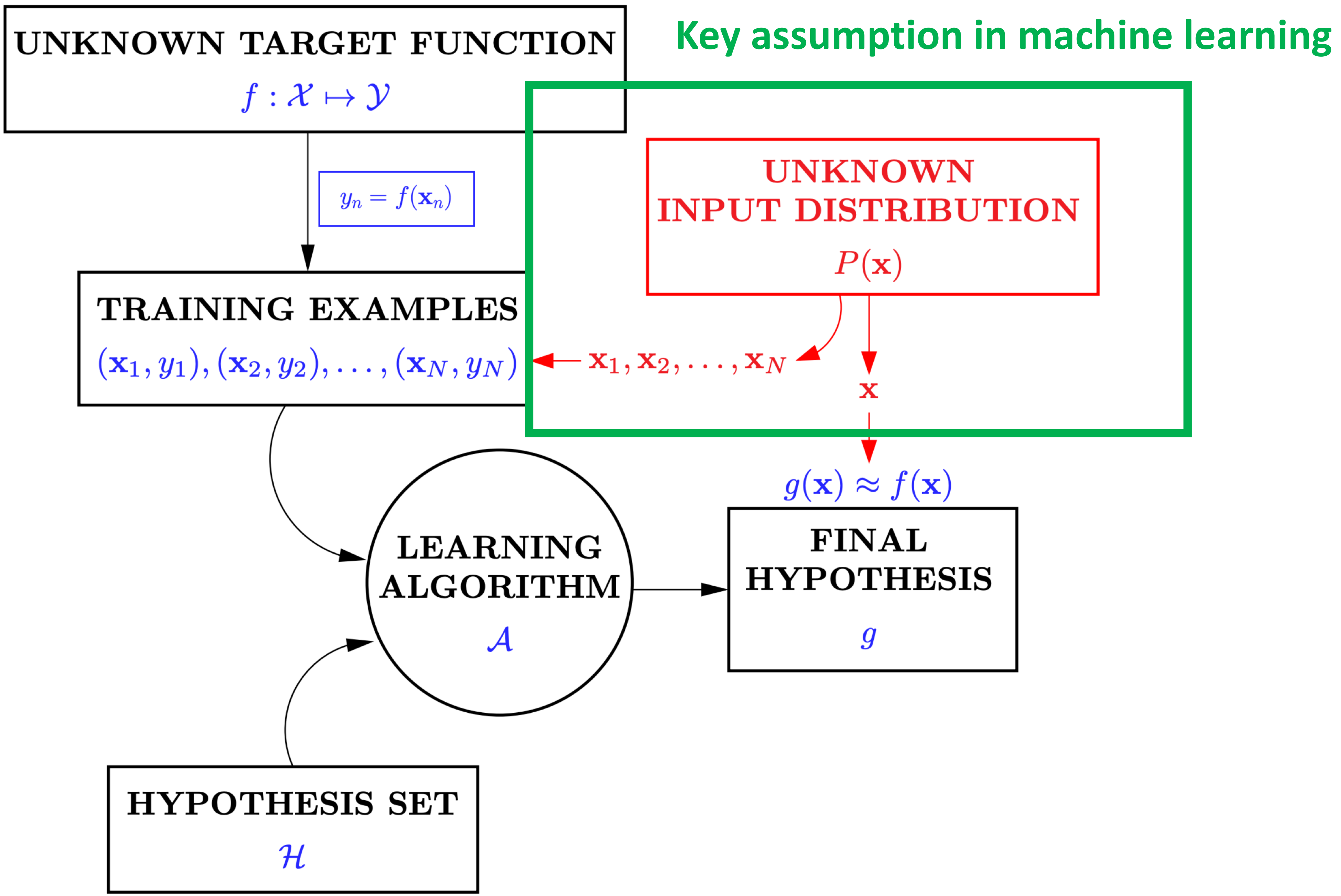
Logistics: Office Hours

- Tentative schedule of TA office hours (starting next week)

Mondays	10:00-11:30AM (Heming)	01:30-03:00PM (Flora)
Tuesdays	01:00-02:30PM (Xinyu)	03:00-04:30PM (Yi)
Wednesdays	10:00-11:30AM (Ziyang)	12:30-2:000PM (Ruoyao)
Thursdays	02:30-04:00PM (Connor)	04:00-05:30PM (Tong)
Fridays	12:30-02:00PM (Brendan)	02:30-04:00PM (Jiahao)
Sundays	05:00-06:30PM (Ina)	

