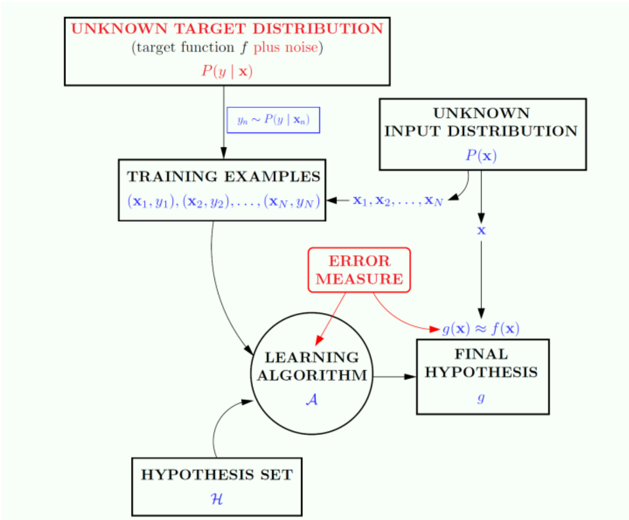
# Until Now




$$\mathcal{H} = \{h | h_{\boldsymbol{w}}(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x}\}$$

$$\mathcal{A} = \{\text{Perceptron, SGD, p} - \text{inv}\}$$

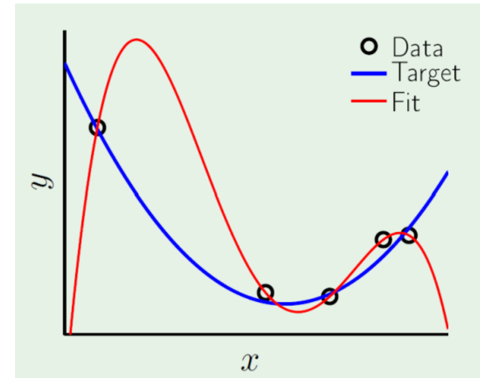
$$\begin{aligned} Loss = \{&(y - h_w(x))^2, \\ &[\![ y \neq sign(h_w(x)) ]\!], \\ &\ln(1 + e^{-y h_w(x)}), \end{aligned}$$

Solutions: $\{(\mathcal{H}, \mathrm{A} \in \mathcal{A}, \mathrm{L} \in Loss)\}$

OVERFITTING

ERM

$$\min_{\boldsymbol{w}} E_{in}(\boldsymbol{w})$$




$$\mathbb{E}_{\mathcal{D}}[E_{out}(g^{(\mathcal{D})})] = \mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\boldsymbol{x}) - \tilde{g}(\boldsymbol{x})\right)^2\right] + (\tilde{g}(\boldsymbol{x}) - f(\boldsymbol{x}))^2$$

Variance bias

**REGULARIZATION -** A thinning cure $\mathcal{H}$

$$\min_{\boldsymbol{w}} E_{in}(\boldsymbol{w}) + \lambda\,\Omega(\boldsymbol{w}), \qquad \lambda > 0$$

variance ⬇ & bias ⬆

$\Omega(\boldsymbol{w}) = \boldsymbol{w}^T\boldsymbol{w}$ Weight Decay

$\Omega(\boldsymbol{w}) = \boldsymbol{w}^T \Gamma^T \Gamma \boldsymbol{w}$ General case

# Until Now

**Validation is a new cure for overfitting from data**

## $E_{out}$ estimate

$\mathcal{D} = \mathcal{D}_{train} + \mathcal{D}_{val}$ (K-size)

$g^-$ is learned from $\mathcal{D}_{train}$

$g$ is learned from $\mathcal{D}$

$$E_{out}(g) \leq E_{out}(g^-) \leq E_{val}(g^-) + \mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$$

**High variance pessimistic estimate**

**Cross-Validation:**

1.- Average estimate
2.- Heuristic approach
3.- Heavy computing

## Model Selection

M-models trained independently can be assesed using their error on $\mathcal{D}_{val}$

$E_{val}(g^-_{m^*})$ is an optimistic (bias) estimator

$g^*_m$ is learned from $\mathcal{D}$

$$E_{out}(g^*_m) \leq E_{out}(g^-_{m^*}) \leq E_{val}(g^-_{m^*}) + \mathcal{O}\left(\sqrt{\frac{\ln M}{K}}\right)$$

Here $E_{val}$ plays the role of $E_{in}$ for the models

# What left?

- We already know how to fit a good model from ERM

$$E_{in} \to 0$$

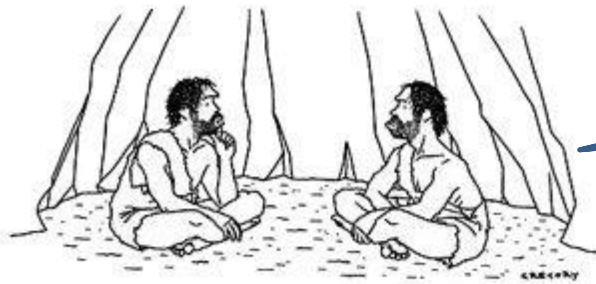
- BUT what  guarantees that by doing ERM provides us information about the population?

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

  – Some inequalities have emerged,

$$E_{out}(g) \leq E_{in}(g) + \ \mathcal{O}(\frac{d_{\mathcal{H}}}{\sqrt{N}}) \quad \text{(classification)}$$

but so far without justification.



"Something's just not right—our air is clean, our water is pure, we all get plenty of exercise, everything we eat is organic and free-range, and yet nobody lives past thirty."

