hnt3lm eh mn el slides de?
1. hal e7na n2dr nt3lm general function a2dr beha a3ml correct prediciton le data ana mshufthash bona2n 3la el data set elly etmrnt 3leha wala laa, lw ah b3ml keda ezay.
2. y3ny a hoeffdeings inequallity, w ezay btshtghl 3la el single hypothsis, and multiple hypotheses.

always remember that our goal is to know -> is learning from data feasible?

Lecture 2: The Learning Theory

DINA ELREEDY

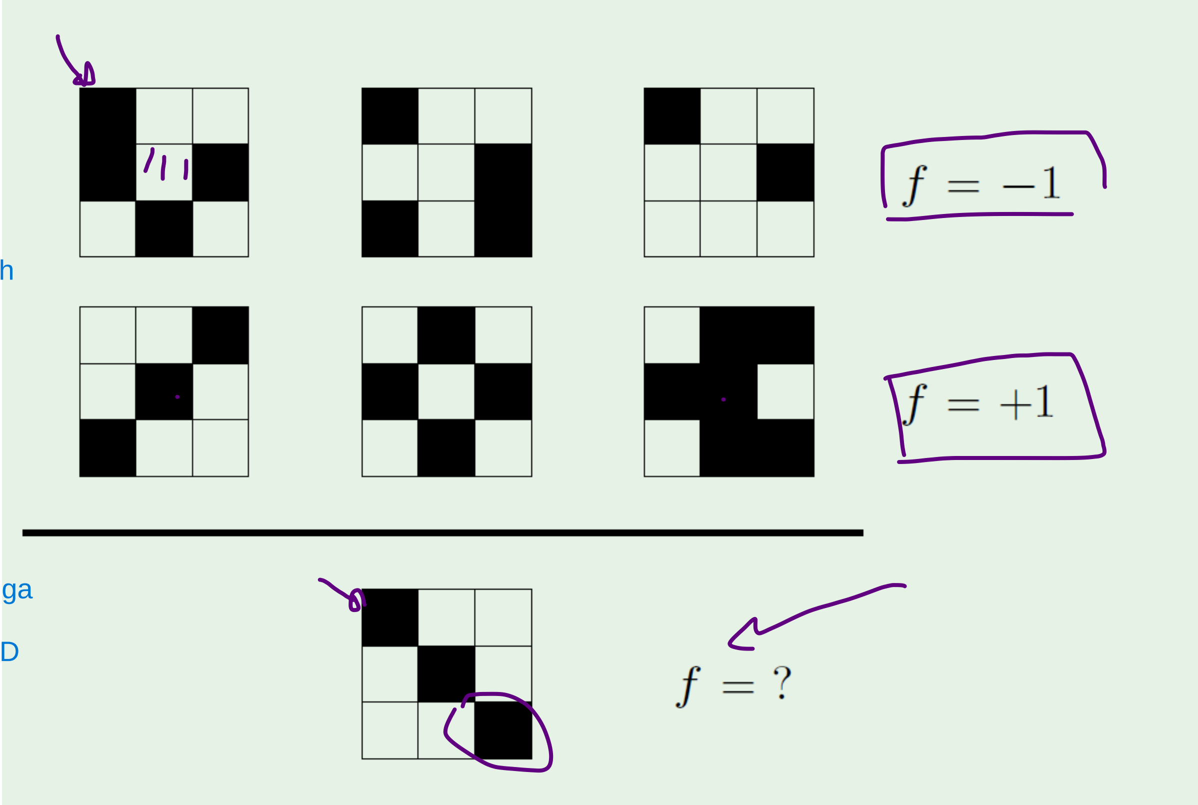
CMP402B-SPRING 2022

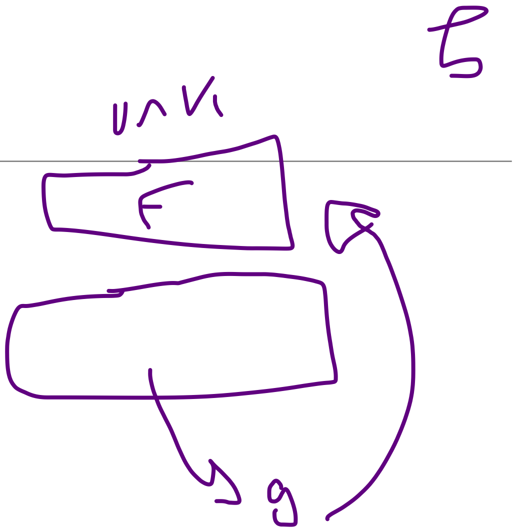
Given that the upper row is classified as -1
and the lower row is classified as 1.
given the test case, can you tell in which class it would be classified?

As we sajd in the last lecture, we have a target function which is completely unknown, and we are trying to learn from data, in other word, we learn from experience, not by applying analysis.
so the obvious question here is, how can a limited data set allow me to correctly generate g function which allows me to accuratly estimate f which is the unknown target function.

