Error :-

Regression :- $e(h(x_n), f(x_n)) = (h(x_n) - f(x_n))^2$

Classification :- $e(h(x_n), f(x_n)) = [\![h(x_n) \neq f(x_n)]\!]$

Pointwise

$$E_{in}(h) = \frac{1}{N} \sum_{n=1}^{N} (e(h(x_n), f(x_n))$$

$$E_{out}(h) = \mathbb{E}_x [e(h(x), f(x))]$$

Fingerprint → M → +1 / −1

① Supermarket / Malls

② IB office / Passport

①	Predicted Class	
Actual Class	Yes	No
Yes		−100
No	−10	

②	Predicted Class	
Actual Class	Yes	No
Yes		−1
No	−10	

Given domain Knowledge

⇒ if not Known Domain Knowledge

① Plausible

② Freindly

Noise :-

→ Unknown Distribution

⇒ P(y|x) ← Points being Generated

Aye

Salary

⋮

→ +1 / −1

Noisy Targets = Det Target + Noise

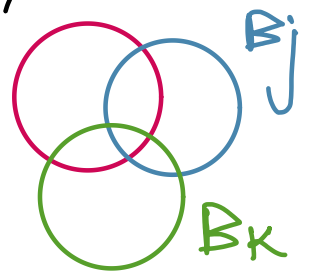
Noise: $y - f(x)$

$f(x) = \mathbb{E}(y|x)$

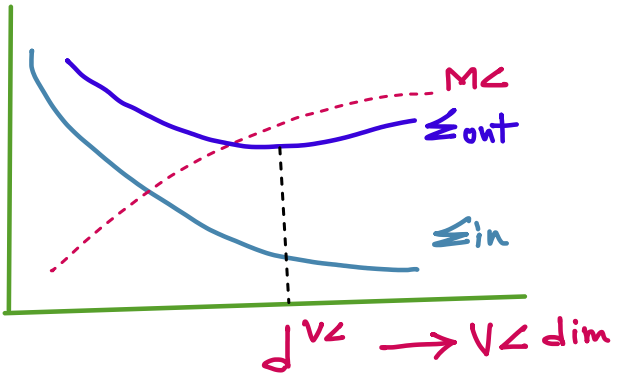
Det Target: $P(y|x) = 0$ except $y = f(x)$

