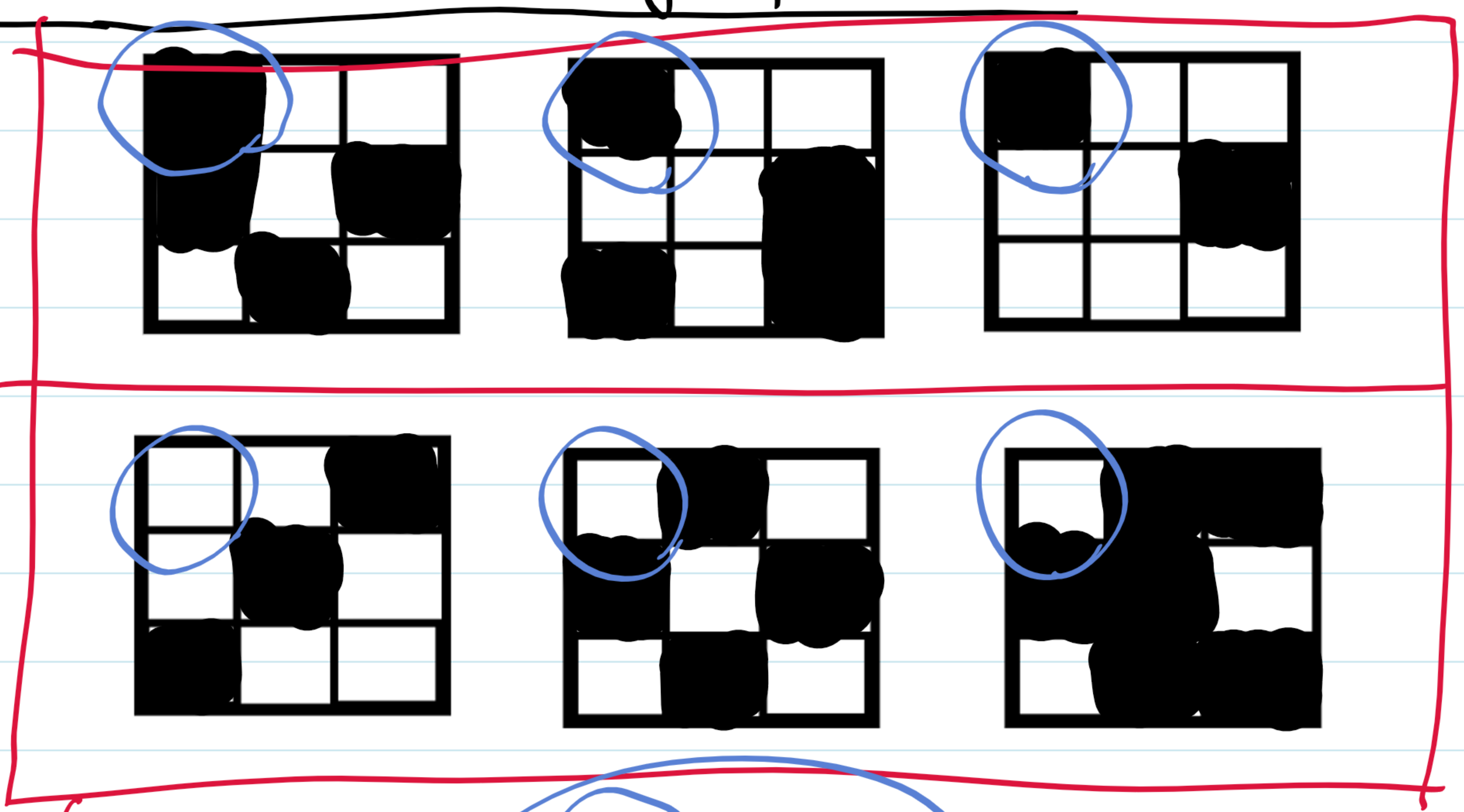
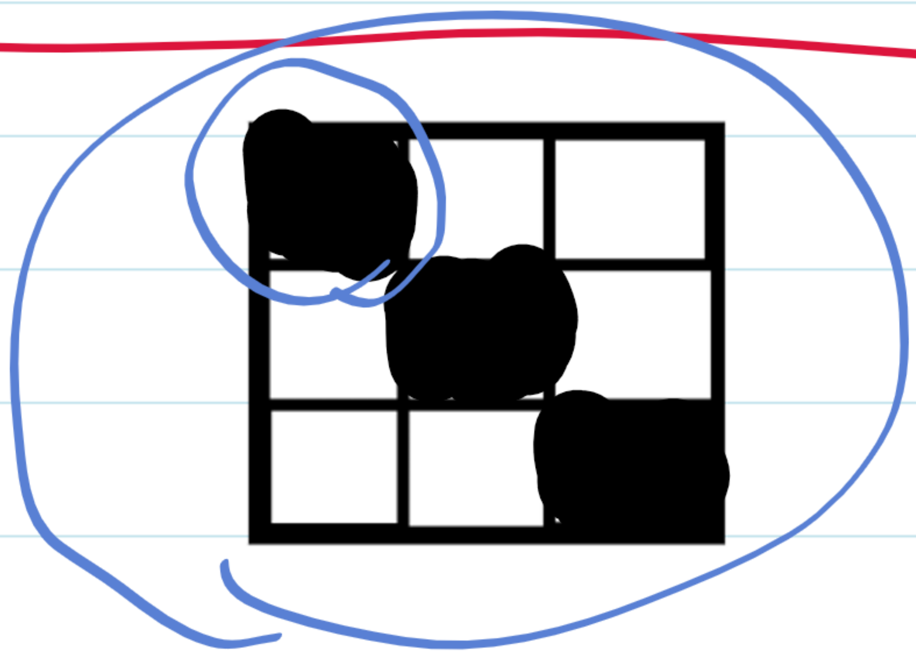
A visual learning problem :

1) $f(x) = -1$

2) $f(x) = +1$



D_{out}



$f = ??$

9-bit vector (3x3 black & white array)

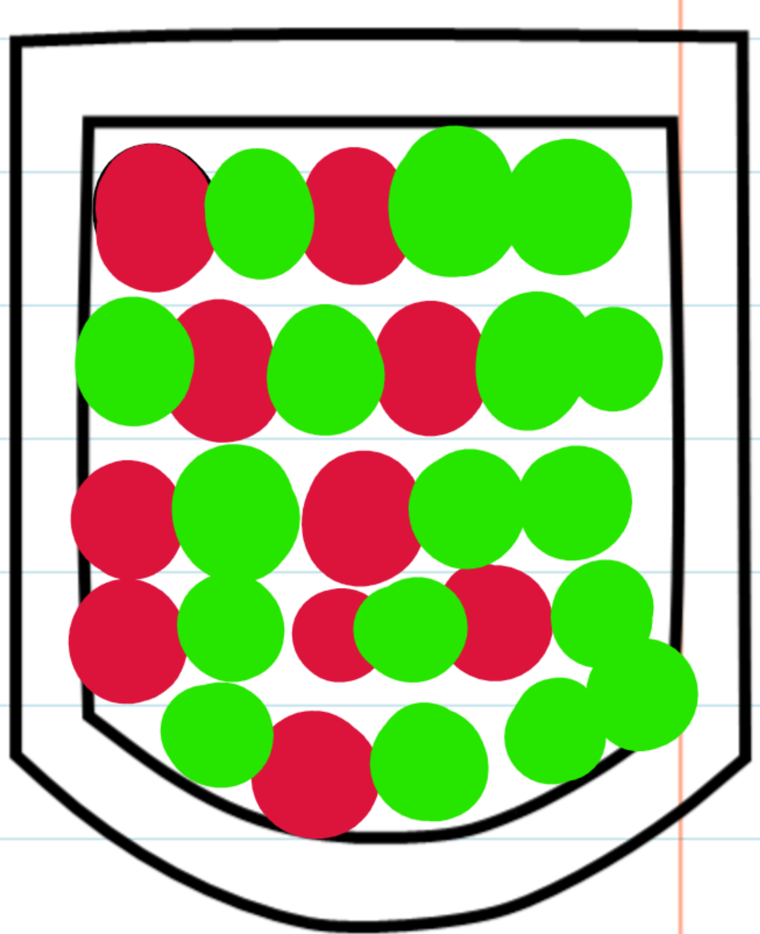
Possibility-1 : $f = +1$, due to symmetry

