Lecture 18: Learning Theory

COSC 480 Data Science, Spring 2017

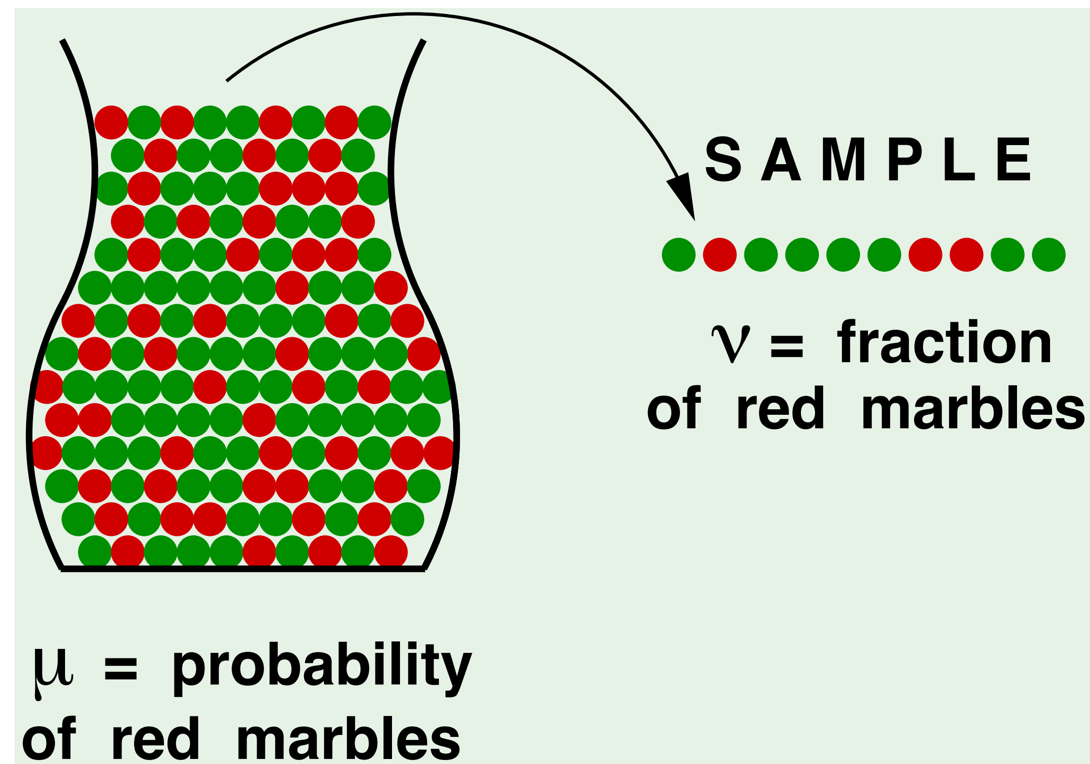
Michael Hay

Logistics

- Lab "6" (Project work) due tonight
  - Be sure to update weekly status
- Lab 7 out
  - Please read before lab tomorrow

Related experiment

- Consider a "bin" with red and green marbles
  - P(picking a red marble) = μ
  - P(picking a green marble) = 1 - μ
- Value of μ is <u>unknown</u> to us
- We pick *n* marbles* independently
- Fraction of red marbles in sample = ν



* sample with replacement

Credit: Yaser Abu Mostafa https://work.caltech.edu/

Does $\boldsymbol{\nu}$ say anything about μ?

What ν says about μ

- In a big sample (large *n*), is "**likely**" to be "**close**" to
- Formally,

"**close**" you can choose what
ε is… 0.1, 0.01, 0.001, etc.

