Lecture 2

Presentation · March 2019

CITATIONS
0

READS
155

1 author:

Alaa Tharwat
Fachhochschule Bielefeld
120 PUBLICATIONS **6,195** CITATIONS

SEE PROFILE

Lecture 2: Feasibility of learning

Alaa Tharwat

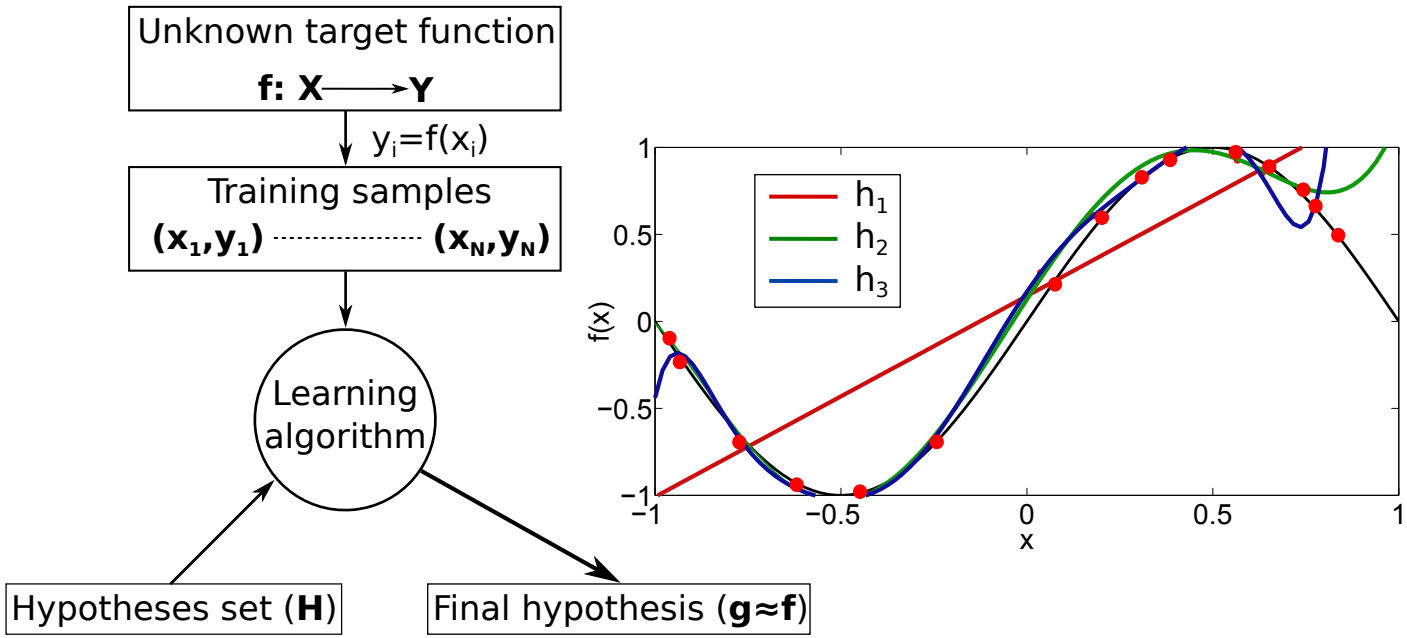
Lecture 2: Feasibility of learning

- Review of Lecture 1
- Hoeffding's inequality
- Connection to the learning problem
- Multiple hypotheses

Lecture 2: Feasibility of learning

- Review of Lecture 1
- Hoeffding's inequality
- Connection to the learning problem
- Multiple hypotheses

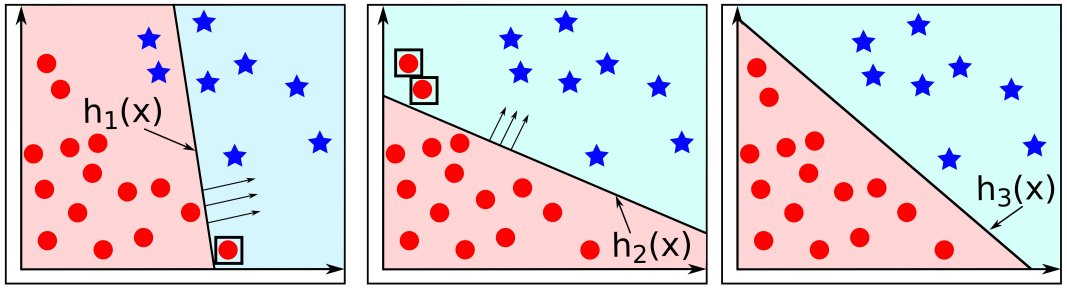
- Definition of machine learning
  - Machine learning is a subset/part of the artificial intelligence science which uses statistical techniques to give computers the ability to *learn* to do a specific task without being explicitly programmed.
- Examples of machine learning
  - Credit approval and fish types
- Components of learning
  - **Input**: $(\mathbf{x} \in \mathcal{R}^d)$, **Output**: $(y \in \{-1, +1\})$, **Target function**: $f : X \to Y$, **Data**: $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N)\}$, where $y_i = f(\mathbf{x}_i)$ and **Hypothesis**: $g : X \to Y$

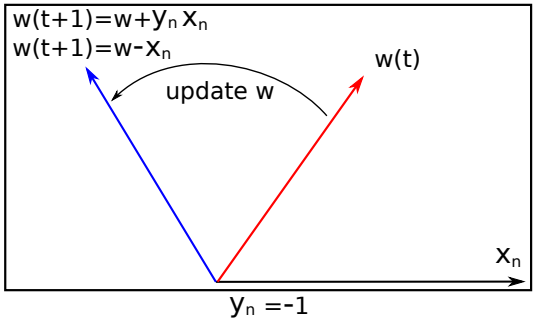


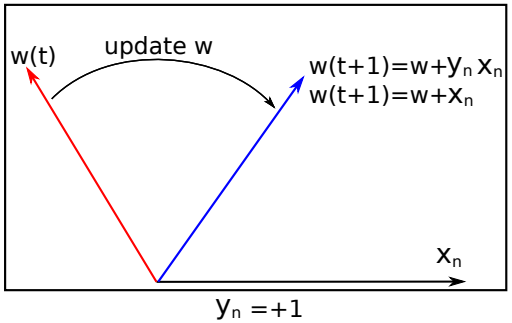
- Simple model (changing w generates many hypotheses)