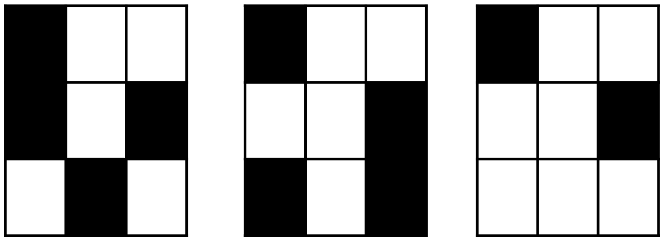


# What is learning?:

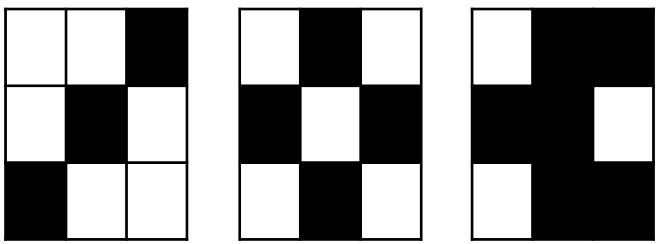
## The Empirical Risk Minimization(ERM) rule

# Is Learning Feasible?

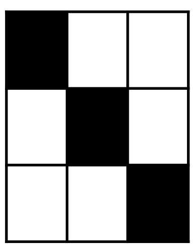
- Let us consider the following two examples:



$f$ = -1



$f$ = 1



$f$ = ?

$f: \{0,1\}^3 \rightarrow \{0,1\}$

We know $f$ only partially in its domain

| | |
|---|---|
| **0 0 0** | **0** |
| **0 0 1** | **1** |
| **0 1 0** | **1** |
| **0 1 1** | **0** |
| **1 0 0** | **1** |
| **1 0 1** | **?** |
| **1 1 0** | **?** |
| **1 1 1** | **?** |

How is $f$ in the last 3 elements?

# It is Not...

| | $g$ | $f_1$ | $f_2$ | $f_3$ | $f_4$ | $f_5$ | $f_6$ | $f_7$ | $f_8$ |
|---|---|---|---|---|---|---|---|---|---|
| 0 0 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 0 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 0 1 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 0 1 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 0 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 1 0 1 | ? | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |
| 1 1 0 | ? | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 |
| 1 1 1 | ? | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 1 |

- We can’t learn this function !
- Try to verify that  the eight solutions are equivalent, this is,  all provide the same error.
  - Fix one of them as true solution and count how many of the others provide one, two o three errors on the unknown values.

# What then ?

- **Inductive Learning** is a hopeless approach:

In a strict sense learning out of the sample is not possible!!

( see Inductivist Turkey (Bertrand Russell ) ☺ )

Is there any hope to know anything about $f$ outside the data set **without making assumptions** about $f$ ?

Yes, if we are willing to give up "for sure".

Try to learn something less exigent than the proper **unknown** function,

i.e. some useful property about the **unknown** function

# Let's try to exploit randomness….

- NEW Hypothesis: items inside $\mathcal{D}$ are i.i.d samples from a probability distribution $\mathcal{P}$

- **Consequences:**
  - $\mathcal{D}$ **is the output of a random variable (vector)**
  - **It is not realistic to expect that every sample** $\mathcal{D}$ represent equally well the distribution $\mathcal{P}$
  - **The function *g* depends on** $\mathcal{D}$, hence its election is also a random process .

- **Where is the novelty ?**
  - **Probability theory shows that there are probabilistic dependencies between a random variable and a sample of it (under conditions).**
    - **Example : Confidence interval for the sample mean** $\boldsymbol{P(|\bar{x}-\mu|<\epsilon)>1-\delta,\qquad \delta(\epsilon)\ll 1}$

# But probability it is not enough!

- MAIN QUESTION:

  There exists a learning algorithm $\mathcal{A}$ and a sample size $m$ such that for every distribution $\mathcal{P}$, if $\mathcal{A}$ receives $m$ i.i.d. samples from $\mathcal{P}$, there is a high chance it output a predictor $g$ with low error?

- No-Free-Lunch (NFL) Theorem (Informal): "For every algorithm there exist a $\mathcal{P}$ on which it fails, even though that $\mathcal{P}$ can be sucessfully learned by another learner. Moreover, all algorithms are equivalent in average on all possible target functions $f$"

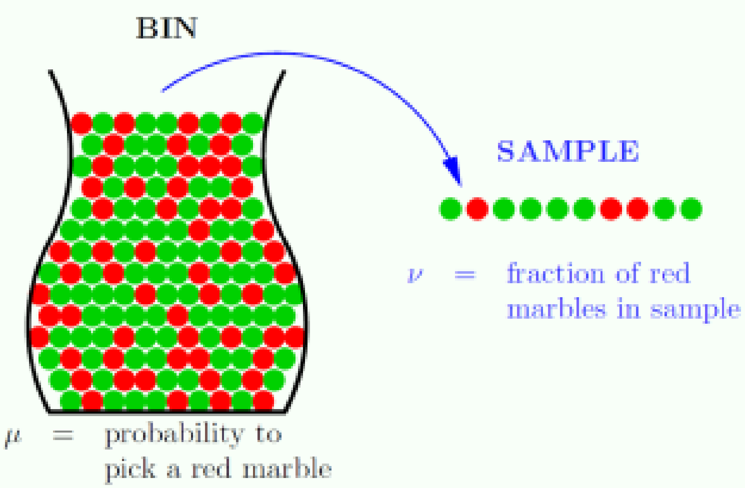
- In order to succeed each learner ( $\mathcal{A}, \mathcal{H}$ ) must be applied on the class of distributions $\mathcal{P}$ that it can learn.

