# Lecture 3

# CS436/536: Introduction to Machine Learning

**Zhaohan Xi**
**Binghamton University**

**zxi1@binghamton.edu**

# Syllabus available on Brightspace

- Please review the syllabus
- Please read the CS department Academic Honesty letter to students
- Please review Watson College and University Academic Integrity policy

# HW1

- Due Monday Feb/06 before the class starts
- -3% points for each day the submission is late
- 0 points, not graded if submitted more than 5 days late
- Late days include weekends or holidays
- To be released by end of the day
- Please watch for announcement on Brightspace
- Submission on Gradescope
- Please follow TA’s instructions

# Recap Quiz Question

The perceptron model can be described mathematically as the set of functions:

$$\mathcal{H} = \left\{ h: h(\boldsymbol{x}) = sign\left( \left( \sum_{i=1}^{d} w_i x_i \right) + w_0 1 \right) \right\}$$

# Hoeffding's Inequality

Hoeffding / Chernoff proved that $\nu$ tends to be close to $\mu$, most of the time

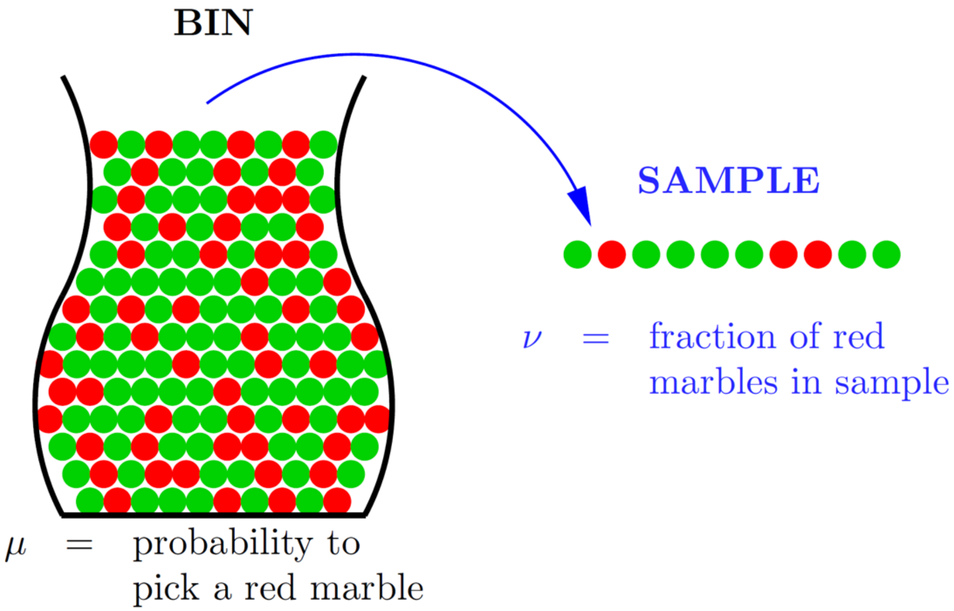
$$\boxed{\mathbb{P}[|\nu - \mu| > \epsilon] \leq 2e^{-2\epsilon^2 N}, \text{ for any } \epsilon > 0}$$