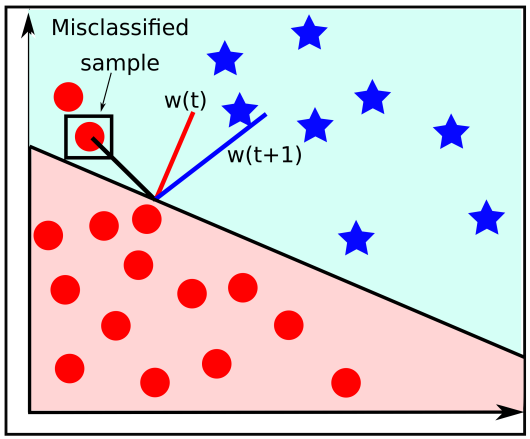
$$h(\mathbf{x}) = \text{sign}(\mathbf{w}^T \mathbf{x}_i)$$

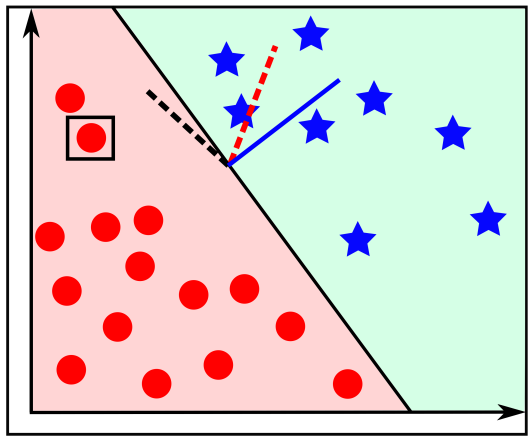
- We want to select $g \in H$ so that $g \approx f$ on the data D
- The weights are updated as, $\mathbf{w}(t+1) \leftarrow \mathbf{w}(t) + \mathbf{x}_n y_n$
- Types of learning
  - Supervised learning
  - Unsupervised learning
  - Reinforcement learning







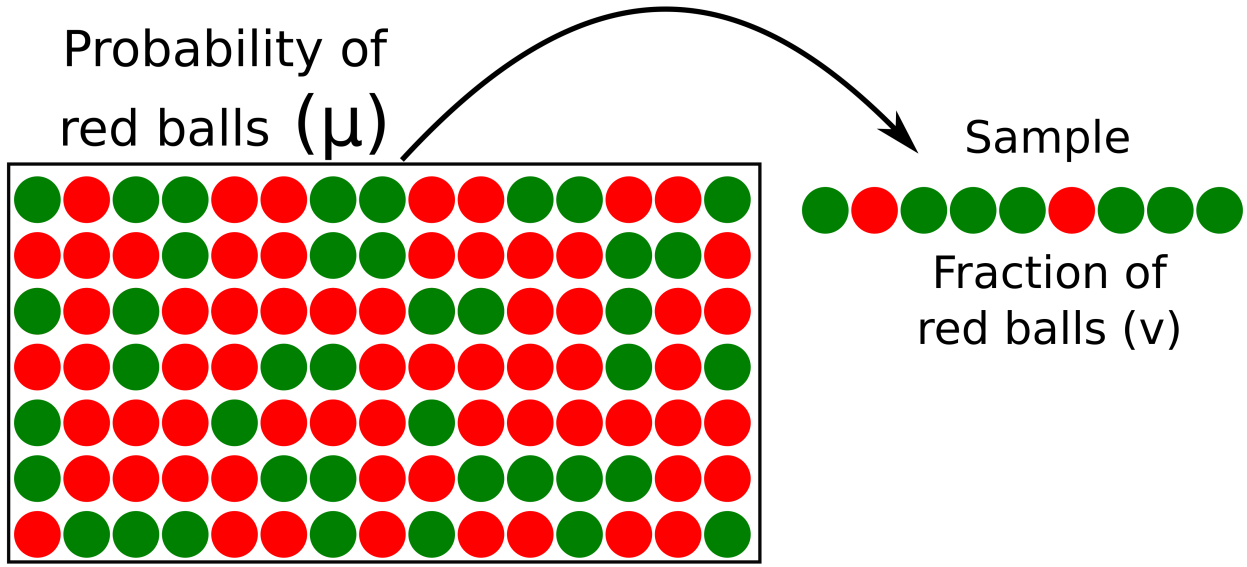




Lecture 2: Feasibility of learning

- Review of Lecture 1
- Hoeffding's inequality
- Connection to the learning problem
- Multiple hypotheses

- Given a box of balls (red and green):
  - Probability of picking red ball is $P[\text{picking red ball}] = \mu$
  - Probability of picking green ball is $P[\text{picking green ball}] = 1 - \mu$
- The value of μ is unknown, e.g. probability of males in a country
- We cannot predict μ; but, we can select/pick N balls/samples from the box independently, the fraction of red ball is v, v **and** μ **are similar or not**?
- The sample is mostly green, while the box is mostly red (biased sample)
- Sample frequency (v) is likely to be close to μ



- In a big sample ($N \gg$), v is probably close to μ, BUT, within constraints as follows:

$$P[|v - \mu| > \epsilon] \leq 2e^{-2\epsilon^2 N} \tag{1}$$

where ϵ is a small number. This is called *Hoeffding's inequality*, and it means that

- The probability of v far from μ (bad event) by ϵ must be within a limit $2e^{-2\epsilon^2 N}$, for all N and ϵ
- This bound/constraint (right hand side) does not depend on μ or the number of samples in the box (good news, why?)
  - Because μ and number of samples in the box are **unknown**

