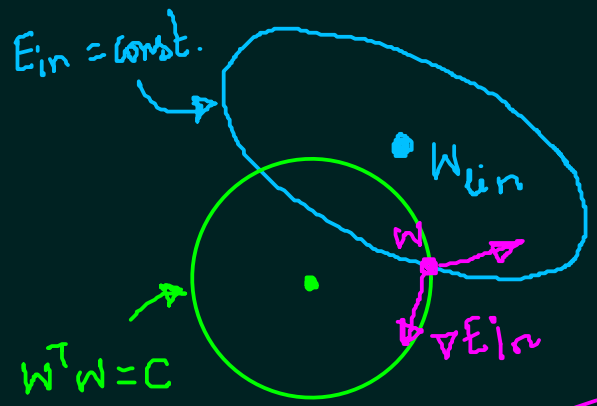
▶ Regularization:

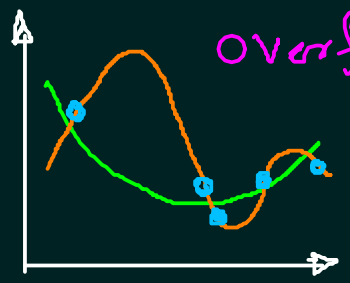
unconstrained ⟶ Constrained

[Reduce overfitting Problem]

SUMMARY

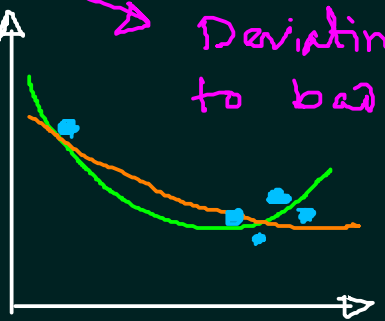


- Minimize $E_{aug}(w) = E_{in}(w) + \frac{\lambda}{N} w^T w$
- Minimize $E_{in}(w)$ subject to $w^T w \leq C$

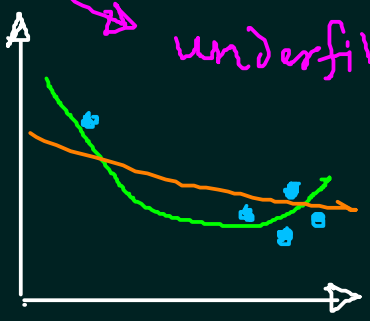


$\lambda = 0$

$\lambda = 0.0001$



$\lambda = 0.1$



$\lambda = 1.0$

Choosing a Regularizer:

$$E_{aug}(h) = E_{in}(h) + \frac{\lambda}{N}\Omega(h)$$

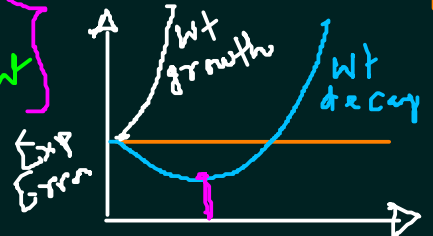
- $\Omega(h)$: heuristic ⟶ smooth, simple h functions
- Popular : weight decay in stochastic gradient descent

$$w(t+1) \leftarrow w(t)\left[1 - 2\eta\frac{\lambda}{N}\right] - \eta\nabla E_{in}(w(t))$$

decay / iterations

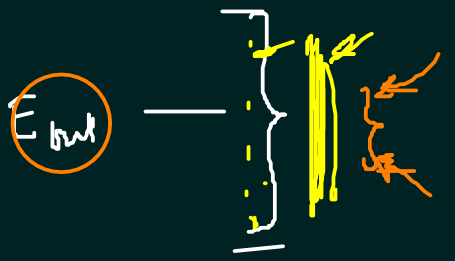
- λ : validation

TODAY



$$E_{out}(h) = E_{in}(h) + \text{Penalty}$$

Penalty → overfit, model noise → stoch., det.



