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

get rid of

give bound

$$\mathbb{P}\left[\,|E_{in}(g) - E_{out}(g)| \le \varepsilon\,\right] \ge 1-\delta .$$

Equivalently, with probability $1-\delta$,

$$E_{in}(g) - \varepsilon \le E_{out}(g) \le E_{in}(g) + \varepsilon$$

$$\text{So } \delta = 2M e^{-2\varepsilon^2 N} \Rightarrow \varepsilon = \sqrt{\frac{1}{2N}\log\frac{2M}{\delta}} .$$

$$E_{in}(g) - \sqrt{\frac{1}{2N}\log\frac{2M}{\delta}} \le E_{out}(g) \le E_{in}(g) + \sqrt{\frac{1}{2N}\log\frac{2M}{\delta}}$$

Generalization bound

→ N is training sample. (the more, the better)

→ δ : probability tolerance level, $1-\delta$: "confidence".
small δ : 'Conservative', So needs to have large N to compensate for $\log 1/\delta$.

→ M : Model complexity, Large M : Complicated model.
So we need large N to compensate for $\log M$.

Upper limit :

$$E_{out}(g) \le E_{in}(g) + \sqrt{\frac{1}{2N}\log\frac{2M}{\delta}}$$

→ E_{out} Can not be worse than $E_{in} + \varepsilon$.

→ Performance guarantee : $E_{in} + \varepsilon$ is the worst you will have.

Lower limit :

$$E_{in}(g) - \sqrt{\frac{1}{2N}\log\frac{2M}{\delta}} \le E_{out}(g) .$$

→ E_{out} can not be better than $E_{in} - \varepsilon$.

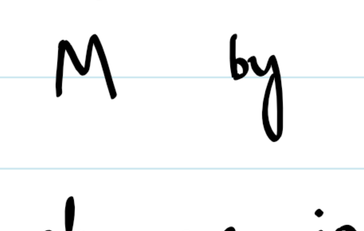
N, M, $1-\delta$

Growth function :

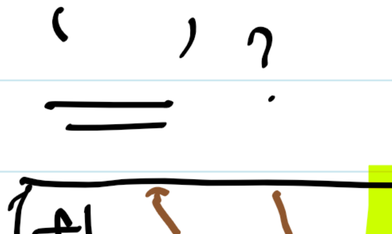
Let B_m be the (bad) event that "$|E_{in}(h_m) - E_{out}(h_m)| > \varepsilon$"

Then $\mathbb{P}[B_1 \text{ or } B_2 \text{ or } \dots \text{ or } B_M] \leq \mathbb{P}[B_1] + \mathbb{P}[B_2] + \dots + \mathbb{P}[B_M]$.

Two cases : weakly overlapping



strongly overlapping



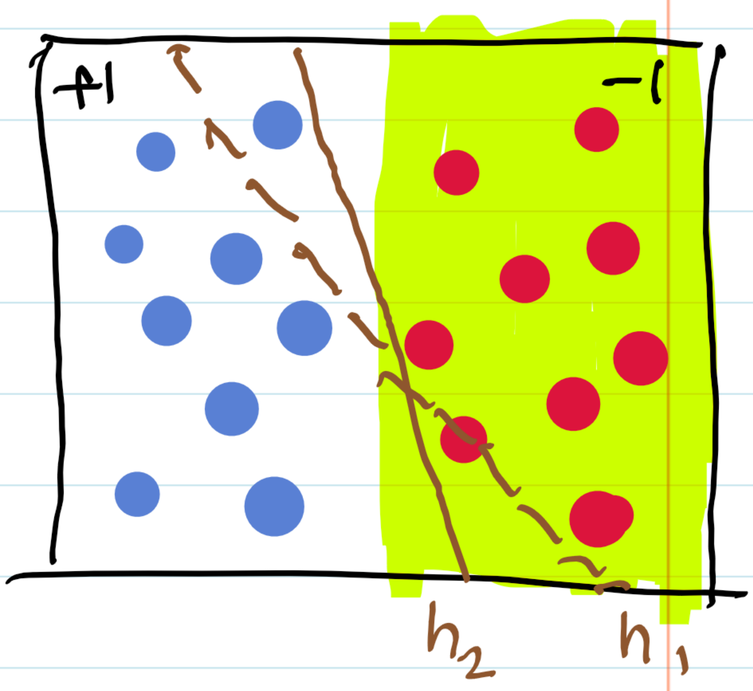
gives a gross over-estimate.