Learning Puzzle

Actually we can apply many analytical approaches:
1. if symmetric, then we may say it is 1.
2. if the top left corner is black, then it should be -1
3. etc...
however, we will see that there are no enough data to precisly judge the test case.





We know what we have already seen, but
that is not learning, that is memorizing -> basmga

Does the data set D tell us anything outside of D
that we didn't know before? if yes -> learning
if no -> them learning is not feasible.

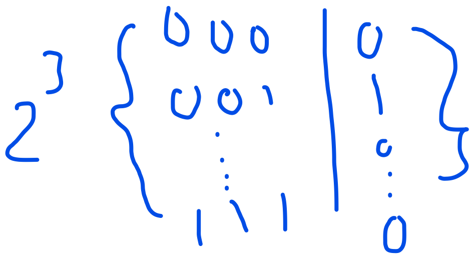
And we will discuss in the following slides that this is not a special case, usually small data set can not give me a general rule to be able to judge correctly.

0.1
1
4
h_1
w
0.1
1
4
h_2
0.5
0.6
3
Σ h

Is learning feasible?-Example

$X=\{0,1\}^3$

$Y=\{0,1\}$



in how many ways we can have distinct y vector?
since repeation is allowed,
since we have 8 samples,
since each sample have 2 options.
therefore # of ways = 2 ^ 8.

2^8

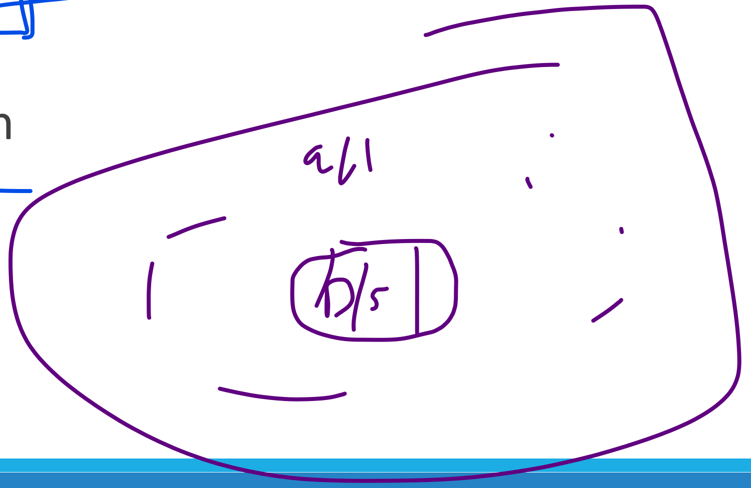
Hypothesis set H:All Boolean functions X→Y