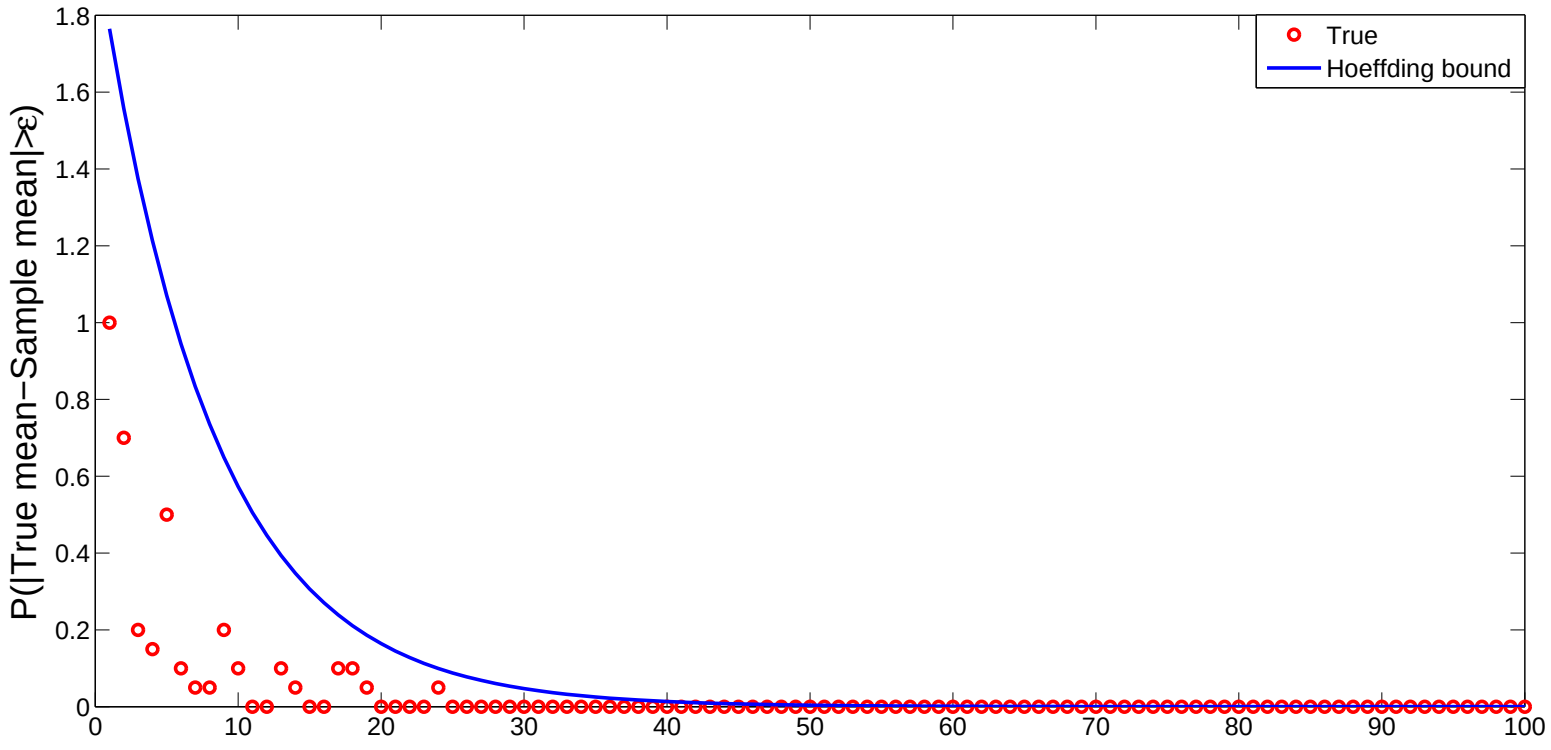
$$P[|v-\mu| > \epsilon] \leq 2e^{-2\epsilon^2 N}$$
$$= P[|v-\mu| \leq \epsilon] \geq 1 - 2e^{-2\epsilon^2 N} \quad (2)$$

- $N \uparrow \infty \Rightarrow 2e^{-2\epsilon^2 N} \approx 0$; hence, $P[|v-\mu| \leq \epsilon] \geq 1 - 2e^{-2\epsilon^2 N} \Rightarrow P[|v-\mu| \leq \epsilon] \geq 1$, ($v \approx \mu$ with high probability, *but not sure*).
- On the other hand, $N \downarrow \Rightarrow e^{-2\epsilon^2 N} \approx 1 \Rightarrow$ $P[|v-\mu| \leq \epsilon] \geq 1 - 2e^{-2\epsilon^2 N} \Rightarrow P[|v-\mu| \leq \epsilon] \geq 0$; hence, v far from μ
- Small ϵ means ($v \approx \mu$). On the other hand, large ϵ means v far from μ

$$P[|v - \mu| > \epsilon] \leq 2e^{-2\epsilon^2 N} \tag{3}$$

- How far v from μ:
  - Given $N = 1000$ and $\epsilon = 0.05$, this means that $P[|v - \mu| > 0.05] \leq 2e^{-2\times 0.05^2 \times 1000} \Rightarrow \mu - 0.05 \leq v \leq \mu + 0.05$ with probability $\approx 99\%$
  - Given $N = 1000$ and $\epsilon = 0.1$, this means that $P[|v - \mu| > 0.1] \leq 2e^{-2\times 0.1^2 \times 1000} \Rightarrow \mu - 0.1 \leq v \leq \mu + 0.1$ with probability $\approx 99.9999996\%$
  - Given $N = 100$ and $\epsilon = 0.05$, this means that $P[|v - \mu| > 0.1] \leq 2e^{-2\times 0.05^2 \times 100} \Rightarrow \mu - 0.05 \leq v \leq \mu + 0.05$ with probability $\approx 39.4\%$
  - Given $N = 50$ and $\epsilon = 0.05$, this means that $P[|v - \mu| > 0.1] \leq 2e^{-2\times 0.05^2 \times 50} \Rightarrow \mu - 0.05 \leq v \leq \mu + 0.05$ with probability $\approx 22.1\%$
- Tradeoff N, ϵ (large sample size N or looser gab ϵ)

- The figure below shows a binomial distribution ($\mathcal{B}$)
- In a coin flip example, the probability of getting head or tail is $P = 0.5$, $E[X] = 0.5$, and let $\epsilon = 0.25$
- With a small data size, the difference between the expected/empirical mean and the true mean is big
- Some distributions converge to their true means when the data size increased