i.e. $\nu$ is approximately correct most of the time

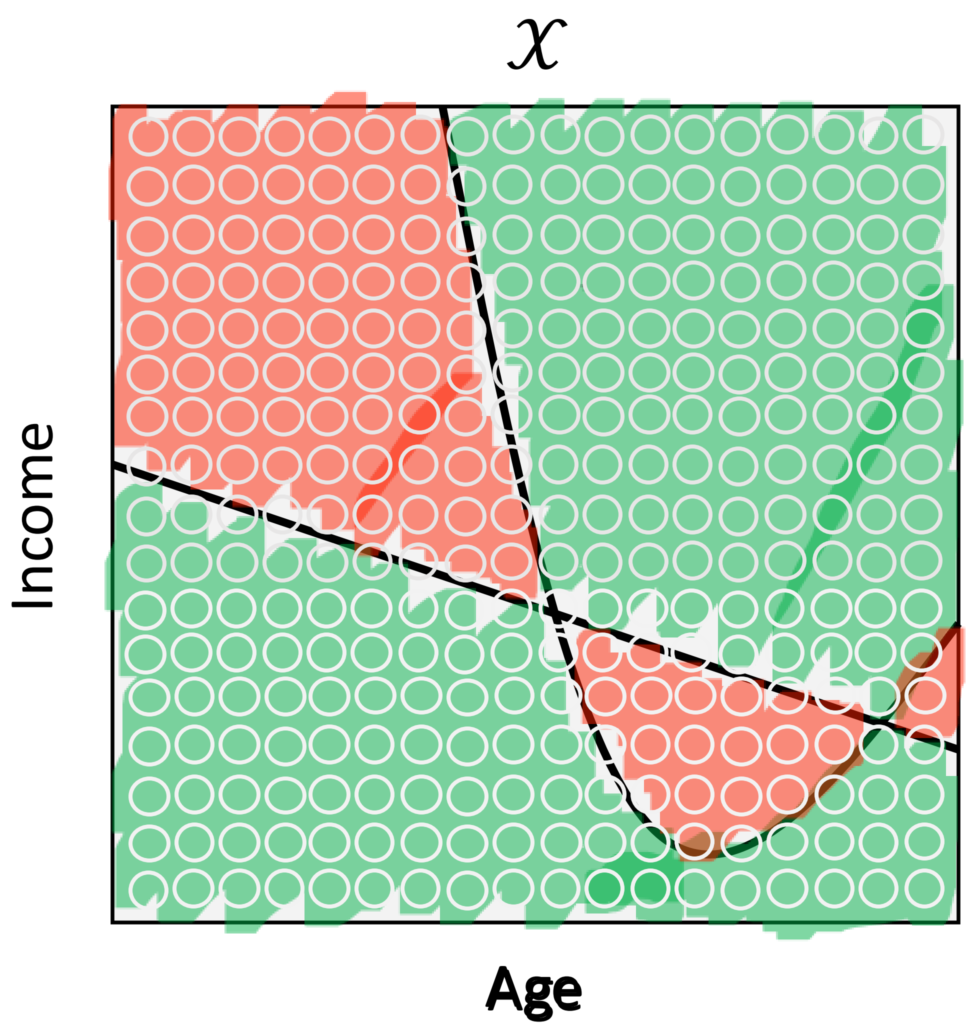
or in other words…

probably approximately correct (PAC)

We can learn *something*!