→ If h_1 is very similar to h_2, then the two events '$|E_{in}(h_1) - E_{out}(h_1)| > \varepsilon$' and '$|E_{in}(h_2) - E_{out}(h_2)| > \varepsilon$' are likely to coincide for most data sets.

→ Replace M by an effective number '$=$' ?

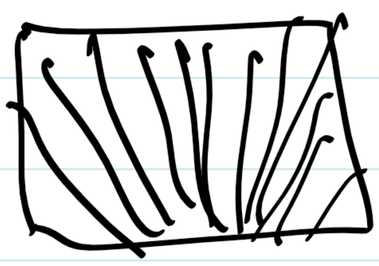
ΔE_{out} = change in the $+1$ & -1 area

ΔE_{in} = chang in the labels of training samples.

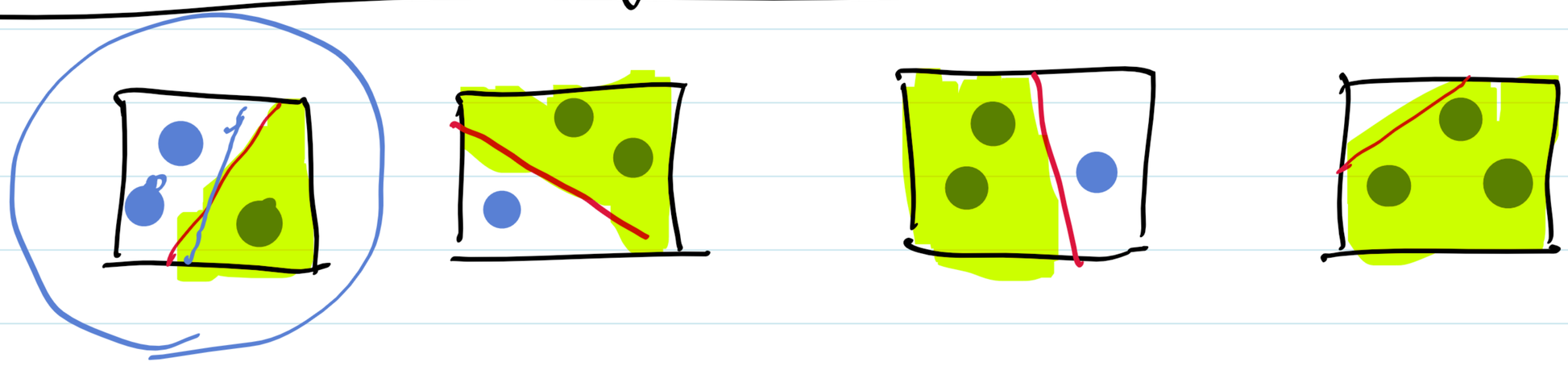


We shoud expect

$$\mathbb{P}[|E_{in}(h_1) - E_{out}(h_1)| > \varepsilon] \simeq \mathbb{P}[|E_{in}(h_2) - E_{out}(h_2)| > \varepsilon].$$



Restrict to training sets only :

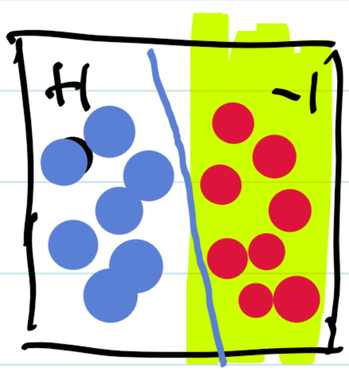


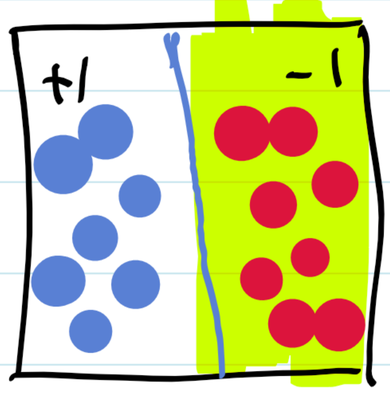
Idea : Just look at the training samples. Don't care until a training sample flip its sign.