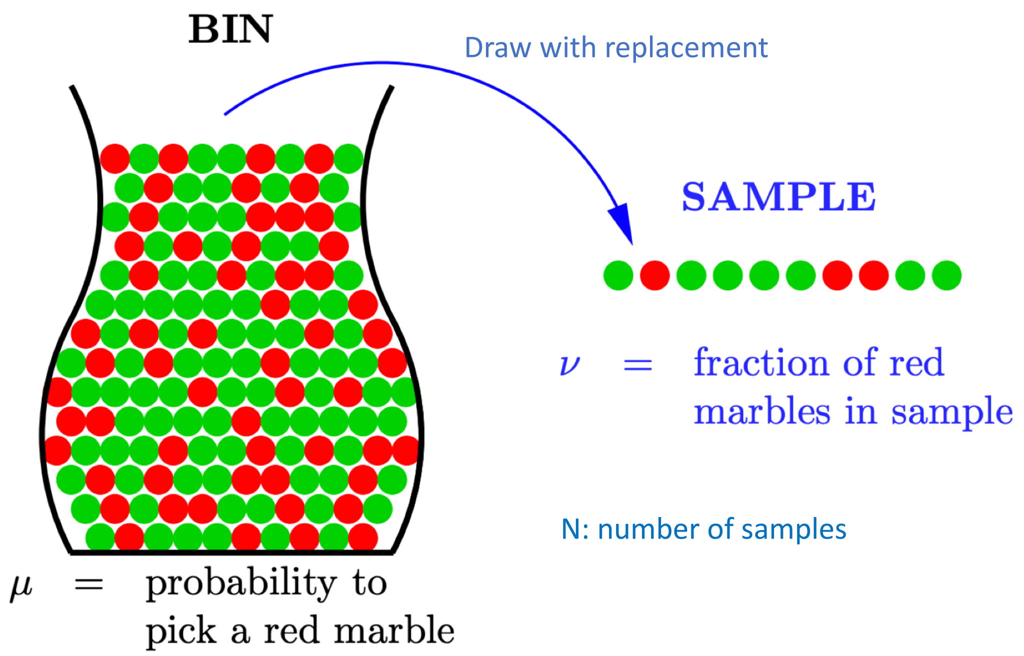
- There might still be changes as we are waiting for the room confirmations.
- Please follow **Piazza** for the announcements.

Recap



Key assumption in machine learning
UNKNOWN TARGET FUNCTION
$f : \mathcal{X} \mapsto \mathcal{Y}$
$y_n = f(\mathbf{x}_n)$
UNKNOWN INPUT DISTRIBUTION
$P(\mathbf{x})$
TRAINING EXAMPLES
$(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N)$
$\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$
$\mathbf{x}$
$g(\mathbf{x}) \approx f(\mathbf{x})$
LEARNING ALGORITHM
$\mathcal{A}$
FINAL HYPOTHESIS
g
HYPOTHESIS SET
$\mathcal{H}$

Hoeffding's Inequality



$$\mathbf{Pr}[|\boldsymbol{\mu} - \boldsymbol{\nu}| > \boldsymbol{\epsilon}] \leq \mathbf{2}\boldsymbol{e}^{-\mathbf{2}\boldsymbol{\epsilon}^{\mathbf{2}}\boldsymbol{N}}$$

Define $\delta = \Pr[|\mu - \nu| > \epsilon]$

- Fix δ, ϵ decreases as N increases
- Fix ϵ, δ decreases as N increases
- Fix N, δ decreases as ϵ increases

Informal intuitions of notations
N: # sample
δ: probability of "bad" event
ϵ: error of estimation

Connection to Learning

- Given dataset $D = \{(\vec{x}_1, y_1), \dots, (\vec{x}_N, y_N)\}$.
- Fix a hypothesis h
  - $\boldsymbol{E_{in}(h)} \stackrel{\text{def}}{=} \frac{1}{N}\sum_{n=1}^{N} \mathbb{I}[h(\vec{x}_n) \neq f(\vec{x}_n)]$ [In-sample error, analogy to ν]
  - $\boldsymbol{E_{out}(h)} \stackrel{\text{def}}{=} \Pr_{\vec{x} \sim P(\vec{x})}[h(\vec{x}) \neq f(\vec{x})]$ [Out-of-sample error, analogy to μ]

- Apply Hoeffding's inequality

$$\boldsymbol{Pr[|E_{out}(h) - E_{in}(h)| > \epsilon] \leq 2e^{-2\epsilon^2 N}}$$

