ERROR DECOMPOSITION

Classification Task

(x, y), $x \in \mathbb{R}^d$, $y \in \{1, 2, \dots K\}$

K-way classification.

x: feature vector in d-dim

y: Class label ground truth

$p(x, y)$ $\boxed{\text{Unknown}}$ joint distribution of i/p x & output y.

$\downarrow$

D sampled from $p(x,y)$ || $D = \{D_{train}, D_{test}\} \sim p(x, y)$

D_{train}: N sample training data $\{(x_i, y_i)\}_{i=1}^{N}$

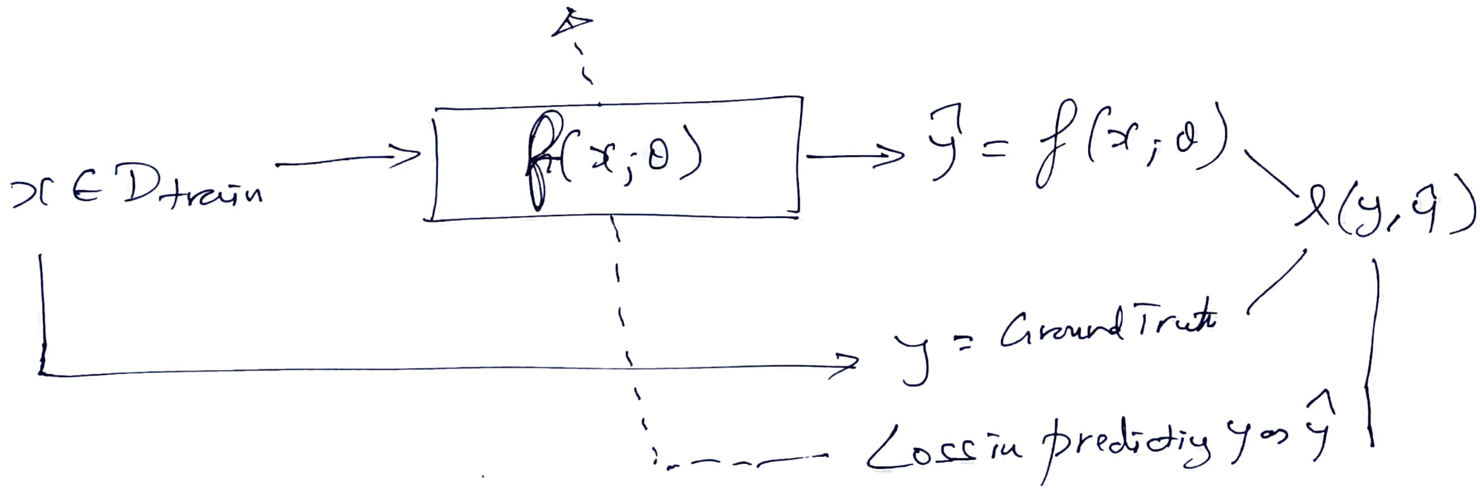
D_{test}: $\{x^{test}\}$

$\exists$ optimal hypothesis/model $f^{\dagger}$: $f^{\dagger}(x) \Rightarrow y$

But within a hypothesis/model space $\mathcal{F}$, ML/DL Learning learns to discover $f^{\dagger}$ by fitting D_{train} & testing on D_{test}.

as

$$f(\cdot\,;\theta) : f(x;\theta) = \hat{y}$$: Predicted Label of x.



Expected Loss/Risk

Given $p(x,y)$

$$R(f) = \int_{x,y} \ell(f(x),y)\, dp(x,y)$$

$$= \mathop{\mathbb{E}}_{p(x,y)} \left[\ell(f(x),y)\right]$$

Expected Loss/Risk Minimizer

$$f^{\dagger} = \operatorname*{argmin}_{f} R(f)$$

Within $\mathcal{F}$

$$f^{*}_{\mathcal{F}} = \operatorname*{argmin}_{f \in \mathcal{F}} R(f)$$

Empirical Loss/Risk

(3)

Given $D_{train} = \{(x_i, y_i)\}_{i=1}^{N}$

$$R_N(f) = \frac{1}{N}\sum_{i=1}^{N} \ell(f(x_i), y_i)$$

Empirical Loss/Risk Minimizer

$$f_N = \operatorname*{argmin}_{f \in \mathcal{F}} R_N(f)$$

To know how well f_N performs on test:

Measure $\underline{\underline{R(f_N)}}$

$p(x, y)$

D_{train} $\quad$ D_{test}

$\{x_i, y_i\}_{i=1}^{N}$

Expected Loss/Risk

$$R(f) = \mathbb{E}_{p(x,y)} \left[l(f(x), y) \right]$$

Empirical Loss/Risk

$$R_N(f) = \frac{1}{N} \sum_{i=1}^{N} l(f(x_i), y_i)$$

optimal

$$\boxed{f^{\dagger} = \underset{f}{\operatorname{argmin}} \, R(f)}$$

- Minimize Expected Risk
- NO CONSTRAINTS on f

Within F

$$\boxed{f^{*}_{F} = \underset{f \in F}{\operatorname{argmin}} \, R(f)}$$

- Minimize Expected Risk.
- $f \in F$
- $p(x, y)$! UNKNOWN.

Within F, Using D_{train}

$$\boxed{f_N = \underset{f \in F}{\operatorname{argmin}} \, R_N(f)}$$

- Minimize empirical risk on $D_{train} \sim p(x,y)$ of N samples
- $f \in F$

$$f^{\dagger} = \underset{f}{\operatorname{argmin}}\ R(f)$$

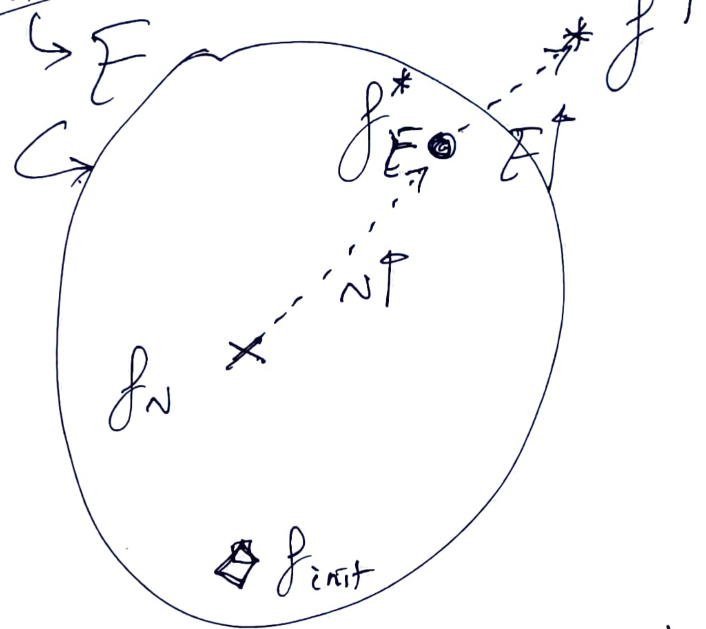
$$f^{*}_{\mathcal{E}} = \underset{f \in \mathcal{F}}{\operatorname{argmin}}\ R(f)$$

$$f_N = \underset{f \in \mathcal{E}}{\operatorname{argmin}}\ R_N(f)$$

where $R(f) = \mathbb{E}_{p(x,y)}\left[\ell(f(x), y)\right]$

& $R_N(f) = \frac{1}{N} \sum_{i=1}^{N} \ell\left(f(x_i), y_i\right)$

Hypothesis / Model space



As $\mathcal{E}\uparrow$: $f^{*}_{\mathcal{E}} \longrightarrow f^{\dagger}$

As $N\uparrow$: $f_N \longrightarrow f^{*}_{\mathcal{E}}$

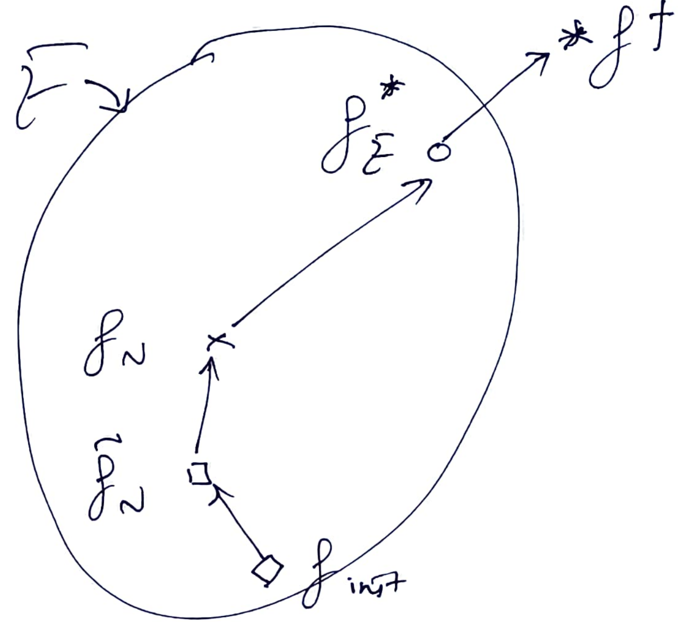
Including "optimization" Process & Corresponding 'Error'

f_N is determined from same f_{init} by an optimization process to yield a sub-optimal $\tilde{f}_N$

i.e.

