Error bar :

$$E_{out}(g) \leq E_{in}(g) + \underbrace{\sqrt{\frac{8}{N}\log\left(\frac{4\left((2N)^{d_{VC}}+1\right)}{\delta}\right)}}_{\text{error bar}}$$

Example : $N = 100$, $\delta = 0.1$ (90% confidence)

$d_{VC} = 1$ (Too simple model)

$$E_{out}(g) \leq E_{in}(g) + 0.848$$

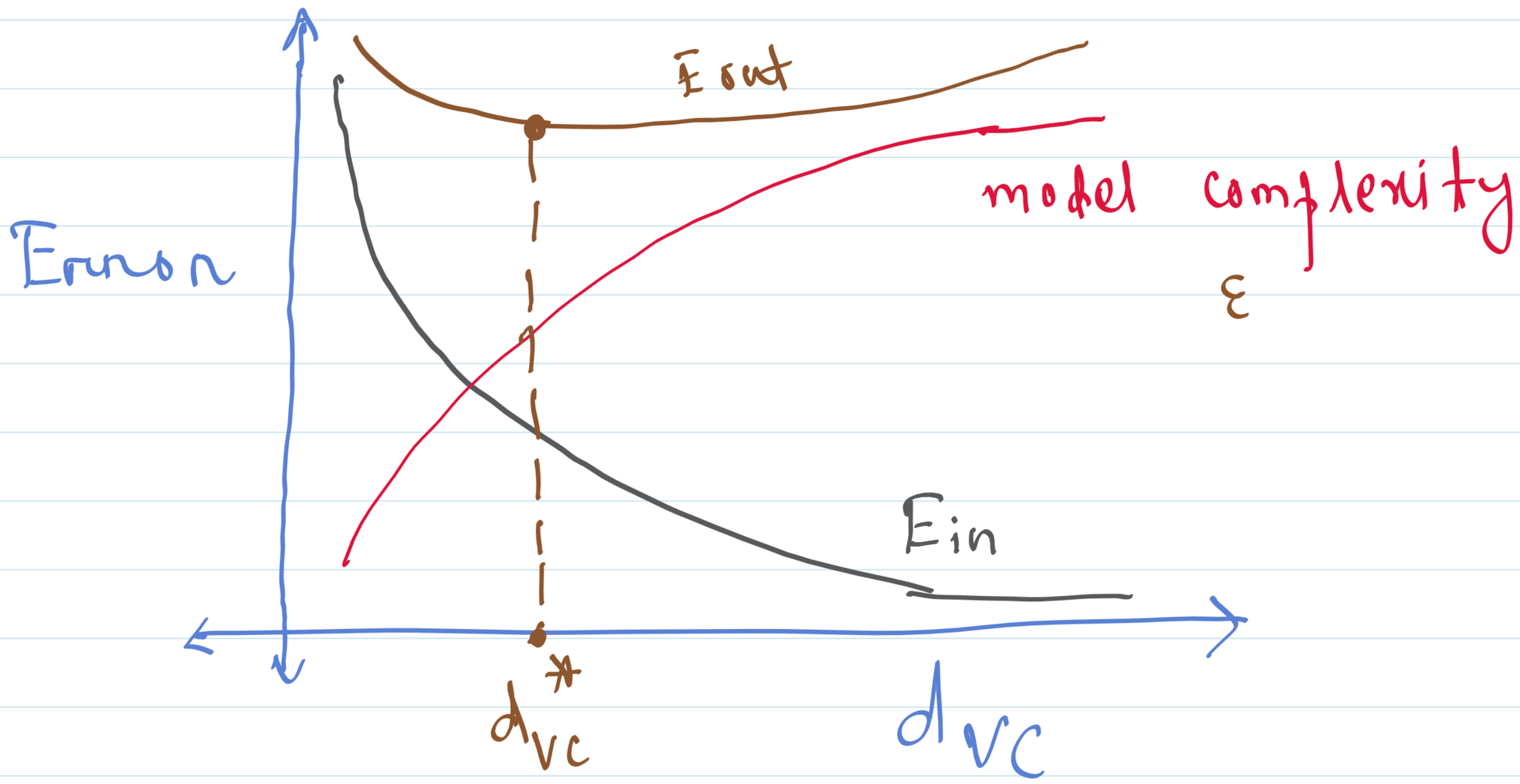
$N = 1000$, $E_{out}(g) \leq E_{in}(g) + 0.301$

Model complexity :

$$\underbrace{\sqrt{\frac{8}{N}\log\left(\frac{4\left(2N\right)^{d_{VC}}+1\right)}{\delta}\Bigg)}}_{\text{penalty of the model complexity}} := \varepsilon(N, \mathcal{H}, \delta)$$

→ If d_{VC} is large, then $\varepsilon(N, \mathcal{H}, \delta)$ is big.