Dichotomies : Means no. of mini-hypothesis.

Hypothesis	Dichotomy
→ $h : X \to \{+1, -1\}$	→ $h : \{x_1, \ldots, x_N\} \to \{+1, -1\}$
→ # population samples	→ for training samples only
→ No. can be infinite	→ No. is atmost 2^N.

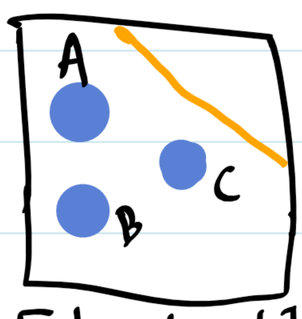
→ Observe : Different hypothesis, same dichotomy.



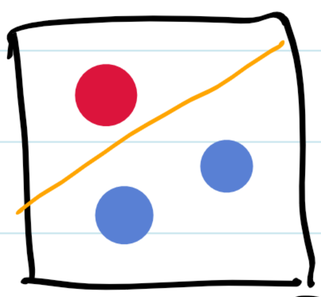


Def^n : Let $x_1, \ldots, x_N \in X$. The dichotomies generated by $\mathcal{H}$ on these points are

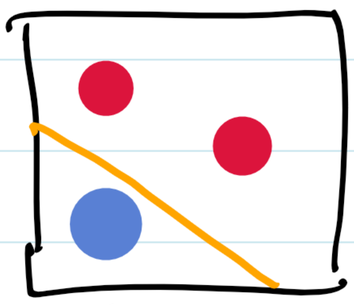
$$\mathcal{H}(x_1, \ldots, x_N) = \{(h(x_1), \ldots, h(x_N)) \mid h \in \mathcal{H}\}$$



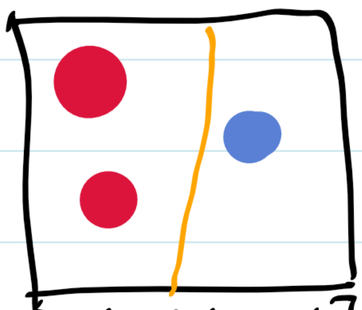
[-1, -1, -1]



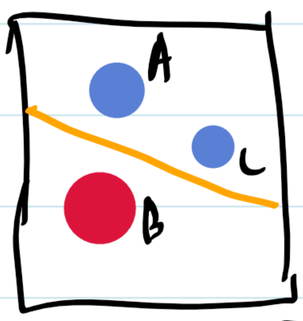
[+1, -1, -1]



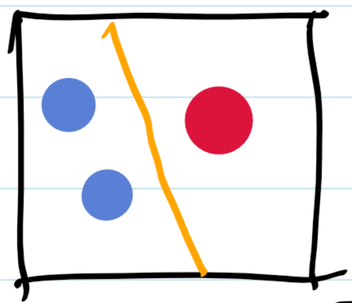
[+1, -1, +1]



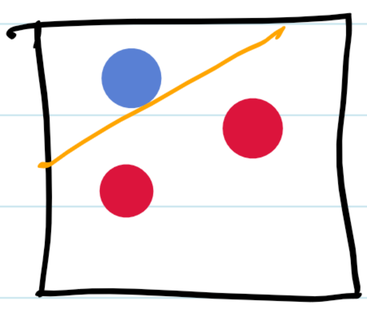
[+1, +1, -1]