$$\tilde{f}_N = \text{opt}\left[\underset{f \in \mathcal{F}}{\arg\min}\; R_N(f)\right]$$

$$f^{\dagger} \xrightarrow[f \in \mathcal{F}]{} f^{*}_{\mathcal{F}} \xrightarrow[D_{train}]{} f_N \xrightarrow[opt]{} \tilde{f}_N$$



Generalization on Test: $p(x,y)$

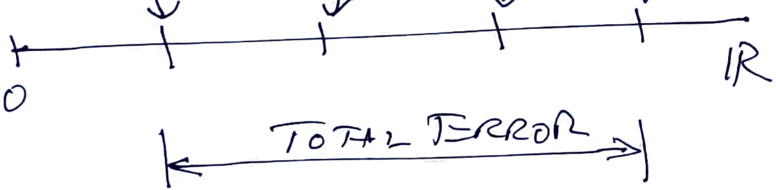
$$R(f^{\dagger}) \le R(f^{*}_{\mathcal{F}}) \le R(f_N) \le R(\tilde{f}_N)$$

$$\boxed{\text{Note: Models evaluated on } R(\cdot)}$$

$$f^{\dagger} \longrightarrow f^{*}_{F} \longrightarrow f_N \longrightarrow \tilde{f}_N$$

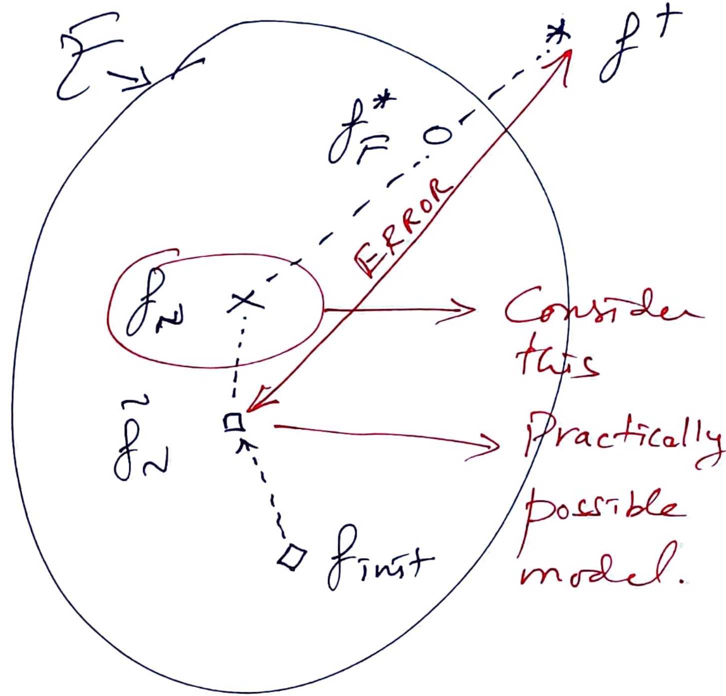
$R(\cdot)$

$$R(f^{\dagger}) < R(f^{*}_{F}) < R(f_N) < R(\tilde{f}_N)$$

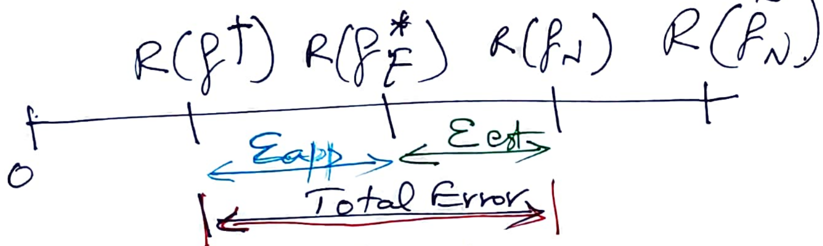


$$= \mathbb{E}\left[R(\tilde{f}_N) - R(f^{\dagger})\right]$$

on samplings of D_{train}.