# The Error Function



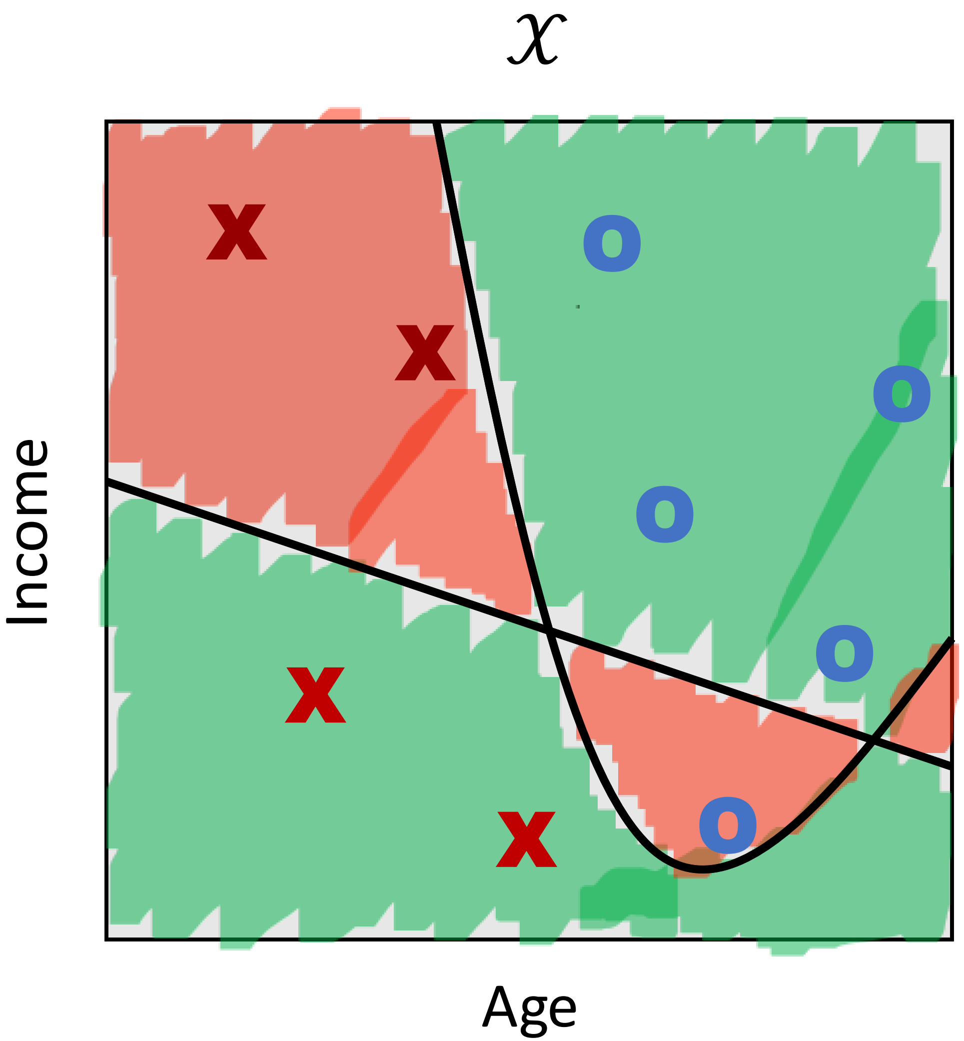

**Green:** $h(\boldsymbol{x}) = f(\boldsymbol{x})$

**Red:** $h(\boldsymbol{x}) \neq f(\boldsymbol{x})$

$$E_{out}(h) = \underbrace{\mathbb{P}_{\boldsymbol{x}}[h(\boldsymbol{x}) \neq f(\boldsymbol{x})]}_{\text{(size of red region)}}$$

But this is UNKNOWN

# The Error Function




**Green:** $h(\boldsymbol{x}) = f(\boldsymbol{x})$

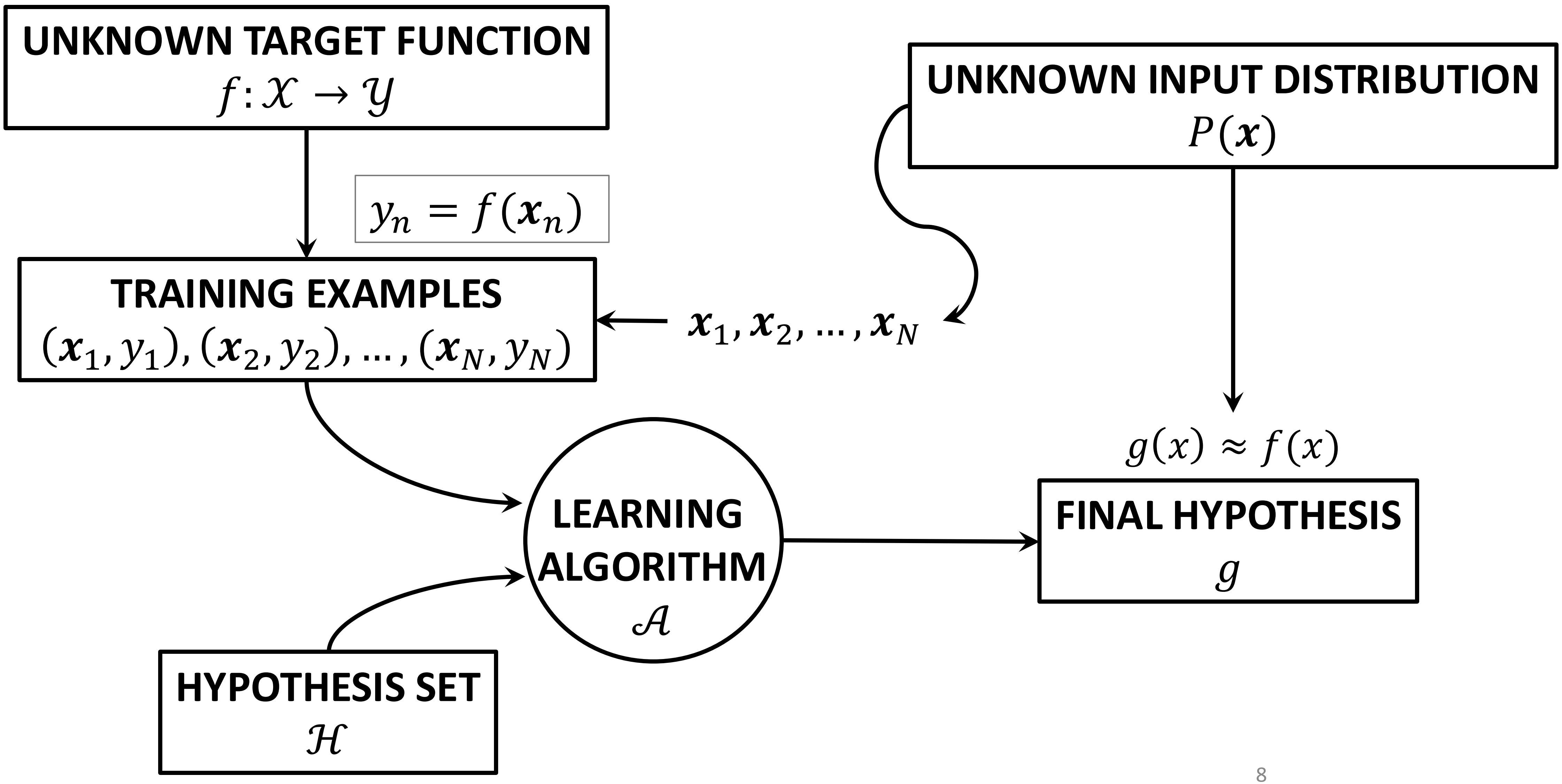
**Red:** $h(\boldsymbol{x}) \neq f(\boldsymbol{x})$