Regularization:

$$E_{aug}(h) = E_{in}(h) + \frac{\lambda}{N}\Omega(h)$$

Validation: $E_{out}(h)\downarrow$

$E_{in} \approx E_{out}$ (goal)

▶ Out-of-sample (x, y) → Error, $e(h(x), y)$

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

$(h(x) - y)^2$ $\quad$ $[\![h(x) \neq y]\!]$

$$\text{var}\left[e(h(x), y)\right] = \sigma^2 \quad \text{(ONE POINT)}$$

▶ Set of points: (x_1, y_1) ------ (x_K, y_K) ← X

$$E_{val}(h) = \frac{1}{K}\sum_{k=1}^{K} e(h(x_k), y_k) \longrightarrow \mathbb{E}\left[E_{val}(h)\right] = E_{out}(h)$$

$$E_{out}(h) = E_{in}(h) \pm O\left(\frac{1}{\sqrt{K}}\right)$$

$$\text{var}\left[E_{val}(h)\right] = \frac{1}{K^2}\sum_{k=1}^{K} \text{var}\left[e(h(x_k), y_k)\right] = \frac{\sigma^2}{K}$$

$\mathcal{D} = (x_1, y_1) \text{ ----- } (x_N, y_N)$

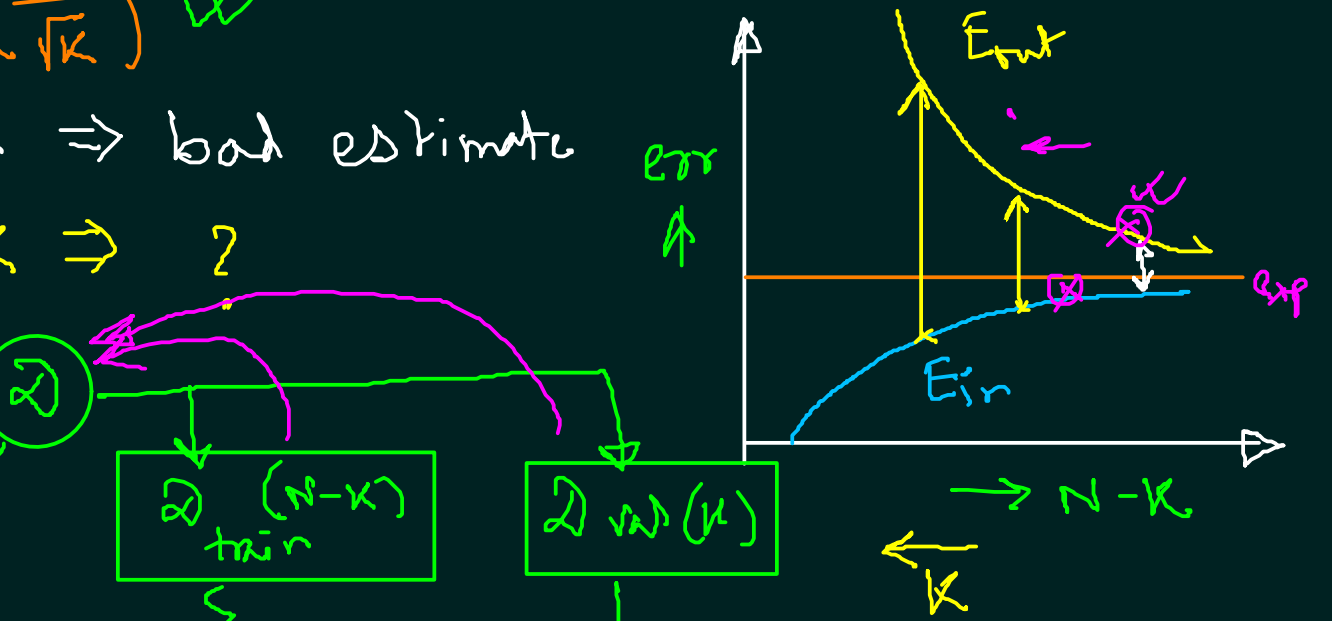
K point $\Rightarrow \mathcal{D}_{val}$?