- This highlights the need for **exploiting problem-specific knowledge** to achieve better than random performance
  - Geometric constraint
  - Class of function with zero or very small $E_{out}$
  - Finite class $\mathcal{H}$
  - Finite VC dimension
  - etc

# Can we infer something outside the data using only $\mathcal{D}$ ?:

# The PAC answer

# Population Mean from Sample Mean




**The BIN Model**

- Bin with red and green marbles.
- Pick a sample of $N$ marbles *independently*.
- $\mu$: probability to pick a red marble.
  $\nu$: fraction of red marbles in the sample.

$$\begin{array}{rcccl} \text{Sample} & \longrightarrow & \text{the data set} & \longrightarrow & \nu \\ \text{BIN} & \longrightarrow & \text{outside the data} & \longrightarrow & \mu \end{array}$$

Can we guarantee anything about ***μ* (outside the data)** after observing ***ν* (the data)**?
ANSWER: No. It is possible for the sample to be all green marbles and the bin to be mostly red.

Then, why do we trust polling (e.g. to predict the outcome of a presidential election).
ANSWER: The bad case is possible, but **not probable**.

# Hoeffding's Inequality

Hoeffding/Chernoff proved that, most of the time, **for a fixed $\mu$**, $\nu$ cannot be too far from $\mu$

$$\mathbb{P}(\mathcal{D}\colon |\mu - \nu| > \epsilon) \le 2e^{-2\epsilon^2 N} \qquad \text{for any } \epsilon > 0$$

$$\mathbb{P}(\mathcal{D}\colon |\mu - \nu| \le \epsilon) \ge 1 - 2e^{-2\epsilon^2 N} \qquad \text{for any } \epsilon > 0$$

Question: What does the value of $\nu$ tell us on $\mu$: $\mu \approx \nu \Leftrightarrow \nu \approx \mu$

**Example**: $N = 1,000$; draw a sample and observe $\nu$.

| | | |
|---|---|---|
| $99\%$ of the time | $\mu - 0.05 \le \nu \le \mu + 0.05$ | $(\epsilon = 0.05)$ |
| $99.9999996\%$ of the time | $\mu - 0.10 \le \nu \le \mu + 0.10$ | $(\epsilon = 0.10)$ |

**What does this mean?** If I repeatedly pick a sample of size 1,000, observe $\nu$ and claim that

$$\mu \in [\nu - 0.05, \nu + 0.05], \qquad \text{(the error bar is } \pm 0.05)$$

I will be right 99% of the time. On any particular sample you **may** be wrong, but not often.

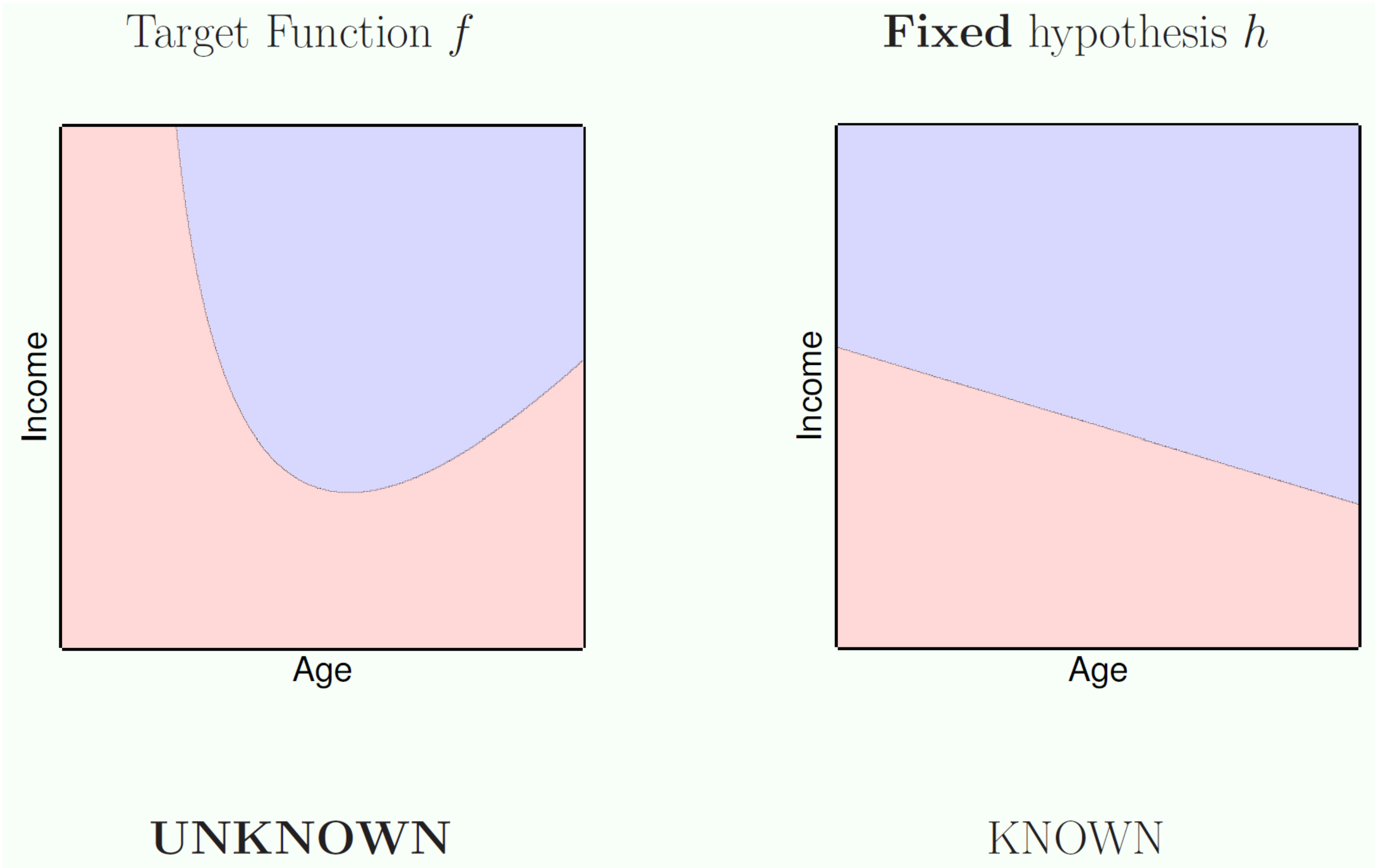
We learned *something*. From $\nu$, we reached outside the data to $\mu$.