$E_{in}(h)$ = fraction of sampled data points in **red** region
i.e. misclassified data points

$$E_{in}(h) = \frac{1}{N} \sum_{n=1}^{N} [\![ h(\boldsymbol{x}_n) \neq f(\boldsymbol{x}_n) ]\!]$$

We know this

# Learning Problem Setup with Probability



# Hoeffding's Inequality for Learning

For a ***fixed*** hypothesis $h$

$$\boxed{\mathbb{P}[|E_{in}(h) - E_{out}(h)| > \epsilon] \leq 2e^{-2\epsilon^2 N}, \text{ for any } \epsilon > 0}$$

- If $E_{in} \approx 0$ then $E_{out} \approx 0$ i.e. $\mathbb{P}_{\boldsymbol{x}}[h(\boldsymbol{x}) \neq f(\boldsymbol{x})]$ with high probability i.e. $f \approx h$ over all of $\mathcal{X}$

Now: Given $h$, we can **verify** whether it is "good"

# Hoeffding's Inequality for ~~Learning~~ Verification

For a ***fixed*** hypothesis $h$

$$\boxed{\mathbb{P}[|E_{in}(h) - E_{out}(h)| > \epsilon] \leq 2e^{-2\epsilon^2 N}, \text{ for any } \epsilon > 0}$$

- If $E_{in} \approx 0$ then $E_{out} \approx 0$ i.e. $\mathbb{P}_{\boldsymbol{x}}[h(\boldsymbol{x}) \neq f(\boldsymbol{x})]$ with high probability

i.e. $f \approx h$ over all of $\mathcal{X}$

Now: Given $h$, we can **verify** whether it is "good"

# What about “Real Learning”?

- Want $g \approx f$ over all of $\mathcal{X}$

In other words: we want $g(\boldsymbol{x}) \approx f(\boldsymbol{x})$ for any $\boldsymbol{x} \in \mathcal{X}$ (even when $\boldsymbol{x} \notin \mathcal{D}$)

Want: $E_{out}(g) \approx 0$

- $E_{in}(g) \approx E_{out}(g)$
- $E_{in}(g)$ is small -- Select $g$ from $\mathcal{H}$ with minimum $E_{in}$ on $\mathcal{D}$

- But Hoeffding’s inequality only applies to a fixed hypothesis selected before seeing $\mathcal{D}$

- Will $E_{out}$ be small?