Possibility-2 : $f = -1$, top left corner

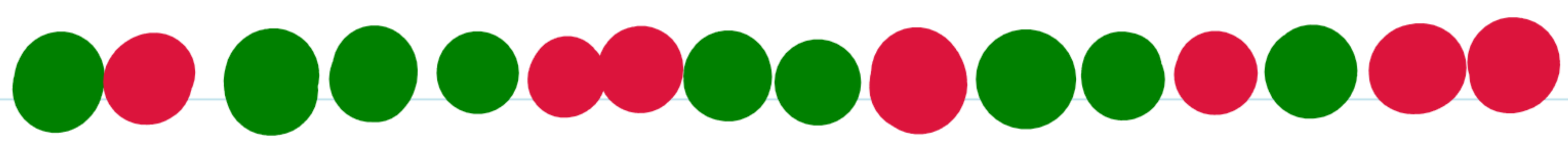
Consider a bin with red and green marbles.

$\mathbb{P}$[Picking a red marble] $= \mu$.

$\mathbb{P}$[Picking a green marble] $= 1 - \mu$.



→ Value of μ is unknown. Then we pick N marbles independently.

→ Sample = {  }.

→ The fraction of red marbles in sample $= \nu$.

Q : Does ν say something about μ ?

In a big sample (large N), ν is probably close to μ (with in ε)

Formally,

$$\mathbb{P}\left[\,|\nu-\mu| > \varepsilon\,\right] \leq 2e^{-2\varepsilon^2 N}$$

Hoeffding inequality

ν: Sample frequency

μ: Bin frequency

In other words, '$\mu = \nu$' P.A.C

P.A.C: Probably Approximately Correct

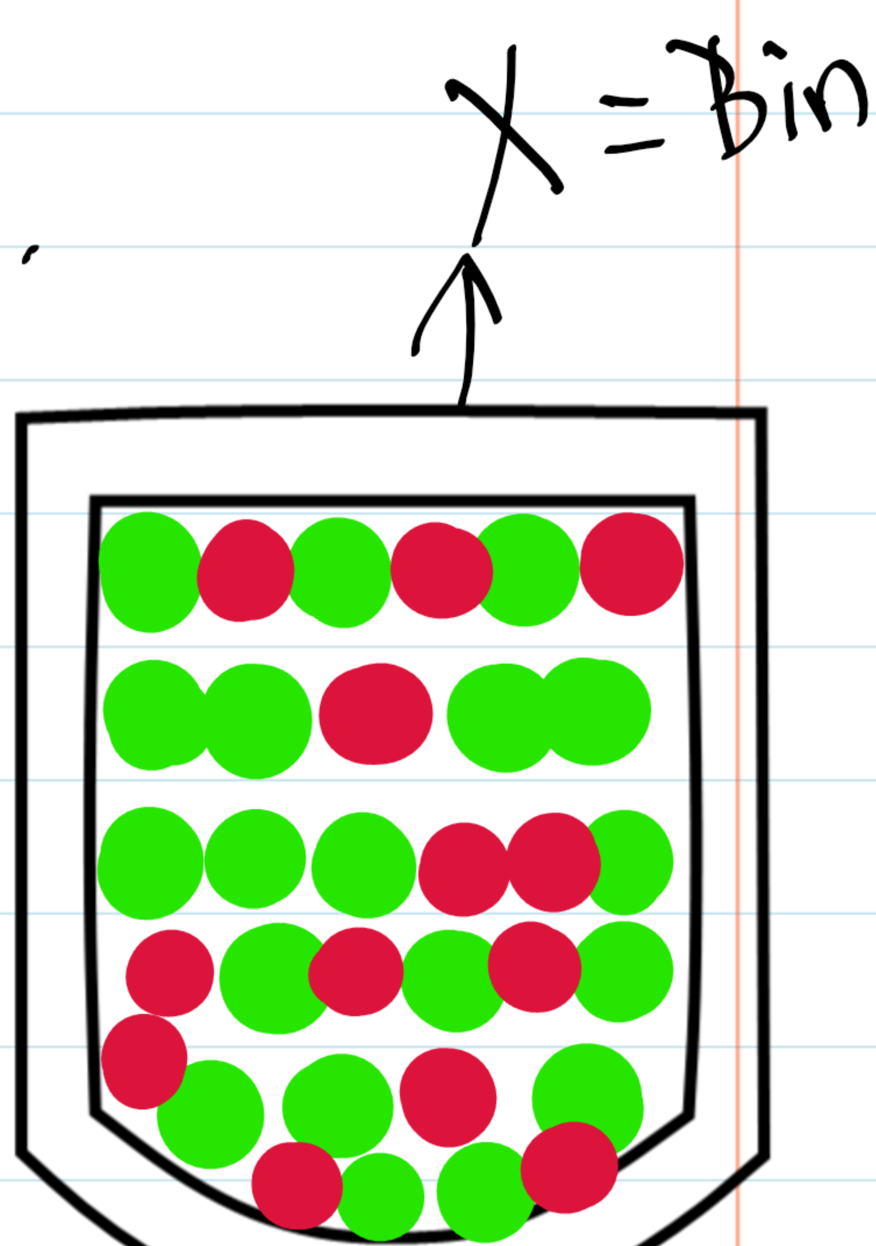
Trade off : N, ε and the bound.

Connection to learning :

Bin : The unknown is a number μ.

Learning : The unknown is an entire function $f : \mathcal{X} \rightarrow \mathcal{Y}$

Every marble 'o' is a point $x \in \mathcal{X}$.

