Lecture-9

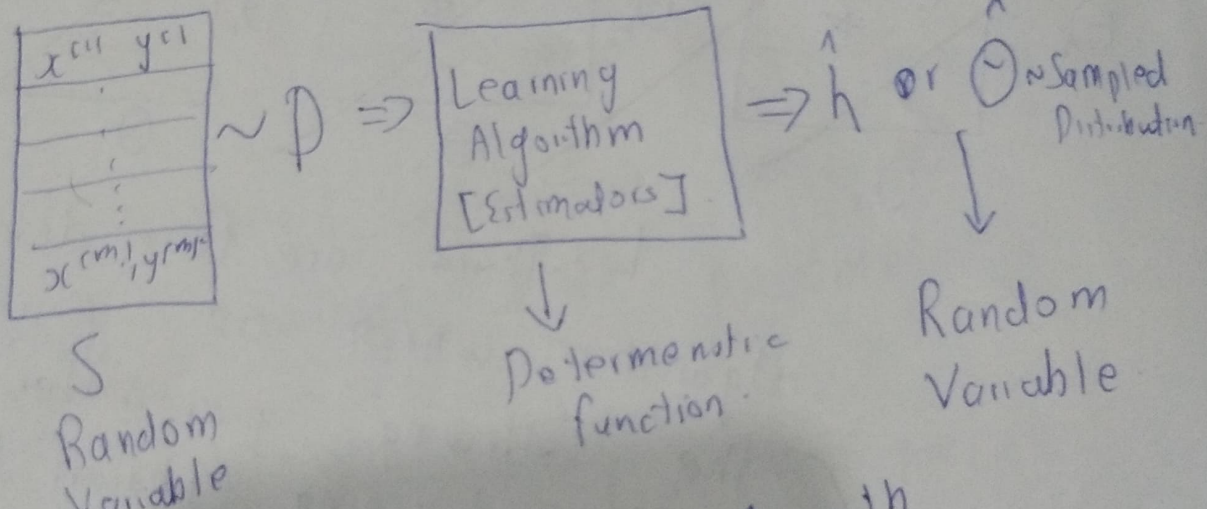
(*) Assumptions:-

(1) Data Distribution : D

$(x, y) \sim D$

↳ There is a data generating distribution

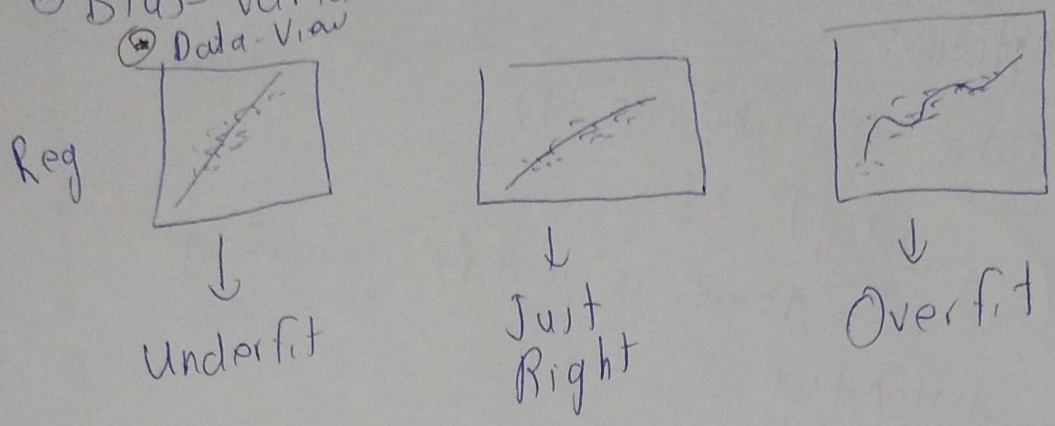
(2) All the samples are sampled independently