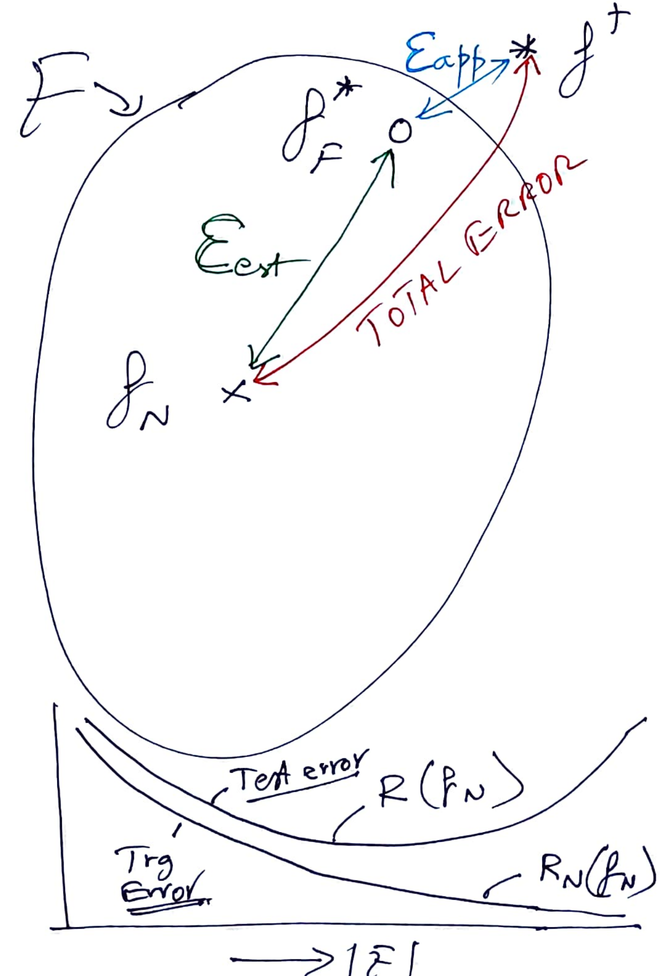
Consider only f_N (& not $\tilde{f}_N$)



Generalization Error

$$= \mathbb{E}\left[R(f_N) - R(f^\dagger)\right]$$

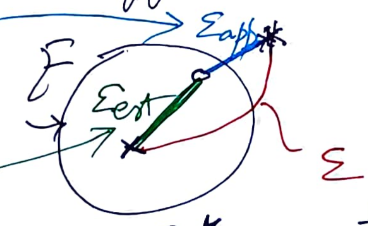
$$= \mathbb{E}\left[R(f_N) - R(f^*_F)\right] + \mathbb{E}\left[R(f^*_F) - R(f^\dagger)\right]$$



$$\text{Error } \mathcal{E} = \mathbb{E}\left[R(f_N) - R(f^*_{\mathcal{F}})\right] \rightsquigarrow \mathcal{E}_{estimation}$$

$$+ \ \mathbb{E}\left[R(f^*_{\mathcal{F}}) - R(f^+)\right] \rightsquigarrow \mathcal{E}_{approximation}$$

$$= \mathcal{E}_{estimation} + \mathcal{E}_{approximation}$$



As $|\mathcal{F}| \to \infty$ or $|\mathcal{F}| \uparrow$ $\mathcal{E}_{app} \longrightarrow 0$ ie $f^*_{\mathcal{F}} \to f^+$

For a given $\mathcal{F}$,

As $N \to \infty$ or $N \uparrow$, $\mathcal{E}_{est} \longrightarrow 0$ ie $f_N \longrightarrow f^*_{\mathcal{F}}$

then

Error (GE) $= \mathbb{E}\left[R(f_N) - R(f_{\mathcal{F}}^*)\right] + \epsilon_{\mathcal{F}}$

where $\epsilon_{\mathcal{F}} = \mathcal{E}_{app}$ [a function of $\mathcal{F}$ only]