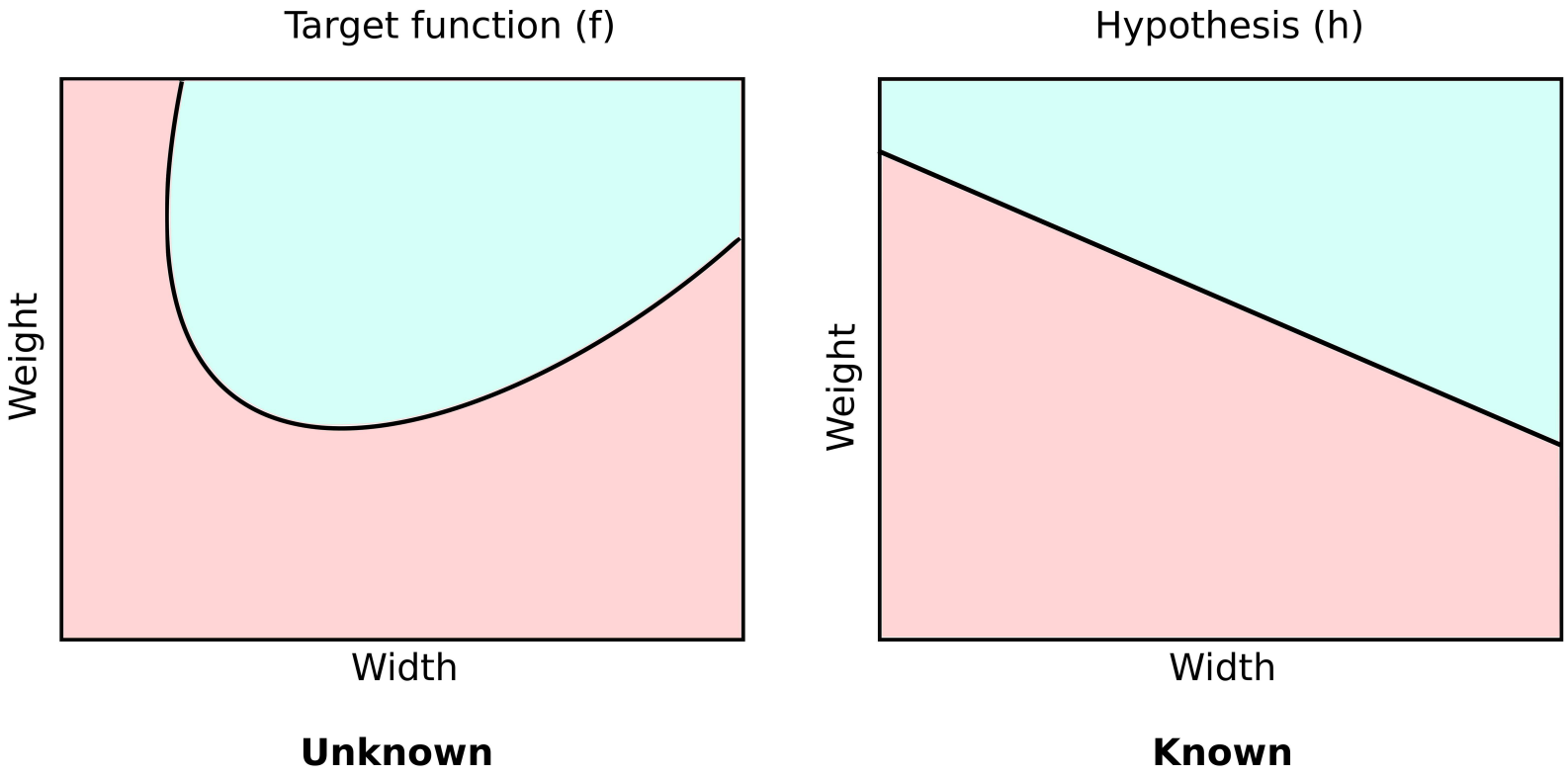


Lecture 2: Feasibility of learning

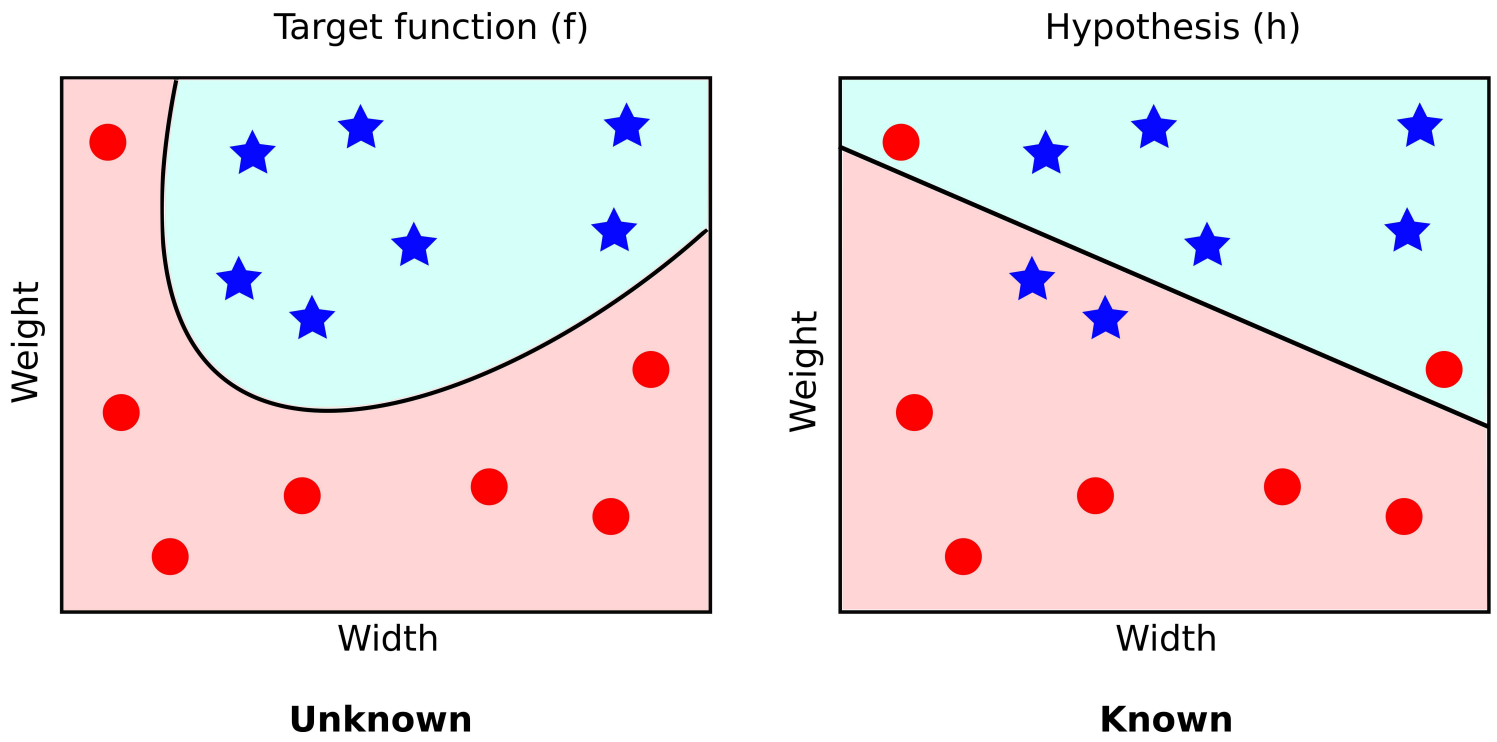
- Review of Lecture 1
- Hoeffding's inequality
- Connection to the learning problem
- Multiple hypotheses