$(x,y) \sim P(y|x)\,P(x)$

h_i B_i B_j B_K

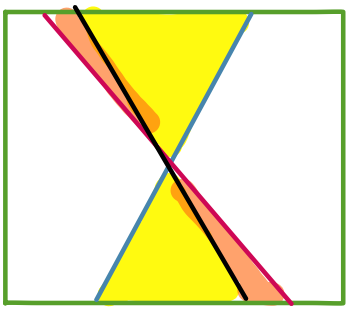


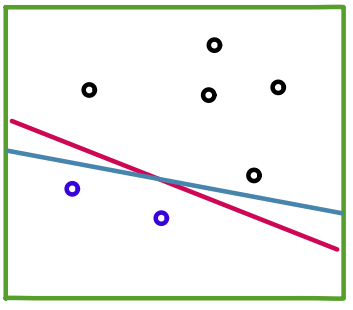
error



$$P(Bad) \leq P(B_1) + \cdots + P(B_M)$$

M factor





$h: X \rightarrow \{+1,-1\}$

$d: \{x_1,\ldots,x_N\} \rightarrow \{+1,-1\}$ $\rightarrow 2^N$

Dichotomy

$$m_H(n) = \max_{x_1,\ldots,x_n \in X} |\mathcal{H}(x_1,\ldots,x_n)|$$

SHATTER

$m_p(1) = 2$

$m_p(2) = 4$

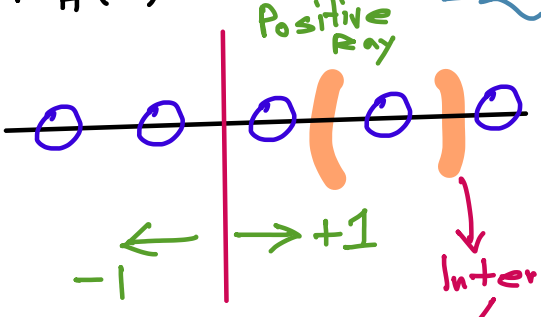
$m_p(3) = 8$

$m_p(4) = 14$

Breakpoint

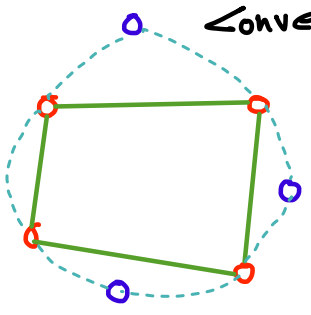
$K \geq 4$

$m_H(N) = N+1$ $\rightarrow$ B.P $= 2$



$$m_H(N) = \binom{N+1}{2} + 1 = \frac{1}{2}(N^2+N)+1$$

Convex functions:-



$$m_H(N) = 2^N$$

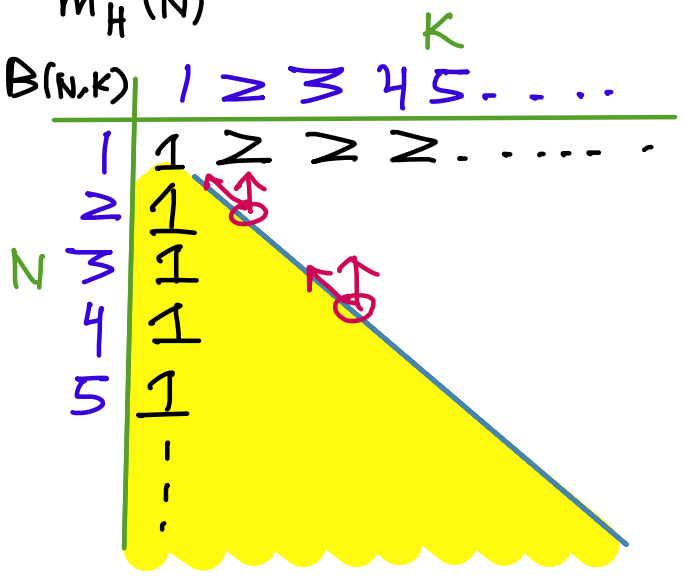
$K = \infty$ (N_0)

▷ If BP = finite, then $m_H(N)$ is poly

$m_H(N) \leq \ldots \leq \ldots \leq$ Poly

$B(N,K)$ = Max number of dichotomy given N points with B.P K.

$m_H(N)$



$\leq \alpha + 2\beta$

$= (\alpha+\beta)+\beta$

$\leq B(N-1,K)$ ← $(\alpha+\beta)$; $\leq B(N-1,K-1)$ ← β

$B(N,K) \leq \sum_{i=0}^{K-1} \binom{N}{i}$

$\sum_{i=0}^{K-1}$ → N^{K-1} → $N^{d_{VC}}$

Perceptron $\leq O(N^3)$

$d_{VC} = K-1$

VC-Inequality

$$Pr\left[|E_{in}(h) - E_{out}(h)| > \epsilon\right] \leq 4 m_H(2N) e^{-\frac{\epsilon^2 N}{8}}$$

PAC

$$-\frac{1}{8}\epsilon^2 N$$

$$\delta = 4 m_H(2N) e^{-\frac{1}{8}\epsilon^2 N}$$

$$P\left[|E_{in} - E_{out}| < \epsilon\right] \geq 1-\delta$$

$$\Rightarrow E_{out} \leq E_{in} + \epsilon$$

$$\epsilon = \sqrt{\frac{8}{N} \log \frac{4 m_H(2N)}{\delta}}$$

Generalization Bound

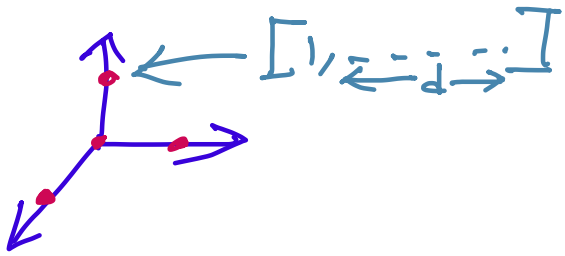
d-dim

$d_{VC} = d+1$

$d_{VC} \leq d+1 \rightarrow$ d+1 can be shattered

$d_{VC} \geq d+1 \rightarrow$ d+2 cannot be "

d=3



$X =$

$$\begin{bmatrix} 1000\text{---}0 \\ 1100\text{---}0 \\ 1010\text{---}0 \\ \text{---} \\ 1000\text{---}1 \end{bmatrix}_{(d+1)\times(d+1)}$$