$$P(|\nu - \mu| > \epsilon) \leq 2\exp(-2\epsilon^2 n)$$

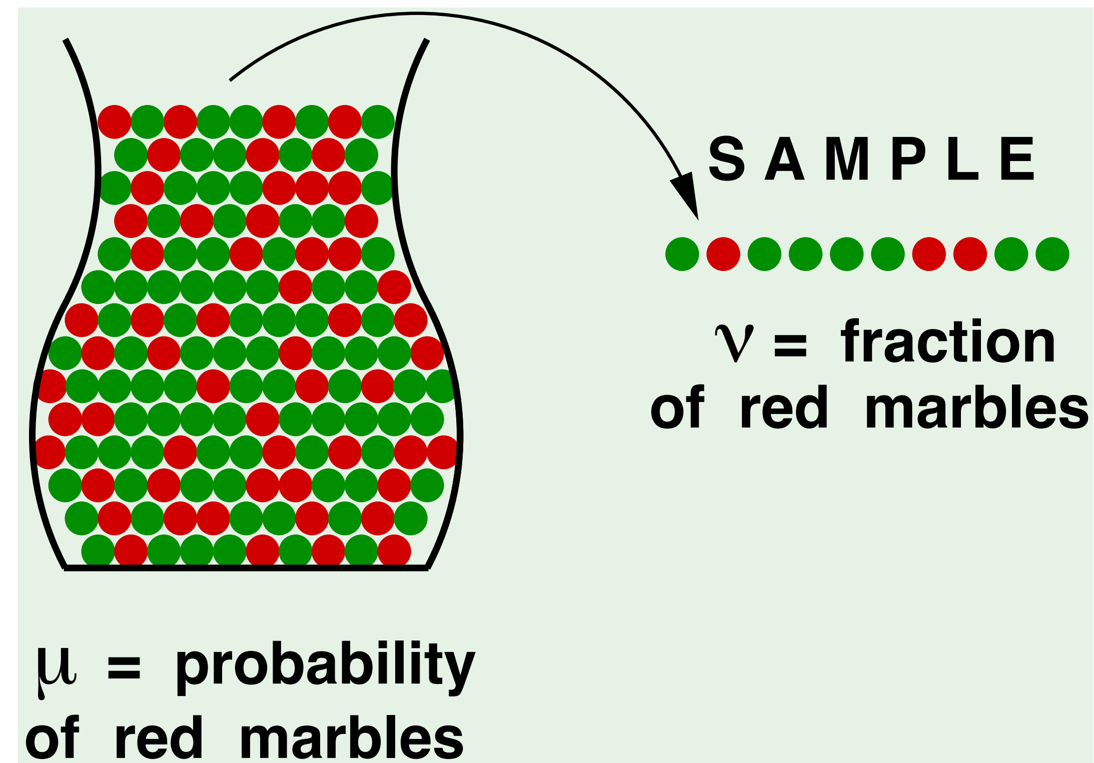
"**likely**" this gets very
small as *n* grows

- This is **Hoeffding's inequality**

Credit: Yaser Abu Mostafa https://work.caltech.edu/

$$P(|\nu - \mu| > \epsilon) \leq 2\exp(-2\epsilon^2 n)$$

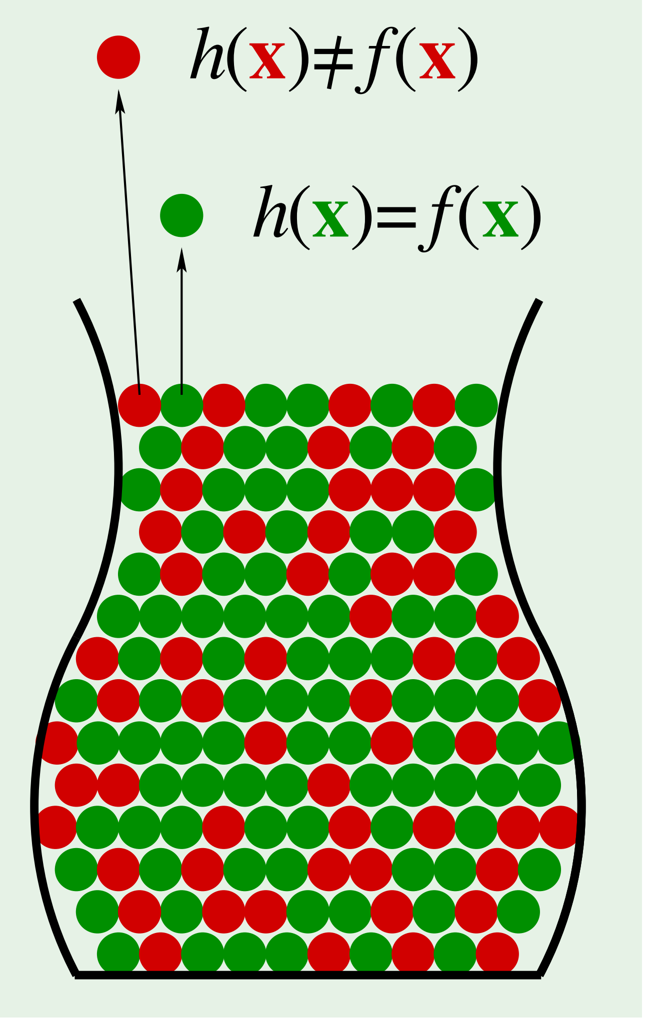
- Valid for all *n* and **ε**
- Bound does not depend on **μ**, **ν**
- Only assumption: samples are [independent]
- Tradeoff: n, **ε**, and probability bound