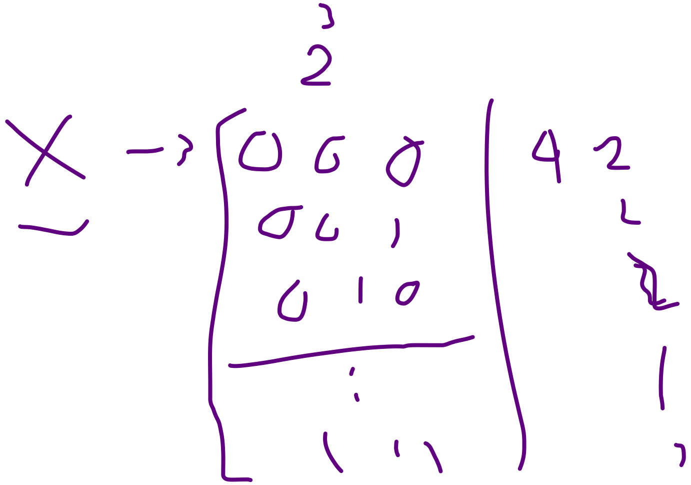
|H|=?

|H|=2^8=256

f: Target function

g





this is not a complete table.

Is learning feasible?-Example

?

x	y	g	f1	f2	f3	f4	f5	f6	f7	f8
0 0 0	0	0	0	0	0	0	0	0	0	0
0 0 1	1	1	1	1	1	1	1	1	1	1
0 1 0	1	1	1	1	1	1	1	1	1	1
0 1 1	0	0	0	0	0	0	0	0	0	0
100	1	1	1	1	1	1	1	1	1	1
101	?	?	0	1	0	1	0	1	0	1
110	?	?	0	0	1	1	0	0	1	1
111	?	?	0	0	0	0	1	1	1	1

Observed Training data

Testing data

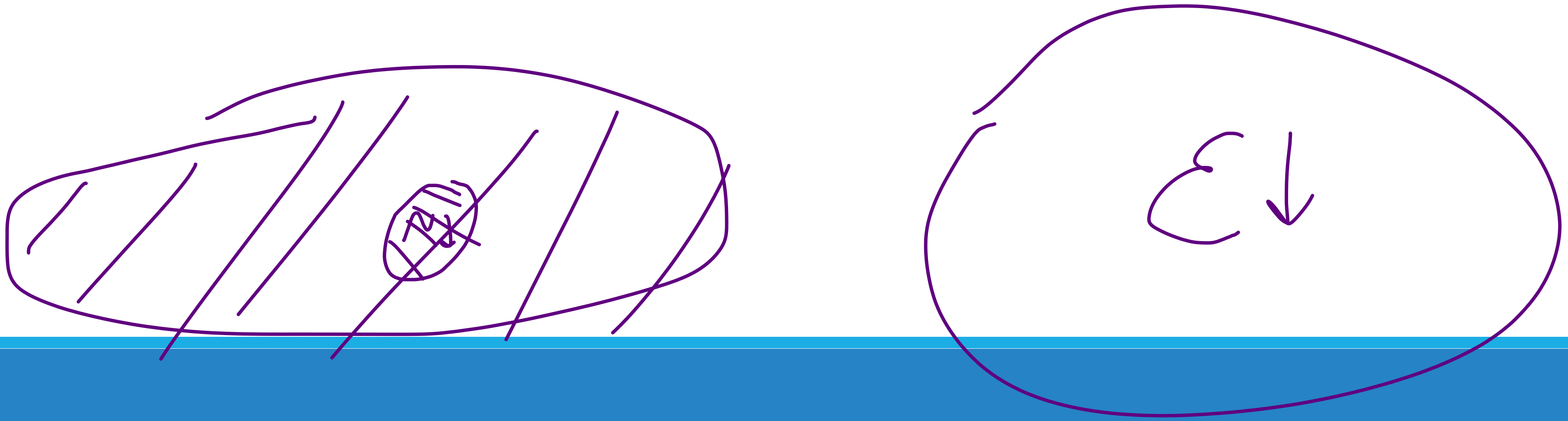
Same values

so does this mean that learning from the data is doomed? if yes, then why we have invented this course y3ny? :D

Nothing to tell us which f is the true function??