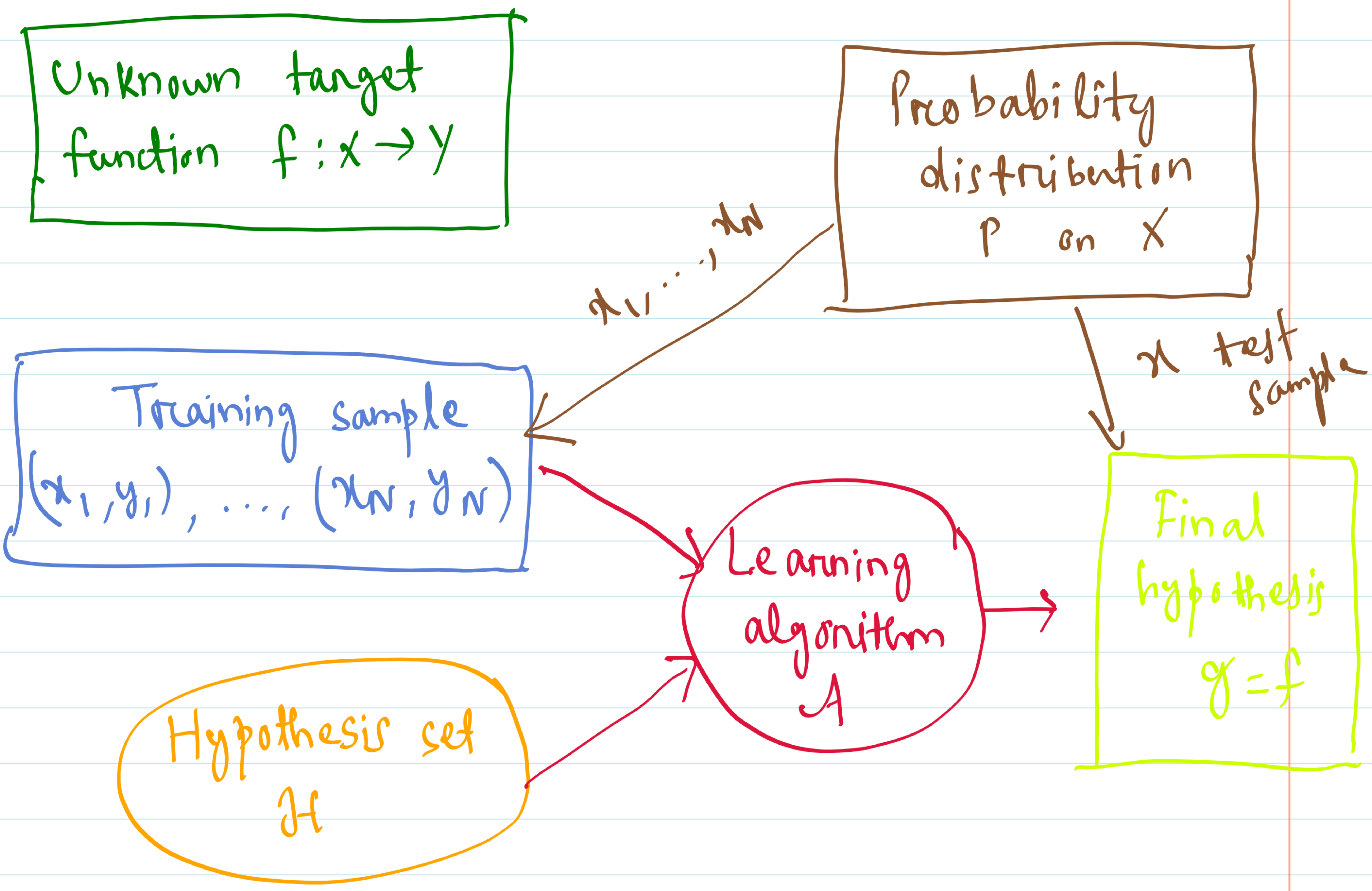


Take any single hypothesis $h \in \mathcal{H}$ & compare it to f at each point $x \in \mathcal{X}$.

- ● (green) ; hypothesis got it right $h(x) = f(x)$
- ● (red) ; hypothesis got it wrong $h(x) \neq f(x)$.

→ $\mathbb{P}$ can be anything

→ I don't need to know what $\mathbb{P}$ is.



→ Both μ and ν depend on h.
ν is 'in-sample' $E_{in}(h)$ & μ is 'out-sample' $E_{out}(h)$.

The Hoeffding inequality :

$$\mathbb{P}\left[\,|E_{in}(h) - E_{out}(h)| > \varepsilon\,\right] \leq 2e^{-2\varepsilon^2 N}.$$

Probability inequalities :

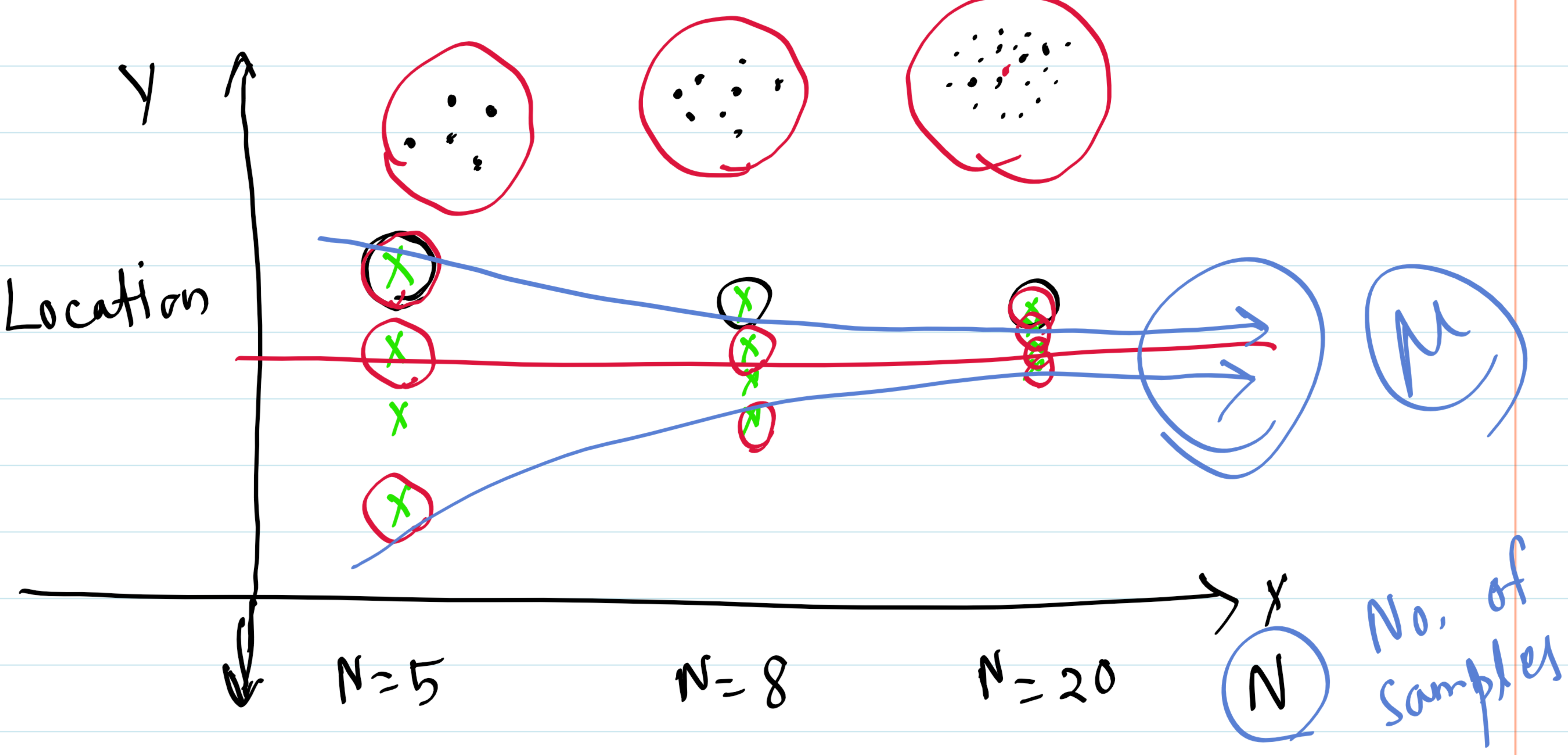
→ Take $X_1, \dots, X_N$ in 1-D, which are i.i.d r.v,

$$\mathbb{E}[X_1] = \dots = \mathbb{E}[X_N] = \mu \quad \text{(population mean)}$$

→ Empirical average $\nu = \frac{1}{N}\sum_{n=1}^{N} X_n$.