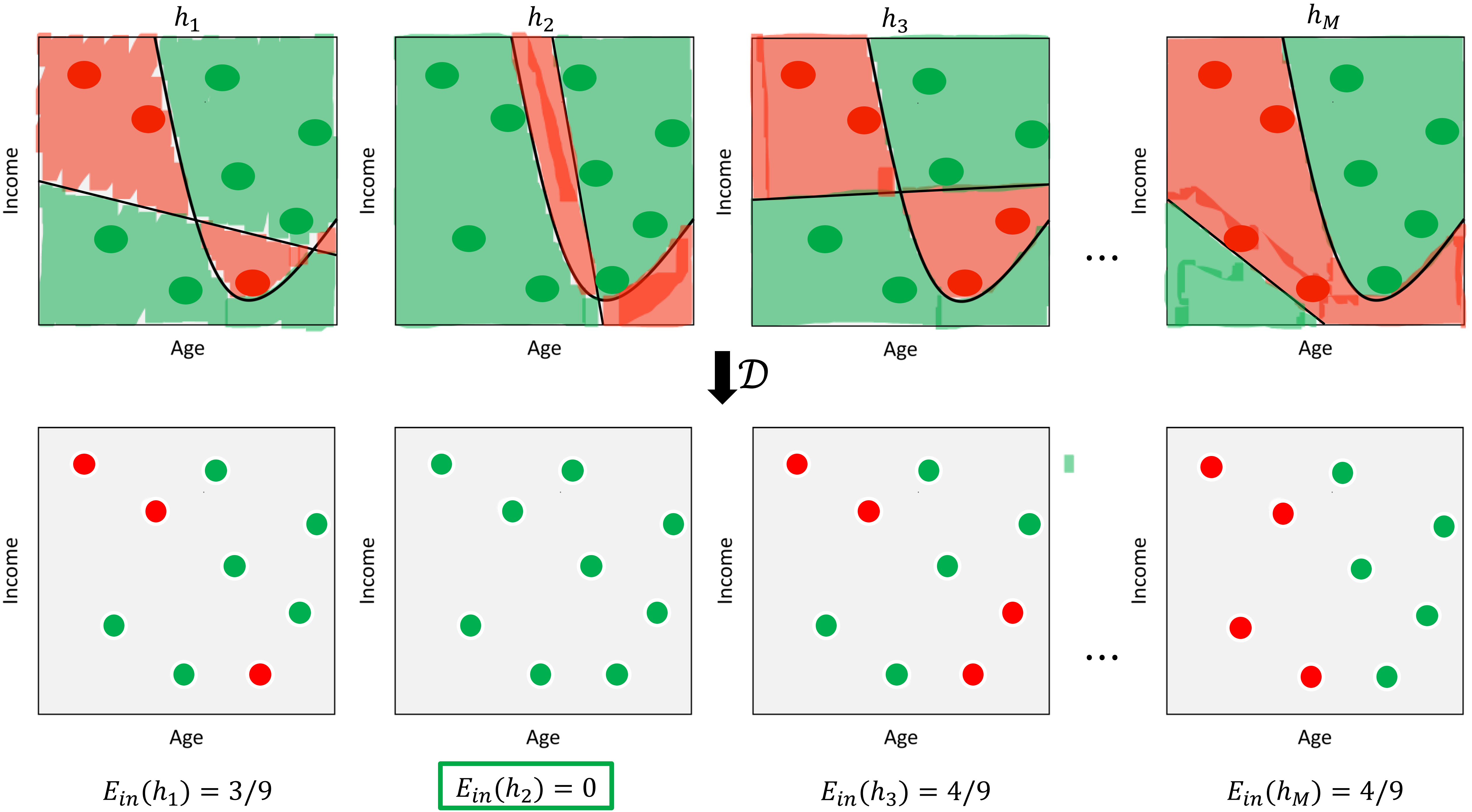
# What is Learning?

- Obtaining $f$
- Result of learning is an approximation of $f$

$$g : \mathcal{X} \rightarrow \mathcal{Y}$$

- Want:

$g \approx f$ i.e. $g(\boldsymbol{x}_*) \approx f(\boldsymbol{x}_*)$ where $\boldsymbol{x}_*$ is the next *test data point*



$h_1$
$h_2$
$h_3$
$h_M$
Income
Age
$\mathcal{D}$
$E_{in}(h_1) = 3/9$
$E_{in}(h_2) = 0$
$E_{in}(h_3) = 4/9$
$E_{in}(h_M) = 4/9$

# Selection Bias Illustrated with Coin Tossing

Statman, find me a coin guaranteed to turn up $Heads$

Run some experiments:

- Say you only have ***one*** coin.