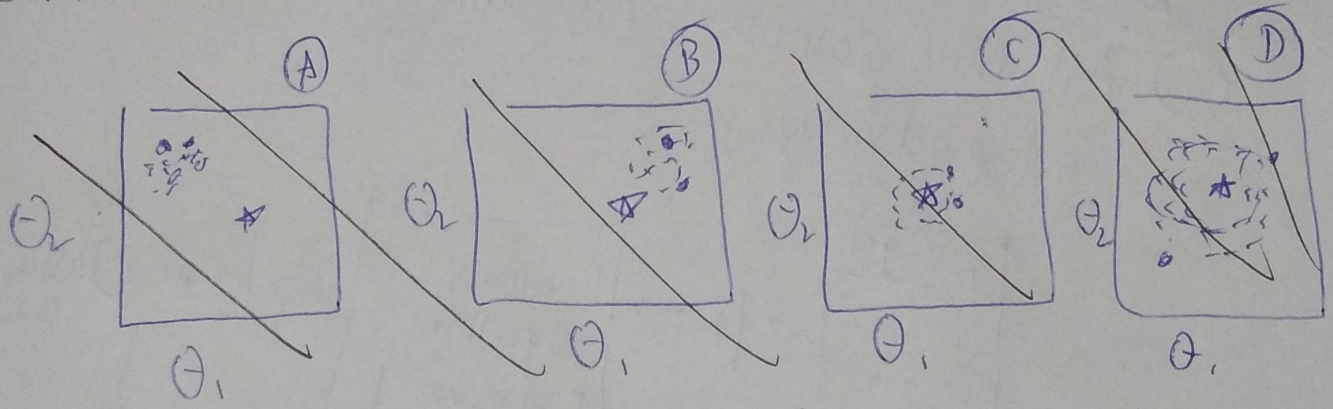
- There is a distribution associated with parameters
  - ↳ There exist some θ^* or h^* which in a sense are true parameter and wish to be the parameters
  - θ^* or h^* → Not random (It is a unknown constant).

⊛ Bias-Variance

⊛ Data-View



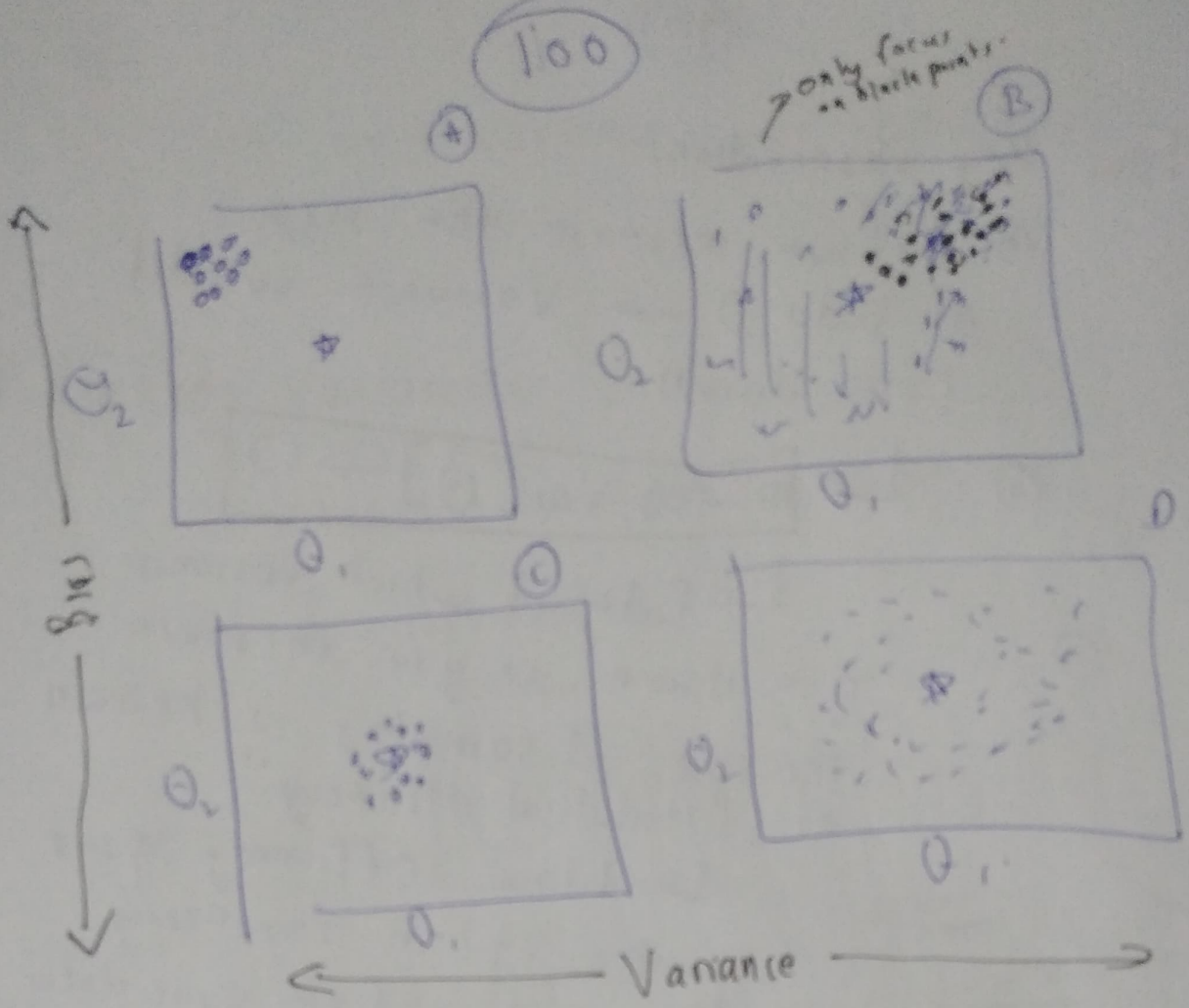
⊛ Parameter View of Bias-Variance.