Is learning feasible?

- Can we generalize to unseen data?
- Yes, it is feasible to learn the unknown function f in a probabilistic way.



30%

$\boxed{3}$

Analogy

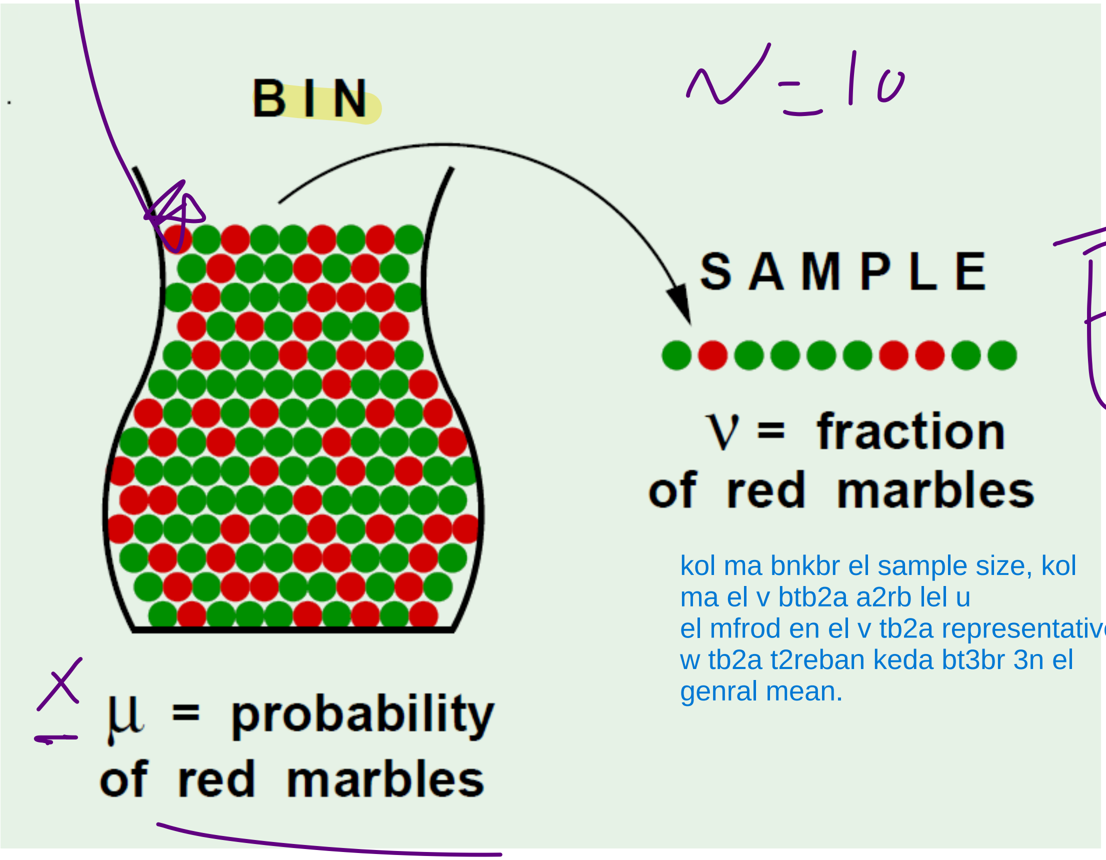
N

- Assume we have a bin of infinitely many red and green marbles.

- Assume we pick a random sample of N marbles out of the bin. with replacement.

- $\mu = \nu$??

- Can ν say anything about μ?



ν = fraction of red marbles

μ = probability of red marbles

kol ma bnkbr el sample size, kol ma el v btb2a a2rb lel u el mfrod en el v tb2a representative w tb2a t2reban keda bt3br 3n el genral mean.

example el 1k dollars lw gbt 10 mrat head forst enk t5sr hya
P(1 = T) * P (2 = T) .....