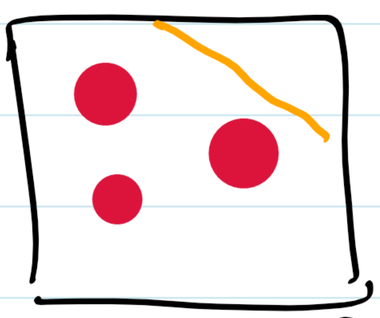
[-1, +1, -1]



[-1, -1, +1]

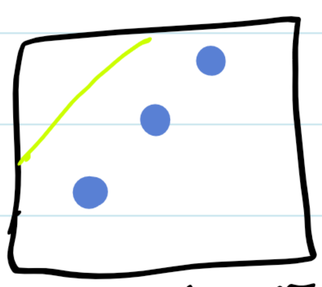


[-1, +1, +1]

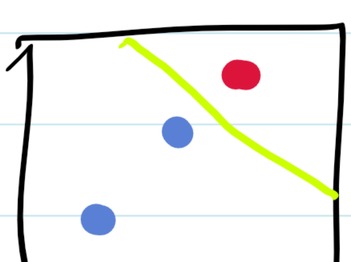


[+1, +1, +1]

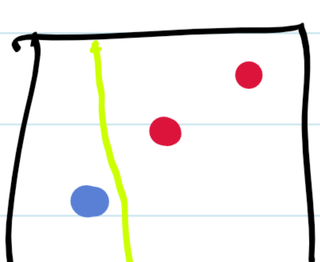
Not allowed :



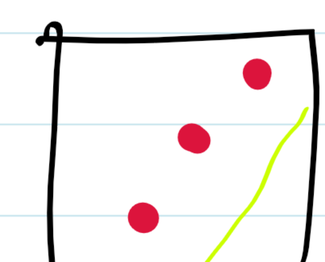
[-1, -1, -1]



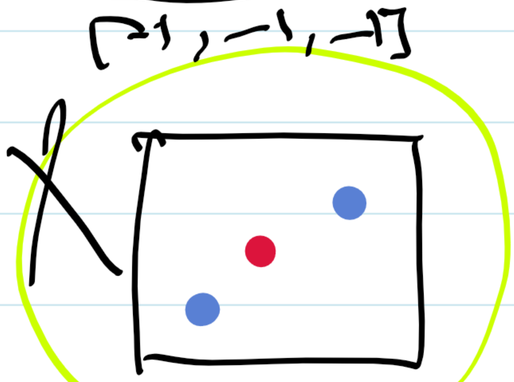
[+1, -1, -1]



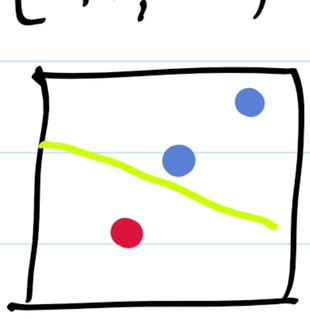
[-1, +1, -1]



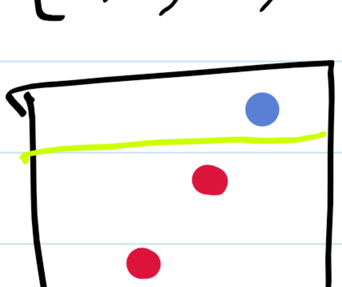
[+1, +1, +1]



[-1, +1, -1]



[-1, -1, +1]



[+1, -1, +1]

[+1, +1, +1]

Candidate to replace M :

Define Growth function

$$m_{\mathcal{H}}(N) = \max_{x_1,\dots,x_N \in X} \left| \mathcal{H}(x_1,\dots,x_N) \right| ,$$

here $|\cdot|$ denotes the cardinality (no. of elements) of a set.

In words, $m_{\mathcal{H}}(N)$ is the maximum number of dichotomies that can be generated by $\mathcal{H}$ on any N points.

→ Large $m_{\mathcal{H}}(N)$ = more complicated $\mathcal{H}$.

→ $m_{\mathcal{H}}(N)$ depends on $\mathcal{H}$ and N, but not on the algorithm and not on the probability distribution $P(x)$

→ To compute $m_{\mathcal{H}}(N)$: Consider all possible choices of N points $x_1, .., x_N$ from X. Then pick the one that gives you the most dichotomies.