# Hoeffding's Inequality: Remarkable facts

- The key ingredient: samples must be i.i.d.
  - If the sample is constructed in some arbitrary fashion, then indeed we cannot say anything.
  - Even with independence, ν can take on arbitrary values; but some values are more likely than others.
  - This is what allows us to learn something – it is likely that ν ≈ μ.
- The bound $2e^{-2\epsilon^2 N}$ does not depend on μ or the size of the bin
  - The bin can be infinite.
  - It's great that it does not depend on μ because μ is unknown; and we mean **unknown**.
- The key player in the bound $2e^{-2\epsilon^2 N}$ is N.
  - If N → ∞, μ ≈ ν with very very very . . . high probabilty, but not for sure.
  - Can you live with $10^{-100}$ probability of error?

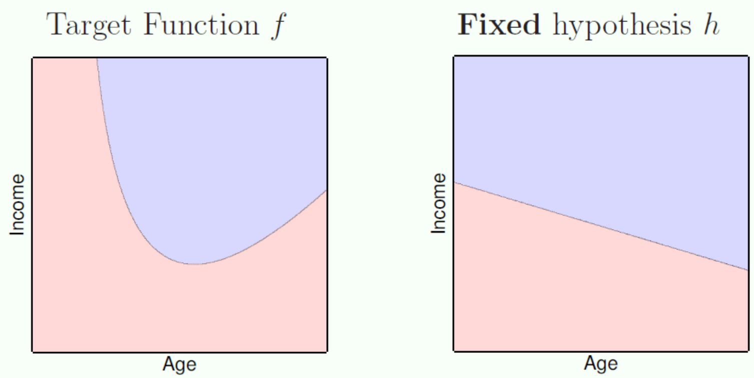
$$\boxed{\mathbb{P}(\mathcal{D}:\ |\mu - \nu| > \epsilon) \leq 2e^{-2\epsilon^2 N}}$$

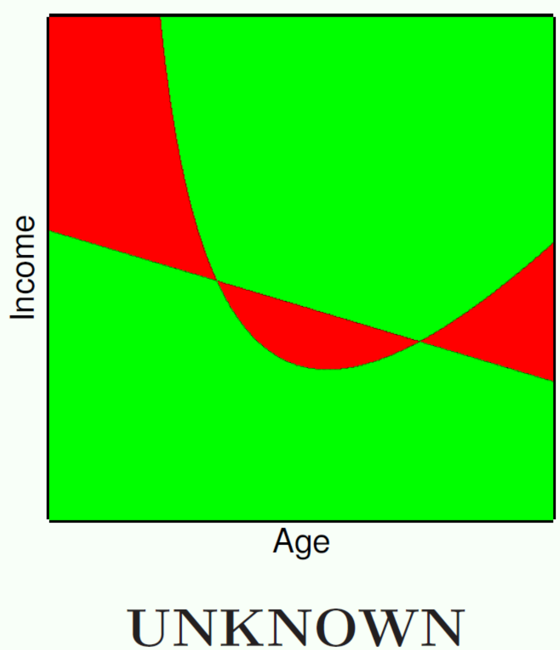
# Learning setup




In learning, the unknown is an entire function *f*; in the bin it was a single number μ.

# The Learning Error Function




green: $h(\mathbf{x}) = f(\mathbf{x})$

red: $h(\mathbf{x}) \neq f(\mathbf{x})$

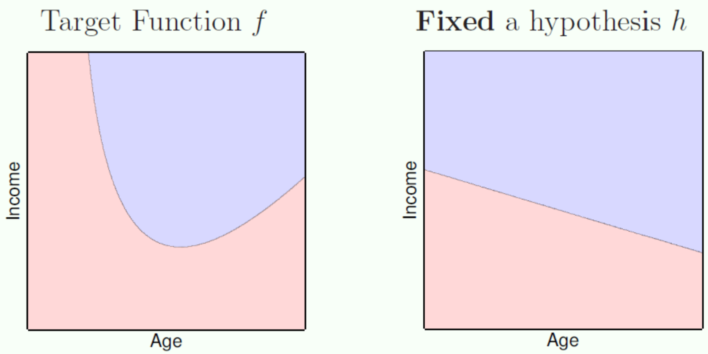
$$E(h) = \mathbb{P}_{\mathbf{x}}[h(\mathbf{x}) \neq f(\mathbf{x})]$$