↳ There are four different algorithm which are generating the following parameter value after training (shown by Blue dots).

↳ Blue dots are samples from each hypothesis distribution.

→ Refer to diagram on next page.



(Y)

- Bias → Whether Parameter are centered around true parameters (First moment)
- Variance → How dispersed is the sampling distribution of parameter from hypothesis sample. (Second moment).

Few Key observations:-

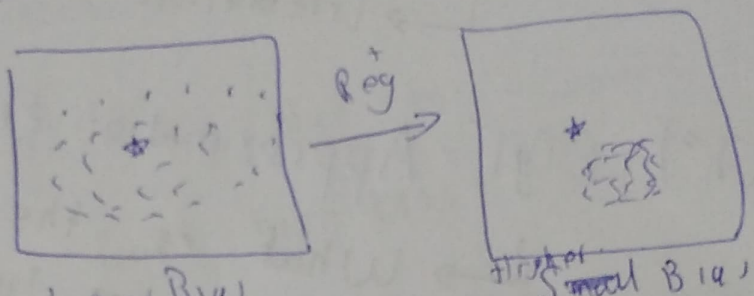
- As we increase the size of Data D → Variance would come down.
- ↳ $m \to \infty, Var[\hat{\theta}] \to 0$
  - ↳ Rate at which variance goes down as you increase the size of samples is known as "statistical efficency".
    - ↳ How efficency your algorithm is in squeezing out information from your data.
- $\hat{\theta} \to \theta^*, m \to \infty$:- You call such algorithm consistent.
- $E[\hat{\theta}] = \theta^* \to$ Unbiased Estimator.
- High Bias won't come down no matter how much data you provide to your algorithm.

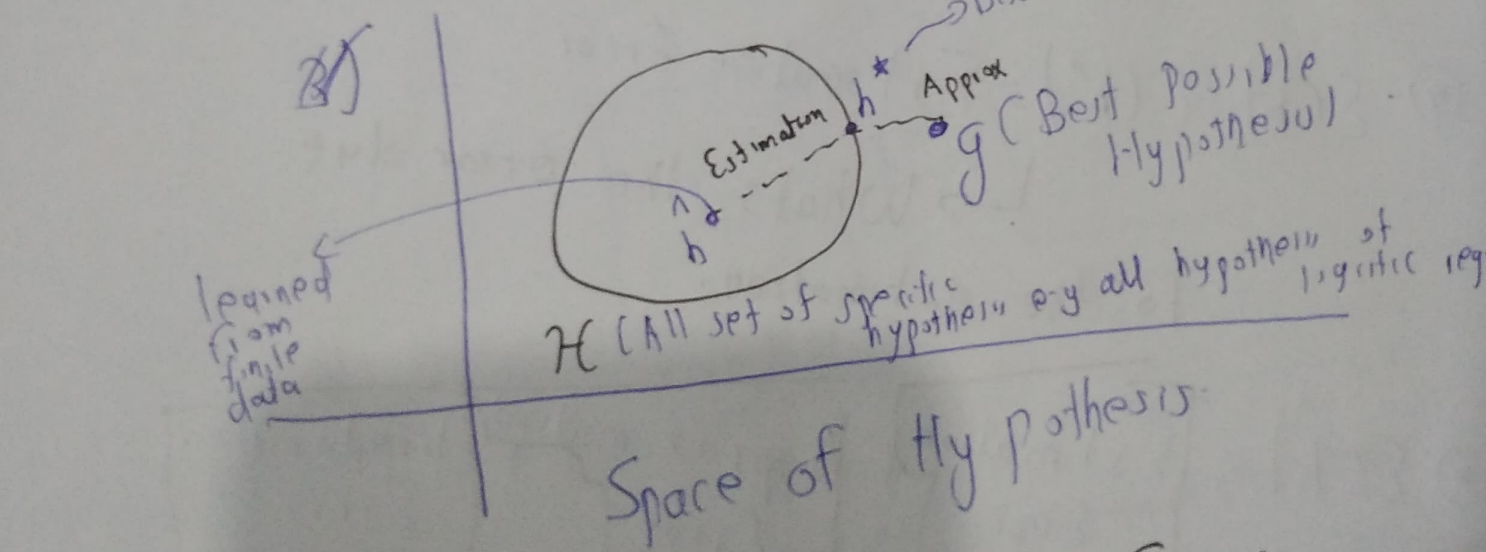
* Bias and Variance are independent of each other.