Σ

2

Hoeffding's Inequality is a statistical concept that provides a bound on the probability that the sum of a large number of random variables deviates significantly from its expected value. It's often used in the context of online learning and machine learning to quantify how much our estimates might vary.

Hoeffdeing’s Inequality

μ: Frequency of red marbles in the bin

ν: Fraction of red marbles in a random sample of size N.

3auzak tukhrug b 3 ma3lomat mn hena:
1. kol ma btkbr el 7gm bta3 el sample, kol ma e7tmalyt enk tkhrug bra el boundry bt2l exponentially.

2. kol ma btsm7 be nsbt error akbur, kol ma brdu el probability bt2l exponentially.

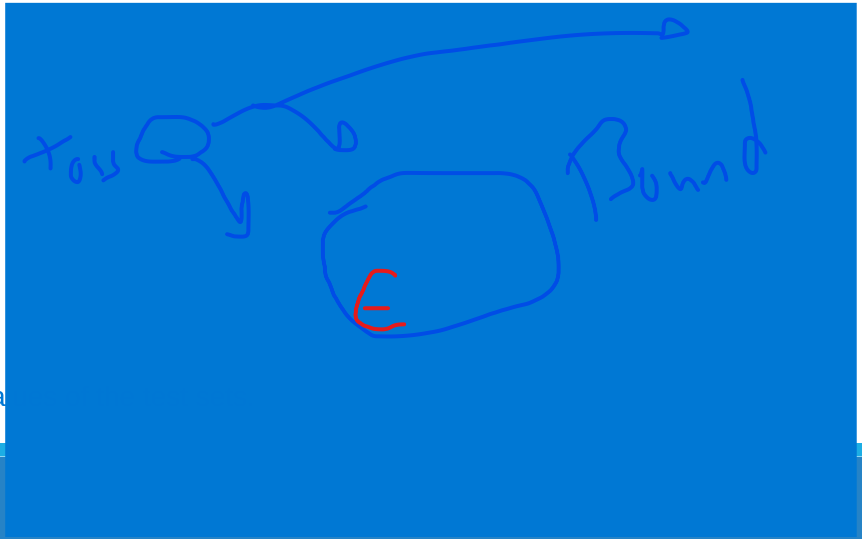
3. el equation de esmaha Hoeffdeing's inequality, w hya hadafha enha t2olak en kol lama bnakhud samples akbur w ngeb el mean bta3hom (v), nesbt en el v ykon b3ed 3n el u bnsba akbur mn e hya nesba olayela.

$$P(|\nu-\mu|>\varepsilon)\le 2e^{-2\varepsilon^2 N}\quad \text{For any } \varepsilon>0$$

e7tmaleyt enk tkhrog bara el bound.

factor in known values, fkeda helw.

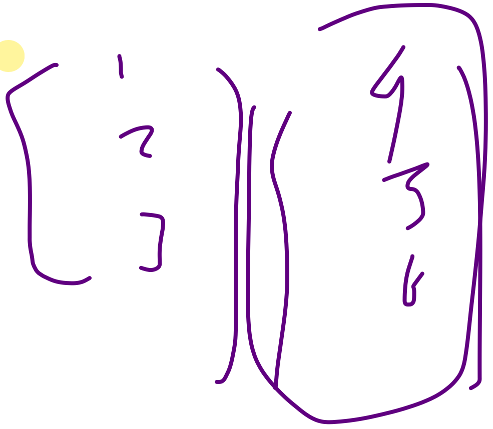
- What happens as sample size N increases?
  - ν has a better approximation of μ.
- Trade-off between N, ε, and the bound.



tb eh el hadaf y3ny? el hadaf eny a2dr a3ml prediction lel data points elly baraa el D data set -> predict the va ezay ne2dr norbot el fo2 b el learning problem bta3tna?