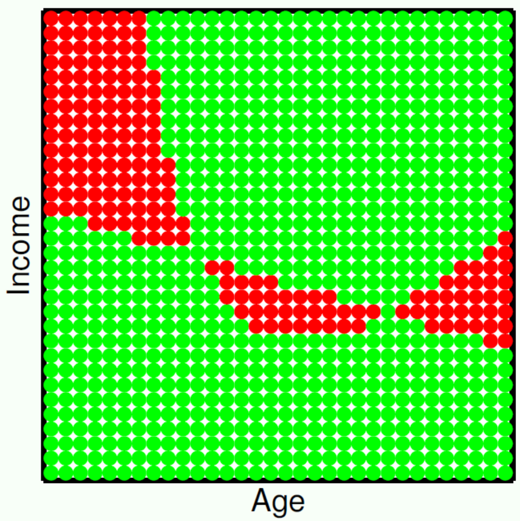
("size" of the red region)

$P(\mathbf{x})$

The function h defines an unknown but fixed error probability E (h)

# Relating the Bin to Learning




green "marble": $h(\mathbf{x}) = f(\mathbf{x})$

red "marble": $h(\mathbf{x}) \neq f(\mathbf{x})$

BIN: $\mathcal{X}$

$$E_{\text{out}}(h) = \mathbb{P}_{\mathbf{x}}[h(\mathbf{x}) \neq f(\mathbf{x})]$$

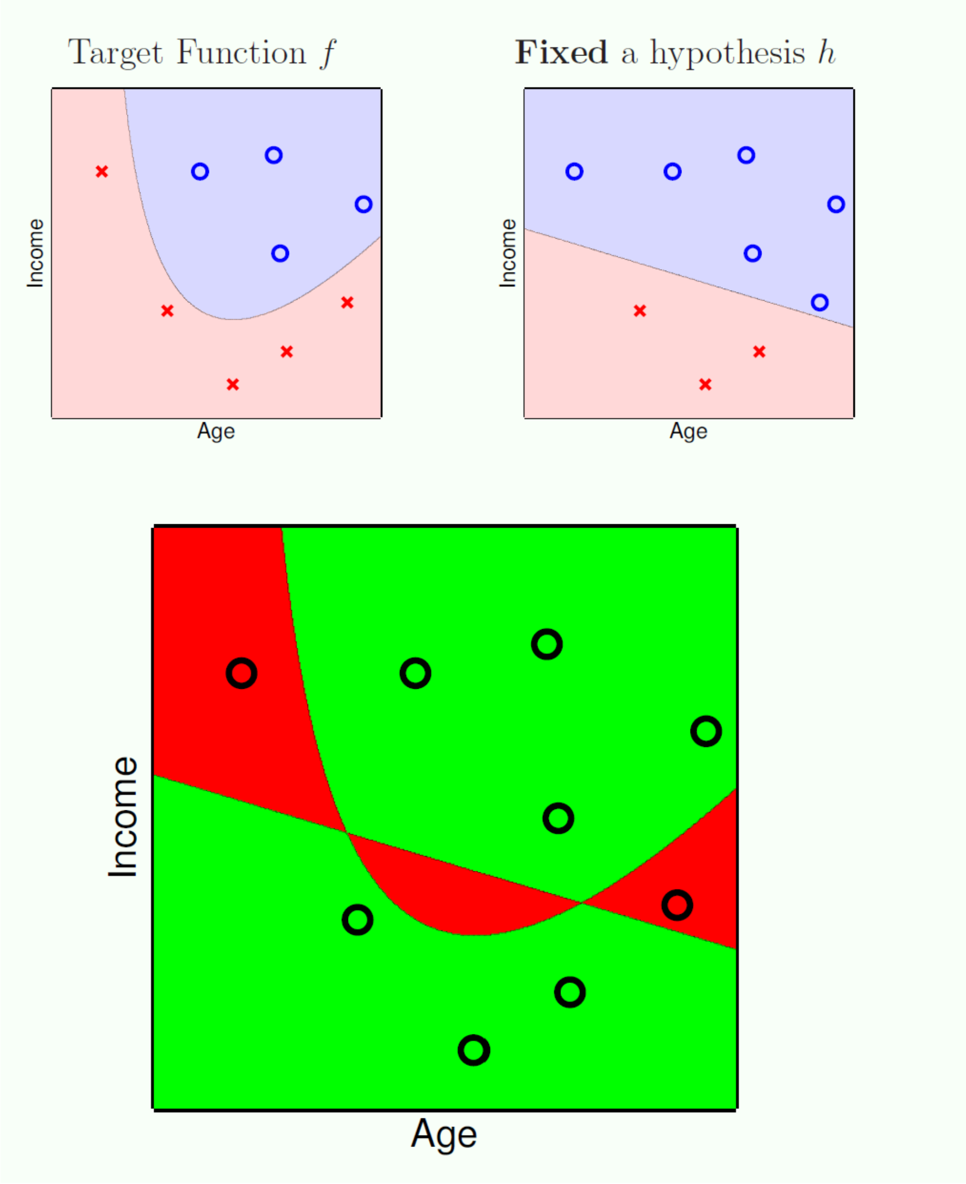
out-of-sample

**UNKNOWN**

Let´s consider all possible sample points

Now a Bin Model is defined by *h* and *f*

# Relating the Bin to Learning - the Data



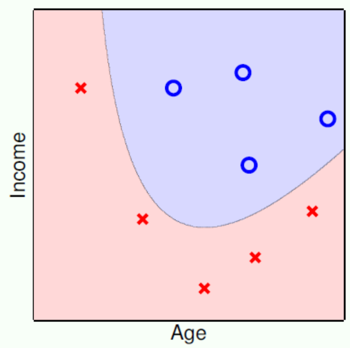

On the same sample, the target function $f$ and the hypothesis $h$ provides us with different labels

We have points in different zones of the error function.