$y = \begin{bmatrix} \pm 1 \\ \pm 1 \\ \vdots \\ \pm 1 \end{bmatrix}$ $Xw = y$

$\Rightarrow w = X^{-1}y$

$\Rightarrow d+2$ points $x_j = \sum_{i \neq j} a_i x_i$ $a_j \neq 0$ for some j

$\Rightarrow a_j \neq 0 \rightarrow \begin{matrix} > 0 \Rightarrow +1 \\ < 0 \Rightarrow -1 \end{matrix}$

$\Rightarrow y_j \rightarrow -1$

$w^T x_j \Rightarrow > 0$ (Contradiction)

$x_j \rightarrow +1 > 0$

$x_j \rightarrow -1 < 0$

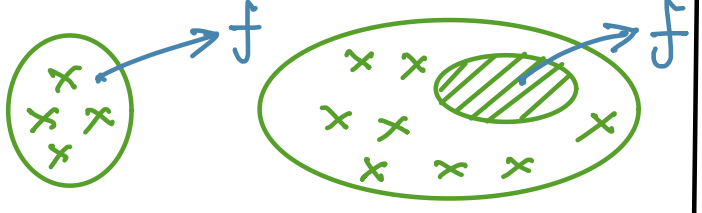
Bias vs Variance Tradeoff

$H = \{h_1, \ldots, h_M\} \rightarrow g \approx f$

① H complex $\Rightarrow$ $|f| \downarrow$

↓

How to get g?



$E_{out}(g)$

$= E_D \left[E_x \left[(g^D(x) - f(x))^2 \right] \right]$

$= E_x \left[E_D \left[(g^D(x) - f(x))^2 \right] \right]$

any hypothesis

$\bar{g}(x) = E_D(g^D(x))$

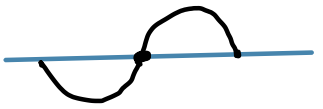
$= E_D \left[(g^D(x) - \bar{g}(x))^2 \right] + (\bar{g}(x) - f(x))^2$

$E_D \left[(g^D(x) - \bar{g}(x))^2 \right]$ = Var(x)

$(\bar{g}(x) - f(x))^2$ = Bias(x)

$E_{out}(g) = E_x \left[Var(x) + Bias(x) \right]$

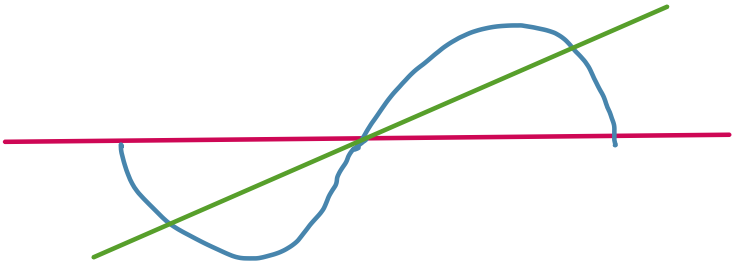
$f : [-1, +1] \to \mathbb{R}$ $f(x) = \sin \pi x$