Analogy to the learning problem

- Bin: Input space X ✓
- Sample:(Observed) Training set D ✓
- Each marble is a point x that belongs to the input space X ✓
- Red marbles: Misclassified points by hypothesis h h(x)≠f(x)
- Green marbles: Correctly classified points by hypothesis ~~h~~ (h(x)=f(x))
- ν: Fraction of misclassified data points by hypothesis h in traning data. (Ein(h))
- ν is the in-sample error of the hypothesis h (Ein(h))
- μ: Fraction of misclassified data points by hypothesis h in the input space. (Eout(h))
- μ is the out-of-sample error of the hypothesis h (Eout(h))



In-sample versus Out-of sample error

$Ein(h) = fraction\ of\ D\ where\ f\ and\ h\ disagree$

$= \frac{1}{N} \sum_{n=1}^{N} [\![h(xn) \neq f(xn)]\!]$ ✓

Expected mean.

$Eout(h) = P[h(xn) \neq f(xn)]$ μ

True mean

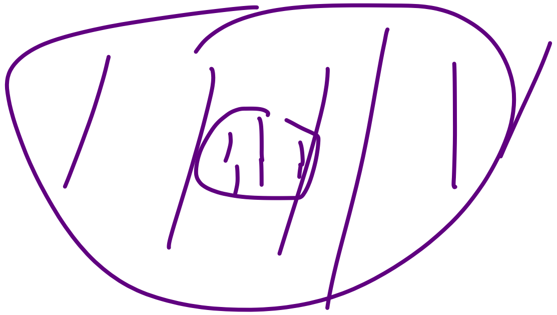
Remeber that: This equation is proved based on an assumption that -> h is fixed before you generate the dataset -> before sampling.
if h was not fixed, this equation would not hold -> eftkrha 34an hn7tagha kman 3 slides kda.

y3ny ana moftrd en msln el function elly btfr2 ben el kowar el 7amra wl khdra -> h -> khlas ana zbtha w tl3tha, fa gahz eny abd2 a3ml sampling w adkhl el data 3leha.
msh el 3ks, eny a3ml sampling, w b3den azbt el h.

Hoeffdeing’s Inequality

- Hoeffdeing’s Inequality to the learning problem:

$$P(|Ein(h)-Eout(h)|>\varepsilon)\leq 2e^{-2\varepsilon^2 N} \quad \text{For any } \varepsilon>0$$