Trade-off curve :



optimum value
of Vc dimension

Generalisation bound for testing :

Testing set : $D_{test} = \{x_1, \dots, x_L\}$

The final hypothesis is already determined.

So we do not need a union bound.

Hoeffding : $\mathbb{P}\left[\left|E_{in}(g) - E_{out}(g)\right| > \varepsilon\right]$

$$\leq 2e^{-2\varepsilon^2 L}.$$

Generalisation bound :

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{1}{2L}\log\frac{2}{\delta}}$$

$L \to \infty \Rightarrow \to 0$

→ If you have big L, then $E_{in}(g)$ will be a good estimate for $E_{out}(g)$.

→ Independent of model complexity, dependent δ & L.

→ VC analysis → based on binary target fun's. (Can be ~~the~~ extended to real valued)

→ Alternative approach → real valued fun's based bias - Variance analysis.

→ VC analysis - E_{in} or E_{out}, $h(x) = f(x)$ or $h(x) \neq f(x)$

From VC analysis to Bias - Variance analysis

VC analysis is a decomposition. As we decompose E_{out} into E_{in} and ε i.e.,

$$E_{out} \leq E_{in} + \sqrt{\frac{8}{N} \log \frac{4((2N)^{d_{VC}} + 1)}{\delta}}$$

E_{in}: training error

penalty of the model complexity

→ Bias - Variance is another decomposition.

1) How well can $\mathcal{H}$ approximate f?
2) How well can we zoom in a good h on $\mathcal{H}$?

→ Vc analysis, $E_{out}(g) = \mathbb{P}[g(x) \neq f(x)]$

Let $B = \{g(x) \neq f(x)\}$ (bad event / $B \in \{0, 1\}$)

$$E_{out}(g) = \mathbb{P}[B = 1] = 1 \cdot \mathbb{P}[B = 1] + 0 \cdot \mathbb{P}[B = 0]$$

$$= \mathbb{E}[B]$$

$$E_{out}(g) = \mathbb{E}_x\left[\mathbb{1}_{\{g(x) \neq f(x)\}}\right], \quad x \in p(x)$$

(0_1 loss)

change the error measure: look at squared loss (Bias-Variance)

$$E_{out}(g) = \mathbb{E}_x\left[(g(x) - f(x))^2\right]$$

Dependancy on training set D:

VC analysis $\boxed{E_{out}(g^{(D)}) = \mathbb{E}_x\left[\mathbb{1}_{\{g^{(D)}(x) \neq f(x)\}}\right]}$

Bias-Variance analysis:

$$E_{out}(g^{(D)}) = \mathbb{E}_x\left[(g^{(D)}(x) - f(x))^2\right].$$

Q Why Not D in VC analysis? As Hoeffding bound is uniform for all D. Not true for (Bias-Variance)

Average over all D

$$\mathbb{E}_D\left[E_{out}(g^{(D)})\right] = \mathbb{E}_D\left[\mathbb{E}_x\left[(g^{(D)}(x) - f(x))^2\right]\right]$$

→ VC trade-off is a "worst-case" analysis, i.e., uniform on every D.

→ Bias-Variance trade-off is an "average" analysis, i.e., average over different D's

Decomposing $E_{out}(g^{(\mathcal{D})})$

$$\mathbb{E}_{\mathcal{D}}\left[E_{out}(g^{(\mathcal{D})})\right] = \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_x\left[\left(g^{(\mathcal{D})}(x) - f(x)\right)^2\right]\right]$$

$$= \mathbb{E}_x\left[\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(x) - f(x)\right)^2\right]\right]$$

$$= \mathbb{E}_x\left[\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(x)\right)^2 - 2\, g^{(\mathcal{D})}(x) f(x) + \left(f(x)\right)^2\right]\right]$$

$$= \mathbb{E}_x\left[\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(x)\right)^2\right] - 2\,\mathbb{E}_{\mathcal{D}}\left[g^{(\mathcal{D})}(x)\right] f(x) + \left(f(x)\right)^2\right]$$

$\bar{g}(x)$ - averaged final hypothesis

The asymptotic limit of the estimate

$$\bar{g}(x) \simeq \frac{1}{K}\sum_{k=1}^{M} g^{(\mathcal{D}_k)}(x)$$

→ $g^{(\mathcal{D}_k)}$ are inside the hypothesis set, but $\bar{g}$ is not necessarily inside.