In the learning framework

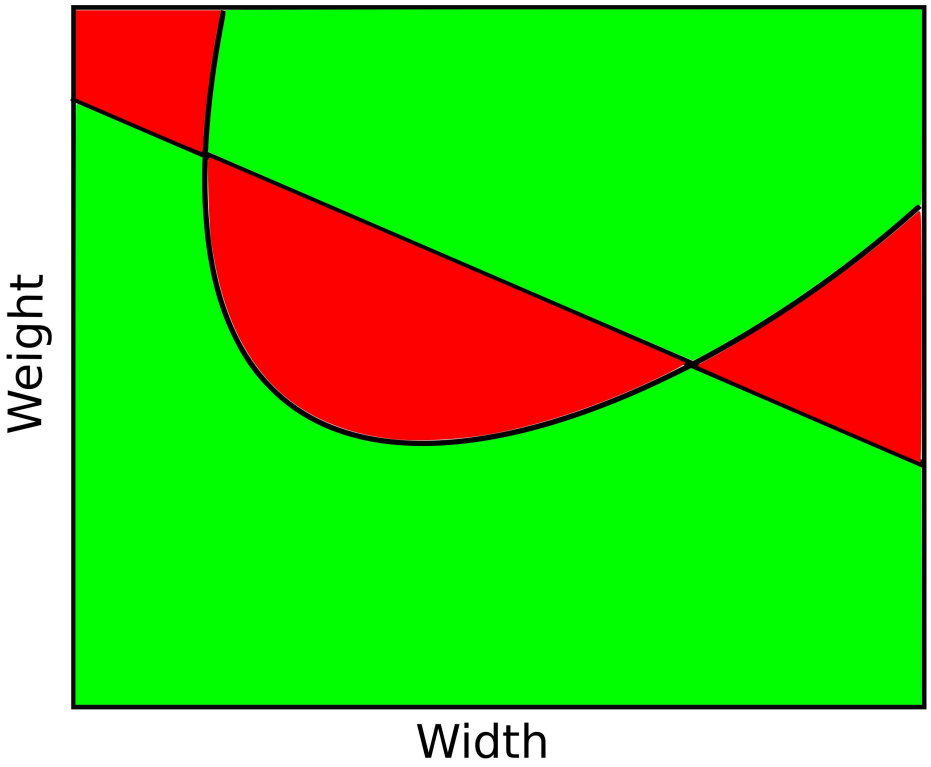
- The target function (f) is unknown as μ. In other words, we can say f is not only the given data D, but, f also outside the data
- Each ball is $\mathbf{x} \in X$, and the sample v is the dataset (D)

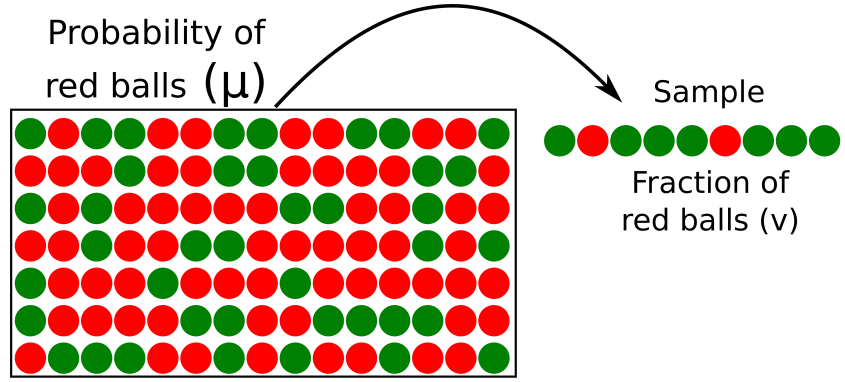


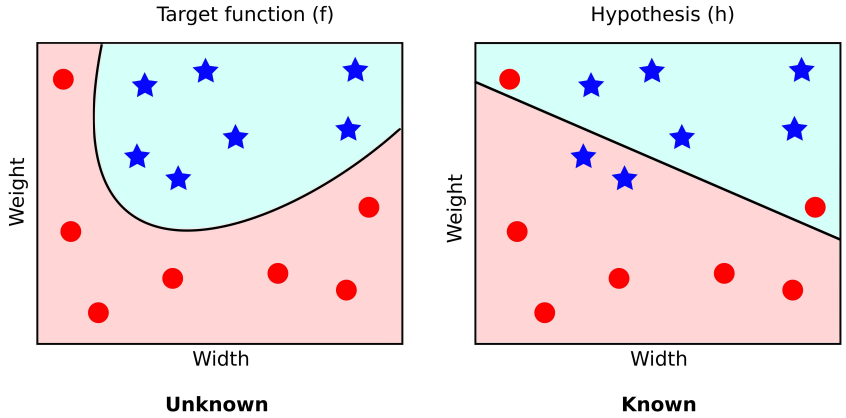
- Some samples are generated/drawn from f, these samples represent D
- The learning model uses the data D for generating a set of hypotheses H, and $h \in H$



- Green ball means the hypothesis is rights: $h(\mathbf{x}) = f(\mathbf{x})$
- Red ball means the hypothesis is wrong: $h(\mathbf{x}) \neq f(\mathbf{x})$
- The error is the size of red region (Error function), $E(h) = P_x[h(\mathbf{x}) \neq f(\mathbf{x})]$