→ For any $\mathcal{H}$, $\mathcal{H}(x_1, .., x_N) \subseteq \{-1, +1\}^N$.

So $m_{\mathcal{H}}(N) \leq 2^N$.

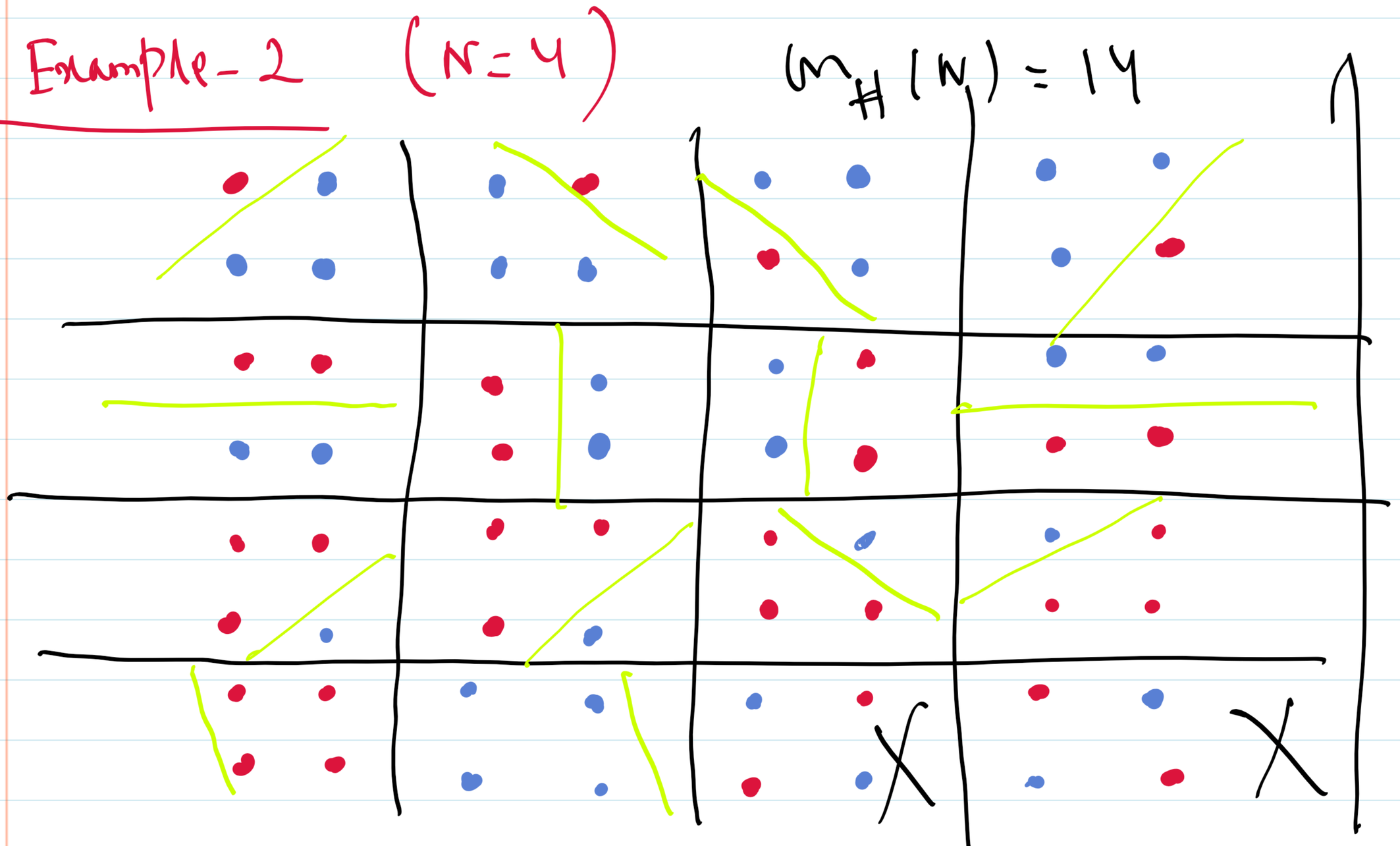
Example - 1 ($N = 3$)