→ How close ν is to μ.

→



→ Markov inequality : $\mathbb{P}[X \geq \varepsilon] \leq \frac{\mathbb{E}[X]}{\varepsilon}$.

→ Chebyshev inequality : Let $X_1, \dots, X_N$ be i.i.d. with $\mathbb{E}[X_n] = \mu$, $Var[X_n] = \sigma^2$, $Y = \frac{1}{N}\sum_{n=1}^{N} X_n$,

Then $\mathbb{P}[|Y - \mu| > \varepsilon] \leq \frac{\sigma^2}{N\varepsilon^2}$.

→ Weak law of large number (WLLN)

Let $X_1, X_2, \dots, X_N$ be i.i.d. r.v. with common μ.

Let $M_N = \frac{1}{N}\sum_{n=1}^{N} X_n$. Then for any $\varepsilon > 0$

$$\lim_{n \to \infty} \mathbb{P}[|M_n - \mu| > \varepsilon] = 0$$

→ Strong law of large number (SLLN)

$$\mathbb{P}\left[\lim_{n \to \infty} |M_n - \mu| > \varepsilon\right] = 0.$$

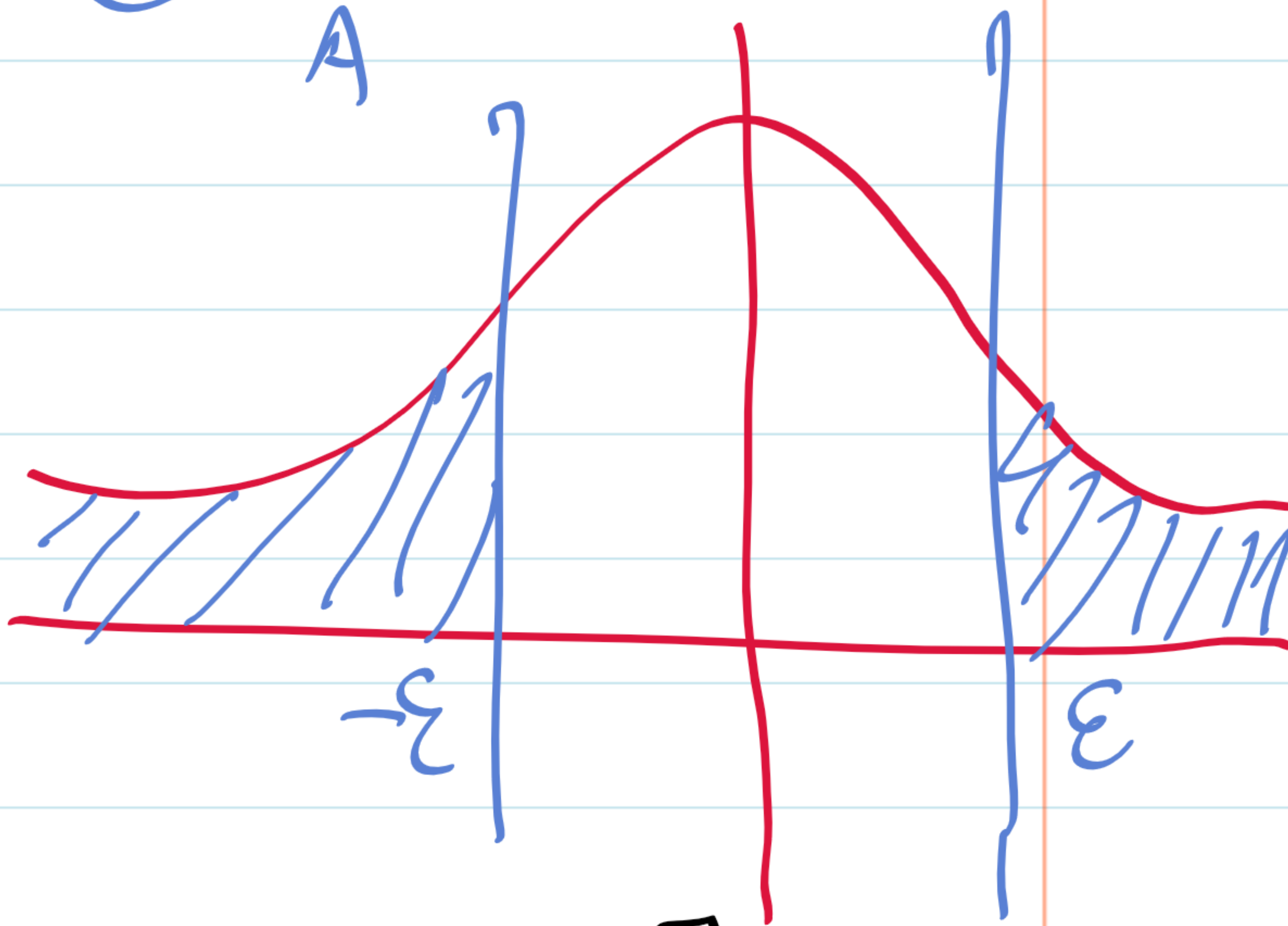
Let us revisit the bad event :

$$\mathbb{P}[|\nu-\mu| \geq \varepsilon] = \mathbb{P}[\nu-\mu \geq \varepsilon \text{ or } \nu-\mu \leq -\varepsilon]$$

$$\leq \underbrace{\mathbb{P}[\nu-\mu \geq \varepsilon]}_{A} + \underbrace{\mathbb{P}[\nu-\mu \leq -\varepsilon]}_{A}$$

$$\leq 2A$$

$A \to e^{-2\varepsilon^2 N}$.



The e-trick :

$$\mathbb{P}[\nu-\mu \geq \varepsilon] = \mathbb{P}\left[\frac{1}{N}\sum_{n=1}^{N} X_n - \mu \geq \varepsilon\right]$$

$$= \mathbb{P}\left[\sum_{n=1}^{N}(X_n-\mu) \geq \varepsilon N\right]$$

(e-trick) $$= \mathbb{P}\left[e^{s\sum_{n=1}^{N}(X_n-\mu)} \geq e^{s\varepsilon N}\right]$$

(Markov) $$\leq \frac{\mathbb{E}\left[e^{s\sum_{n=1}^{N}(X_n-\mu)}\right]}{e^{s\varepsilon N}}.$$

$$\leq \left(\frac{\mathbb{E}\left[e^{s(X_n-\mu)}\right]}{e^{s\varepsilon}}\right)^N.$$

If we let $Z_n = X_n - \mu$. Then

$$\mathbb{E}\left[e^{s(X_n-\mu)}\right] = M_{Z_n}(s) = \text{MGF of } Z_n$$

(moment generating function)

Hoeffding lemma :

If $a \le x_n \le b$, then $\mathbb{E}\left[e^{s(x_n-\mu)}\right] \le e^{\frac{s^2(b-a)^2}{8}}$.