Credit: Yaser Abu Mostafa https://work.caltech.edu/

Connection to learning

- Bin: the unknown is a number μ
- Learning: the unknown is a function *f*
- Each marble is an input x
    - Green marble: hypothesis got it right: $h(x) = f(x)$
    - Red marble: hypothesis got it wrong: $h(x) \neq f(x)$

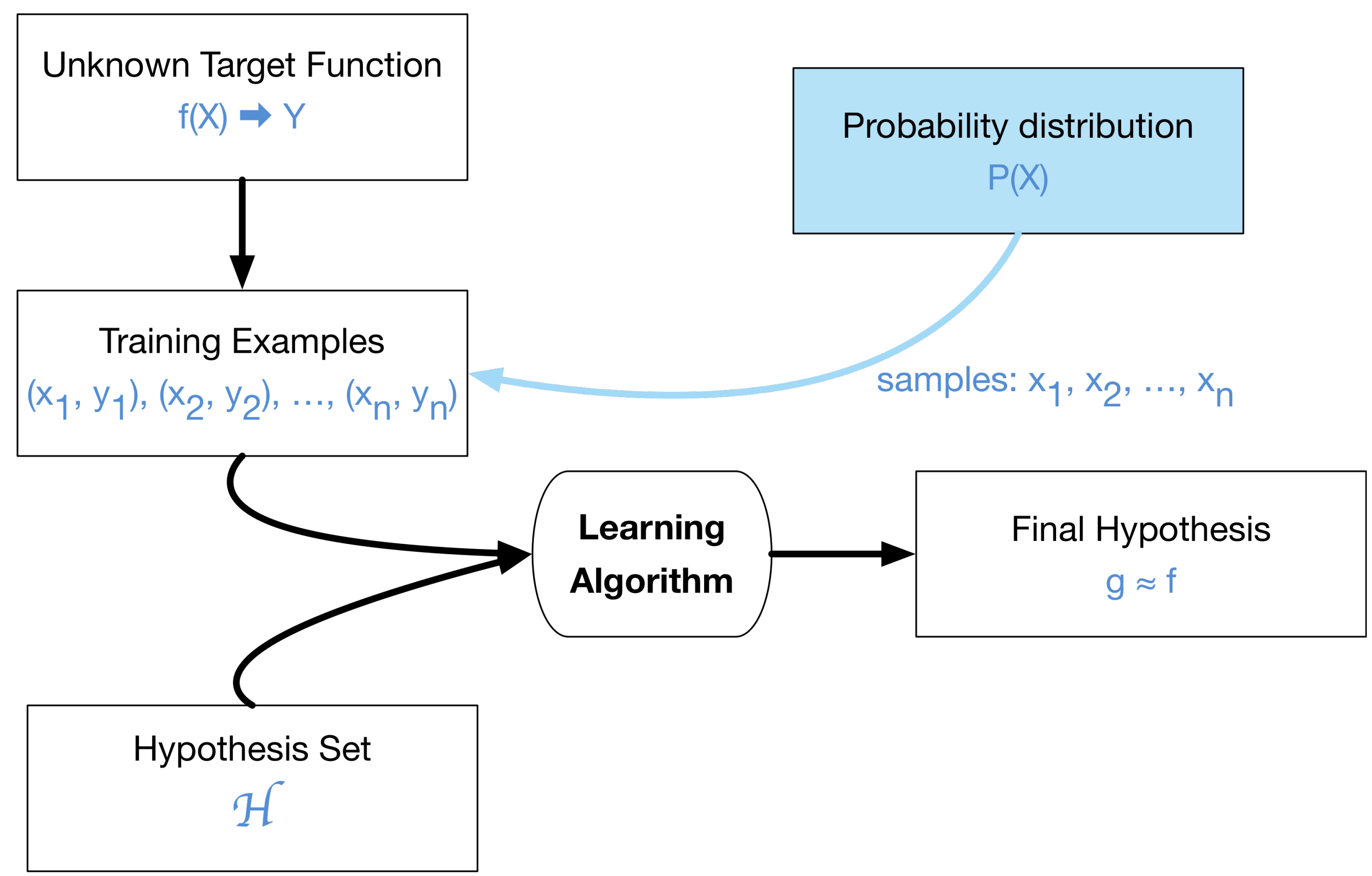


Credit: Yaser Abu Mostafa https://work.caltech.edu/

Learning diagram

Bin analogy:

- Each training data point x_i is a sample from from a "bin" of possible x's
- But what about y_i?
  - Each y_i is generated by applying *f*, as in $y_i = f(x_i)$
  - Small detail: this assumes f is deterministic; story not that different for noisy f

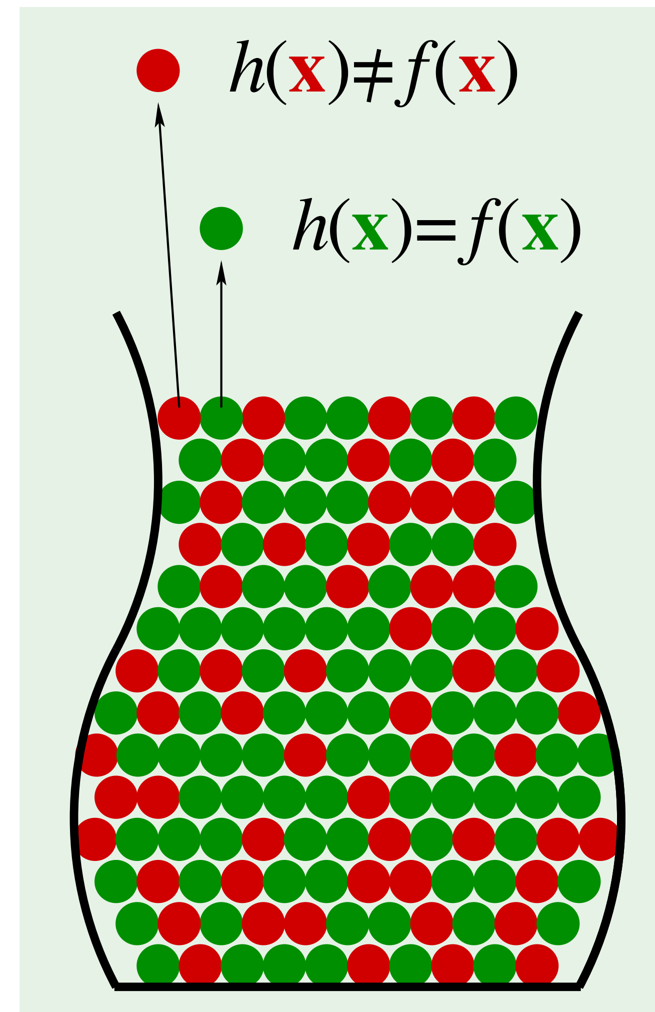


Credit: Yaser Abu Mostafa https://work.caltech.edu/

What we have so far

- Our training data is sampled from unknown distribution over X
- We apply hypothesis *h* on training data and observe *low error*: $\nu \approx 0$
- Are we done? Have we learned a good hypothesis?

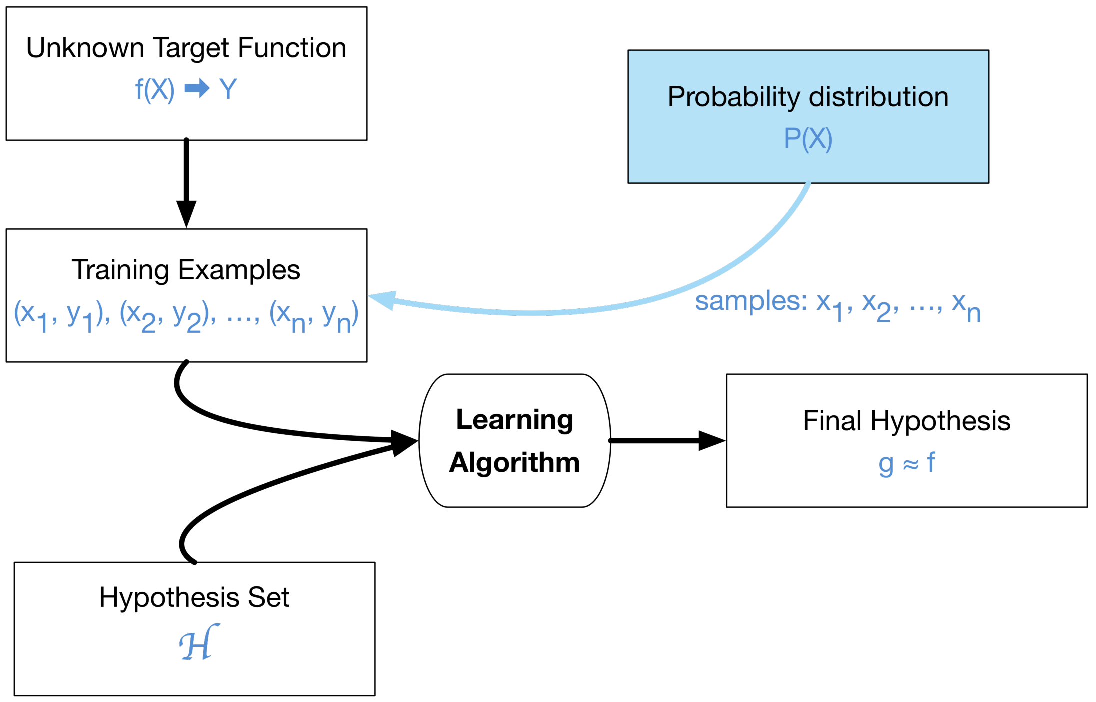
  In other words, is it likely that $\mu \approx 0$?