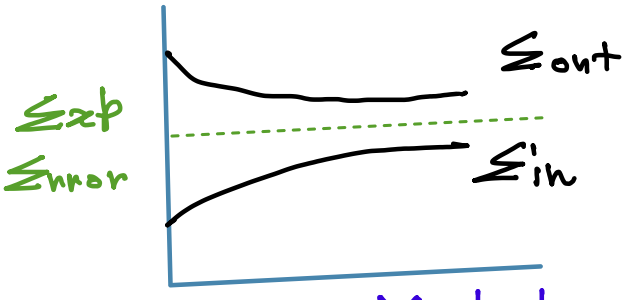


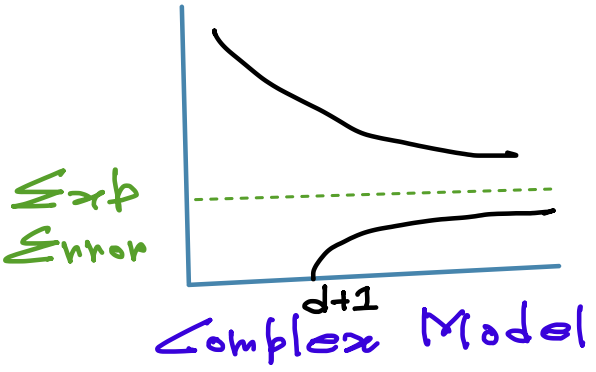
$H_1 : y(x) = b$

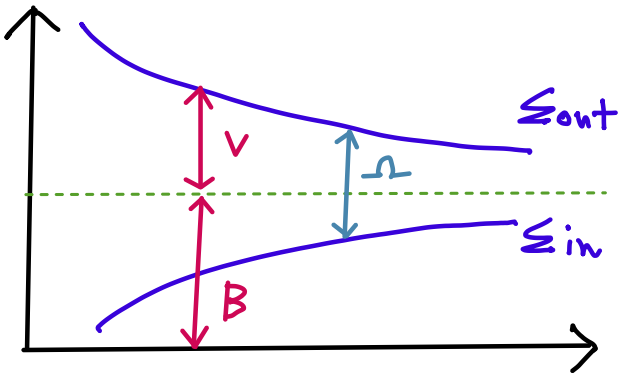
$H_2 : y(x) = ax + b$

$|D| = 2$ points





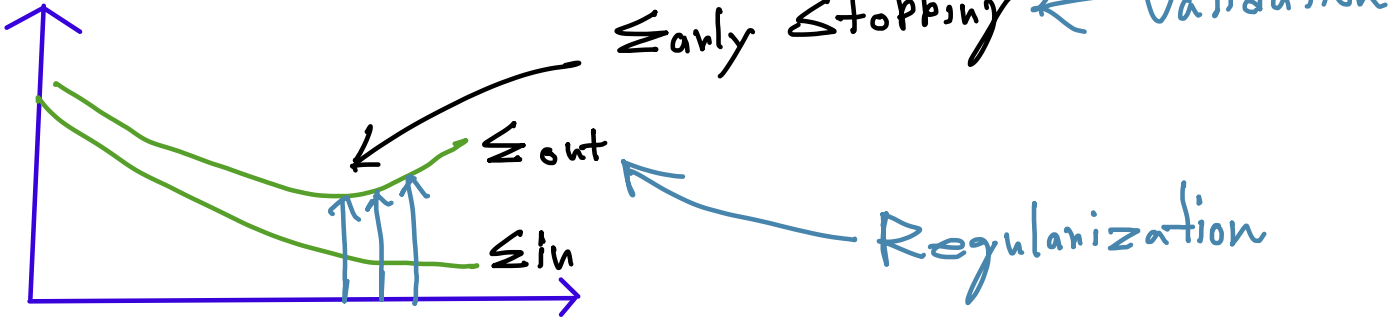




Overfitting $f(x) + \epsilon(x) \rightsquigarrow$

$\hookrightarrow \sum_{q=0}^{\alpha} x^q a_q$ Normalized

$$E_{out}(g) = E_D\left[\left(g^D(x) - \bar{g}(x)\right)^2\right] + \left(\bar{g}(x) - f(x)\right)^2 + E_{\epsilon,D}[\epsilon(x)]$$



Regularization!