* Fighting Variance:-

1) Increase amount of Data.

2) Regularization:





* $E(h)$ — Risk / Generalization Error

↳ $= E_{(x,y)\sim D}[2\{h(x) \neq y\}]$ (Fraction of all your mistakes).

- $\hat{\varepsilon}_S(h)$: Empirical Risk

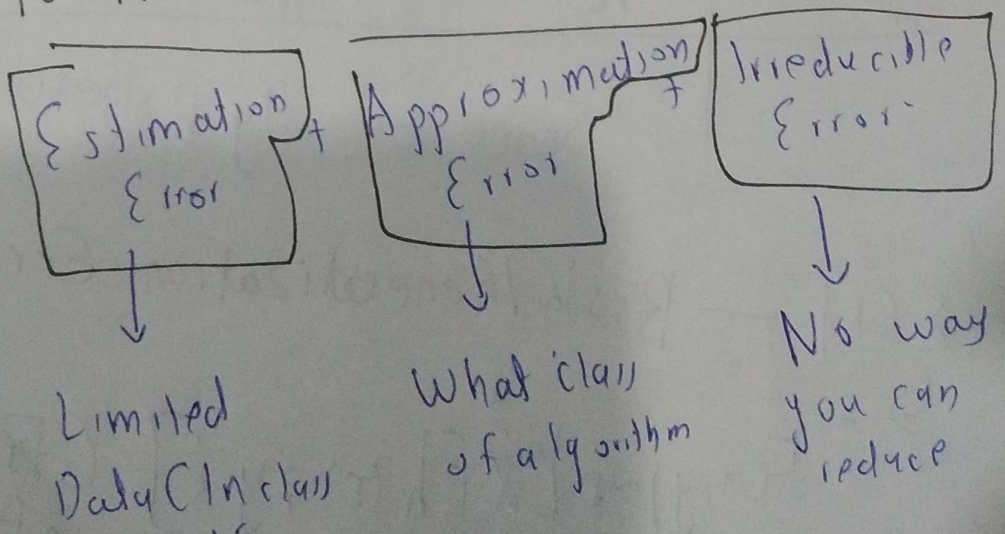
$$\frac{1}{m}\sum_{i=1}^{m} 1\{h(x^{(i)}) \neq y^{(i)}\}.$$

* $\varepsilon(g)$ = Bayes Error (Best optimal Possible hypothesis error).
  - ↳ Irreducible Error.

* $\varepsilon(h^*) - \varepsilon(g)$ = Approximation Error
  - ↳ What is the price we are paying if we are limiting ourselves to one specific class of algorithm.

* $\varepsilon(\hat{h}) - \varepsilon(h^*)$ = Estimation Error.
  - ↳ Whats the error due to the estimation.

* $\varepsilon(\hat{h})$ =



$E[\hat{h}] =$ Estimation Error + Approximation Error + Irreducible Error

Estimation Variance → Variance + Estimation bias → Bias + Irreducible.