or

$$R(f_N) - R(f_{\mathcal{F}}^*) = GE - \mathcal{E}_{\mathcal{F}}$$

For finite $\mathcal{F}$, $\mathcal{E}_{app}$ (or $\epsilon_{\mathcal{F}}$) > 0

then

$$R(f_N) - R(f_{\mathcal{F}}^*) < GE$$

or

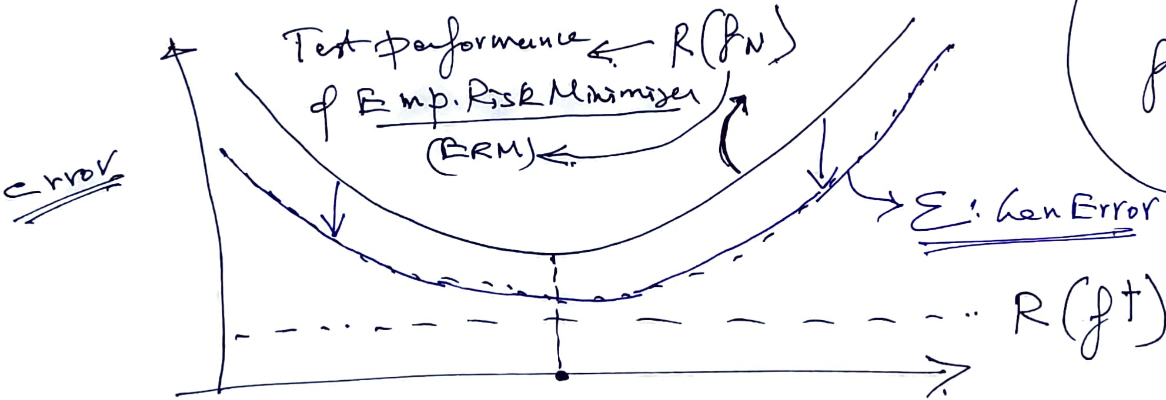
$$\boxed{R(f_N) < R(f_{\mathcal{F}}^*) + GE}$$

Error

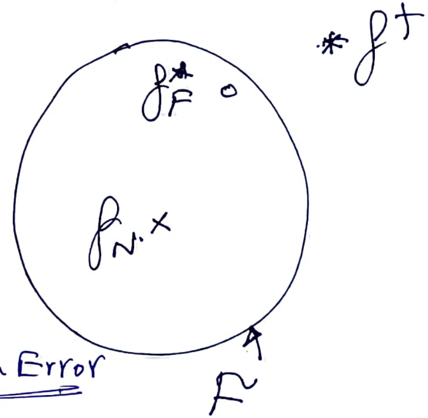
$$\mathcal{E} = \mathcal{E}_{est} + \mathcal{E}_{app}$$

$$\mathcal{E}_{est} = \mathbb{E}\left[R(f_N) - R(f_F^*)\right] \qquad \mathcal{E}_{app} = \mathbb{E}\left[R(f_F^*) - R(f^\dagger)\right]$$

$$\mathcal{E} = \mathbb{E}\left[R(f_N) - R(f^\dagger)\right]$$



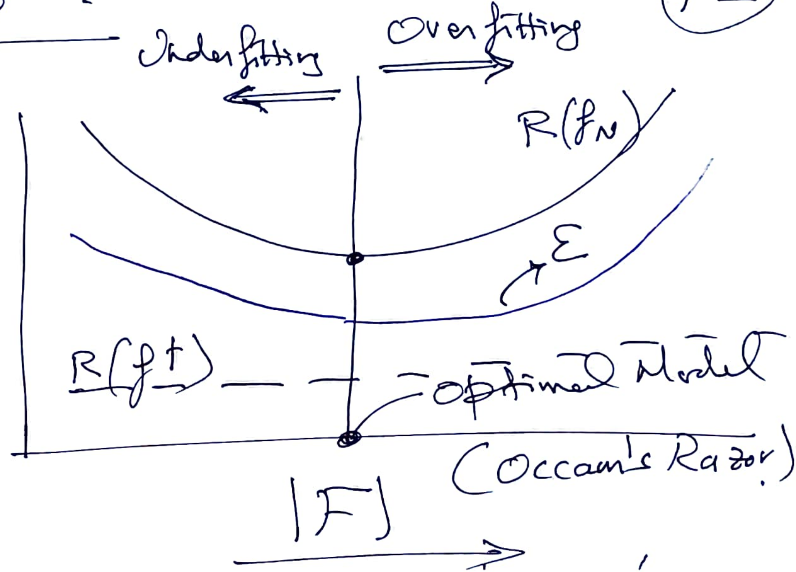
Model Complexity $|F|$

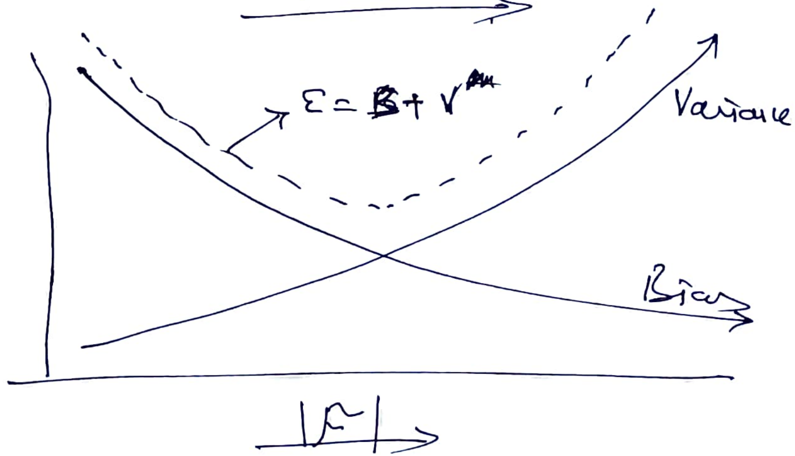


9.2

Why does $R(f_N)$ behave so ?