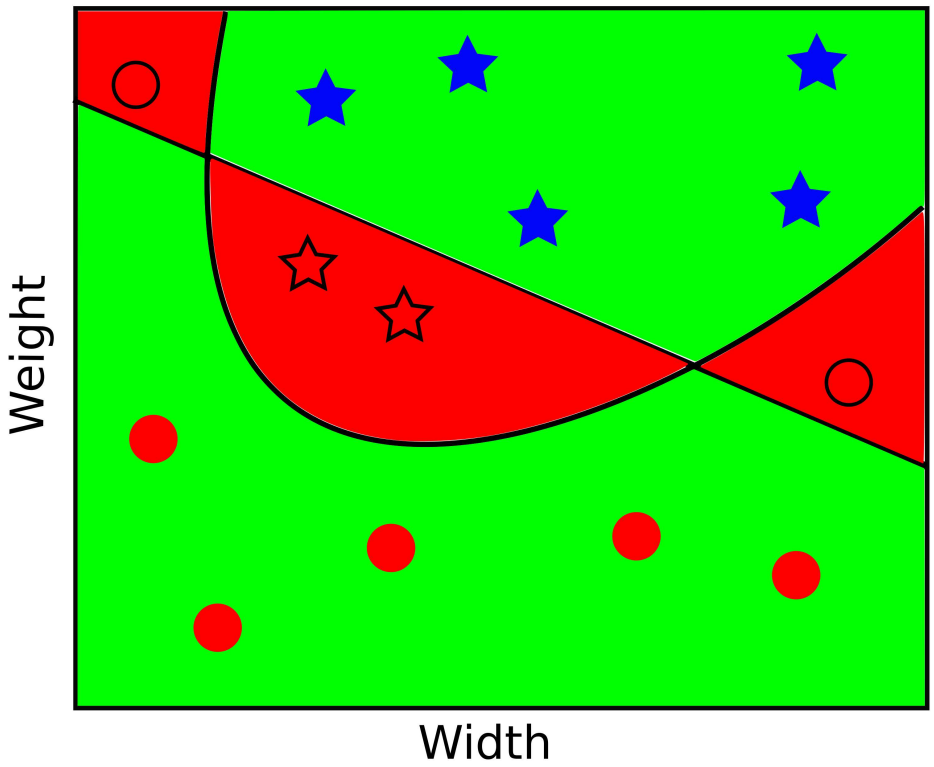




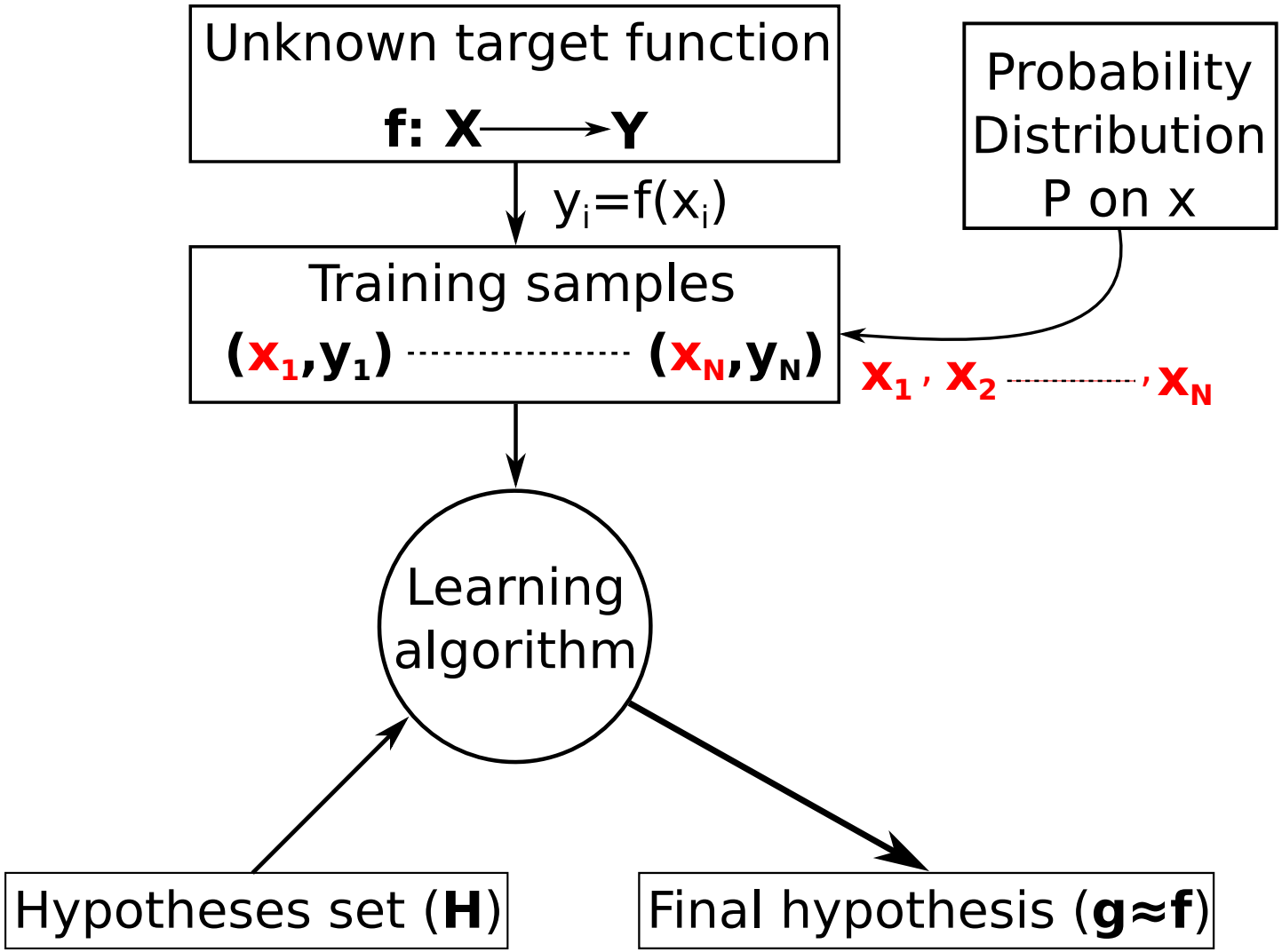


There are two types of errors:

- In-sample Error ($E_{in}(h)$): how many samples in the *dataset* are misclassified, $E_{in}(h) = \frac{1}{m}\sum_{i=1}^{m}[h(\mathbf{x}_i) \neq f(\mathbf{x}_i)]$, m is the number of samples in the dataset (here is $E_{in}(h) = \frac{4}{14}$)
- Out-of-sample Error ($E_{out}(h)$): red regions, $E_{out}(h) = P_x[h(\mathbf{x}) \neq f(\mathbf{x})]$ (we cannot calculate it)



Learning Model	Box Model
Input space X	Box of balls
$\mathbf{x}$, where $h(x) = f(\mathbf{x})$ (h is right)	Green ball
$\mathbf{x}$, where $h(\mathbf{x}) \neq f(\mathbf{x})$ (h is wrong)	Red ball
$P(x)$	Randomly pick balls
dataset $D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ with i.i.d. $\mathbf{x}_n$	Sample of N balls of i.i.d. balls
Out-of-sample (E_{out})	$\mu \rightarrow$ probability of picking red ball in the box
In-sample (E_{in})	$v \rightarrow$ probability of picking red ball in a sample

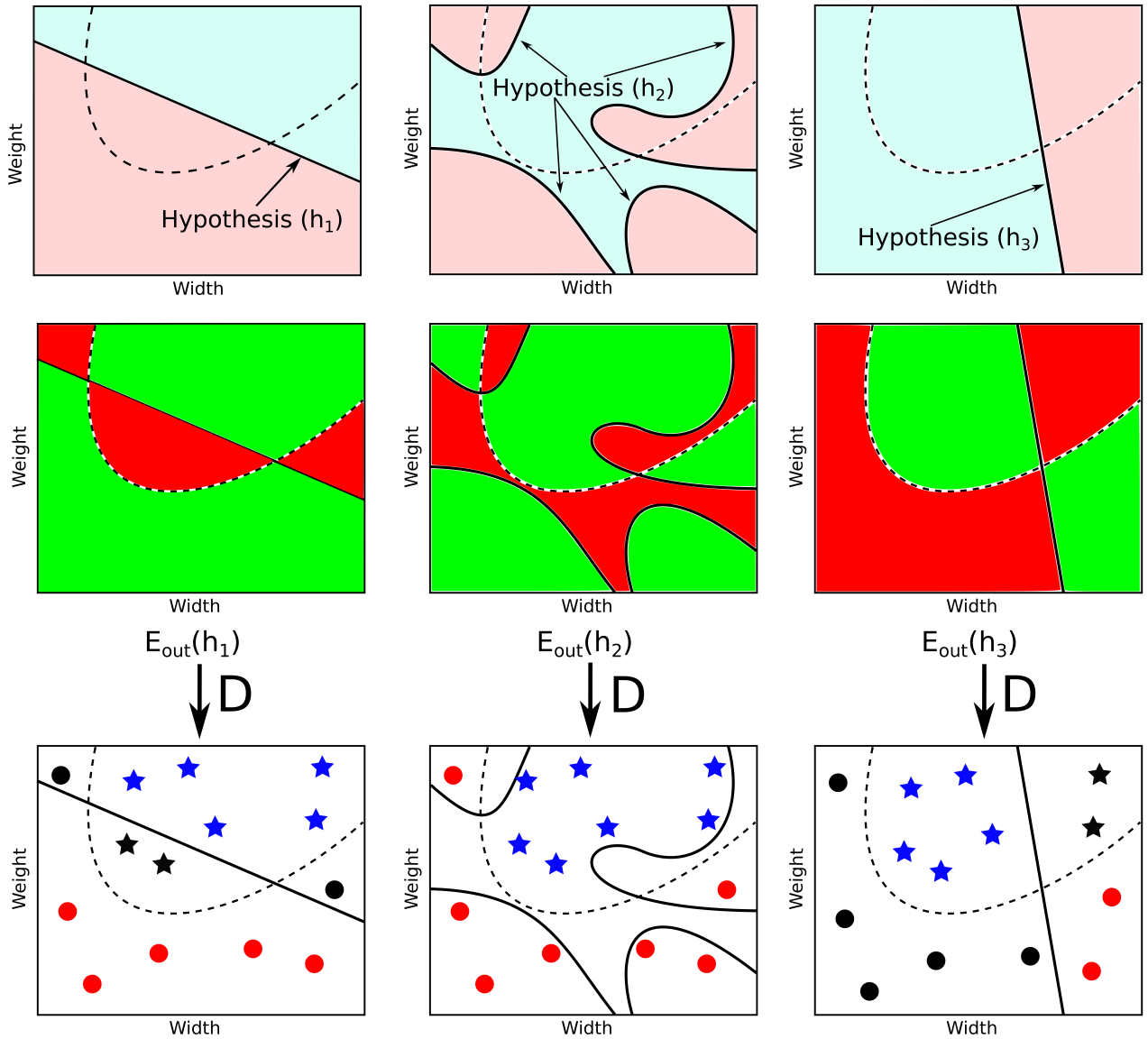

Unknown target function
f: X ⟶ Y
$y_i=f(x_i)$
Training samples
(x_1,y_1) ---------- (x_N,y_N)
Probability Distribution P on x
x_1, x_2 ---------- $, x_N$
Learning algorithm
Hypotheses set (H)
Final hypothesis (g≈f)

Back to learning:

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

- We can calculate E_{in}, and it is random; but, we can **not** calculate E_{out}, and it is fixed
- If $E_{in} \approx 0 \Rightarrow E_{out} \approx 0$ (with a high probability). This means $P_x[h(\mathbf{x}) \neq f(\mathbf{x})] \approx 0$; hence, We have learned something about the entire f, and $f \approx h$ over X (outside D)
- If E_{in} (out of luck). This means we have still learned something about the entire; but, $f \neq h$

- Given $h \in H$, the samples (i.e. dataset) can verify whether h is good (i.e. E_{in} is small) or bad (i.e. E_{in} is large).