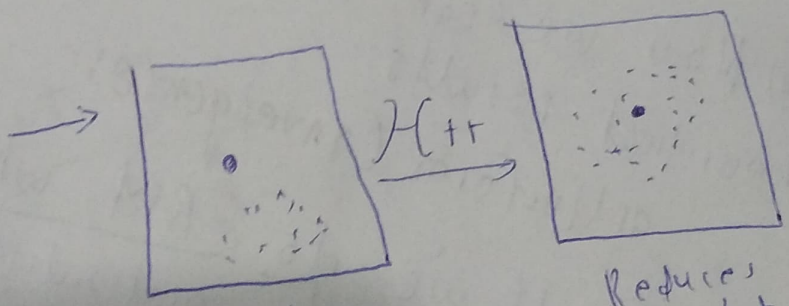
→ Bias" why $\hat{h}$ is far g :- either due to class of algorithm or some other reasons.

→ Variance is generally due to estimation error.

→ How do you fight high bias?

→ Make $\mathcal{H}$ bigger

↳ More complex algorithms.



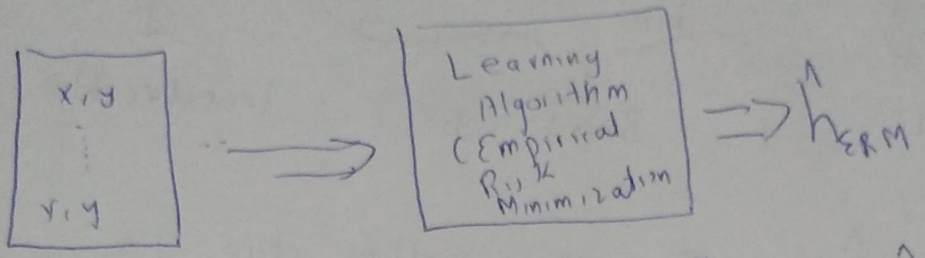
high bias
Some variance

Reduces Bias but also increase your variance.

→ Regularization ~~decreases~~ Increase decrease would effectively make $\mathcal{H}$ larger which will reduce the bias.

Empirical Risk Minimization:- (Minimizer)

* Is a learning algorithm



$$\hat{h}_{ERM} = \underset{h \in \mathcal{H}}{\arg\min} \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}\{h(x^{(i)}) \neq y\}.$$

↳ ERM is an hypothesis that tries to find a hypothesis from class of hypothesis that minimizes the training error. (Increasing the training accuracy).

* Now we can come up with more theoretical results:-
  a) Uniform Convergence:-
  1) If we do ERM what does it say about generalization error:-
     * $\hat{\varepsilon}(h)$ vs $\varepsilon(h)$
  2) How does the generalization error of our learned hypothesis compare to the best possible generalization error in that class:-
     * $\varepsilon(\hat{h})$ vs $\varepsilon(h^*)$.

For above two questions we are going to use two tools:-

(1) Union Bound:-

$A_1, \ldots, A_k$ (Needs not be Independent)

~~$P(A_1 \cup A_2 \cup A_3 \cup A_k) \leq P(A_1) + P($~~

$P(A_1 \cup A_2 \cdots \cup A_k) \leq P(A_1) + P(A_2) + \cdots$

$P(A_k)$.

(2) Hoeffding Inequality

Let $Z_1, Z_2 \ldots Z_m \sim$ Bernoulli[ϕ]