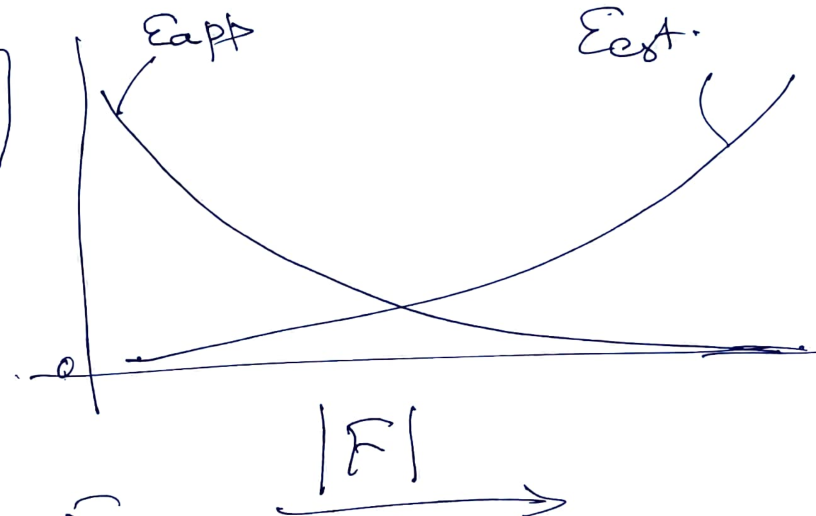
- Model Complexity is ↓ (Low)
  - No learning of complex patterns in data
  - High Bias, Low Variance.
- Model Complexity is ↑ (High)
  - Learns complex patterns
  - Memorizes Trg Data.
  - Performs poorly on Unseen data
  - Low Bias, High Variance.





• $\mathcal{E}_{est} = \mathbb{E}\left[R(f_N) - R(f_F^{*})\right]$

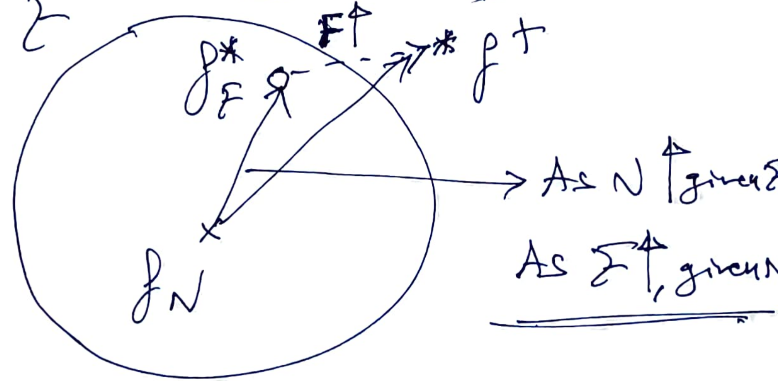
• $\mathcal{E}_{app} = \mathbb{E}\left[R(f_F^{*}) - R(f^{\dagger})\right]$



$\mathcal{E}_{app}$

as $|\mathcal{F}| \uparrow \rightarrow \left[f_F^{*} \longrightarrow f^{\dagger}\right]$