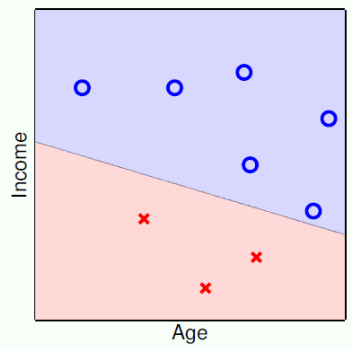
If the sample is draw independently acccording to $P$, each point will be red with probability $\mu$ and green with probability $1- \mu$

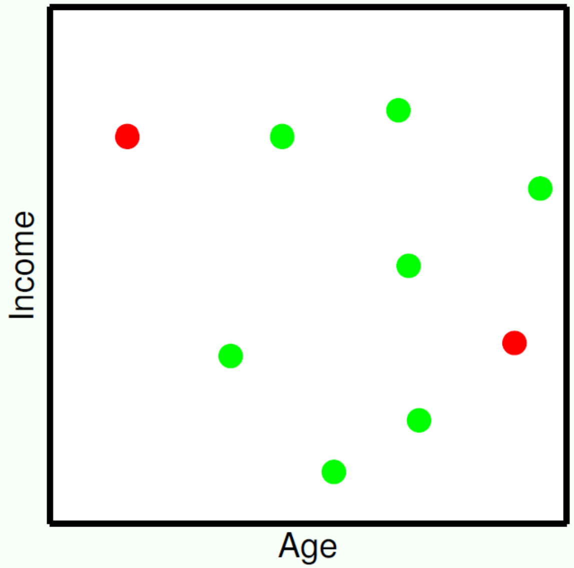
# Relating the Bin to Learning - the Data

Target Function $f$




**Fixed** a hypothesis $h$




**KNOWN!**

green data: $h(\mathbf{x}_n) = f(\mathbf{x}_n)$

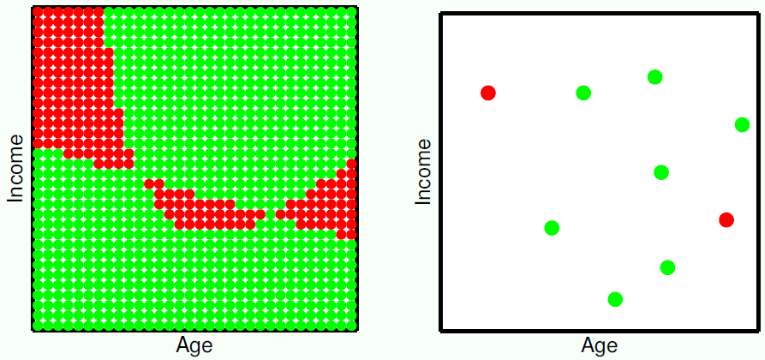
red data: $h(\mathbf{x}_n) \neq f(\mathbf{x}_n)$

$E_{\text{in}}(h) =$ fraction of red data

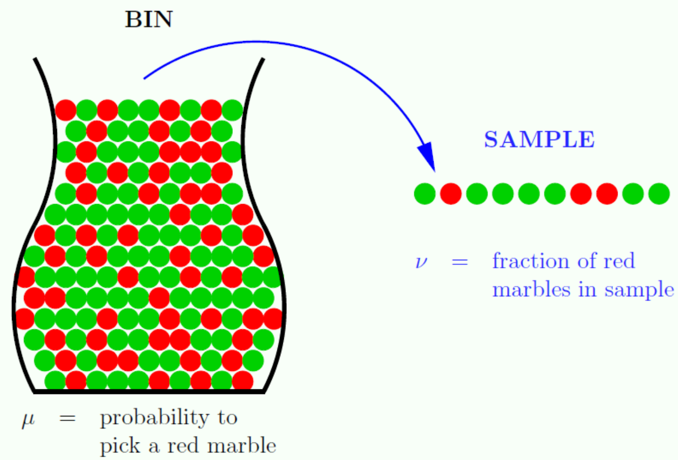
in-sample

misclassified

# Bin Model and Learning




Unknown $f$ and $P(\mathbf{x})$, **fixed** $h$




## Learning

input space $\mathcal{X}$

$\mathbf{x}$ for which $h(\mathbf{x}) = f(\mathbf{x})$

$\mathbf{x}$ for which $h(\mathbf{x}) \neq f(\mathbf{x})$

$P(\mathbf{x})$

data set $\mathcal{D}$

**Out-of-sample Error:** $E_{\text{out}}(h) = \mathbb{P}_{\mathbf{x}}[h(\mathbf{x}) \neq f(\mathbf{x})]$

**In-sample Error:** $E_{\text{in}}(h) = \frac{1}{N}\sum_{n=1}^{N} [\![h(\mathbf{x}) \neq f(\mathbf{x})]\!]$

## Bin Model

Bin

green marble

red marble

randomly picking a marble

sample of $N$ marbles

$\mu$ = probability of picking a red marble

$\nu$ = fraction of red marbles in the sample

# Hoeffding inequality in Learning

- Let's consider $\mathcal{H}$={h}, only one function, and $f$(x) the unknown true function.
- Let's $[\![f(x) = h(x)]\!]$ and $[\![f(x) \neq h(x)]\!]$ represent new binary variables in the population. Now $\mu = \Pr([\![f(x) \neq h(x)]\!])$
- For any training sample $\mathcal{D}$ of size N , $\nu = \text{Fraction}([\![f(x) \neq h(x)]\!])$ on $\mathcal{D}$
- Now $\mu$ and $\nu$ represent the population and sample error respectively.
- Let's denote by $\mathrm{E}_{out}(h) = \mu$ *and* $\mathrm{E}_{in}(h) = \nu$ the h's global and sample error respectively
- The Hoeffding inequality can be rewritten as:

$$P(\mathcal{D}: |\mathrm{E}_{out}(h) - \mathrm{E}_{in}(h)| > \epsilon) \leq 2e^{-2\epsilon^2 N} \text{ for any } \epsilon > 0$$

This is called a "Probably Aproximately Correct (PAC)" result

- IMPORTANT: Note that **h** is fixed before knowing the data sample

# Hoeffding says that $E_{\text{in}}(h) \approx E_{\text{out}}(h)$

$$\mathbb{P}(\mathcal{D}\colon |\mu - \nu| > \epsilon) \leq 2e^{-2\epsilon^2 N} \iff \mathbb{P}(\mathcal{D}\colon |E_{\text{out}}(h) - E_{\text{in}}(h)| > \epsilon) \leq 2e^{-2\epsilon^2 N}$$

$E_{\text{in}}$ **is random, but known**; $E_{\text{out}}$ **fixed, but unknown**.

- If $E_{\text{in}}(h) \approx 0 \overset{\text{N >> 1}}{\Longrightarrow} E_{out}(h) \approx 0$ (with high probability), i.e. $\mathbb{P}_{\mathcal{X}}[h(\mathbf{x}) \neq f(\mathbf{x})] = 0$
  - We have learned something about the *entire* $f$: $f \approx h \; over \; \mathcal{X}$ (outside $\mathcal{D}$)
- If $E_{\text{in}} \gg 0$, we're out of luck.
  - But, we have still learned something about the *entire* $f$: $f \not\approx h \; over \; \mathcal{X}$ ; it is not very useful though.

Questions:

1. Suppose that $E_{\text{in}}$ = 1, have we learned something about the entire $f$ that is useful?
2. What is the worst $E_{\text{in}}$ for inferring about $f$ ?

# Understanding PAC results

$$P(\mathcal{D}: |\mathrm{E}_{out}(h) - \mathrm{E}_{in}(h)| > \epsilon) \leq 2e^{-2\epsilon^2 N} \text{ for any } \epsilon > 0$$

- Let's consider $\delta = 2e^{-2\epsilon^2 N}$ then

$$P(\mathcal{D}: |\mathrm{E}_{out}(h) - \mathrm{E}_{in}(h)| > \epsilon) \leq \delta \Leftrightarrow P(\mathcal{D}: |\mathrm{E}_{out}(h) - \mathrm{E}_{in}(h)| \leq \epsilon) \geq 1 - \delta$$

- Or equivalently:

$$\mathrm{E}_{out}(h) \leq \mathrm{E}_{in}(h) + \epsilon \text{ , } \textit{with probability at least } 1 - \delta \text{ on } \mathcal{D}$$

- Let's write $\epsilon$ as a function of $N$ and $\delta$, then

$$\mathrm{E}_{out}(h) \leq \mathrm{E}_{in}(h) + \sqrt{\frac{1}{2N} \log \frac{2}{\delta}} \quad \textit{with probability at least } 1 - \delta \textit{ on } \mathcal{D}$$

- The higher N the narrow the interval ( The sample size is important !!)
- The smaller $\delta$ the larger the interval ( The higher guarantee the lesser accuracy)

# The Hoeffding inequality for multiple hypothesis

THINGS ARE DIFFERENT: In Hoeffding's inequality the h is fixed before knowing the data, BUT in REAL PROBLEMS the chosen hypothesis, *g*, from $\mathcal{H}$ is identified using the data.

SEARCH CAUSES SELECTION BIAS: THE COIN ANALOGY

**Question:** if toss a **fair coin** ten times, what is the probability that you will get ten heads ?

**Answer:** ≈ **0.1** (try it)

**Question:** if toss 1000 **fair coins** ten times, what is the probability that some coin will get ten heads ?

**Answer:** ≈ **0.63** (try it)

**Identifying coins with functions: the higher the size of $\mathcal{H}$ the higher the probability of having a hypothesis with $E_{in} \approx 0$ error , BUT can we expect $E_{out}$ to be small ?**

# Then what?

- Adapting the Hoeffding's inequality to the case of finite $\mathcal{H}$
  1. The hypothesis solution $\boldsymbol{g}$ **should be fixed before knowing the data sample**. (MANDATORY CONDITION)
  2. Nevertheless, the Learning Algorithm uses the training data to search for $\boldsymbol{g}$ .