Another example:

- $X = \{0,1\}^3$, $Y = \{O, X\}$

x_n	$y_n = h(x_n)$
0 0 0	O
0 0 1	X
0 1 0	X
0 1 1	O
1 0 0	X

- The shaded region is D
- $g \approx f$ inside D (perfectly), but, not outside D

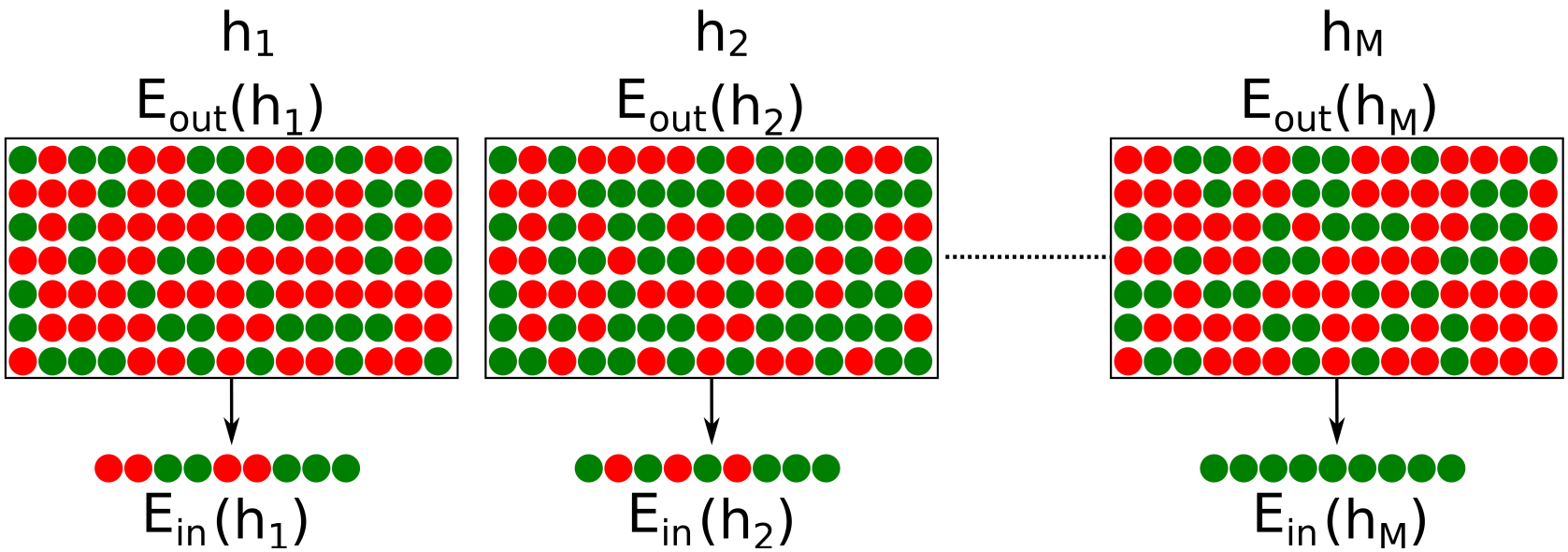
x_n	y_n	g	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
0 0 0	○	○	○	○	○	○	○	○	○	○
0 0 1	X	X	X	X	X	X	X	X	X	X
0 1 0	X	X	X	X	X	X	X	X	X	X
0 1 1	○	○	○	○	○	○	○	○	○	○
1 0 0	X	X	X	X	X	X	X	X	X	X
1 0 1		?	○	○	○	○	X	X	X	X
1 1 0		?	○	○	X	X	○	○	X	X
1 1 1		?	○	X	○	X	○	X	○	X

Lecture 2: Feasibility of learning

- Review of Lecture 1
- Hoeffding's inequality
- Connection to the learning problem
- Multiple hypotheses

- v and μ depends on the hypothesis h
- The learning model generates multiple hypotheses H (not only one)
- Some balls are green in one box and red in the other box, why?
  - Answer: in a box i, green ball means that this hypothesis (h_i) correctly classify this ball/sample, and a red ball indicates that this ball/sample is misclassified
- The learning model scans all hypotheses to find $g \approx f \Rightarrow$ minimum $E_{in}(h_i)$

Hoeffding does not apply to multiple hypotheses!!!



- Given a fair coin, what is the probability of getting 10 heads if toss the coin ten times?
  - Answer: $\approx 0.1\%$[1]
  - The coin is fair, and the ten heads are no indication about the real probability
- Given 1000 fair coins, what is the probability of getting 10 heads if toss the coins ten times?
  - Answer: $\approx 63\%$[2], why?
    - Answer: simply, you tried many times (Hoeffding is applied in each time)

[1] $P = \frac{1}{2^N}$

[2] $1 - (1 - \frac{1}{2^N})^{1000}$

- Bad sample: E_{in} is far from E_{out} (remember Hoeffding). For example $E_{in} = 0$ and $E_{out} = 0.5$, getting 10 heads, or picking green balls from a box which has red and green balls
- Bad data for one h means E_{out} is big (h is far from the target function f) and E_{in} is small ($h(\mathbf{x}_n) = f(\mathbf{x}_n)$ on most samples)

Given M hypotheses (many h's)

$$P[\text{Bad } D] = P[\text{ Bad } D \text{ for } h_1 \text{ or Bad } D \text{ for } h_2 \text{ or } \ldots \text{ Bad } D \text{ for } h_M]$$
$$\leq P[\text{Bad } D \text{ for } h_1] + P[\text{Bad } D \text{ for } h_2 + \ldots P[\text{Bad } D \text{ for } h_M]$$

Some Basic Probability

- $P(A \text{ or} B) = P(A \cup B) \leq P(A) + P(B)$
- If $(A \subseteq B)$ then $P(A) \leq P(B)$

$g \in H$ and hence $g = h_1$, or $h_2, \ldots,$ or h_M

$$\begin{aligned} P[|E_{in}(g) - E_{out}(g)| > \epsilon] \leq P[&|E_{in}(h_1) - E_{out}(h_1)| > \epsilon \text{ or} \\ &|E_{in}(h_2) - E_{out}(h_2)| > \epsilon \text{ or} \\ &\vdots \\ &|E_{in}(h_M) - E_{out}(h_M)| > \epsilon] \end{aligned} \tag{5}$$

The worst case is to make union hence,

$$\begin{aligned} P[|E_{in}(g) - E_{out}(g)| > \epsilon] &\leq \sum_{i=1}^{M} P[|E_{in}(h_i) - E_{out}(h_i)|] \\ &\leq \sum_{i=1}^{M} 2e^{-2\epsilon^2 N} \\ &\leq 2Me^{-2\epsilon^2 N} \end{aligned} \tag{6}$$