- This is *verification*, not *learning*

Connection to "Real" Learning

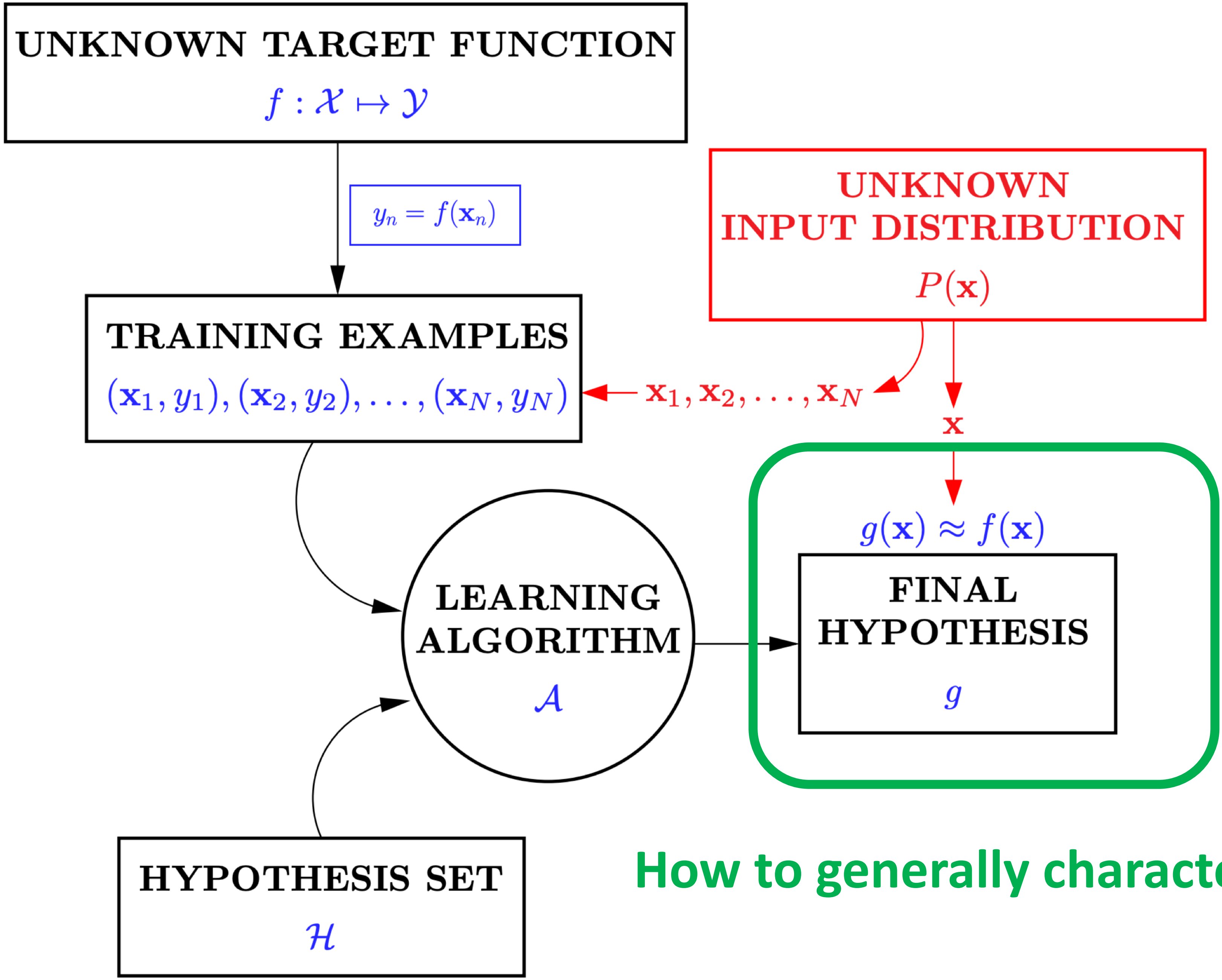
- Given a finite hypothesis set $H = \{\mathrm{h}_1, \ldots, \mathrm{h}_\mathrm{M}\}$
- Apply some learning algorithm on D, output a $g \in H$

- What can we say about $E_{out}(g)$ from $E_{in}(g)$?

$$\boldsymbol{Pr[|E_{out}(g) - E_{in}(g)| > \epsilon] \leq 2Me^{-2\epsilon^2 N}} \quad \textbf{for any } \boldsymbol{\epsilon > 0}$$

- [Will discuss more about the interpretations/intuitions today]