The probability of $N$ Heads after $N$ tosses is $\frac{1}{2^N}$

- Now, suppose you toss $100$ coins, and at least one coin shows $N$ $Heads$ after $N$ tosses
  - Should we select the coin and conclude that $\mathbb{P}[Heads] \approx 1$ for the coin?

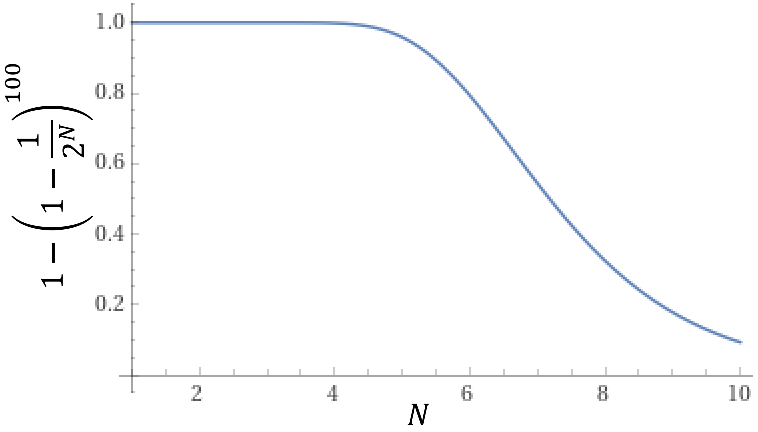
The probability that at least one among 100 coins turns up $N$ $Heads$ is $1-\left(1-\frac{1}{2^N}\right)^{100}$

# Selection Bias Illustrated with Coin Tossing

- Say you only have ***one*** coin.

The probability of $N$ Heads after $N$ tosses is $\frac{1}{2^N}$

The probability of $< N$ Heads after N tosses is $1 - \frac{1}{2^N}$




- Now, suppose you toss $100$ coins, and all coins show $< N\ Heads$ after $N$ tosses

$$\left(1 - \frac{1}{2^N}\right)^{100}$$

- Also the probability that none of the coins shows $N$ heads in $N$ tosses
- The probability that one among 100 coins turns up $N\ Heads$ is $1 - \left(1 - \frac{1}{2^N}\right)^{100}$

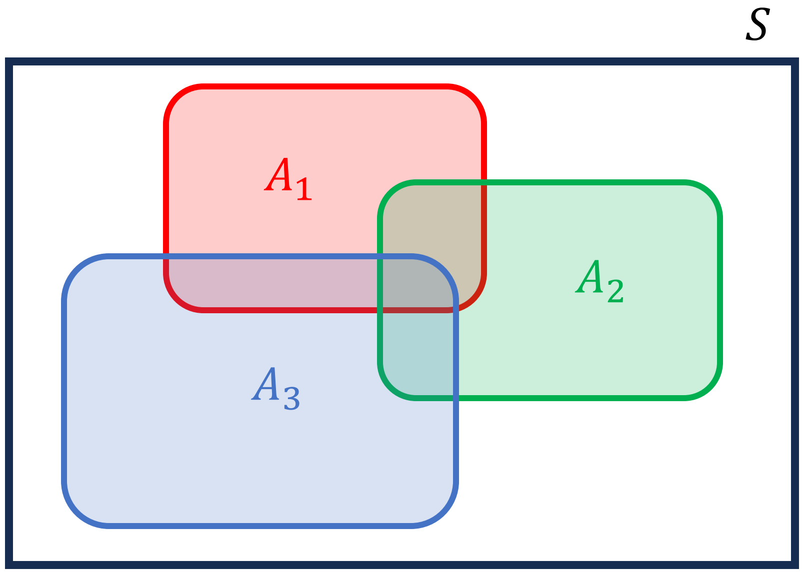
**This is Selection Bias. Search causes Selection Bias.**

# The Union Bound

For any random events $A_1, A_2, \ldots, A_n$

$$\Pr(A_1 \cup A_2 \cup \cdots A_n) \leq \Pr(\mathrm{A}_1) + \Pr(\mathrm{A}_2) + \cdots + \Pr(A_n)$$

Also accepted: $\Pr(A_1 \cup A_2 \cup \cdots A_n) = \sum_{i=1}^{n} \Pr(A_i) - \sum_{i<j} \Pr\left(A_i \cap A_j\right)$
$+ \sum_{i<j<k} \Pr\left(A_i \cap A_j \cap A_k\right)$
$- \cdots + (-1)^n \Pr(\cap_{i=1}^{n} A_i)$