Let $a=0$, $b=1$.

$$\mathbb{P}[\nu-\mu \ge \varepsilon] \le \left(\frac{\mathbb{E}\left[e^{s(x_n-\mu)}\right]}{e^{s\varepsilon}}\right)^N$$

$$\le \boxed{e^{s^2N/8 - s\varepsilon N}} \quad \forall s>0, \ (\text{function of } s)$$

minimize : s^*

$$f(s) = \frac{s^2N}{8} - s\varepsilon N.$$

$$\frac{d}{ds}\left(\frac{s^2N}{8} - s\varepsilon N\right) = \frac{N}{4}s - N\varepsilon = 0 \Rightarrow \boxed{s^* = 4\varepsilon}$$

So $\boxed{\mathbb{P}(\nu-\mu \ge \varepsilon)} \le e^{N/8\, s^{*2} - s^*\varepsilon N}$

$$= e^{N/8(16\varepsilon^2) - 4\varepsilon^2 N} = e^{-2\varepsilon^2 N}.$$

$$\boxed{\mathbb{P}(|\nu-\mu| \ge \varepsilon) \le 2e^{-2\varepsilon^2 N}}$$ Hoeffding inequality.

Compare between Hoeffding & chebyshev :

chebyshev : $\mathbb{P}(|\nu-\mu| \ge \varepsilon) \le \frac{\sigma^2}{N\varepsilon^2}$

Hoeffding : $\mathbb{P}(|\nu-\mu| \ge \varepsilon) \le 2e^{-2\varepsilon^2 N}$.

} $\mathbb{P}[(\nu)-(\mu) \ge \varepsilon] \le (\delta)$

Equivalent to: For probability at least $1-\delta$, we have

$$\mu - \varepsilon \le \nu \le \mu + \varepsilon$$

(μ: testing error (unknown); ν: training error)

→ Chebyshev: $\delta = \frac{\sigma^2}{N\varepsilon^2} \Rightarrow \varepsilon = \frac{\sigma}{\sqrt{\delta N}}$

Hoeffding: $\delta = 2e^{-2\varepsilon^2 N} \Rightarrow \varepsilon = \sqrt{\frac{1}{2N}\log\left(\frac{2}{\delta}\right)}$

Example: Chebyshev: For probability at least $1-\delta$, we have

$$\mu - \frac{\sigma}{\sqrt{\delta N}} \le \nu \le \mu + \frac{\sigma}{\sqrt{\delta N}}.$$

Hoeffding: For probability at least $1-\delta$, we have

$$\mu - \sqrt{\frac{1}{2N}\log\frac{2}{\delta}} \le \nu \le \mu + \sqrt{\frac{1}{2N}\log 2/\delta}.$$

You ask: I have data $x_1, \dots, x_N$. I want to estimate μ. How many data points N do I need?

Other person: How much δ can you tolerate?

You say: I only have limited no. of data points. Then how good my estimate is? (ε)

Other person: How many data points N do you have.

Example : Let $\delta = 0.01$ (1% error), $N = 10000$
$\sigma = 1$ (r.v.s). How much tolerance you can afford?

$$\varepsilon = \frac{\sigma}{\sqrt{\delta N}} = 0.1$$
(Chebyshev)

$$\varepsilon = \sqrt{\frac{1}{2N} \log \frac{2}{\delta}} = 0.016$$
(Hoeffding) → lot tighter

Change your question: probability of error $\delta \leq 0.01$,

and $\varepsilon = 0.01$, $\sigma = 1$, then

$$N \geqslant \frac{\sigma^2}{\varepsilon^2 \delta} = 1{,}000{,}000$$
(Chebyshev)

and

$$N \geqslant \frac{\log 2/\delta}{2\varepsilon^2} \approx \boxed{26{,}500}$$
(Hoeffding)

PAC

PAC learning framework

Recall the eq^n : $\mathbb{P}\left[|E_{in}(h) - E_{out}(h)| > \varepsilon \right] \leq 2e^{-2\varepsilon^2 N}$.

One can bound $E_{out}(h)$ using $E_{in}(h)$.

$E_{out}(h)$ is something that you don't know; $E_{in}(h)$ which is something that you know.

- → RHS is independent of h and $p(x)$. (universal bound)
- → This works for any $\mathcal{A}$, $\mathcal{H}$, f & $p(x)$.
- → $\delta = 2e^{-2\varepsilon^2 N}$ confidence : $1-\delta$.
- → $\varepsilon = \sqrt{\frac{1}{2N} \log \frac{2}{\delta}}$, accuracy : $1-\varepsilon$.
- → One can write : $\mathbb{P}\left[|E_{in}(h) - E_{out}(h)| > \varepsilon \right] \leq \delta$, which is equivalent to

$$\mathbb{P}\left[|E_{in}(h) - E_{out}(h)| \leq \varepsilon \right] \geq 1-\delta .$$

Probably Approximately Correct (PAC) framework,

Probably: Quantify error using probability

$$\mathbb{P}\left[|E_{in}(h) - E_{out}(h)| \leq \varepsilon \right] \geq 1-\delta$$

Approximatel correct : In-sample error is an approximation of the out-sample error.

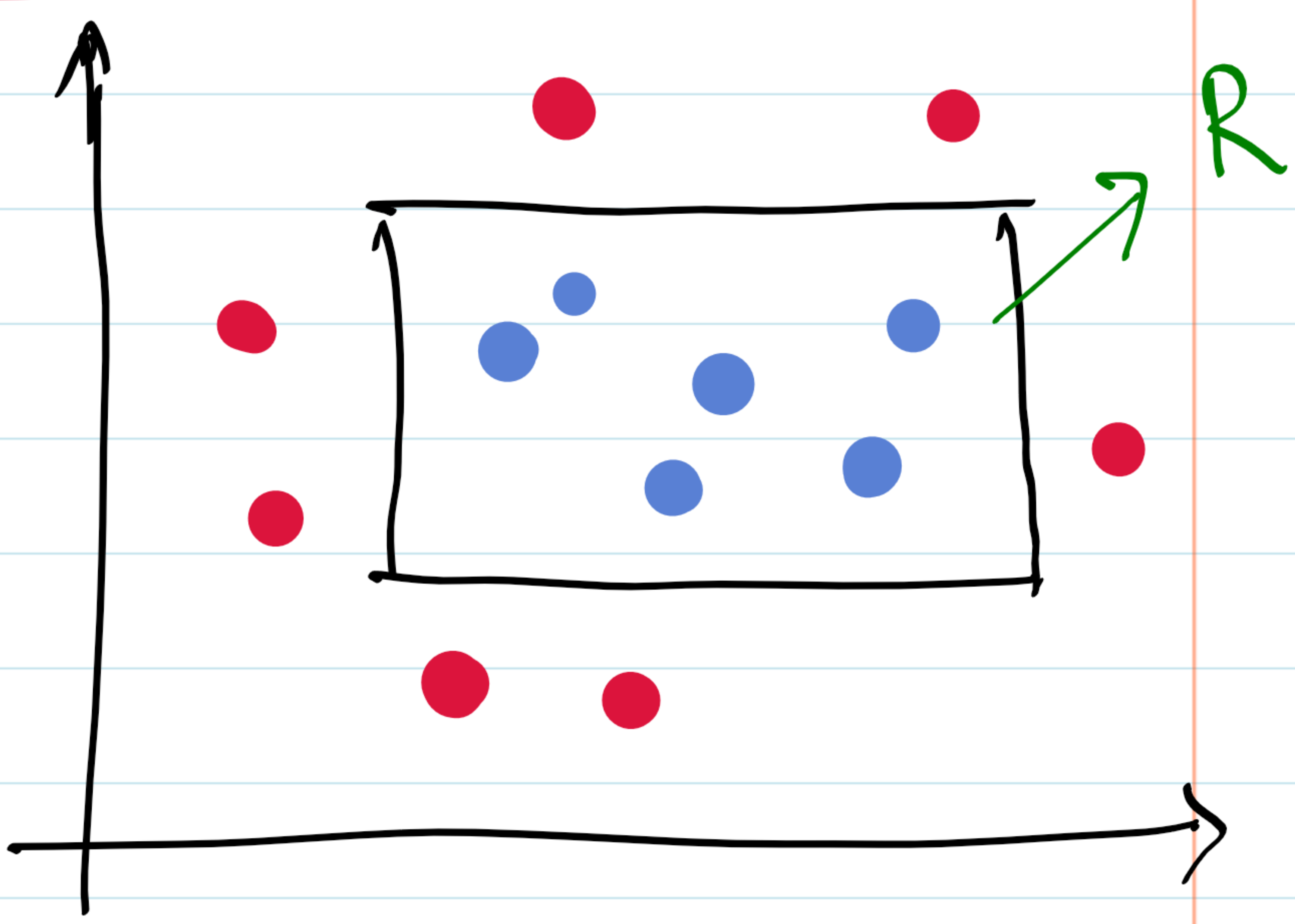
$$\mathbb{P}\left[|E_{in}(h) - E_{out}(h)| \leq \varepsilon \right] \geq 1-\delta$$