Bias - Variance decomposition :

$$\mathbb{E}_{\mathcal{D}}\left[E_{out}(g^{(\mathcal{D})})\right] = \mathbb{E}_x\left[\mathbb{E}_{\mathcal{D}}\left[g^{(\mathcal{D})}(x)^2\right] - 2\,\bar{g}(x) f(x) + f(x)^2\right]$$

$$= \mathbb{E}_x\left[\mathbb{E}_{\mathcal{D}}\left[g^{(\mathcal{D})}(x)^2\right] - \bar{g}(x)^2 + \bar{g}(x)^2 - 2\,\bar{g}(x) f(x) + f(x)^2\right]$$

$$\mathbb{E}_x\left[\mathbb{E}_D\left[g^{(D)}(x)^2 - \bar{g}(x)^2 + \bar{g}(x)^2 - 2\bar{g}(x)f(x) + f(x)^2\right]\right]$$

$$\mathbb{E}_D\left[\left(g^{(D)}(x) - \bar{g}(x)\right)^2\right]$$

$$\left(\bar{g}(x) - f(x)\right)^2$$

$\mathbb{E}_D\left[\left(g^{(D)}(x) - \bar{g}(x)\right)^2\right]$: Var(x)

Random

deterministic

$\left(\bar{g}(x) - f(x)\right)^2$: bias(x)

how far is $\bar{g}$ (averaged hypothesis) from target func° f

Overall,

$$\text{bias} = \mathbb{E}_x\left[\text{bias}(x)\right] = \mathbb{E}_x\left[\left(\bar{g}(x) - f(x)\right)^2\right]$$

$$\text{Var} = \mathbb{E}_x\left[\text{var}(x)\right] = \mathbb{E}_x\left[\mathbb{E}_D\left[\left(g^{(D)}(x) - f(x)\right)^2\right]\right]$$

Finally,

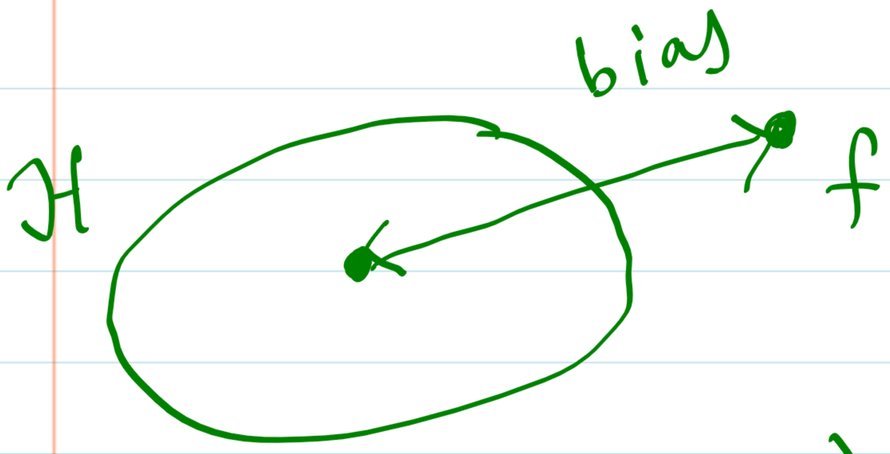
$$\mathbb{E}_D\left[E_{out}\left(g^{(D)}\right)\right] = \mathbb{E}_x\left[\text{bias}(x) + \text{var}(x)\right]$$