$\mathcal{H}$ = linear models in 2D $N = 3$

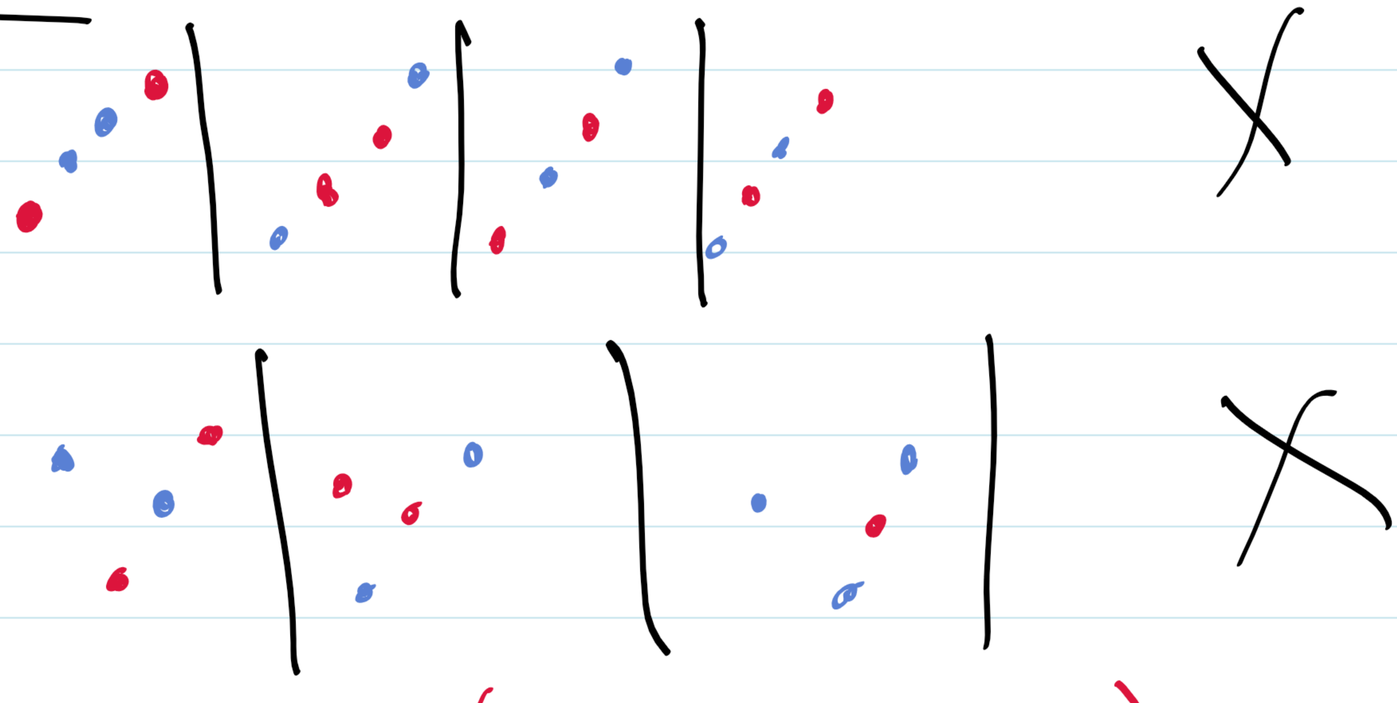
We have 8 dichotomies. (max..) Growth fn ≤ 8.