PAC-learnable : If you can find an algorithm $\mathcal{A}$ such that for any ε and $\delta > 0$, there exists an N which can make the above inequality holds, then we say that the target function is PAC-learnable.

Example : Rectangular classifier :

→ Consider a set of 2D data points.

→ The target function R is a rectangle.



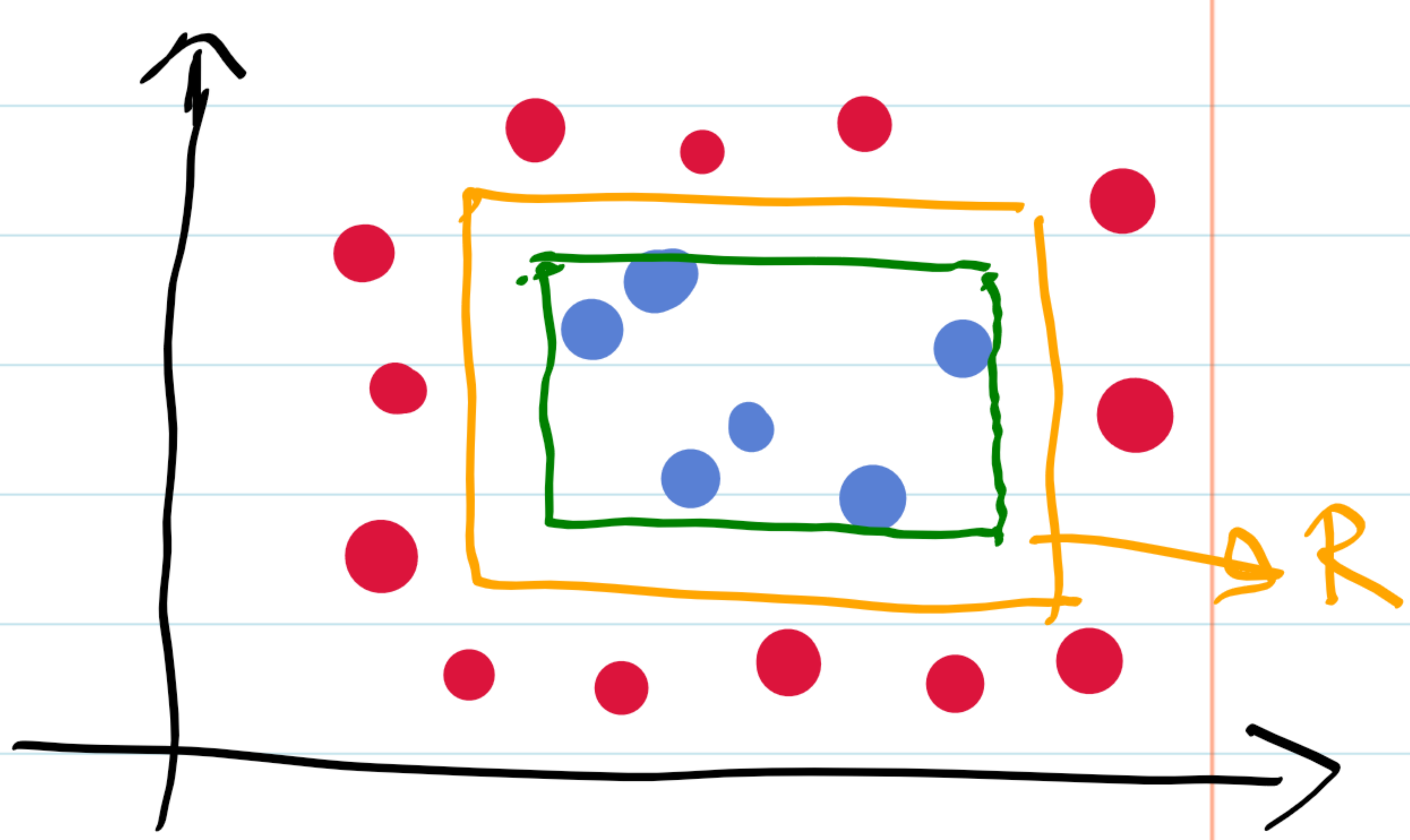
→ Inside R : blue
Outside R : red } Data is separable

Question : Is this problem PAC-learnable?