$$= \text{bias} + \text{Var}$$

bias(x): How close is the average func° $\bar{g}$ compared to the target

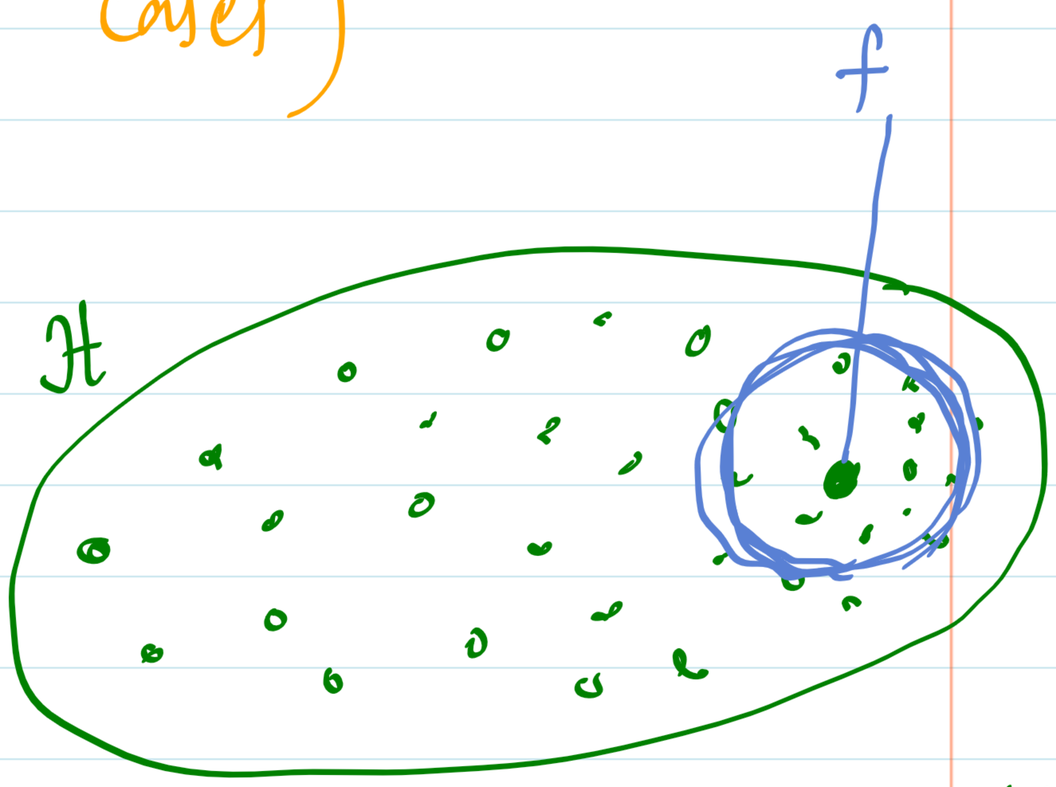
Var(x): How much uncertainty you have around $\bar{g}$.

Picture (Two extreme cases)



(very small model)

(If you have a simple $\mathcal{H}$ then large bias, but small variance)



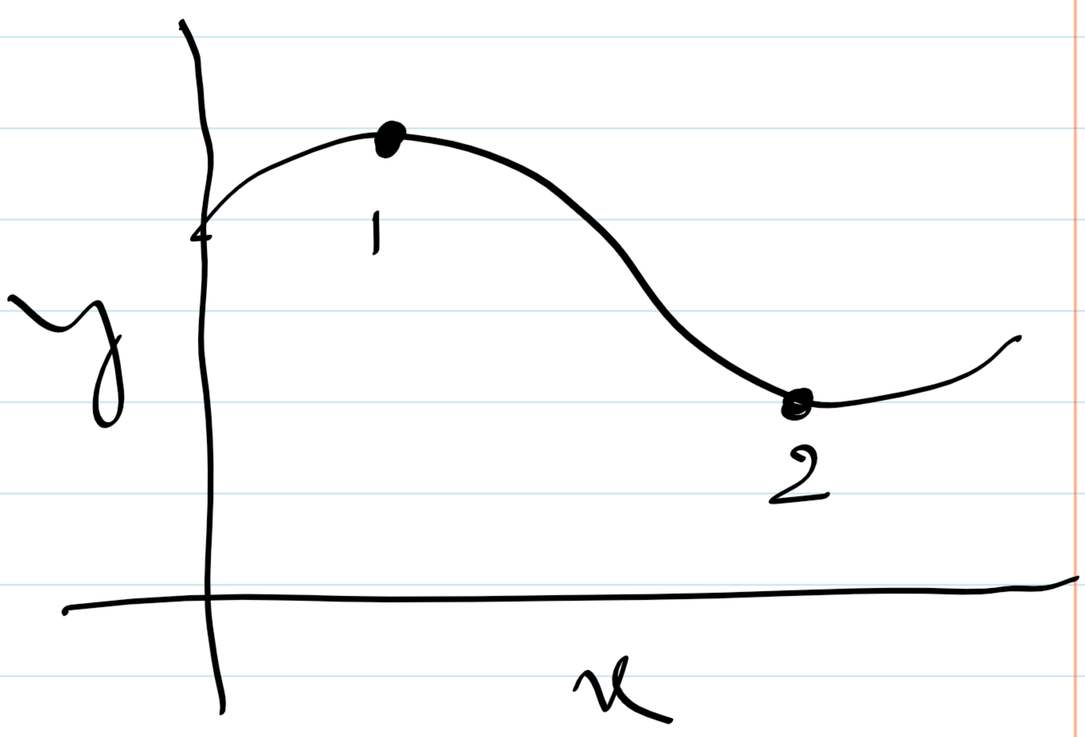
(Very large model)