Credit: Yaser Abu Mostafa https://work.caltech.edu/

Are we done?

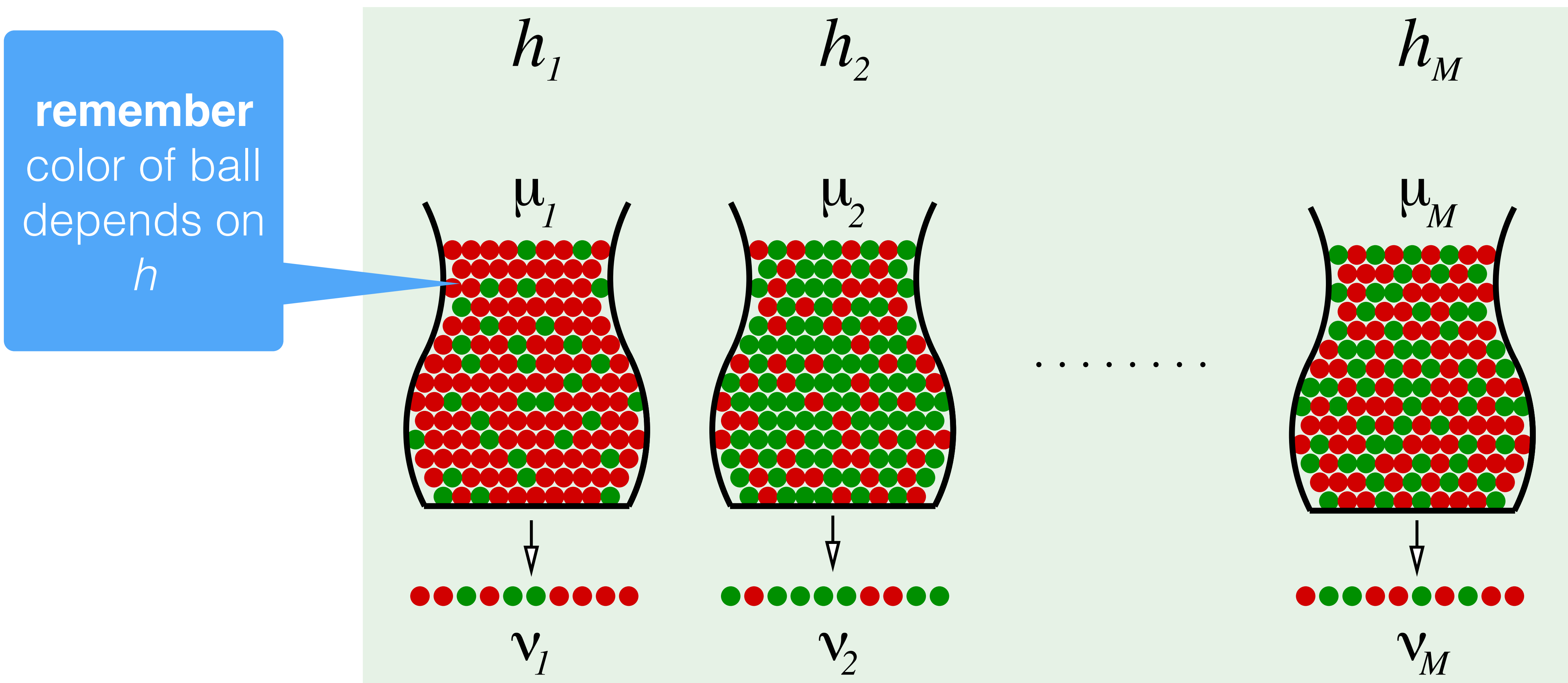
- Not so fast!
- Previous analysis treats hypothesis *h* is fixed.
- What's missing?
    - Learning! We start with $\mathcal{H}$ and **choose** a *single* hypothesis out of *many*.