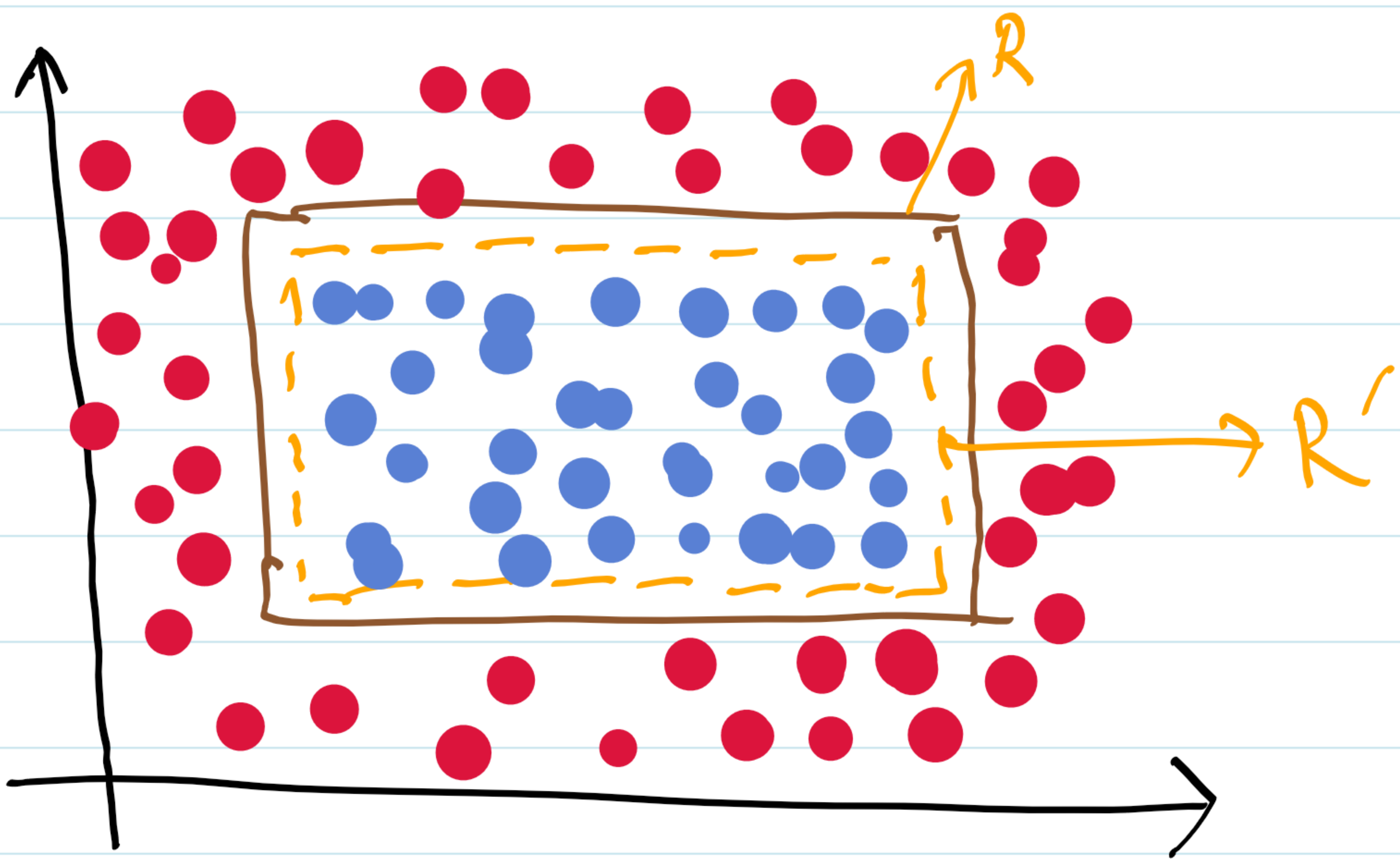
→ Mathematically we need to propose an algorithm $\mathcal{A}$ which takes the training data and returns R′, such that for any $\varepsilon > 0$ & $\delta > 0$, there exist an N (which is a func. of ε & δ) with

$$\mathbb{P}\left[\,|E_{in}(R') - E_{out}(R')| > \varepsilon\,\right] \leq \delta .$$

→ Proposed algorithm : Give me the set of data points, find the tightest rectangle that covers the blue dots.



Intuition : As N grows, we can find R′ which is getting closer and closer to R.



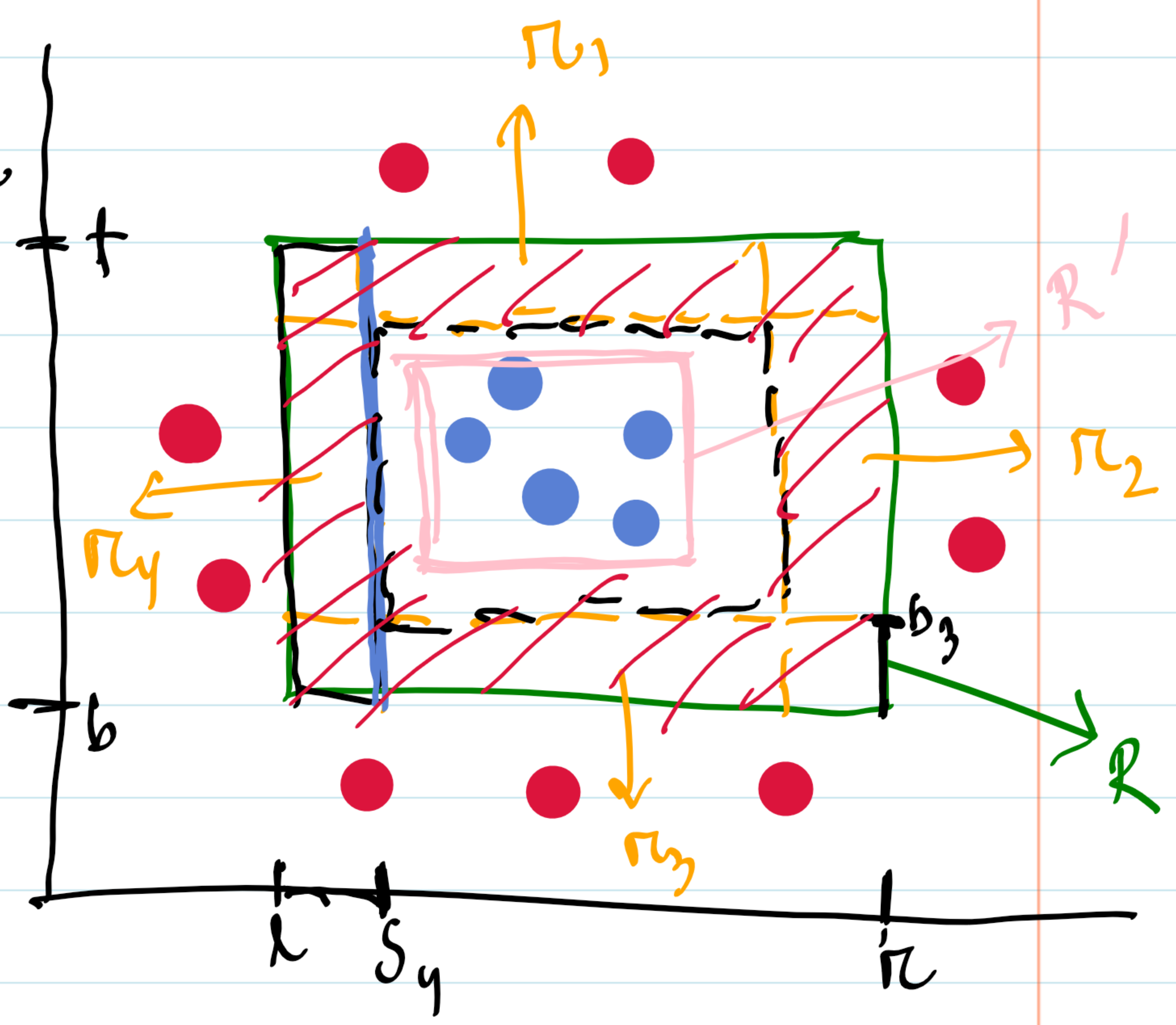
So for any $\varepsilon > 0$, $\delta > 0$ it is possible that as long as N is large enough, we will be able to make training error close to testing error.

Proof - Let R be the target fun^c. Fix $\varepsilon > 0$.

Let $\mathbb{P}[R]$ denote the probability mass of the region defined by R. i.e., the probability that a point randomly drawn according to distribution P falls within R.

We assume $\mathbb{P}[R] > \varepsilon$.

We can define 4 rectangular regions r_1, r_2, r_3 & r_4 along the sides of R, each with probability atleast $\varepsilon/4$.



Let l, r, b & t be the four real values defining R: $R = [l, r] \times [b, t]$.

r_4 is defined by $r_4 = [l, s_4] \times [b, t]$, with

$$s_4 = \inf\{ s : \mathbb{P}[[l, s] \times [b, t]] \geq \varepsilon/4 \}.$$

$\bar{r}_4$ = excluding the right most, $\mathbb{P}[\bar{r}_4] \leq \varepsilon/4$.