Legendre Poly

$$y(x) = w^T x$$

$$E_{in}(w) = \frac{1}{N}\sum_{n=1}^{N}(w^T x_n - y_n)^2$$

Minimize $= \frac{1}{N}(Xw-y)^T(Xw-y)$

$$\frac{\partial E_{in}(w)}{\partial w} = 0 \Rightarrow w_{in} \leftarrow (X^{-1}X)^T X y \quad \text{s.t } w^T w \le C$$

$C \uparrow$ V.C

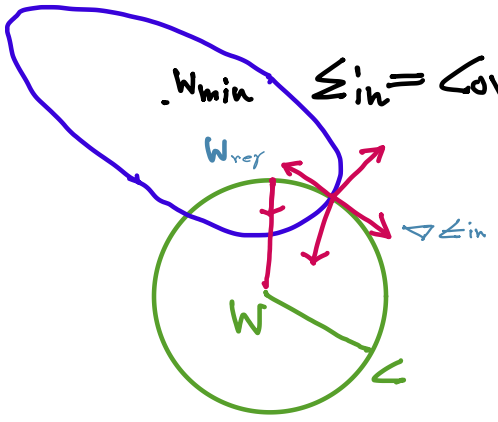
Hard Constraint

Coeff$(q>2) \to 0$

Soft Constraint

$w^T w \le C$

$\nabla E_{in}(w_{reg}) \propto -w_{reg}$



$$\Rightarrow \nabla E_{in}(w_{reg}) = -2\frac{\lambda}{N} w_{reg}$$

$$\Rightarrow \nabla E_{in}(w_{reg}) + 2\frac{\lambda}{N} w_{reg} = 0$$

Dual $\underset{\lambda \ge 0}{\text{Min}}\left[E_{in}(w_{reg}) + \frac{\lambda}{N} w^T w\right]$

$$\frac{\partial}{\partial w}\left[\frac{1}{N}(Xw-y)^T(Xw-y) + \frac{\lambda}{N} w^T w\right] = 0 \rightarrow E_{aug}(w)$$

$$\Rightarrow X^T(Xw-y) + 2\lambda w = 0$$

$\lambda \uparrow \; C \downarrow$

$\lambda \downarrow \; C \uparrow$

$$w = (X^T X + 2\lambda I)^{-1} X^T y$$

$$w^{t+1} \leftarrow w^t - \eta \nabla E_{aug}(w)$$ ← Instead of $\nabla E_{in}(w)$ for reg