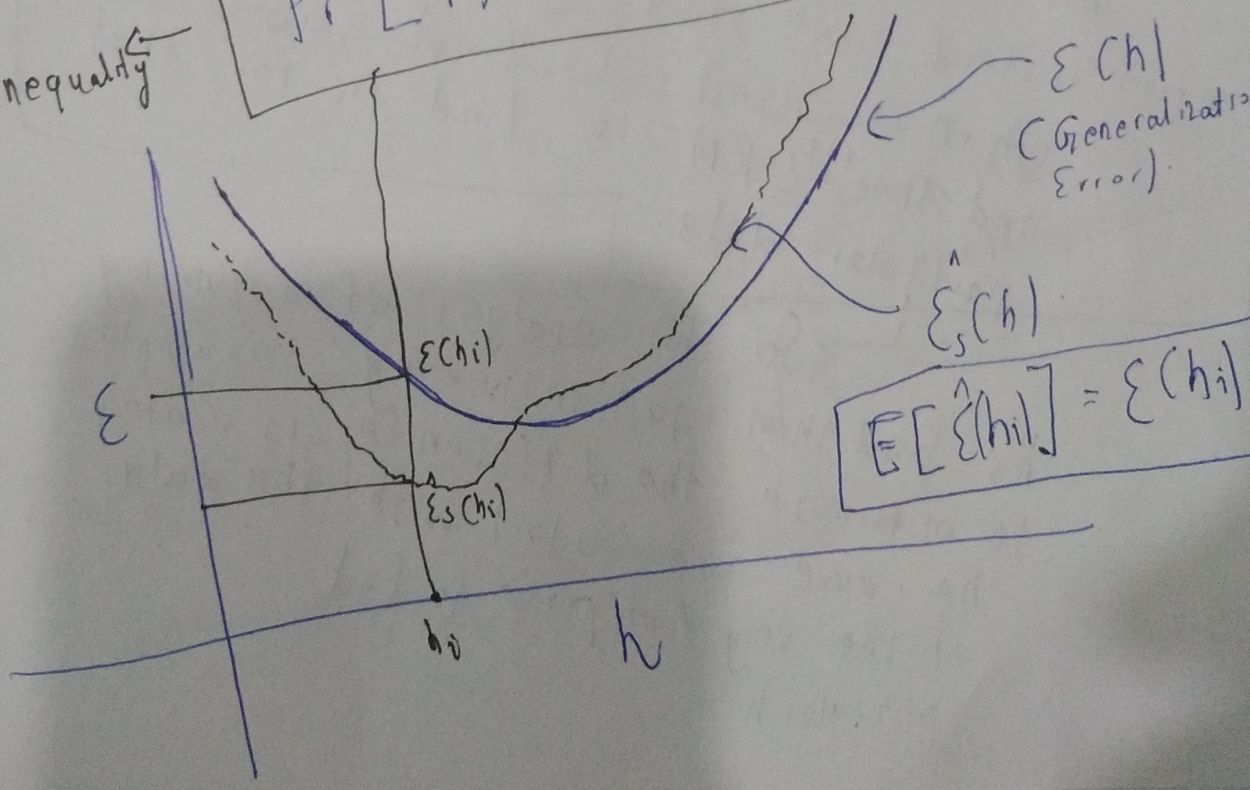
Sample of 0,1

$\hat{\phi} = \frac{1}{m} \sum_{i=1}^{m} Z_i$ (Estimator)

Let $\gamma > 0$ [margin].

$$P_r[|\hat{\phi} - \phi| > \gamma] \leq 2 \exp(-2\gamma^2 m$$

Inequality



$$E[\hat{\varepsilon}(h_i)] = \varepsilon(h_i)$$

⊛ Now Applying Hoeffdings inequality:-

⊛ $Pr[|\hat{\varepsilon}_s(h_i) - \varepsilon(h_i)| > \gamma] \leq 2\exp(-2\gamma^2 m)$

↳ the gap between generalization error (Error of hypothesis h_i on all data) and empirical error (Error of hypothesis h_i on subsample of data) is bounded by this inequality (expression).

↳ As m increases empirical and generalization are going to come near to each other i.e their gap would reduce.

⊛ However, the problem is we don't start with some hypothesis h_i and then average across, start and then but in practice, instead we start with some data ~~which is used to find some h_i~~ and then ERM to find h_i for that particular data.

↳ So this analysis of assuming h_i and then just increasing examples to minimize the difference is useless because we start with the data at the very first place to find

↳ To fix the above problem, we would generalize the above inequality to ~~all classes of hypothesis~~ to account to for all $\mathcal{H}$.

↳ We will look at the difference of the whole class $\mathcal{H}$ rather than particular h_i of $\mathcal{H}$.

↳ This is called uniform convergence because we want to see how empirical risk curve converge uniformly across all class to generalization risk curve.

→ Now the results would look different for two cases for all $\mathcal{H}$ if we are to show the same bound that we show for particular h_i.