# Implication Rule

For any random events $A, B$,

If $A$ implies $B$ $(A \Rightarrow B)$, then

$$\Pr(A) \qquad \leq \qquad \Pr(B)$$

# Hoeffding's Inequality for Learning (from finite $\mathcal{H}$)

Bound $\mathbb{P}[|E_{in}(g) - E_{out}(g)| > \epsilon]$ no matter which $g$ is picked from $\mathcal{H}$

$$
\begin{aligned}
|E_{in}(g) - E_{out}(g)| > \epsilon \Rightarrow \quad & |E_{in}(h_1) - E_{out}(h_1)| > \epsilon \\
\text{or} \quad & |E_{in}(h_2) - E_{out}(h_2)| > \epsilon \\
\cdots \quad & \\
\text{or} \quad & |E_{in}(h_M) - E_{out}(h_M)| > \epsilon
\end{aligned}
$$

Implication Rule: If $A \Rightarrow B$, then $\mathbb{P}[A] \le \mathbb{P}[B]$

$$\mathbb{P}[|E_{in}(g) - E_{out}(g)| > \epsilon] \quad \le \quad \mathbb{P}[\mathrm{OR}_{m=1}^{M} |E_{in}(h_m) - E_{out}(h_m)| > \epsilon]$$

# Hoeffding's Inequality for Learning (from finite $\mathcal{H}$)

Union Bound: $\mathbb{P}[A \text{ } or \text{ } B] = \mathbb{P}[A \cup B] \leq \mathbb{P}[A] + \mathbb{P}[B]$

So long as $g$ is picked from $\mathcal{H}$, where $|\mathcal{H}| = M$:

$$\begin{aligned}
\mathbb{P}[|E_{in}(g) - E_{out}(g)| > \epsilon] \quad &\leq \quad \mathbb{P}[\mathrm{OR}_{m=1}^{M} |E_{in}(h_m) - E_{out}(h_m)| > \epsilon] \\
&\leq \quad \sum_{m=1}^{M} \mathbb{P}[|E_{in}(h_m) - E_{out}(h_m)| > \epsilon] \\
&\leq \quad M 2 e^{-2\epsilon^2 N}
\end{aligned}$$

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

# Interpreting Hoeffding's Bound for Finite $\mathcal{H}$

So long as $g$ is picked from $\mathcal{H}$,

**Theorem.** With probability at least $1 - \delta$,

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

where $\delta = 2|\mathcal{H}|e^{-2\epsilon^2 N}$

# Real Learning is Feasible

$$E_{out}(g) \le E_{in}(g) + \sqrt{\frac{1}{2N}\log\frac{2|\mathcal{H}|}{\delta}} \qquad = O\left(\sqrt{\frac{\log|\mathcal{H}|}{N}}\right)$$

If $N \gg \log|\mathcal{H}|$, then $E_{out}(g) \approx E_{in}(g)$

- No matter how $g$ is selected
- Does not depend on $\mathcal{X}, P(\boldsymbol{x})$, or the target function $f$
- Only requires that the data set $\mathcal{D}$ and the test point can be generated *independently* from $P(\boldsymbol{x})$

# Achieving Learning: $E_{out} \approx 0$

2 Conditions:

(1) $E_{in}(g) \approx E_{out}(g)$ $\Rightarrow$ $E_{out}(g) \approx 0$

(2) $E_{in}(g) \approx 0$

How to ensure that (1) is satisfied? We cannot compute $E_{out}(g)$

- Must be ensured theoretically (e.g. using Hoeffding's inequality)