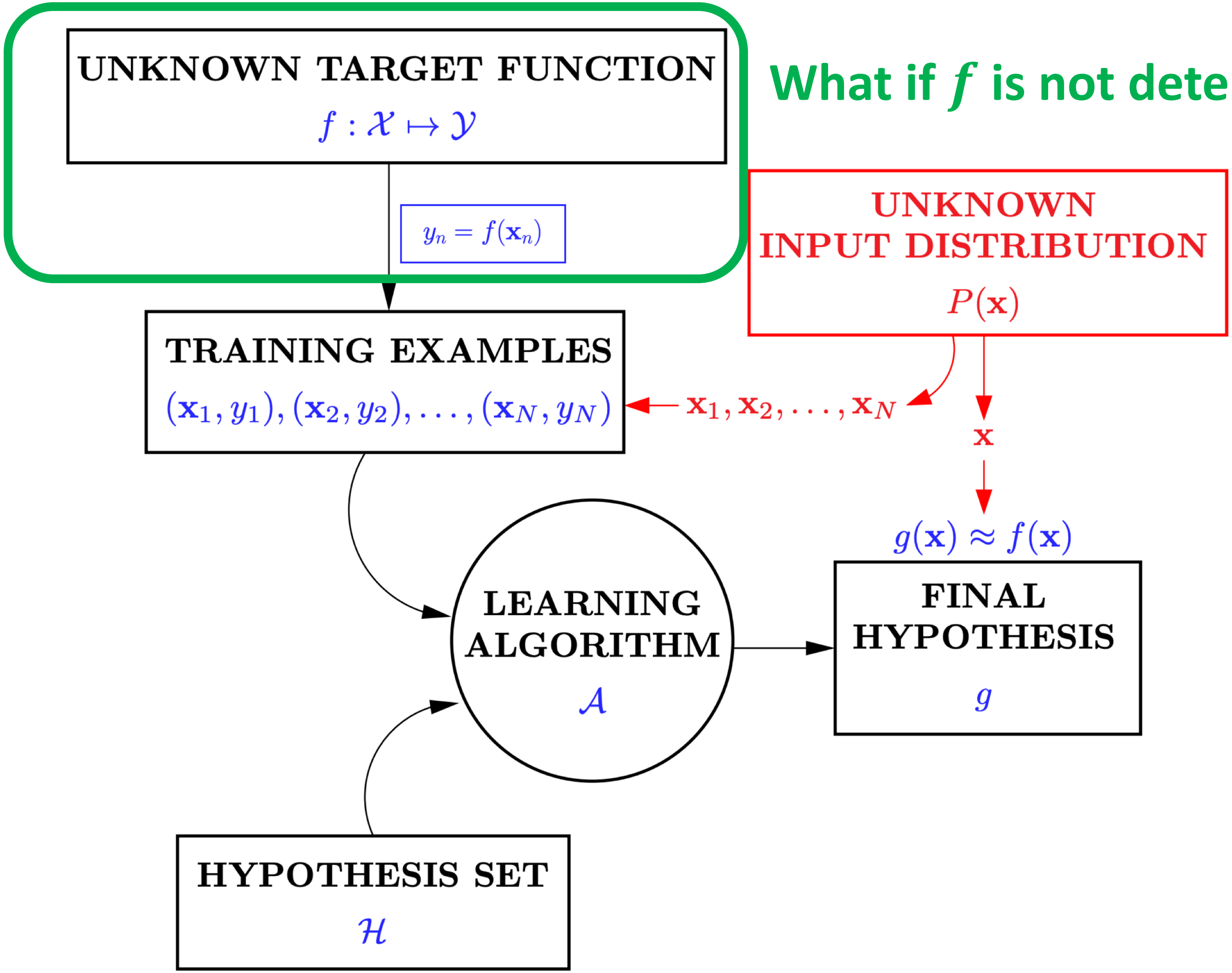
Revisit the learning problem



How to generally characterize $g \approx f$

Goal: $g \approx f$

- A general approach:
  - Define a error function $E(h, f)$ that quantify how far away g is to f
  - Choose the one with the smallest error
  - For example: $g = \underset{h \in \mathcal{H}}{\operatorname{argmin}} E(h, f)$
- E is usually defined in terms of a pointwise error function $e(h(\vec{x}), f(\vec{x}))$
  - Binary error (classification):$e(h(\vec{x}), f(\vec{x})) = \mathbb{I}[h(\vec{x}_n) \neq f(\vec{x}_n)]$ (What we have discussed so far)
  - Squared error (regression):$e(h(\vec{x}), f(\vec{x})) = \left(f(\vec{x}) - h(\vec{x})\right)^2$
- In-sample and out-of-sample errors
  - $E_{in}(h) = \frac{1}{N} \sum_{n=1}^{N} e(h(\vec{x}_n), f(\vec{x}_n))$
  - $E_{out}(h) = \mathbb{E}_{\vec{x}}[e(h(\vec{x}_n), f(\vec{x}_n))]$