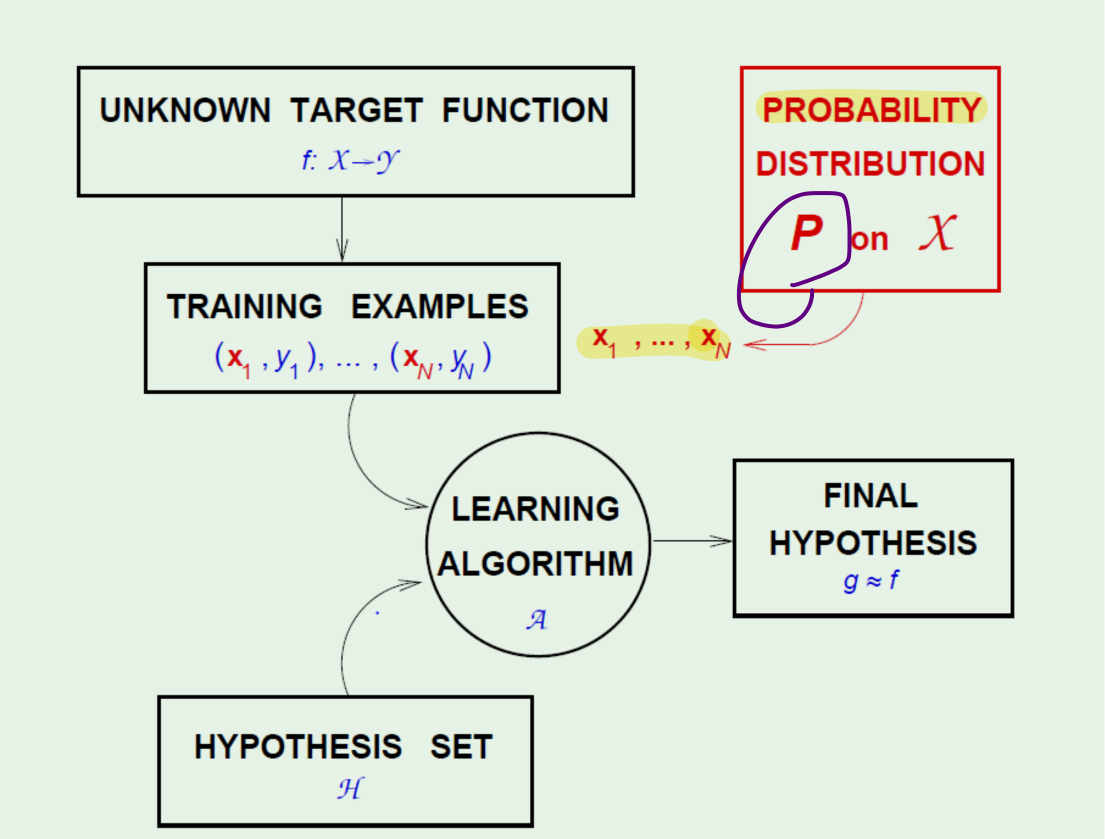
The probability that h’s performance on points not in the training set deviates from h’s performance on points in the training set is low.

Learning Setup-updated

hwa da elly b3ml beh el sampling, in other words, generating inputs in random way.

- Important Note:

A fundamental assumption made in this analysis is that the sample is taken randomly and training and testing data have the same distribution.



Source: Learning from data book

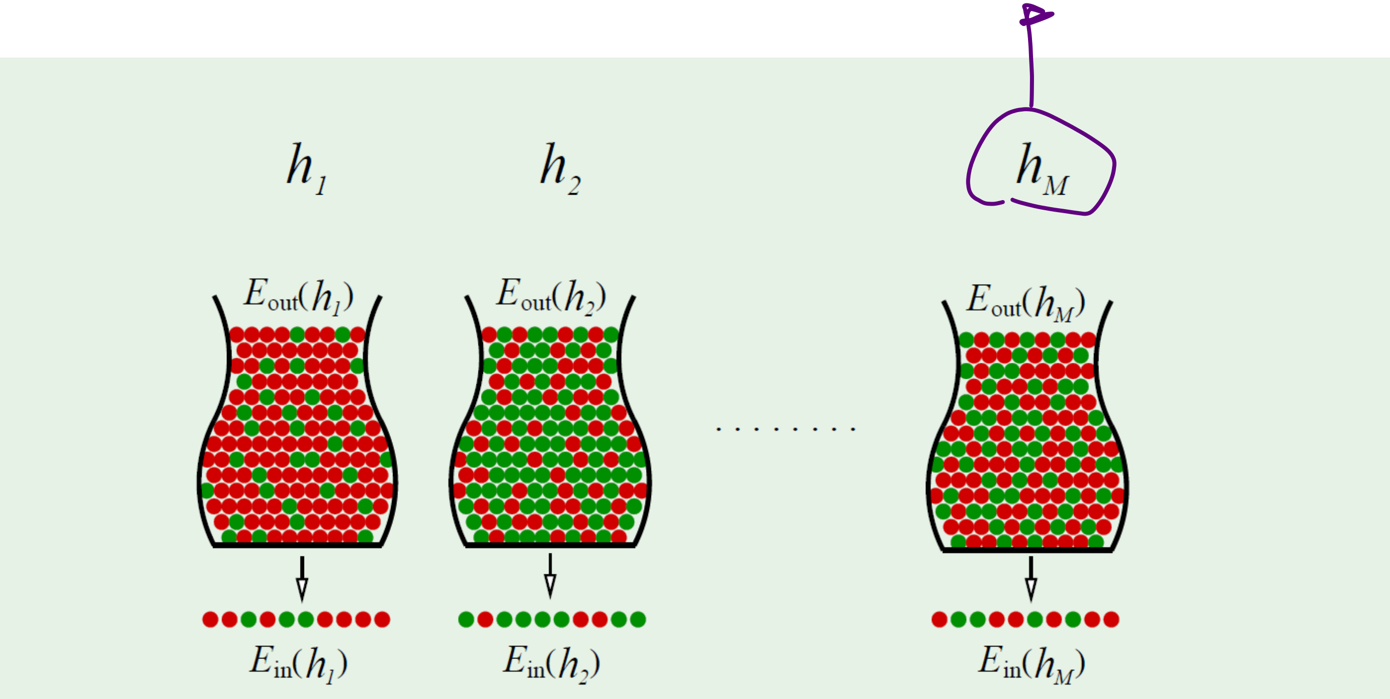
usually we don't have a control over v, however, usually we have a set of Hypotheses H, from which we select one hypothesis, which has the least error.