Credit: Yaser Abu Mostafa https://work.caltech.edu/

Multiple bins

- Extending the bin analogy to more than one hypothesis



Credit: Yaser Abu Mostafa https://work.caltech.edu/

$$I[s] = \begin{cases} 1 & \text{if } s \text{ is "true"} \\ 0 & \text{if } s \text{ is "false"} \end{cases}$$

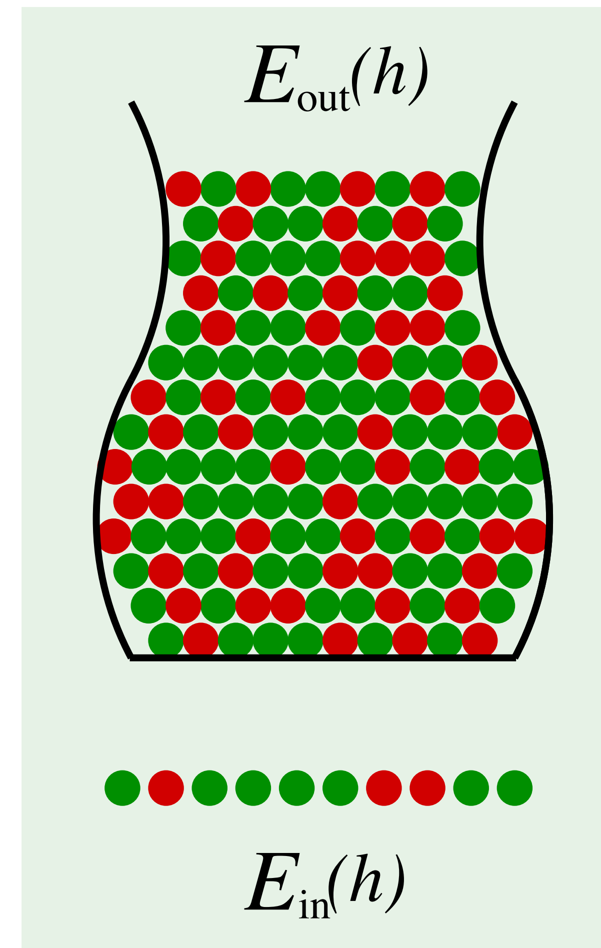
Notation

- Both ν and μ depend on which hypothesis *h*
  - ν is "in sample" error, denoted