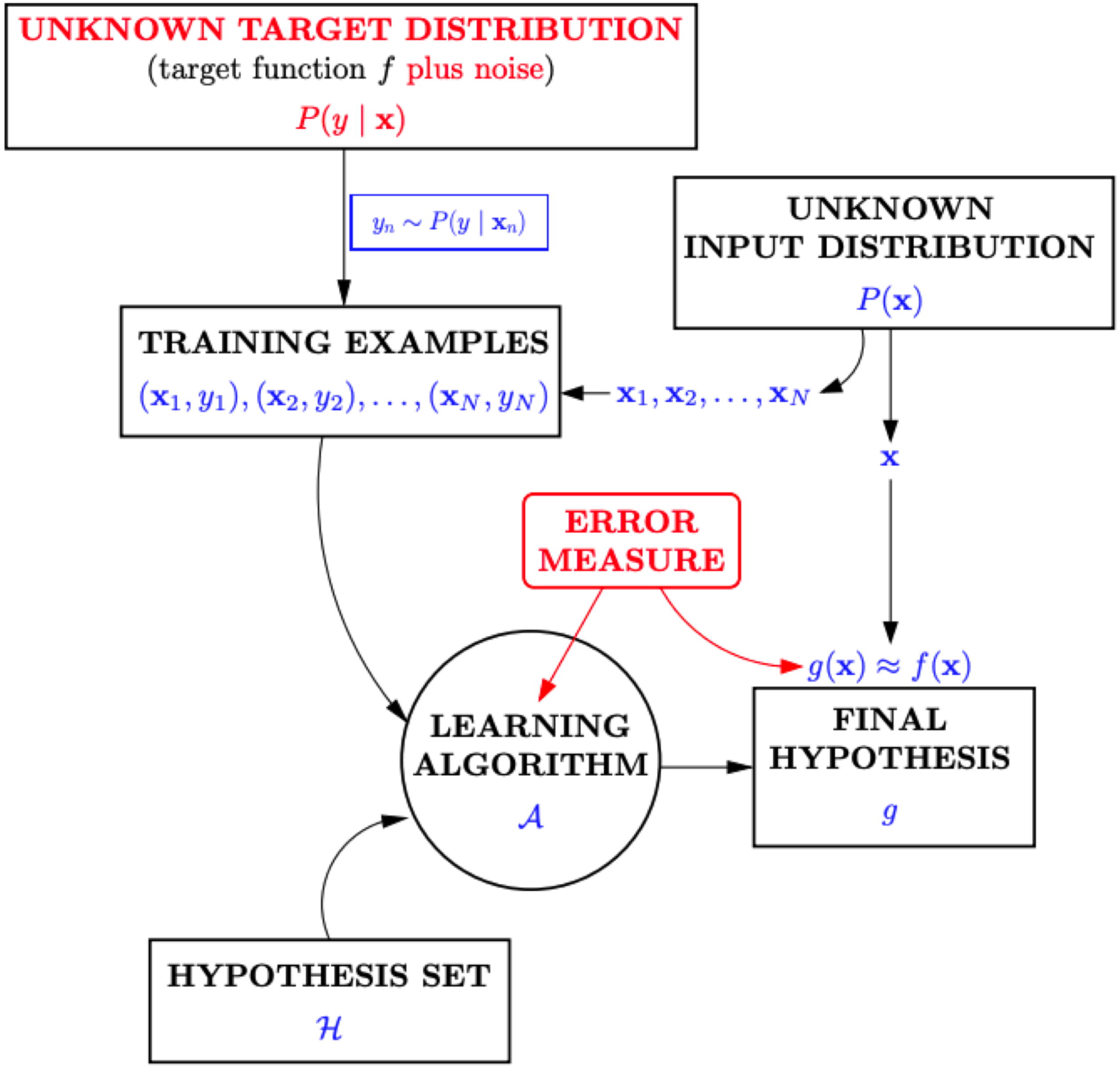
What if f is not deterministic?