If you have a Complex $\mathcal{H}$, then small bias but large variance

Example: Consider a $\sin(\cdot)$ func.

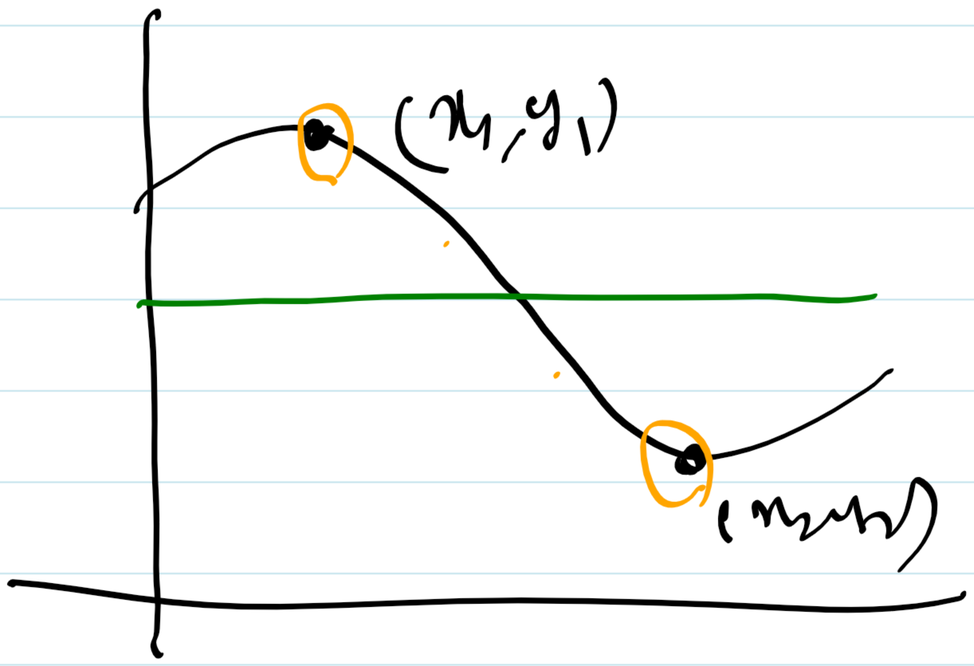
$$f(x) = \sin(\pi x).$$

$N=2$ training samples (x_1, y_1) & (x_2, y_2) by sampling uniformly on $[-1, 1]$