$$= w^t - \eta\left(\nabla E_{in}(w) + \frac{2\lambda}{N} w^t\right)$$

$$= w^t\left(1 - \frac{2\lambda}{N}\right) - \eta \nabla E_{in}(w)$$

Weight Decay

$E_{out}(g) = \mathbb{E}_x[e(g(x), y)] \quad \sigma^2 = Var[e(g(x), y)]$

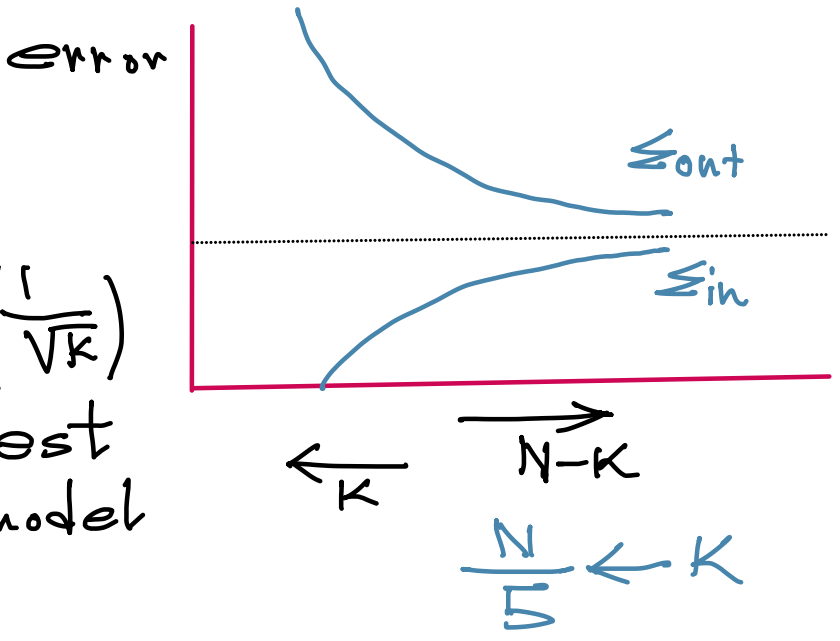
$(x_1, y_1) \; - - - \; (x_K, y_K)$

$\frac{1}{K}\sum_{k=1}^{K} \mathbb{E}_{x_k}(e(g(x_k), y)) = E_{out}(g)$

$Var = \frac{1}{K^2}\sum_{k} Var(e(g(x_k), y)) = \frac{\sigma^2}{K}$

$D_{Train} \rightarrow N-K$

$D_{Val} \rightarrow K$

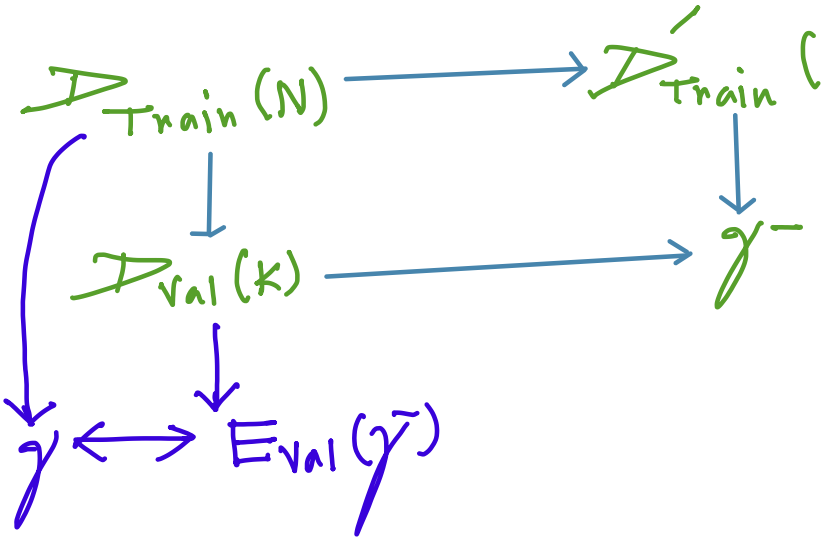


$E_{out}(g) \approx E_{val}(g) \pm O\left(\frac{1}{\sqrt{K}}\right)$

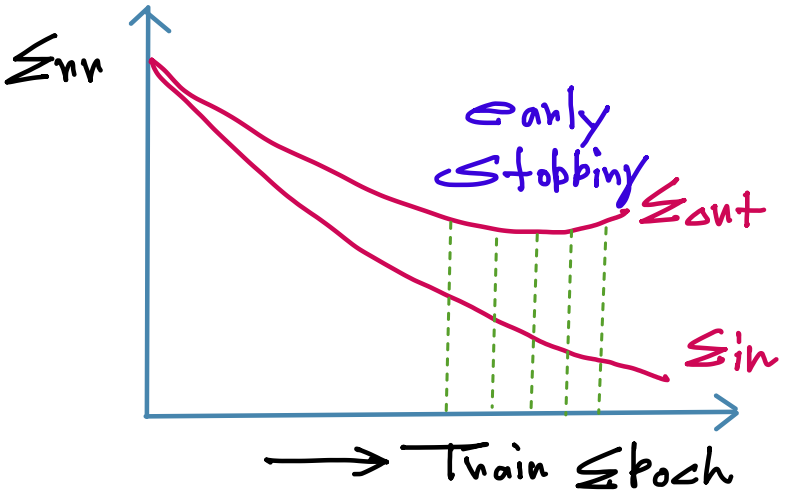
Small $K \rightarrow$ bad est

Large $K \rightarrow$ bad model

