Feasibility of Learning

Starfly
starfly3119@gmail.com

Beihang University — February 14, 2020

Introduction

We have four sections. The first section tells us **absolutely no free lunch outside** D; the second section tells us **probably approximately correct outside** D; the third section tells us **verification possible if** $E_{in}(h)$ **for fixed** h; and the fourth section tells us **learning possible if** $|H|$ **finite and** $E_{in}(g)$ **small**

1 Learning is impossible?

Learning from D (to infer something outside D) is doomed if any 'unknown' f can happen.

2 Probability to the Rescue

2.1 Inferring Something Unknown

Hoeffding's Inequality

- In big sample (N large), υ is probably close to μ (within ϵ)

$$P[|\upsilon - \mu| > \epsilon] \leq 2\exp(-2\epsilon^2 N) \tag{1}$$

where μ = orange probability in bin, υ = orange fraction in sample, N is the sample of size.

The statement '$\upsilon = \mu$' is probably approximately correct (PAC)

- valid for all N and ϵ
- does no depend on μ, no need to 'know' μ
- larger sample size N or looser gap $\epsilon \Rightarrow$ higher probability for '$\upsilon \approx \mu$'

If large N, can probably infer unknown μ by knwon υ

3 Connection to Learning

3.1 Learning?

bin

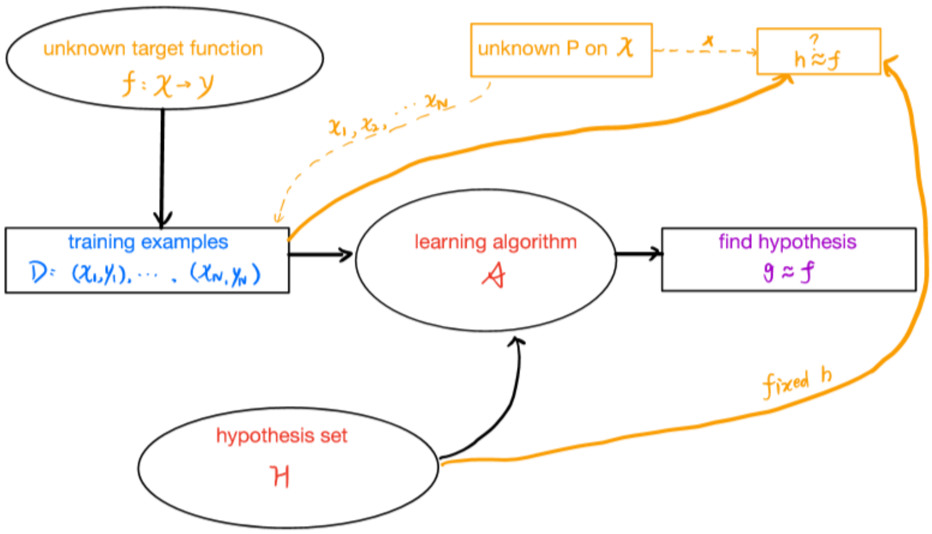
- unknown orange prob. μ
- marble $\bullet \in$ bin

- orange •
- green •
- size-N sample from bin of i.i.d marbles

learning

- fixed hypothesis $h(x) \stackrel{?}{=}$ target $f(x)$
- $x \in X$
- h is wrong $\Leftrightarrow$ $h(x) \neq f(x)$
- h is right $\Leftrightarrow$ $h(x) = f(x)$
- check h on $D = \{(x_n, y_n)\}$ with i.d.d x_n

If large N & orange i.d.d. x_n, can probably infer unknown $[h(x) \neq f(x)]$ probability by knwon $[h(x_n) \neq y_n]$ fraction



For any fixed h, can probably infer unknown

$$E_{out}(h) = \epsilon_{x \sim P}[h(x) \neq f(x)] \tag{2}$$

by colororange known

$$E_{in}(h) = \frac{1}{N} \sum_{n=1}^{N} [h(x_n) \neq y_n] \tag{3}$$

For any fixed h, in 'big' data (N) large, in-sample error $E_{in}(h)$ is probably close to out-of-sample error $E_{out}(h)$ (within ϵ)

$$P[|E_{in}(h) - E_{out}(h)| > \epsilon] \leq 2\exp(-2\epsilon^2 N) \tag{4}$$

- valid for all N and ϵ
- does not depend on $E_{out}(h)$, no need to 'know' $E_{out}(h)$
  -f and P can stay unknown
- '$E_{in}(h) = E_{out}(h)$' is probably approximatedly correct (PAC)