Observe here if R' meets all of these four regions r_i, $i = 1, 2, 3, 4.$, because R' is a rectangle, it will have one side in each of these regions.

$$E_{out}(R') \leq \varepsilon = \left(4 \cdot \varepsilon/4\right)$$

$A_i = \{ R' \cap r_i \neq \phi \ \forall i \}$, $B = \{ E_{out}(R') \leq \varepsilon \}$

$$\bigcap_{i=1}^{4} A_i \subset B \Rightarrow B^c \subset \left(\bigcap_{i=1}^{4} A_i \right)^c = \bigcup_{i=1}^{4} A_i^c$$

$$\Rightarrow \mathbb{P}(B^c) \leq \mathbb{P}\left(\bigcup_{i=1}^{4} A_i^c\right)$$

$$\Rightarrow \mathbb{P}\left(E_{out}(R') > \varepsilon\right) \leq \mathbb{P}\left(\bigcup_{i=1}^{4} \{R' \cap r_i = \emptyset\}\right)$$

$$\leq \sum_{i=1}^{4} \mathbb{P}(\{R' \cap r_i = \emptyset\})$$

$(\mathbb{P}[r_i] \geq \varepsilon/4)$ $\quad \leq 4\,(1-\varepsilon/4)^{m}$ — m → no. of data points

$\begin{pmatrix} 1-x \leq e^{-x} \\ \forall x \in \mathbb{R} \end{pmatrix}$ $\quad \leq 4 \exp(-m\varepsilon/4)$

Thus for any $\delta > 0$, we have

$$\mathbb{P}[E_{out}(R') > \varepsilon] \leq \delta$$

$$4\exp(-\varepsilon m/4) \leq \delta \Rightarrow m \geq \frac{4}{\varepsilon}\log\frac{4}{\delta}$$

$$\mathbb{P}[|E_{in}(R') - E_{out}(R')| > \varepsilon] \leq \delta.$$

($E_{in}(R') = 0$)

Thus for any $\varepsilon > 0$ & $\delta > 0$, if the sample size m is greater than $4/\varepsilon \log 4/\delta$, then

$$\mathbb{P}[E_{out}(R') > \varepsilon] \leq \delta. \quad \text{(PAC-learnable)}$$

Guarantee VS Possibility :

Basically we see the difference between the deterministic and probabilistic learning.

Deterministic : Can D tell us something **Certain** about f outside of D ? — **NO**

Probabilistic : "Can $\mathcal{D}$ tell us something possibly about f outside of $\mathcal{D}$?" — YES

One hypothesis vs final hypothesis :

In Hoeffding, $\mathbb{P}\left[\,|E_{in}(h) - E_{out}(h)| > \varepsilon\,\right] \leq 2e^{-2\varepsilon^2 N}$.

hypothesis h is fixed.

We need to choose h before we look at the data set, not after.

When we need to choose g from $h_1, \ldots, h_M$,

we need to repeat Hoeffding M times.

The factor 'M' :

We can $|E_{in}(g) - E_{out}(g)| > \varepsilon$

$\Rightarrow$ $|E_{in}(h_1) - E_{out}(h_1)| > \varepsilon$

or $|E_{in}(h_2) - E_{out}(h_2)| > \varepsilon$

or . . . .

. . . . .

$|E_{in}(h_M) - E_{out}(h_M)| > \varepsilon$

} possible cases

$$\mathbb{P}\left(|E_{in}(g) - E_{out}(g)| > \varepsilon\right) \leq \sum_{m=1}^{M} \mathbb{P}\left(|E_{in}(h_m) - E_{out}(h_m)| > \varepsilon\right).$$

→ If $A \Rightarrow B$, $P(A) \leq P(B)$

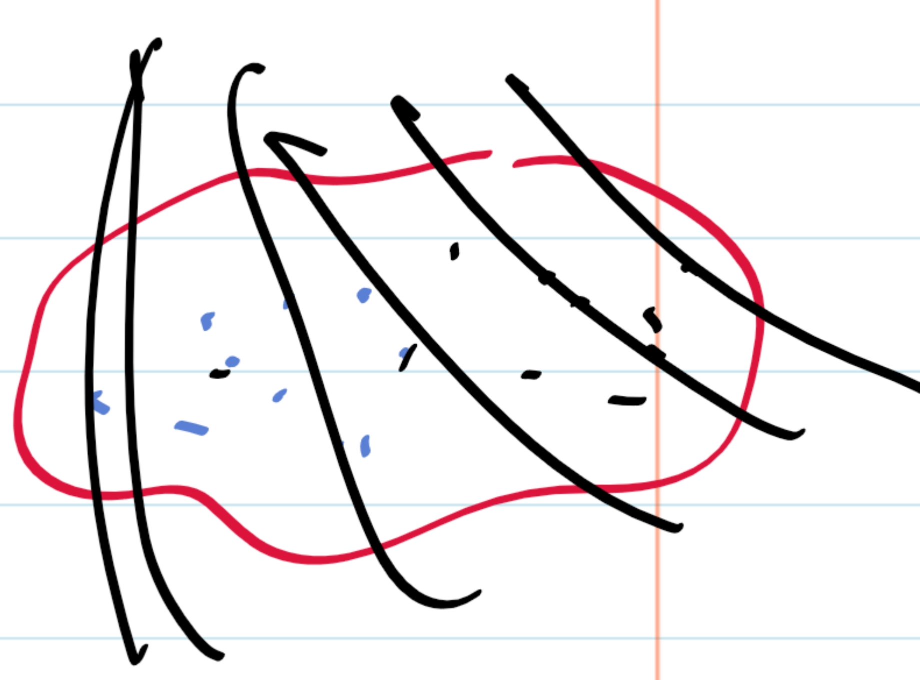
→ Union bound $P(A \text{ or } B) \leq P(A) + P(B)$.

Hoeffding : $\mathbb{P}\left(|E_{in}(g) - E_{out}(g)| > \varepsilon\right) \leq 2Me^{-2\varepsilon^2 N}$.

M is a constant.

Bad : M can be large, or even ∞,

Good : We can bound this M.



Learning goal : $E_{out}(g) \approx 0$.

To achieve this : $E_{out}(g) \approx E_{in}(g) \approx 0$

($E_{out}(g) \approx E_{in}(g)$: Hoeffding inequality; $E_{in}(g) \approx 0$: Training error)

Complex $\mathcal{H}$:

Hoeffding : $\mathbb{P}\left[\,|E_{in}(g) - E_{out}(g)| > \epsilon\,\right] \le 2M e^{-2\epsilon^2 N}$

→ If $\mathcal{H}$ is complex, M will be large, so the approximation will worsen.

→ If $\mathcal{H}$ is complex, you have more options during training. So training error is improved.

→ Trade-off :

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

($E_{out}(g) \approx E_{in}(g)$: worse if $\mathcal{H}$ is complex; $E_{in}(g) \approx 0$: good if $\mathcal{H}$ is complex)

→ We can not use a very complex model. Simple model generalise better.

Complex f (target fu^{n}c) :

→ good : Hoeffding is not affected by f.

→ Bad : If f is complex, then it will be very hard to train, so training error ~~be~~ cannot be small.

→ Trade-off :

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

($E_{out}(g) \approx E_{in}(g)$: not affected; $E_{in}(g) \approx 0$: worse if f is complex)