(N-K) point $\Rightarrow \mathcal{D}_{train}$ ✓

$E_{out}(h) = E_{in}(h) \pm O\left(\frac{1}{\sqrt{K}}\right)$ ✓

Small K $\Rightarrow$ bad estimate

large K $\Rightarrow$?



$\mathcal{D} = \mathcal{D}_{train} \cup \mathcal{D}_{val}$

N $\quad$ N-K $\quad$ K

N $\mathcal{D}$ $\rightarrow$ $\mathcal{D}_{train}^{(N-K)}$ $\quad$ $\mathcal{D}_{val}(K)$

$\mathcal{D}_{train}^{(N-K)} \rightarrow g^-$

$\mathcal{D}_{val}(K) \rightarrow g$

E_{val} ? $\leftarrow g$

$g^- \rightarrow ? \rightarrow E_{val}(g^-)$ (validation)

K is large

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

g: final $\quad$ $E_{val}(g^-)$: est

$K = \frac{N}{5}, \frac{N}{6}$

Val $\leftarrow$ optimistic bias

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

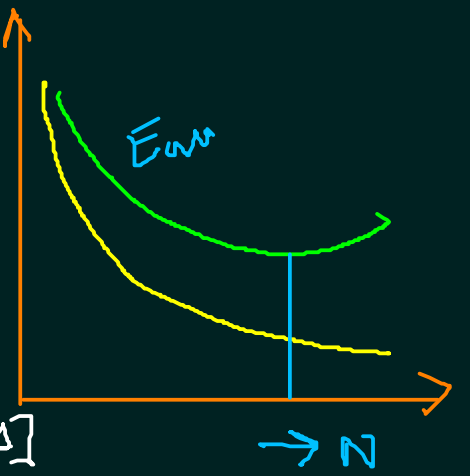
$E(e) < 0.5$

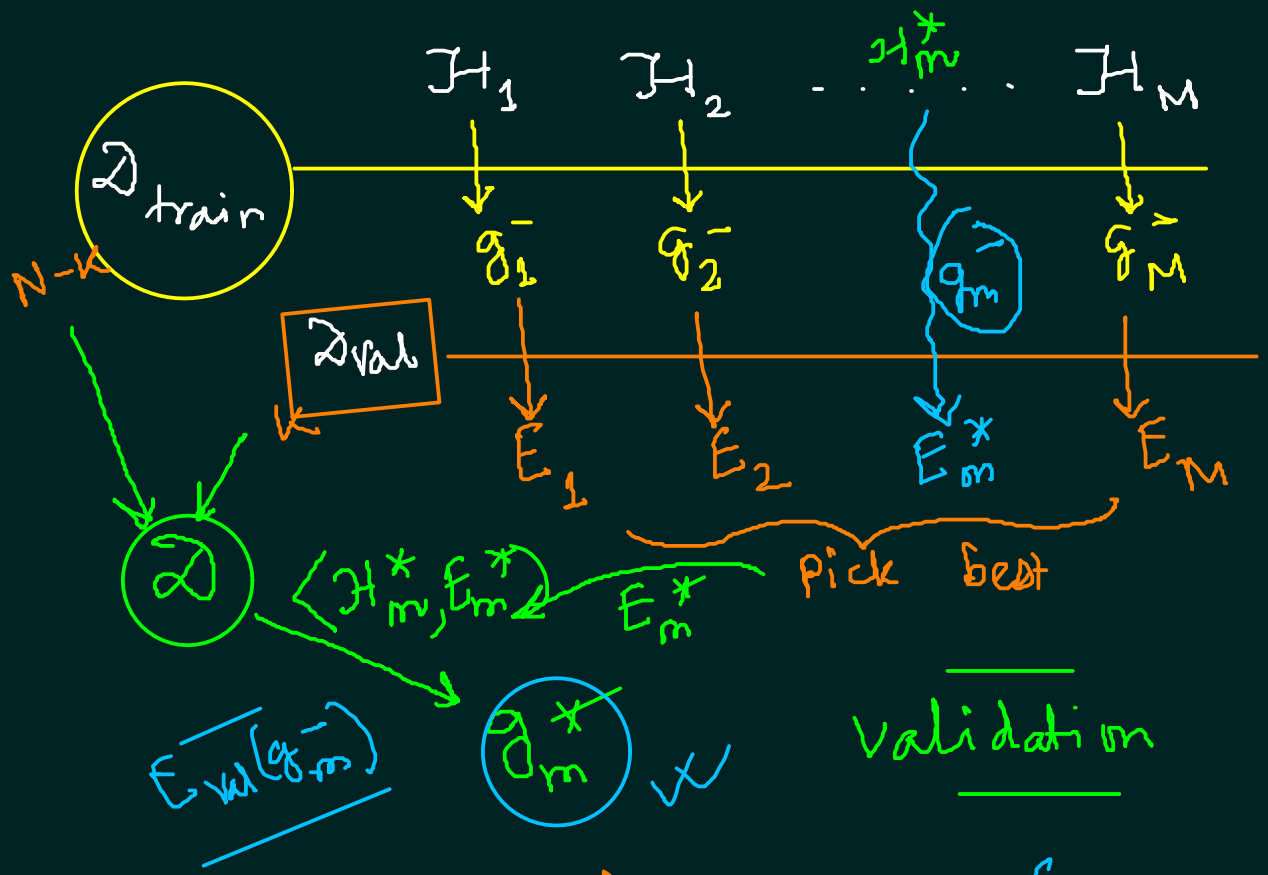
K

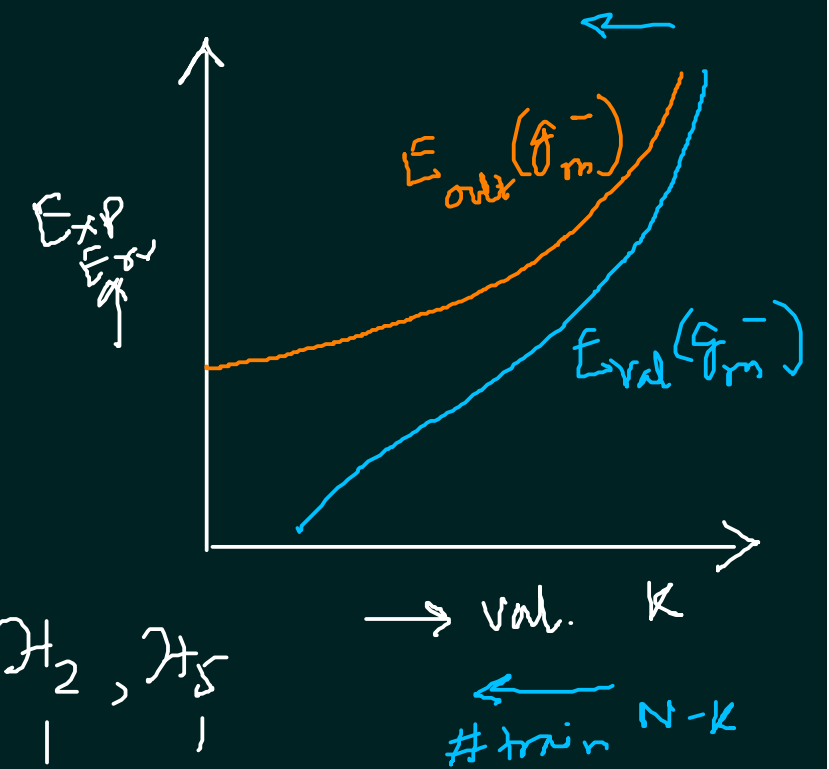
$h_1 \quad h_2$

$E_{out}(h_1) = E_{out}(h_2) = \frac{1}{2}$

$e_1 = e_2$ Unifom $[0,1]$







$$\hookrightarrow \mathcal{H}_{val} = \{ g_1^-, g_2^-, \cdots, g_M^- \}$$

$$E_{out}(g_{m^*}^-) \leq E_{val}(g_{m^*}^-) + O\left(\sqrt{\frac{\ln M}{K}}\right)$$