For any fixed h, when data large enough,

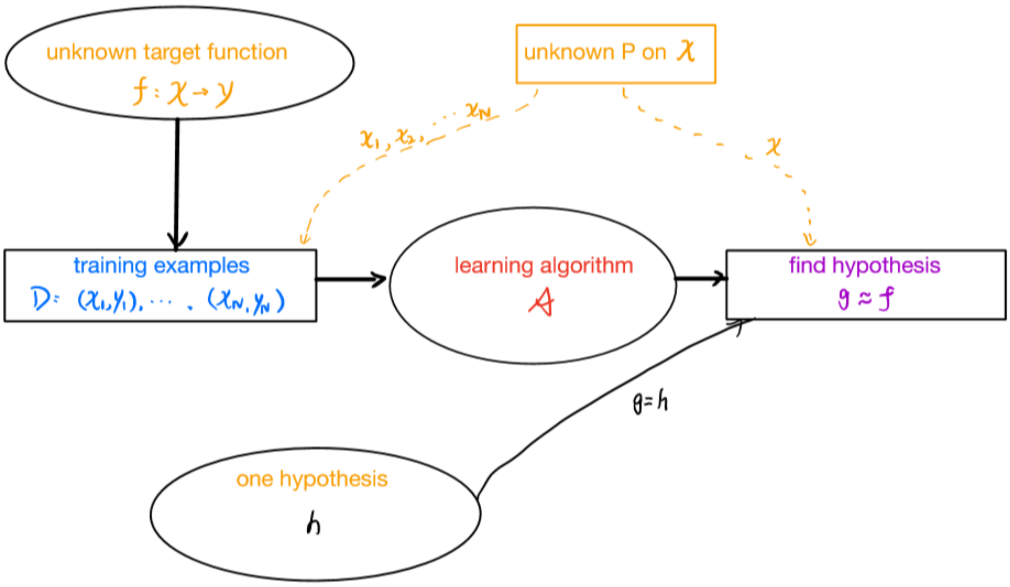
$$E_{in}(h) \approx E_{out}(h) \tag{5}$$

Question 1 *Can we claim 'good learning' ($g \approx f$)*

- Yes!
  if $E_{in}(h)$ small for the fixed h and A pick the h as g
  $\Rightarrow' g = f'$ PAC
- No! if A forced to pick THE h as g
  $\Rightarrow E_{in}(h)$ almost always not small
  $\Rightarrow' g \neq f'$ PAC

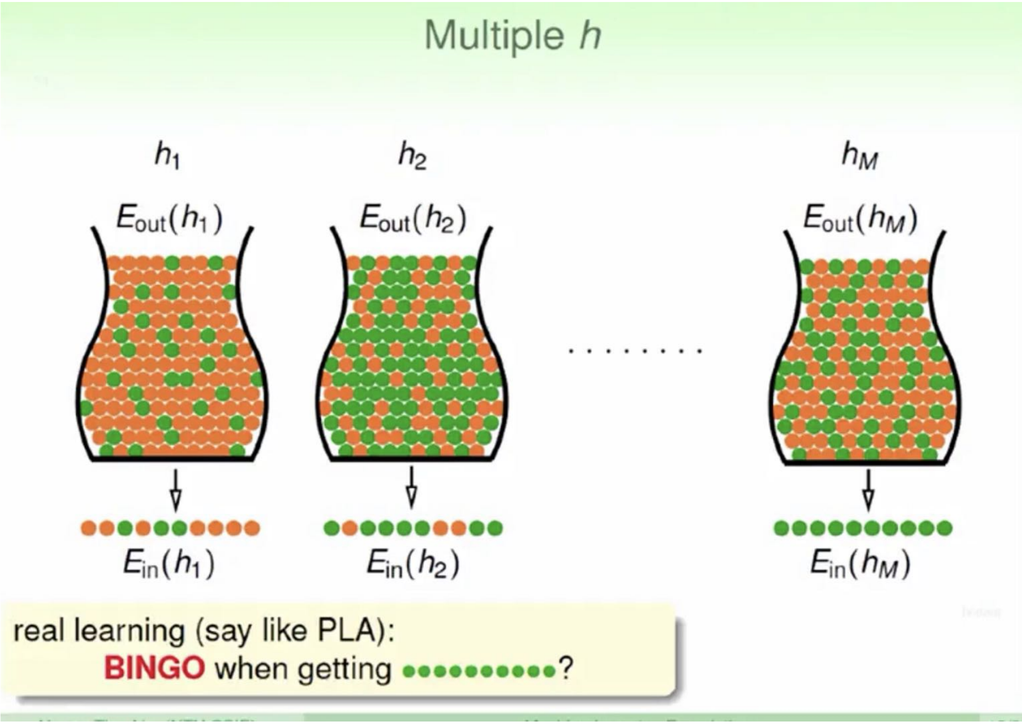
real learning: A shall make choices $\in H$ (like PLA), rather that being forced to pick one h