Example - 2 ($N = 4$)

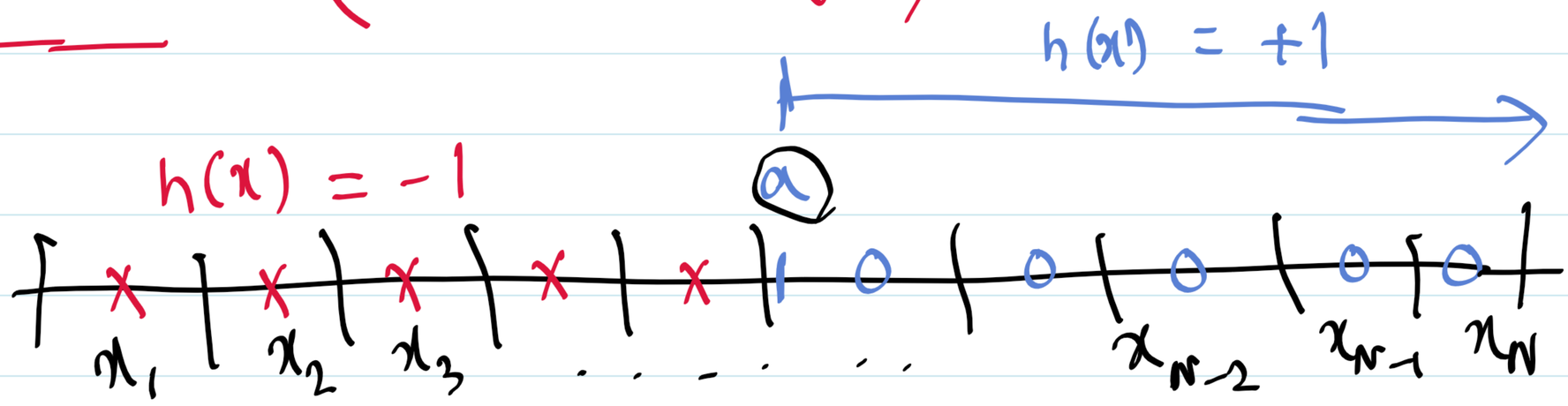
$m_{\mathcal{H}}(N) = 14$



N = 4 case



Example - 3 (Positive rays)



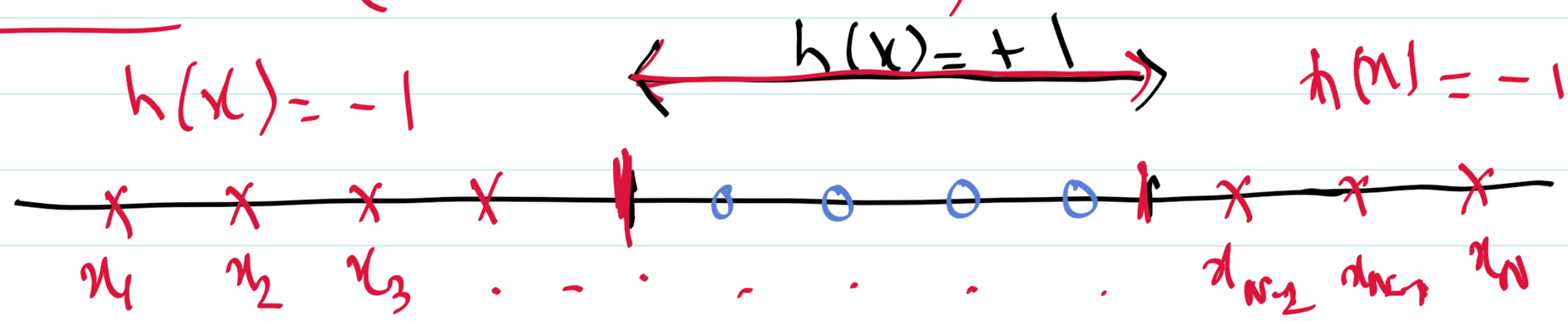
→ $\mathcal{H}$ consists of $h : \mathbb{R} \to \{+1, -1\}$.

→ $h(x) = \text{Sign}(x - a)$

→ Cut the line in to two halves.

→ $m_{\mathcal{H}}(N) = N + 1$

Example - 4 (Positive intervals)



$$m_{\mathcal{H}}(N) = \binom{N+1}{2} + 1 = \frac{N^2}{2} + \frac{N}{2} + 1$$