$$E_{in}(h) = \frac{1}{n}\sum_{i=1}^{n} I[h(x_i) \neq y_i]$$

  - μ is "out of sample" error, denoted

$$E_{out}(h) = \sum_{x \in \mathcal{X}} P(x) I[h(x) \neq f(x)]$$



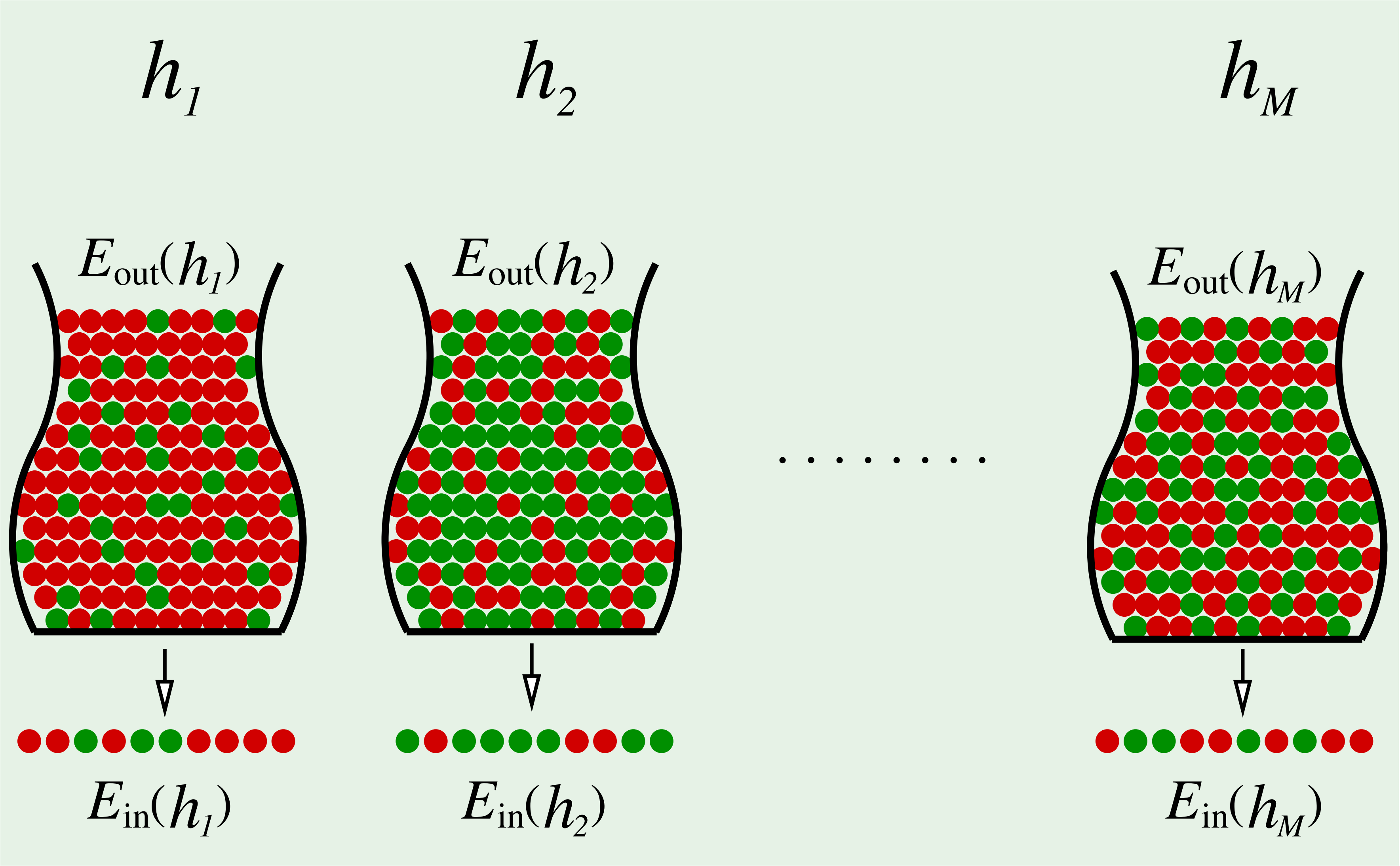
- Hoeffding inequality becomes

$$P(|E_{in}(h) - E_{out}(h)| > \epsilon) \leq 2\exp(-2\epsilon^2 n)$$

Credit: Yaser Abu Mostafa https://work.caltech.edu/

Notation with multiple bins

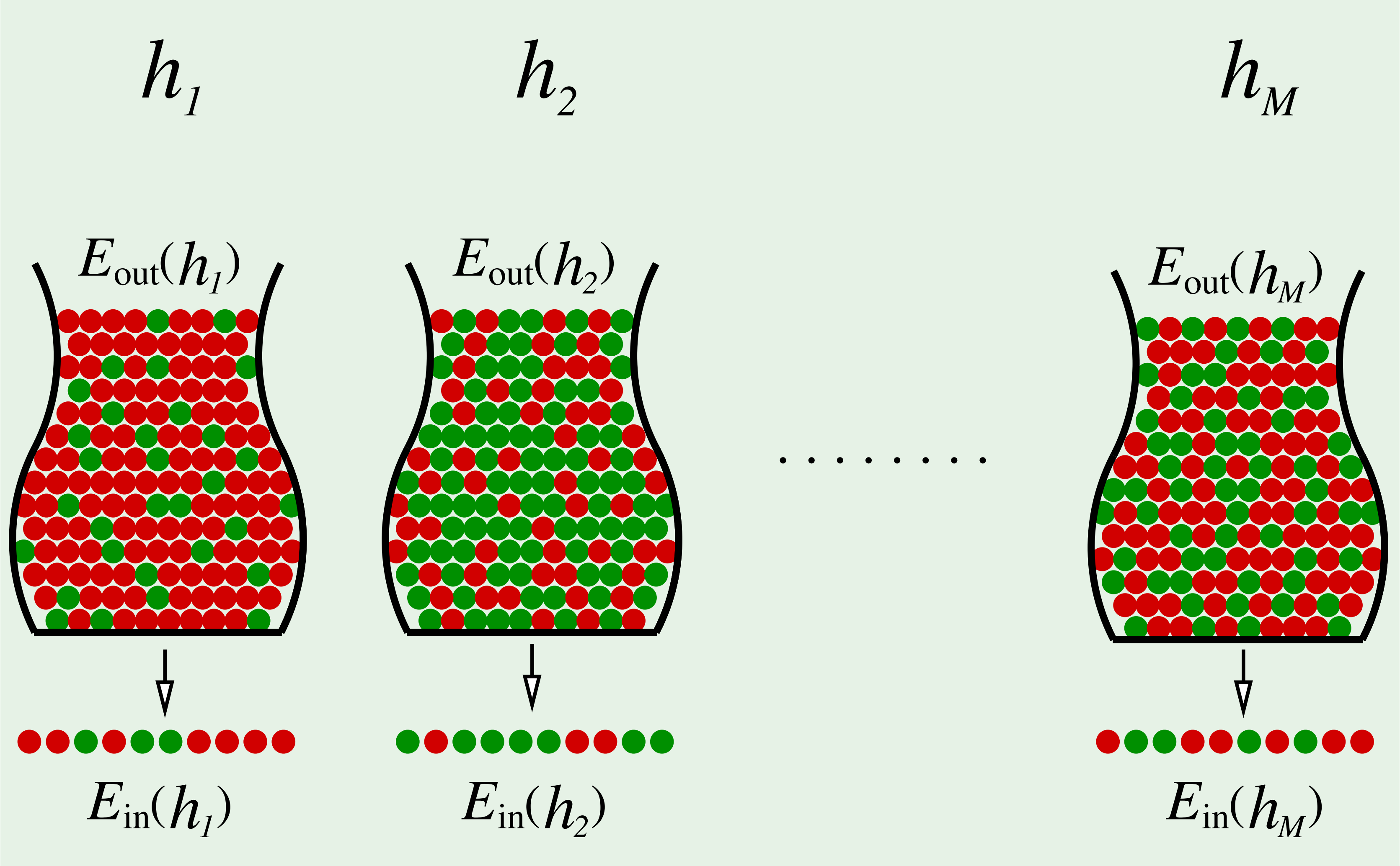
- Extending the bin analogy to more than one hypothesis