- A simple **solution** is to consider an event **valid for all functions** in $\mathcal{H}$.
  - Let $g$ denote a generic hypothesis solution then,

$$\{\mathcal{D}: |E_{in}(g) - E_{out}(g)| > \epsilon\} = \bigcup_{h_i \in H} (\mathcal{D}: |E_{in}(h_i) - E_{out}(h_i)| > \epsilon)$$

  - Using $P\left(\bigcup_{i=1:|\mathcal{H}|} B_i\right) \leq \sum_{i=1}^{|\mathcal{H}|} P(B_i)$

$$P(\mathcal{D}: |\mathrm{E}_{in}(g) - \mathrm{E}_{out}(g)| > \epsilon) < 2|\mathcal{H}|e^{-2\epsilon^2 N} \text{ for any } \epsilon > 0$$

# PAC Learning in finite classes

$$P(\mathcal{D}: |\mathrm{E}_{in}(h) - \mathrm{E}_{out}(h)| > \epsilon) \leq 2|\mathcal{H}|e^{-2\epsilon^2 N} \quad \text{for any } \epsilon > 0$$

- Let's denote $\delta = 2|\mathcal{H}|e^{-2\epsilon^2 N}$ , writing $\epsilon$ as a function of $N$, $\delta$ and $|\mathcal{H}|$

$$\mathrm{E}_{out}(h) \leq \mathrm{E}_{in}(h) + \sqrt{\frac{1}{2N}\log\frac{2|\mathcal{H}|}{\delta}} \;\; \textit{with probability al least } 1-\delta \textit{ on } \mathcal{S}$$

- Once again,
  - The higher N the narrow the interval ( The sample size is important !!)
  - The smaller $\delta$ the larger the interval ( The higher guarantee the lesser accuracy)

- BUT now the size of $\mathcal{H}$ matters too

- For a given $\delta$, the complexity of the sample necessary to learn with a fixed accuracy grows lineal with the log of the size of the set $\mathcal{H}$

- This inequality is meaningless in infinite classes !!!
- **Is this a serious drawback** ??

# Feasibility of Learning vs Complexity

- Learning is only possible in a probabilistic setting (under conditions):
  - Samples from $\mathcal{X}$ **must be i.i.d**
  - Same probability distribution in training and test

- To be succesful in learning means to find a function $g$, s.t. $\mathrm{E}_{out}(g) \approx 0$
- Nevertheless, we are only able to guarantee,

$$P(\mathcal{D}\colon |\mathrm{E}_{in}(g) - \mathrm{E}_{out}(g)| > \epsilon) < 2|\mathcal{H}|e^{-\epsilon^2 N} \text{ for any } \epsilon > 0$$

- Feasibility of Learning must answer two questions:
  1. Can we make sure that $\mathrm{E}_{out}(g)$ is close enough to $\mathrm{E}_{in}(g)$ ?
  2. Can we make $\mathrm{E}_{in}(g)$ small enough?

- What is the relationship between Feasibility of Learning and the complexity of $\mathcal{H}$ and $f$?

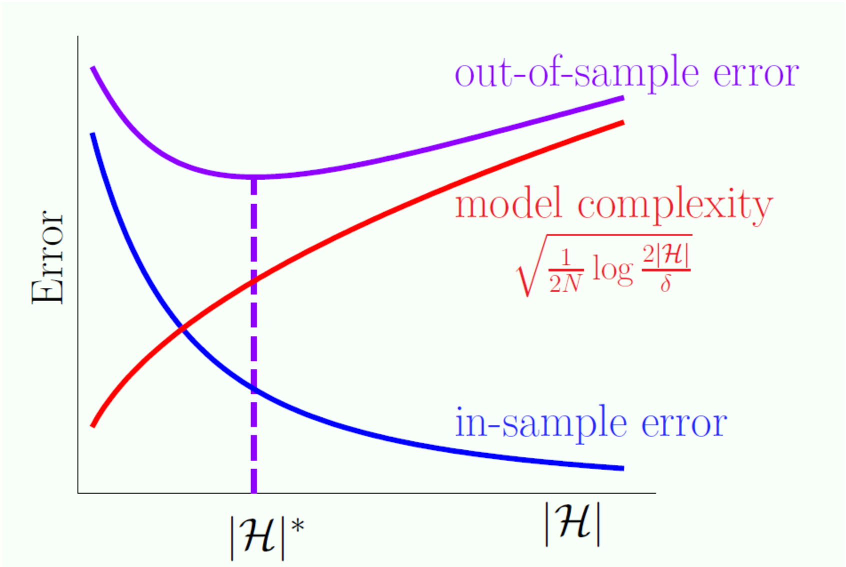
# Feasibility of learning : $E_{out} \approx 0$

Two conditions:

(1) $E_{in} \approx E_{out}$ $\longrightarrow$ Is verified thank to the Hoeffding's inequality

(2) $E_{in} \approx 0$ $\longrightarrow$ Is achieved through the learning algorithm

Together, these ensure $E_{out} \approx 0$




**BUT there is a tradeoff on $\mathcal{H}$:**

- **Small $|\mathcal{H}|$** $\Rightarrow E_{in} \approx E_{out}$
- **Large $|\mathcal{H}|$** $\Rightarrow E_{in} \approx 0$ is more likely

**What about the complexity of $f$ :**

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

# Feasibility of Learning (finite $\mathcal{H}$): Summary

- Out of $\mathcal{D}$, nothing about *f* can be guaranteed

- If $\mathcal{D}$ is an independent sample from $\mathbb{P}(\mathbf{x})$.

  $E_{out} \approx E_{in}$ ($E_{in}$ can reach outside the data set to $E_{out}$).

- But, what we want is $E_{out} \approx 0$.
- The two step solution. We trade $E_{out} \approx 0$ for 2 goals:

  (i) $E_{out} \approx E_{in}$

  (ii) $E_{in} \approx 0$.

  We know $E_{in}$, not $E_{out}$, but we can *ensure* (i) if $|\mathcal{H}|$ is small.

**Any ERM rule is a succesful PAC learner for finite classes $\mathcal{H}$**