Noisy Target

- What if there doesn't exist f such that $y = f(\vec{x})$?
  - f is stochastic instead of deterministic

- Common approach
  - Instead of a target function, define a target **distribution**
  - Instead of $y = f(\vec{x})$, y is drawn from a conditional distribution $P(y|\vec{x})$
  - $y = f(\vec{x}) + \epsilon$ where $\epsilon \sim N(0, \sigma^2)$

General Setup of (Supervised) Learning



Brief Lecture Notes Today

The notes are not intended to be comprehensive. They should be accompanied by lectures and/or textbook.

Let me know if you spot errors.

Revisit the "Multi-Hypothesis" Bound

- Given a finite hypothesis set $H = \{\mathrm{h}_1, \ldots, \mathrm{h}_\mathrm{M}\}$
- Apply some learning algorithm on D, output a $g \in H$

- What can we say about $E_{out}(g)$ from $E_{in}(g)$?

$$\boldsymbol{Pr[|E_{out}(g) - E_{in}(g)| > \epsilon] \leq 2Me^{-2\epsilon^2 N}} \quad \textbf{for any } \boldsymbol{\epsilon > 0}$$

Interpreting $\Pr[|E_{out}(g) - E_{in}(g)| > \epsilon] \leq 2Me^{-2\epsilon^2 N}$

- Playing around with the math
  - Define $\delta = \Pr[|E_{out}(g) - E_{in}(g)| > \epsilon]$
  - We have $\delta \leq 2Me^{-2\epsilon^2 N}$ => $\epsilon \leq \sqrt{\frac{1}{2N}\ln\frac{2M}{\delta}}$

- This means, with probability at least $1 - \delta$
  - $E_{out}(g) \leq E_{in}(g) + \epsilon \leq E_{in}(g) + \sqrt{\frac{1}{2N}\ln\frac{2M}{\delta}}$

Discussion/Interpretation on the learning bound

- With probability at least $1 - \delta$

$$E_{out}(g) \leq (g) + \sqrt{\frac{1}{2N} \ln \frac{2M}{\delta}}$$

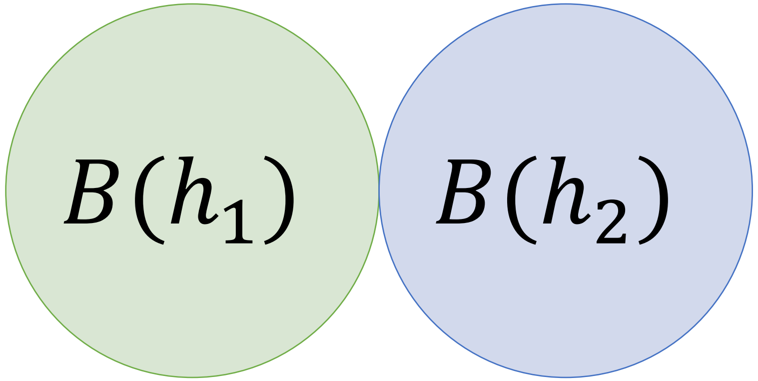
Consider M as a proxy measure on the “complexity” of H

- Our ultimate goal is to have a small $E_{out}(g)$
  - There is a tradeoff of choosing M (what “learning model” to use)
    - Increase M -> Smaller $E_{in}(g)$ (more hypothesis to “fit” the training data)
    - Increase M -> Larger ϵ
  - It also depends on N, the number of data points you have
    - A small number of data points => use simple models (e.g., linear models)
    - Complex models (e.g., deep learning) work when you have a lot of data

What if M is infinite?