Credit: Yaser Abu Mostafa https://work.caltech.edu/

Now are we done?

- Not so fast! Hoeffding bound does not apply to multiple hypotheses.



Credit: Yaser Abu Mostafa https://work.caltech.edu/

Question

Instructions: ~1 minute to think/ answer on your own; then discuss with neighbors; then I will call on one of you

Where *could* we apply the Hoeffding bound? Hint: usually when we do ML we divide our data into a training dataset and a test dataset.

Question

Instructions: ~1 minute to think/answer on your own; then discuss with neighbors; then I will call on one of you

Probability review

1. Given a fair coin, what is probability that you will get 10 heads? (Write a math expression.)

2. Given 1000 fair coins, what is probability that some coin will get 10 heads? (Write a math expression.)

Coin analogy

- Question: given a fair coin, what is probability that you will get 10 heads?
- Answer: ≈ 0.1%
- Question: given 1000 fair coins, what is probability that <u>some</u> coin will get 10 heads?
- Answer: ≈ 63%

Key point: if we try 1000 mediocre hypotheses, one could look good simply by chance.

This is very similar to the multiple hypothesis testing issue that arises in statistics ("p hacking")

Credit: Yaser Abu Mostafa https://work.caltech.edu/

Solution

- Use a *union* bound (from probability lecture)
- Let *g* be the hypothesis we choose from $\mathcal{H}$
- Let *M* be number of hypotheses in $\mathcal{H}$
- End result (details on board):

$$\begin{aligned} P(|E_{in}(g) - E_{out}(g)| > \epsilon) &\leq \sum_{j=1}^{M} P(|E_{in}(h_j) - E_{out}(h_j)| > \epsilon) \\ &\leq \sum_{j=1}^{M} 2\exp(-2\epsilon^2 n) \\ &= 2M\exp(-2\epsilon^2 n) \end{aligned}$$

Another perspective

- (Shown on board) With probability at least 1 - δ, we have

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{1}{2n} \ln\left(\frac{2M}{\delta}\right)}$$

- In words, the **"true" error** of *g* will be **close** to the **error on training data**.
- We can get closer…
  - with larger *n* (more training data)
  - with smaller *M* (smaller hypothesis set)
- Note: δ is generally "fixed" to something small, say 1/1000

Tradeoffs

- We want to find a hypothesis that looks good to us (E_{in} is low)
    - This is more likely if hypothesis set is large
- We want to ensure that hypothesis we find will be good on future inputs (E_{out} is close to E_{in})
    - This is less likely if hypothesis set is small

Question

Instructions: ~1 minute to think/answer on your own; then discuss with neighbors; then I will call on one of you

Suppose we use a perceptron learning algorithm to find a hypothesis that performs well on a spam training dataset. The training dataset was hand-crafted from the personal email of an ML researcher in 2003. The error on the training data is 0.05%.

Can we apply this bound to confidently assert that error on future emails will also be low?

If not, why not?

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{1}{2n} \ln\left(\frac{2M}{\delta}\right)}$$