it sounds harder than what it is actually, fkr feha zy hagat el neuarl network el khdnaha, el set of Hypotheses de hya kol el possible values bta3t el weights, w msh e7na bno3od net3lm l7d ma nwsl le weight vector ydena a2l error? hwa da el final hyposis, kol el ablo homa another hypotheses gowa el set of H.

kol el fat, bfterad lw 3ndy hypothesis wahda.

Multiple Hypotheses ?

- In learning, we have multiple hypotheses to pick from.
- Finite Hypothesis set H={h_1,h_2,…h_M}
- H is analogous to multiple bins.
- The final hypothesis g is selected by the learning algorithm.



Hoeffdeing's Inequality for final hypothesis g

w_1

w_2

w_2

$$\underline{\mathbb{P}[}\ |E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon\ ] \ \le\ \underline{\mathbb{P}}[\quad |E_{\text{in}}(h_1) - E_{\text{out}}(h_1)| > \epsilon$$

$$\textbf{or}\ |E_{\text{in}}(h_2) - \underline{E_{\text{out}}(h_2)| > \epsilon}$$

$$\cdots$$

$$\textbf{or}\ |E_{\text{in}}(h_{\underline{M}}) - E_{\text{out}}(h_M)| > \epsilon\ ]$$

SU

$$\le \sum_{m=1}^{M} \mathbb{P}\left[\underline{|E_{\text{in}}(h_m) - E_{\text{out}}(h_m)| > \epsilon}\right]$$

2^{10}

H 10K

T - 1RS

P(1T) P(2T) P(3T) ... = P(10T)

very upper bound, because we have neglected all the intersections, and assumed that all of the hypothesis, are independent, and that is not true, however, this equation still holds, but it is a very upper bound.

Hoeffdeing’s Inequality for final hypothesis g

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

$$\leq 2e^{-2\epsilon^2 N}$$

$$\leq \sum_{m=1}^{M} \underline{2e^{-2\epsilon^2 N}}$$

constant.

$$\sum_{n=1}^{n} 1$$

$$\underline{M}$$

$$\mathbb{P}[|E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon] \leq \underline{2Me^{-2\epsilon^2 N}}$$

As M (<u>the hypothesis set</u> size) increases, the <u>bound gets looser</u>.

kol ma 3dd el hypotheses zad, kol ma ana kbrt el bound, we de haga msh kwysa, l2n ana hdfy awsl le bound soghyr.

Feasibility of Learning

Two main questions:

1. **Can we make sure that Eout(g) is close enough to Ein(g)?**
   Hoeffdeing's Inequality.
2. **Can we make Ein(g) small?**

   Learning models