Key Intuitions in the Multi-Hypothesis Analysis

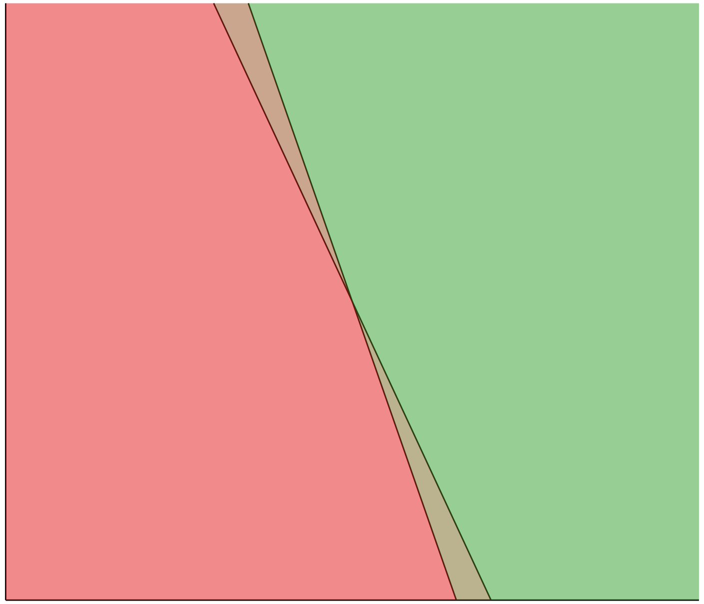
- Define ”bad event of h” $B(h)$ as $|E_{out}(h) - E_{in}(h)| > \epsilon$
- If g is selected from $\{h_1, h_2\}$
  - $B(g) \subseteq B(h_1) \cup B(h_2)$
  - $\Pr[B(g)] \leq \Pr[B(h_1) \text{ or } B(h_2)] \leq \Pr[B(h_1)] + \Pr[B(h_2)]$ (Union Bound)



- Union bound considers the worst case: Bad events don’t overlap

Do Bad Events Overlap?

- Oftentimes, they overlap a lot!



The two linear separators on the left make the same predictions for most points.

If it's a bad event for one, it's likely to be a bad event for the other.

Recall: Informally, you can interpret "bad event of h" as the event that we draw a "unrepresentative dataset D" that makes the in-sample errors of h to be far away from out-of-sample error of h

Effective Number of Hypothesis

- Dichotomy
  - Informally, consider it as “data-dependent” hypothesis
  - Characterized by both H and N data points $(\vec{x}_1, \dots, \vec{x}_N)$

$$H(\vec{x}_1, \dots \vec{x}_N) = \{h(\vec{x}_1), \dots, h(\vec{x}_N) | h \in H\}$$

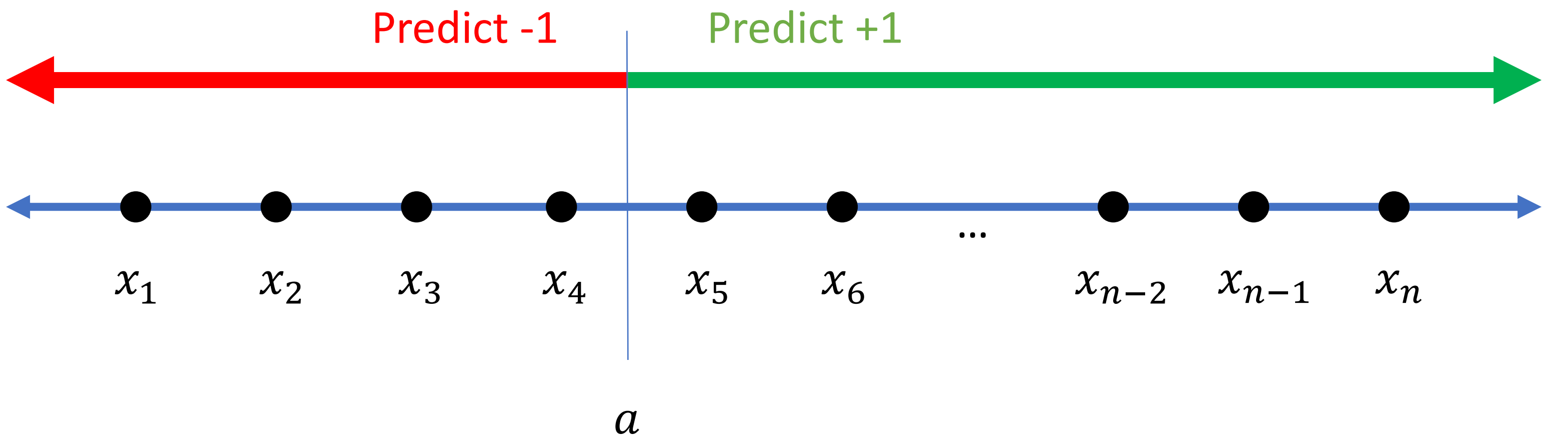
  - The set of possible prediction combinations $h \in H$ can induce on $\vec{x}_1, \dots, \vec{x}_N$

- Growth function
  - Largest number of dichotomies H can induce across all possible data sets of size N

$$m_H(N) = \max_{(\vec{x}_1, \dots, \vec{x}_N)} |H(\vec{x}_1, \dots, \vec{x}_N)|$$