① Finite $\mathcal{H}$ hypothesis classes:-

⊙ $|\mathcal{H}| = k$

① $P_r[\exists h \in \mathcal{H} \mid |\hat{\varepsilon}_s(h) - \varepsilon(h)| > \gamma] \leq k \cdot 2\exp(-2\gamma^2 m)$ but done through two tails

Many steps are missing

$P_r[\forall h \in \mathcal{H} \mid |\hat{\varepsilon}_s(h) - \varepsilon(h)| < \gamma] \geq 1 - \frac{2k\exp(-2\gamma^2 m)}{\delta}$

② $\Rightarrow P_r[\forall h \in \mathcal{H} \mid$

~~Let $\gamma = 2k\exp(-2\gamma^2 m)$~~

Let $\delta = 2k \exp(-2\gamma^2 m)$

Relation →

δ — Prob of error

γ → Margin of error

m → Sample size.

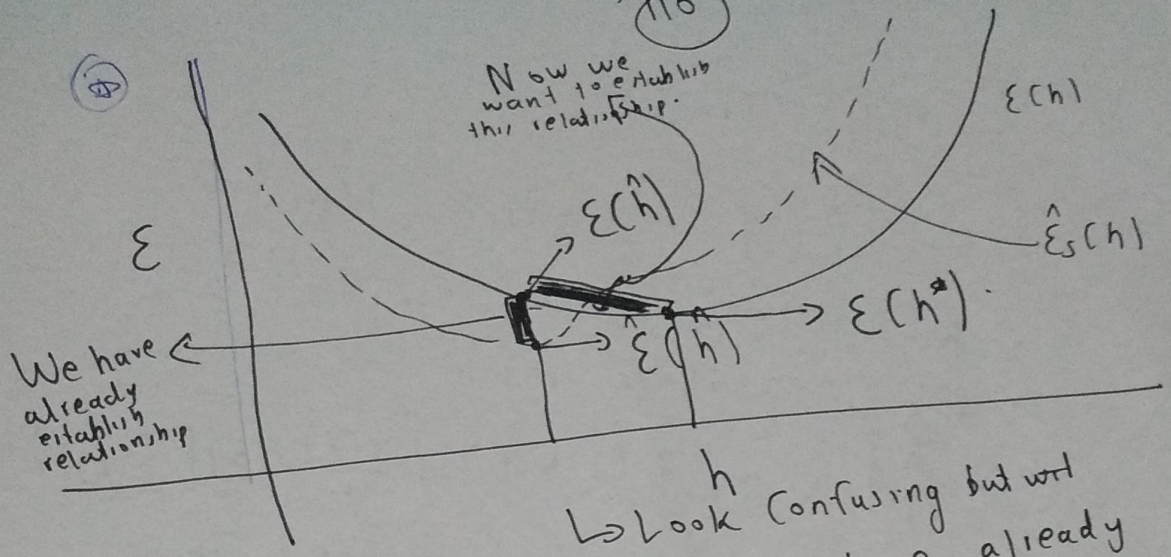
→ What can we do with this relationship is that we can fix any two and solve for the third.

Example: Fix $\gamma, \delta > 0$

◎ $m \geq \frac{1}{2\gamma^2} \log \frac{2k}{\delta}$ (Sample complexity desired).

◎ This shows that the Probability (for all classes of hypothesis) that difference between ~~true~~ empirical risk and generalization risk is less than γ is bounded by $1-\delta$ as long as number of traing examples are greater than expression shown

◎ As you increase m empirical risk and generalization risk becomes closer for $\mathcal{H}$.



↳ Look Confusing but not

basically we have already established relationship between Generalization and empirical error. Now we want to establish relationship between generalization error of a hypothesis ↔ generalization error of best possible hypothesis from $\mathcal{H}$.

of specific

$$\varepsilon(\hat{h}) \leq \hat{\varepsilon}(\hat{h}) + \gamma \quad (\text{Already know}).$$
$$\leq \hat{\varepsilon}(h^*) + \gamma \quad (???).$$
$$\leq \varepsilon(h^*) + 2\gamma. \quad (\text{Relationship established.})$$

$\Rightarrow$ With Probability $1-\delta$, training size m.

$$\varepsilon(\hat{h}) \leq \varepsilon(h^*) + 2\sqrt{\frac{1}{2m}\log\frac{2k}{\delta}}$$

↓

Can ge
this f
previou
& ex

(*) VC-Dimension:-

VC(H) = Some Number which tell how expressive the hypothesis is.

$$\varepsilon(\hat{h}) \leq \varepsilon(h^*) + O\sqrt{\frac{VC(H)}{m}\log\left[\frac{m}{VC(H)} + \frac{1}{m}\right]}$$

↳ Margin for infinite classes h in H.