$\Rightarrow \mathcal{E}_{app} \longrightarrow 0$



$\mathcal{E}_{est}$ $\mathbb{E}\left[R(f_N) - R(f_F^{*})\right]$

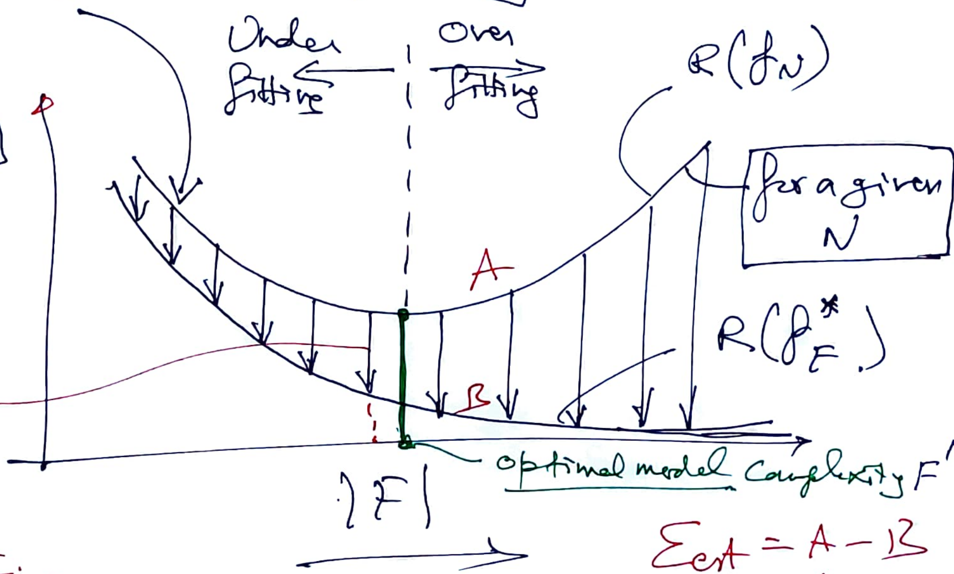
as $|\mathcal{F}| \uparrow \rightarrow R(f_F^{*}) \rightarrow R(f^{\dagger})$

~~faster than $R(f_N) \rightarrow R($~~

Analyse $\mathcal{E}_{est} = \mathbb{E}\left[R(f_N) - R(f^*_F)\right]$ (9.4)

- A decreases faster than B till $F = F'$ [optimal model]
- $A\uparrow$ for $F > F'$

R(·) error



$\mathcal{E}_{est}(F)$

$= A - B$

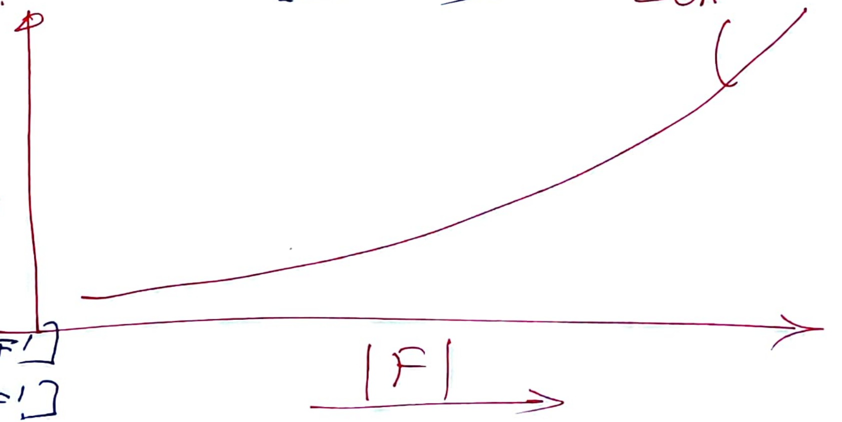
$\mathcal{E}_{est} = A - B$

- Given N
  - $F\uparrow \rightarrow \left[f^*_F \rightarrow f^+\right]$ R(·) error

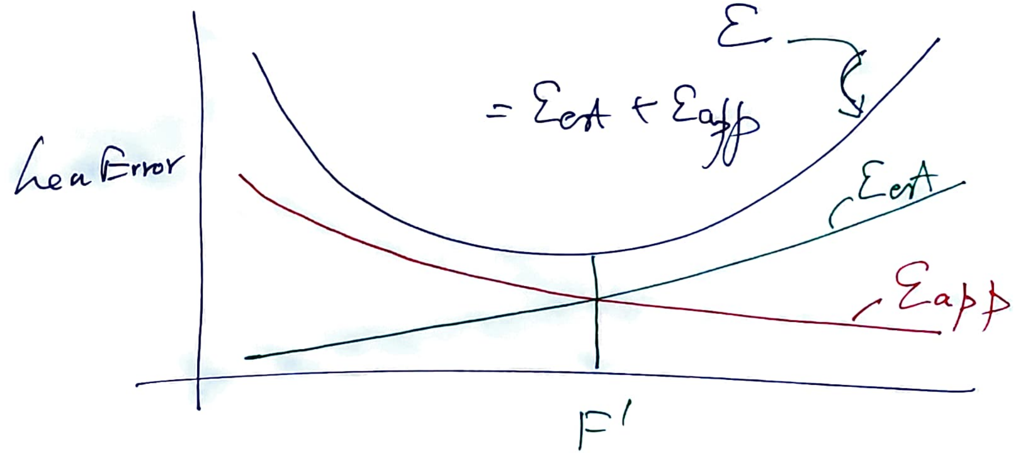
    unaffected by N.
- $F\uparrow \rightarrow f_N$ affected / fixed