$\frac{N}{5} \leftarrow K$



$E_{val}(g) \approx E_{val}(g^-)$



h_1, h_2

$\mathcal{E}_{out}(h_1) = \mathcal{E}_{out}(h_2) = 0.5$

$h_{final} = h_1$ or h_2

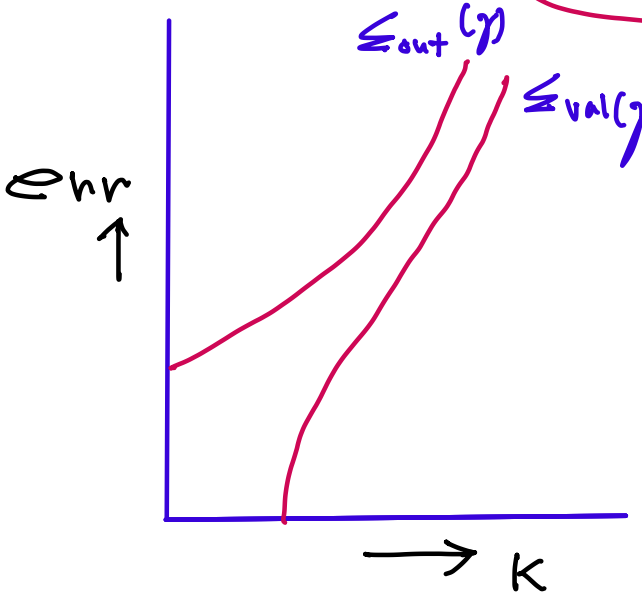
Consider $e = \min\{e_1, e_2\}$

$\mathcal{D}_{Train}$ $(N-K)$ → H_1 H_2 ——— H_M

$\bar{h_1}$ $\bar{h_2}$ $\bar{h_M}$

$\mathcal{D}_{Val}$ (K) →

$\mathcal{E}_{val}(\bar{h_1})$ $\mathcal{E}_{val}(\bar{h_2})$ $\mathcal{E}_{val}(\bar{h_M})$



$(h_m^*, \mathcal{E}_m^*)$

$H_{SF} = \{\bar{h_1}, \ldots, \bar{h_M}\}$

$E_{out}(g) \approx E_{val}(g^*) + O\left(\sqrt{\frac{\ln m_H}{K}}\right)$

$\mathcal{E}_{out}(g) \approx \mathcal{E}_{out}(g^*) \approx \mathcal{E}_{val}(\bar{g})$

Small K ↔ Large K

Leave One Out!