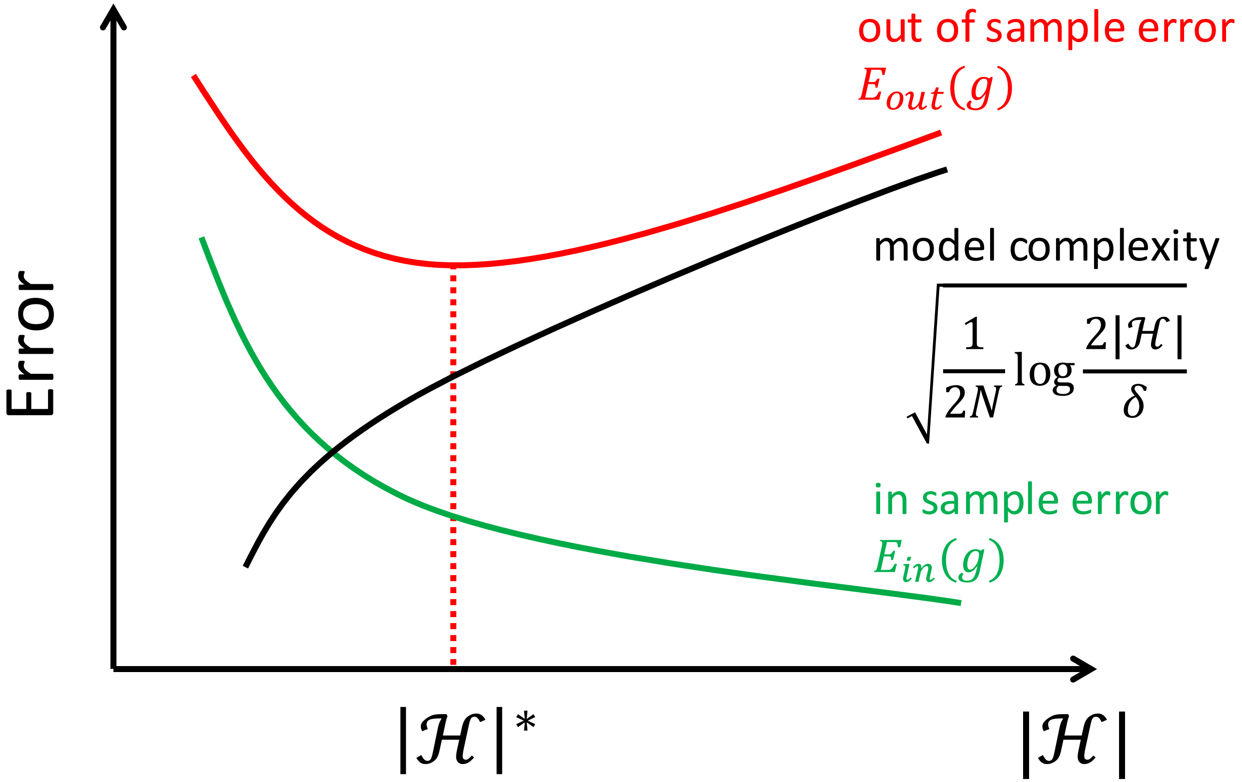
How to ensure (2) is satisfied? Use a good learning algorithm (e.g. PLA)

# But... There is a Tradeoff: The Complexity of $\mathcal{H}$

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

- Smaller $|\mathcal{H}|$ $\Rightarrow E_{out}(g) \approx E_{in}(g)$
- Larger $|\mathcal{H}|$ $\Rightarrow E_{in}(g) \approx 0$

For Fixed $N$, $\delta$:



# Feasibility of Learning (with Finite Models)

- No Free Lunch: Cannot learn $f$ *exactly* from $\mathcal{D}$ over all $\mathcal{X}$
- But, Can learn $f$ with high probability due to Hoeffding, if:

  - $\mathcal{D}$ and the test data point are drawn i.i.d. from $P(\boldsymbol{x})$

  - $\mathcal{H}$ is fixed and $g$ is selected from $\mathcal{H}$

To achieve learning: i.e. select $g$ from $\mathcal{H}$ so that $E_{out}(g) \approx 0$, we must ensure:

(Step 1) $E_{out}(g) \approx E_{in}(g)$ -- Ensure $|\mathcal{H}|$ is small

**Theorem.** With probability at least $1 - \delta$, $E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{1}{2N} \log \frac{2|\mathcal{H}|}{\delta}}$

(Step 2) $E_{in}(g) \approx 0$ -- Learning algorithm $\mathcal{A}$

# The complexity of $f$

More complex target functions are harder to learn

- Simple $f$ $\Rightarrow$ can use small $\mathcal{H}$ to get $E_{in}(g) \approx 0$ using smaller $N$
- Complex $f$ $\Rightarrow$ need large $\mathcal{H}$ to get $E_{in}(g) \approx 0$ and need larger $N$