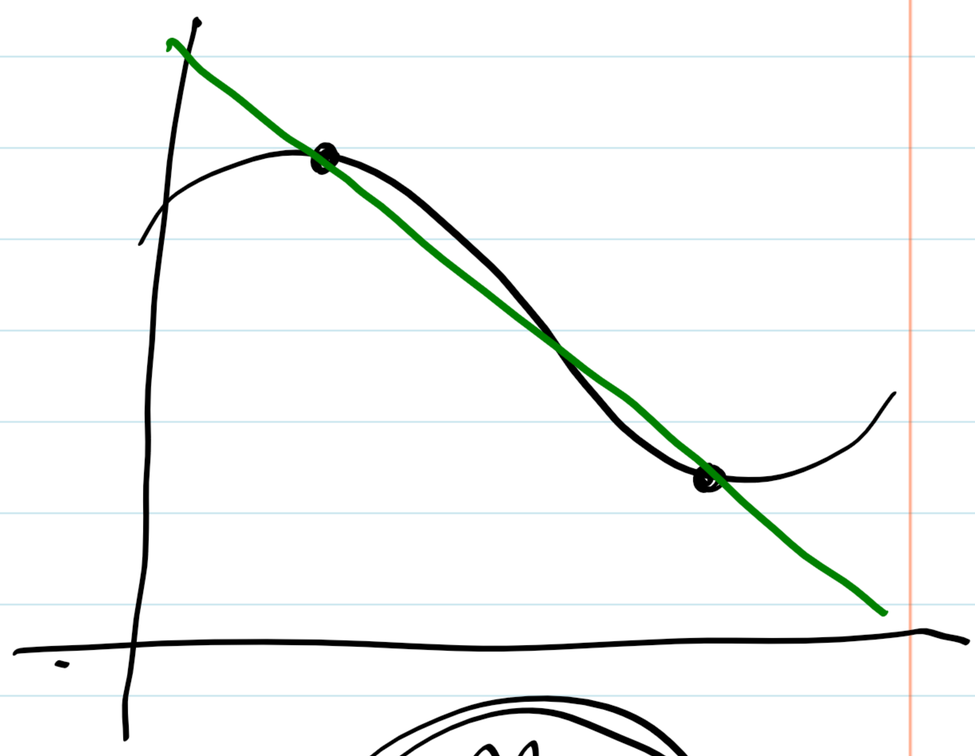


Hypothesis set 0 : $\mathcal{H}_0$ = set of all lines of the form $h(x) = b$

Hypothesis set 1 : $\mathcal{H}_1$ = Set of all lines of the form $h(x) = ax + b$.