$(x_1, y_1) \cdots (x_K, y_K) \cdots (x_N, y_N)$

(x_K, y_K): val

$N-1$

$N-1$ for train

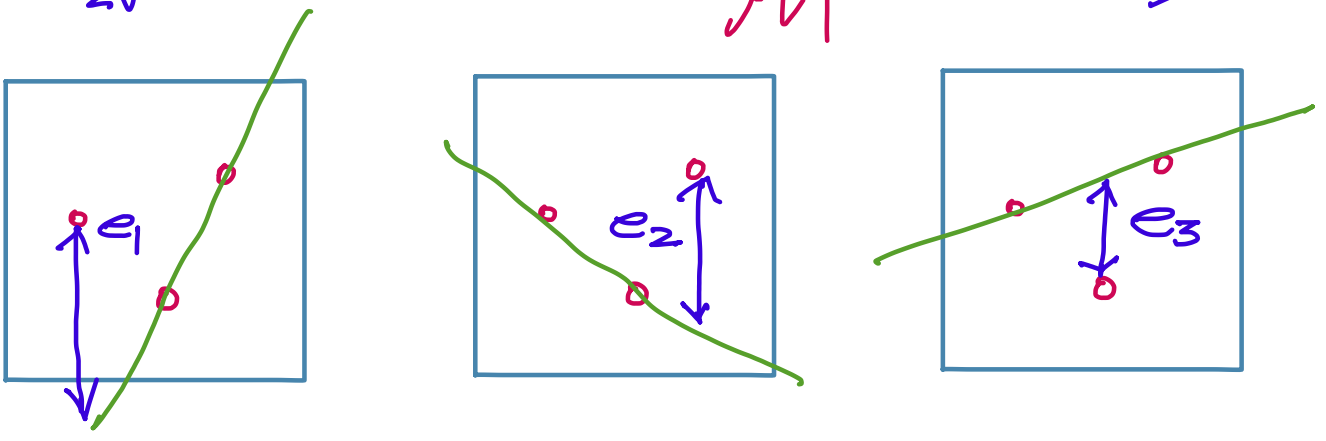
1 for Validation

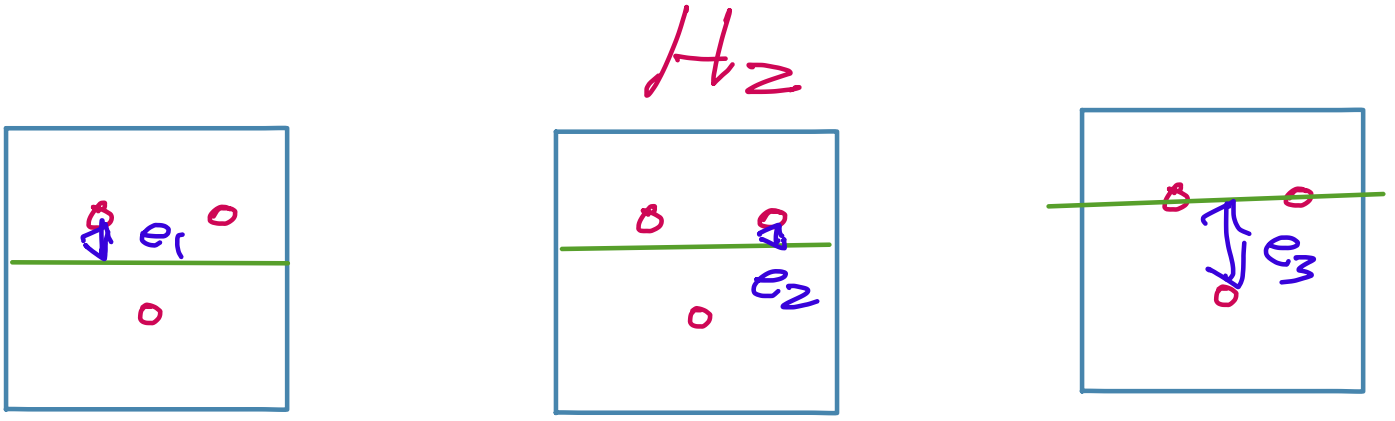
$E_{val} = \frac{1}{N}\sum_{n=1}^{N} E_{val}^{n}$

$\mathcal{E}_{CV}$

H_1

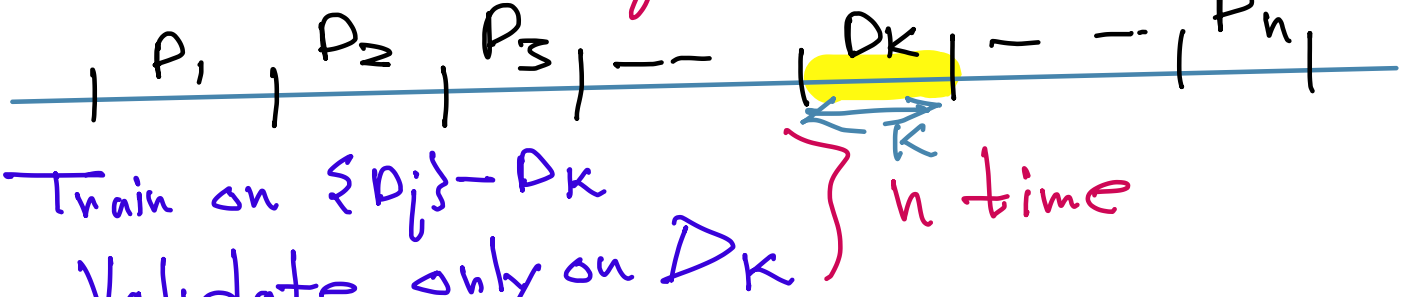
$\mathcal{E}_{CV} = \frac{1}{3}(e_1 + e_2 + e_3)$





$$\varepsilon_{CV} = \frac{1}{3}(e_1 + e_2 + e_3)$$

$\Rightarrow$ N can be large



Train on $\{D_j\} - D_K$

Validate only on D_K

N/K Training Sessions // (N-K) points to train

K-fold Cross Validation