Examples: H = Positive Rays

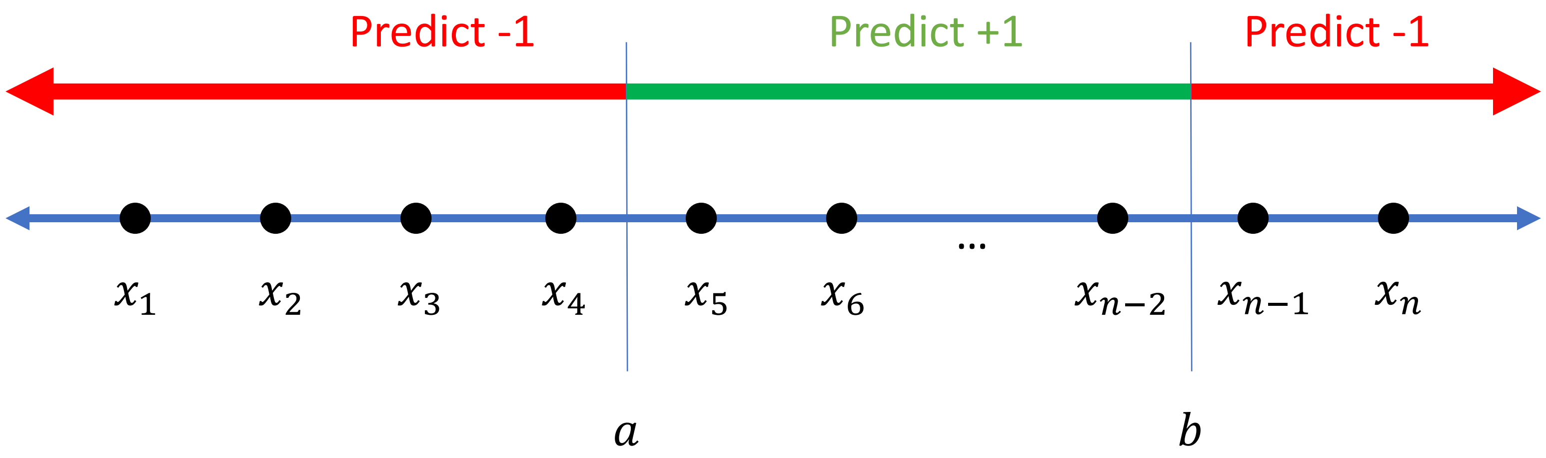
- Data points are in one-dimensional space
- Positive rays: $h(x) = sign(x - a)$



- What is $m_H(N)$?
  - $m_H(N) = N + 1$

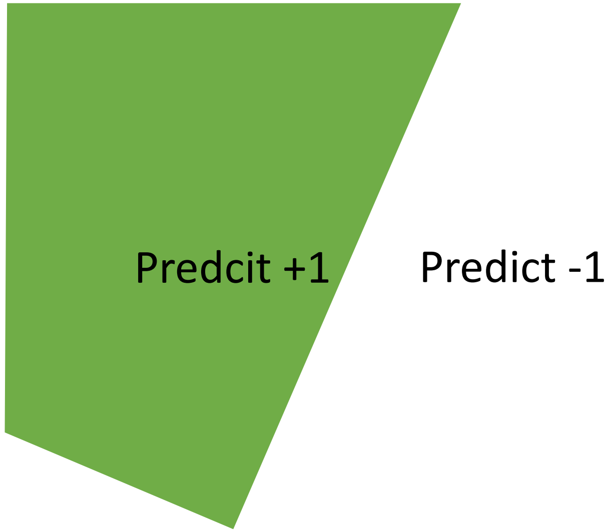
Examples: H = Positive Intervals

- What is $m_H(N)$?
  - $m_H(N) = \binom{N+1}{2} + 1 = \frac{N^2}{2} + \frac{N}{2} + 1$



Example: H = Convex Sets

- What is $m_H(N)$?
  - $m_H(N) = 2^N$



Note:
$m_H(N) \leq 2^N$ for all H and all N
(There are only 2^N possible label combinations for N $points$)

Why Growth Function?

- Growth function $m_H(N)$
  - Largest number of "effective" hypothesis H can induce on N data points
  - A more precise "complexity" measure for H
  - Goal: Replace M in finite-hypothesis analysis with $m_H(N)$
    - With prob at least $1-\delta, E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{1}{2N} ln \frac{2M}{\delta}}$

- Theorem: VC Inequality (1971)

  With prob at least $1-\delta$

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{8}{N} ln \frac{4 m_H(2N)}{\delta}}$$