$\mathcal{H}_0$

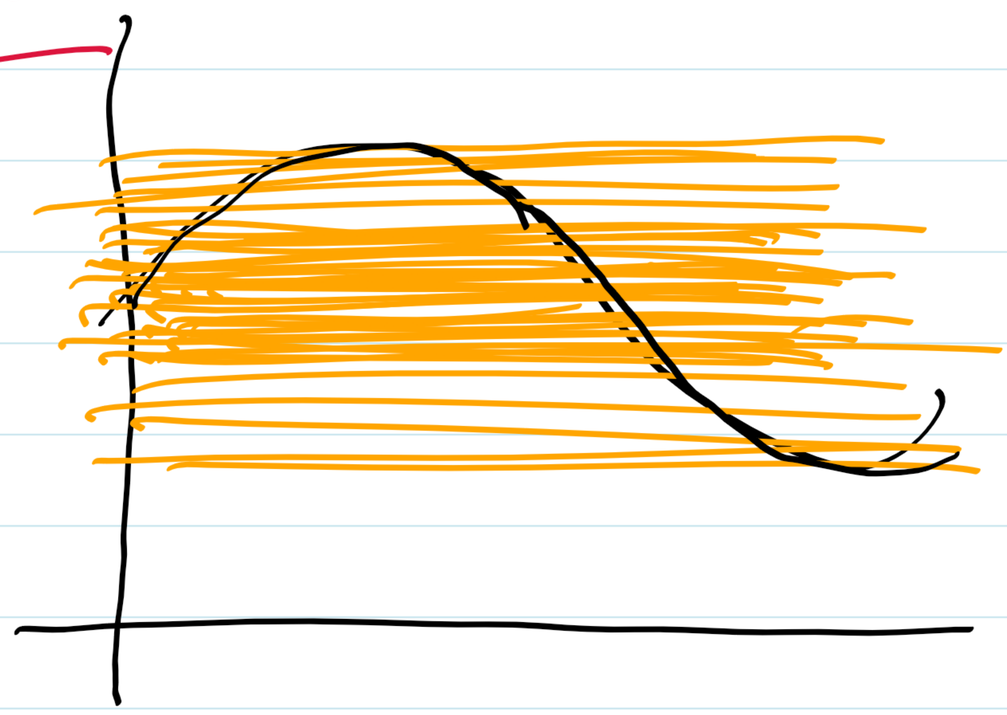


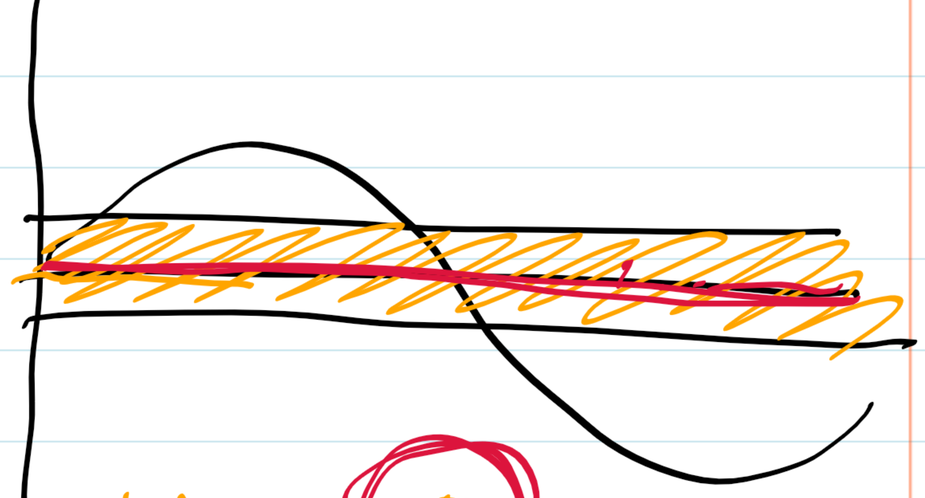
$\mathcal{H}_1$

For $\mathcal{H}_0$: $h(x) = \dfrac{y_1 + y_2}{2}$

For $\mathcal{H}_1$: $h(x) = \underbrace{\left(\dfrac{y_2 - y_1}{x_2 - x_1}\right)x + (y_1 x_2 - y_2 x_1)}$ (better)

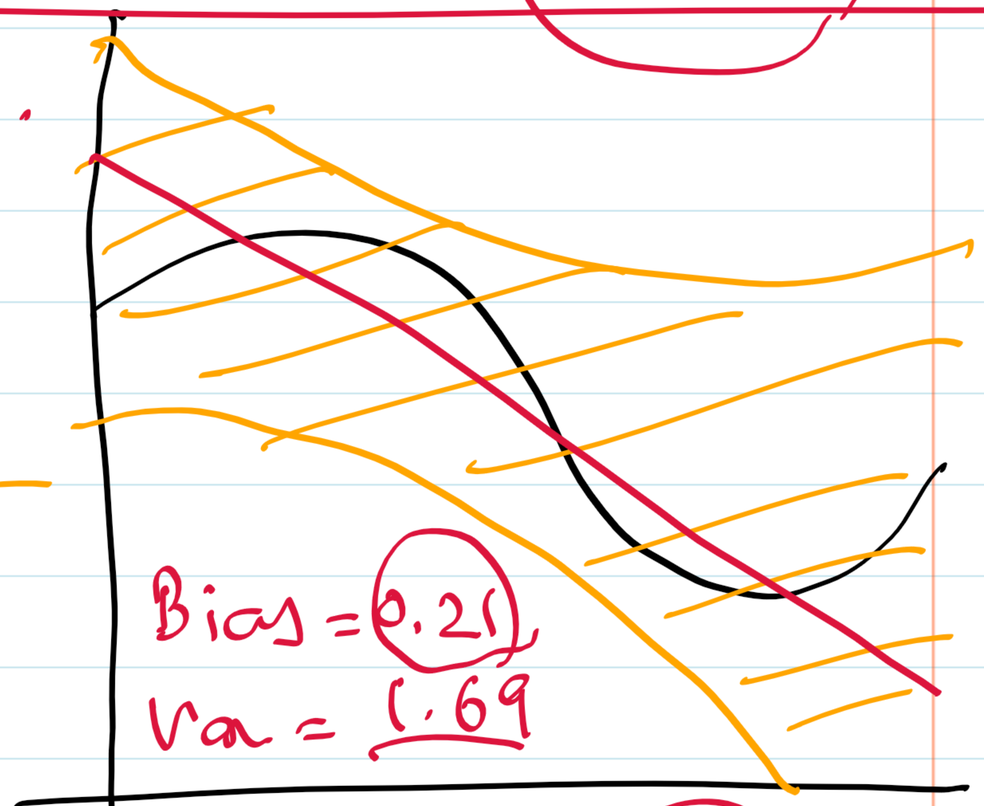
$\mathcal{H}_0$





$\mathcal{H}_0$ Eout = 0.75

$\mathcal{H}_1$



$\mathcal{H}_1$ = Eout = 1.9