3.2 Verification Flow

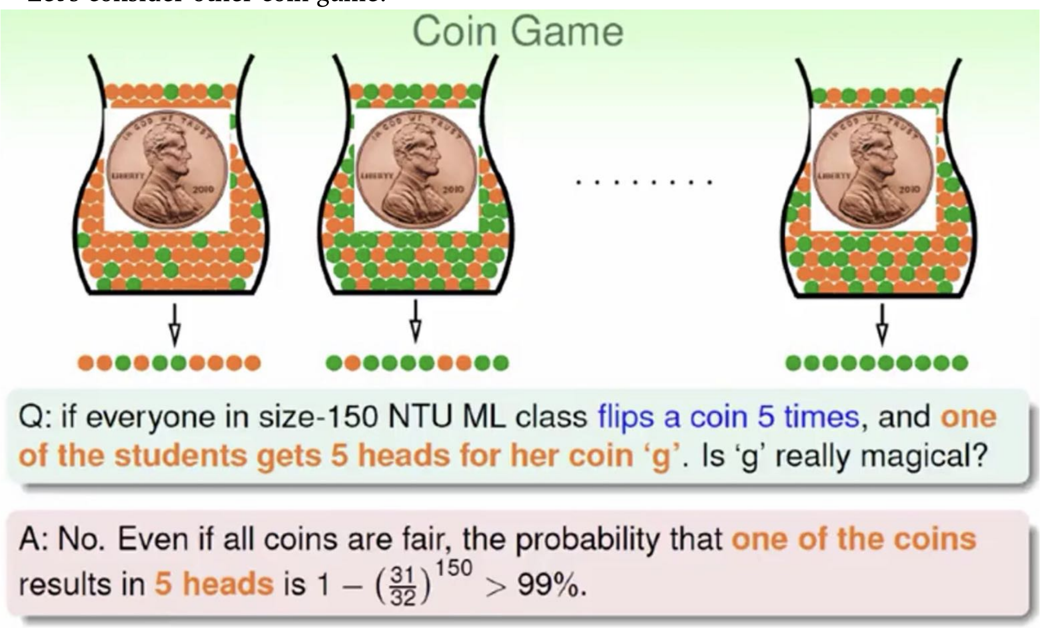


Can now use 'historical records'(data) to verify 'one candidate formula' h

4 Connection to Real Learning



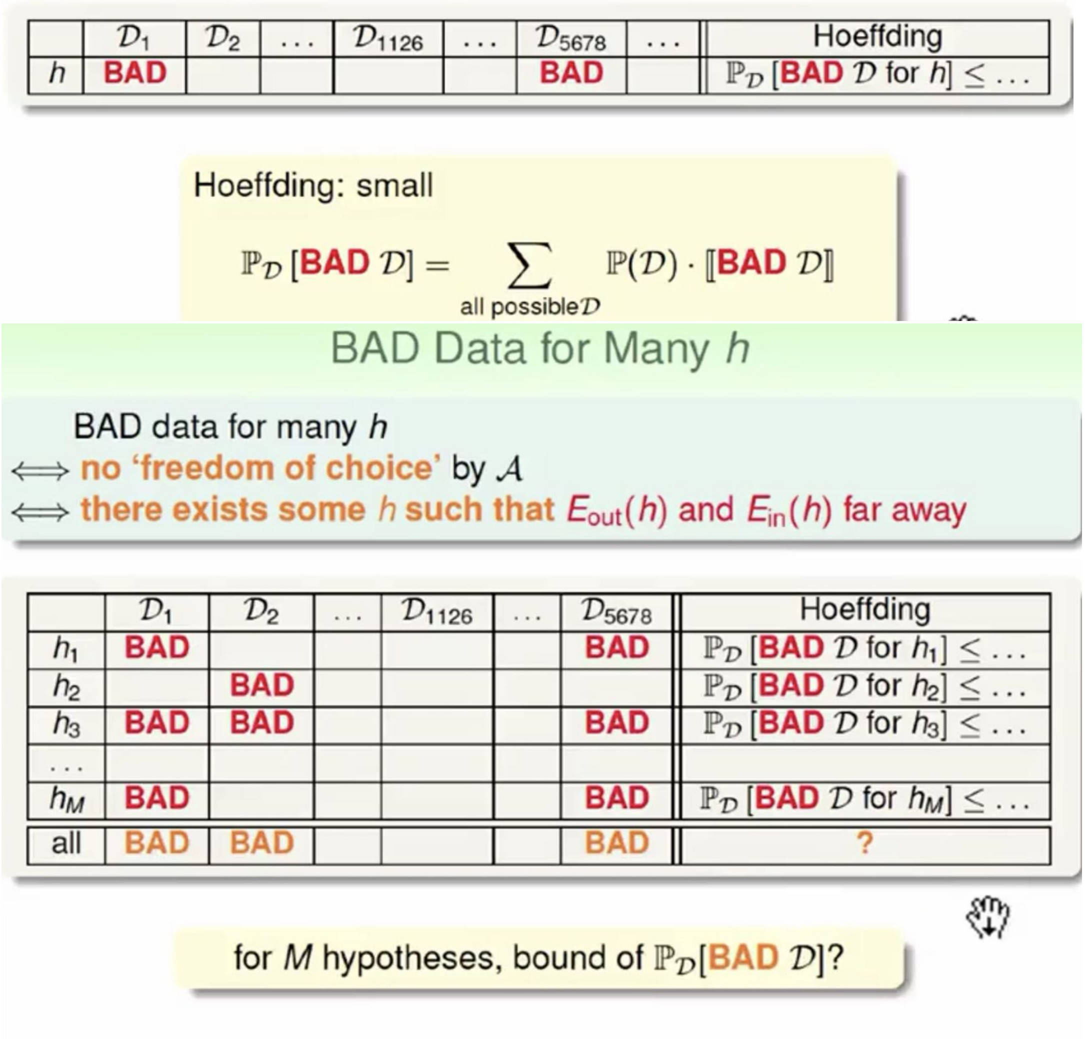
Let's consider other coin game:



- Bad sample: E_{in} and E_{out} far away - can get worse when involving 'choice'
  e.g., $E_{out} = \frac{1}{2}$, but getting all heads $(E_{in} = 0)$
- Bad data for one h: $E_{out}(h)$ and $E_{in}(h)$ far away
  e.g., E_out big (far from f), but E_{in} small(correct on most examples)

4.1 Bad data

D_i is different sample sets with size N



	$\mathcal{D}_1$	$\mathcal{D}_2$	...	$\mathcal{D}_{1126}$	...	$\mathcal{D}_{5678}$	...	Hoeffding
h	BAD					BAD		$\mathbb{P}_{\mathcal{D}}$ [BAD $\mathcal{D}$ for h] $\le \ldots$

Hoeffding: small

$$\mathbb{P}_{\mathcal{D}}[\textbf{BAD } \mathcal{D}] = \sum_{\text{all possible} \mathcal{D}} \mathbb{P}(\mathcal{D}) \cdot [\![\textbf{BAD } \mathcal{D}]\!]$$

BAD Data for Many h

BAD data for many h
$\Longleftrightarrow$ no 'freedom of choice' by $\mathcal{A}$
$\Longleftrightarrow$ there exists some h such that $E_{\text{out}}(h)$ and $E_{\text{in}}(h)$ far away

	$\mathcal{D}_1$	$\mathcal{D}_2$	...	$\mathcal{D}_{1126}$	...	$\mathcal{D}_{5678}$	Hoeffding
h_1	BAD					BAD	$\mathbb{P}_{\mathcal{D}}$ [BAD $\mathcal{D}$ for h_1] $\le \ldots$
h_2		BAD					$\mathbb{P}_{\mathcal{D}}$ [BAD $\mathcal{D}$ for h_2] $\le \ldots$
h_3	BAD	BAD				BAD	$\mathbb{P}_{\mathcal{D}}$ [BAD $\mathcal{D}$ for h_3] $\le \ldots$
...							
h_M	BAD					BAD	$\mathbb{P}_{\mathcal{D}}$ [BAD $\mathcal{D}$ for h_M] $\le \ldots$
all	BAD	BAD				BAD	?

for M hypotheses, bound of $\mathbb{P}_{\mathcal{D}}$[BAD $\mathcal{D}$]?

Bound of Bad Data

$$
\begin{aligned}
P_D[BAD\ D] &= P_D[BAD\ D\ for\ h_1 \textbf{ or } BAD\ D\ for\ h_2 \textbf{ or } ...or\ BAD\ D\ for\ h_M] \\
&\le P_D[BAD\ D\ for\ h_1] + P_D[BAD\ D\ for\ h_2] + ... + P_D[BAD\ D\ for\ h_M] \\
&\le 2\exp(-2\epsilon^2 N) + 2\exp(-2\epsilon^2 N) + ... + 2\exp(-2\epsilon^2 N) \\
&= 2M\exp(-2\epsilon^2 N)
\end{aligned} \tag{6}
$$

- finite-bin version of Hoeffding, valid for all M, N and ϵ
- does not depend on any $E_{out}(h_m)$, no need to 'know' $E_{out}(h_m)$
  - f and P can stay unknown
- '$E_{in}(g) = E_{out}(g)$' is PAC, regardless of A

'most reasonable' A (like PLA/pocket): pick the h_m with lowest $E_{in}(h_m)$ as g

4.2 The 'Statistical' Learning Flow

- if $|H| = M$ finite, N large enough, for whatever g picked by A, $E_{out}(g) \approx E_{in}(g)$

- if A finds one g with $E_{in}(g) \approx 0$, PAC guarantee for $E_{out}(g) \approx 0 \Rightarrow$ learning possible!