  $\rightarrow R(f_N) \rightarrow$ UF $[F < F']$

  $\rightarrow$ OV $[F > F']$



Overall Gen Error $= \mathcal{E}_{est} + \mathcal{E}_{app}$



In Summary

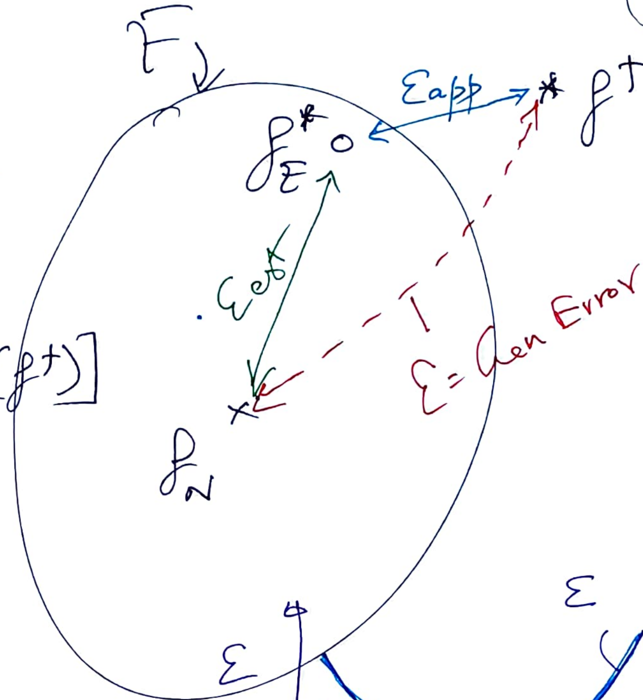
For a given $|D_{train}| = N$

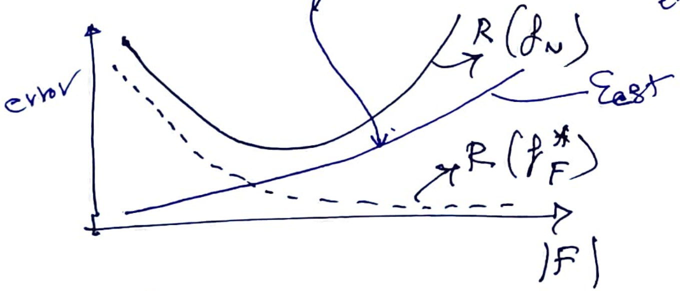
$\mathcal{E}$ = Generalization Error

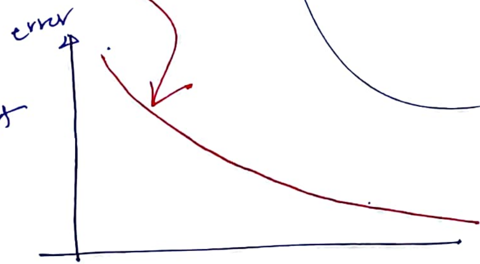
$$= \mathbb{E}\left[R(f_N) - R(f^{\dagger})\right]$$

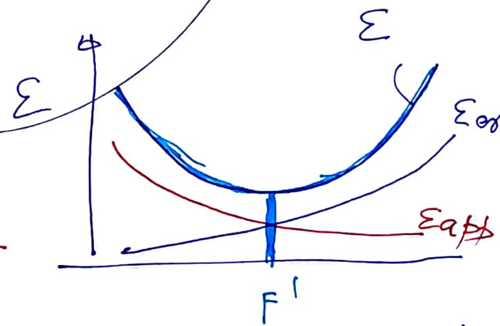
$$= \mathbb{E}\left[R(f_N) - R(f^*_{\mathcal{F}})\right] + \mathbb{E}\left[R(f^*_{\mathcal{F}}) - R(f^{\dagger})\right]$$

$$\boxed{\mathcal{E} = \mathcal{E}_{est}(\mathcal{F}, N) + \mathcal{E}_{app}(\mathcal{F})}$$









For finite $\mathcal{F}$, $\mathcal{E}_{app}(\mathcal{F}) > 0$

$\therefore$ G.E. $\mathcal{E} > R(f_N) - R(f^*_{\mathcal{F}})$

$$\boxed{R(f_N) < R(f^*_{\mathcal{F}}) + \underbrace{\mathcal{E}}_{\text{Generalization Error}}}$$