E_{in}
E_{val}
E_{out}

small K [with $(1-\delta)$ prob]

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

$$\sqrt{\frac{1}{2K}\ln\left(\frac{2M}{\delta}\right)}$$

$\frac{N}{K}$

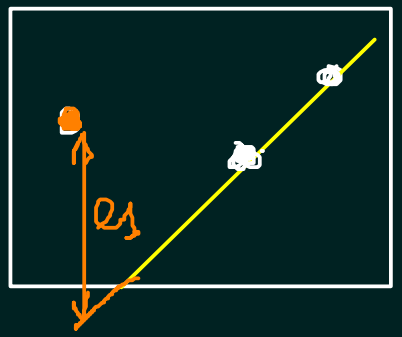
(larger K)

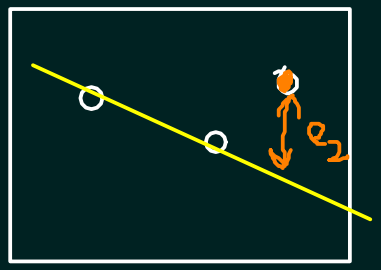
$\mathcal{D}_n = (x_1, y_1) \; \cdots \; (x_n, y_n) \; \cdots \; (x_N, y_N)$

$N-1$ for train

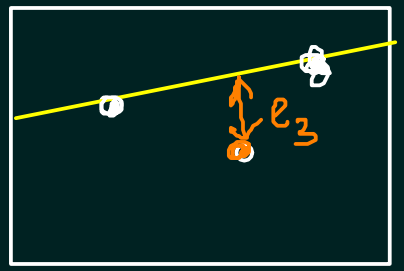
$e_n = E_{val}(g_n^-) = e(g_n^-(x_n), y_n)$ ⇐ 1 pair for validation

Cross validation error $= \frac{1}{N}\sum_{n=1}^{N} e_n$ ⇐ E_{CV}





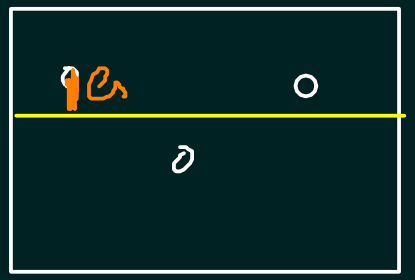
$\mathcal{H}_1$

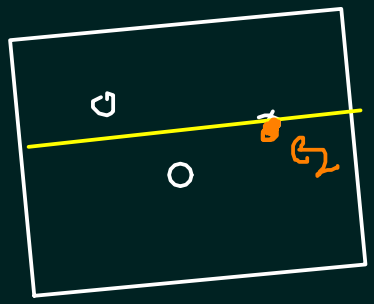


$E_{CV} = \frac{1}{3}(e_1 + e_2 + e_3)$

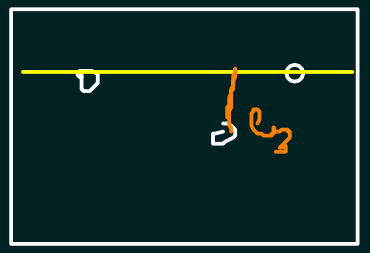
linear

11

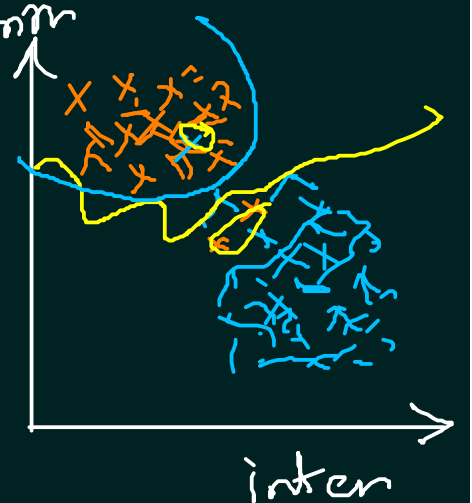




$\mathcal{H}_2$



const

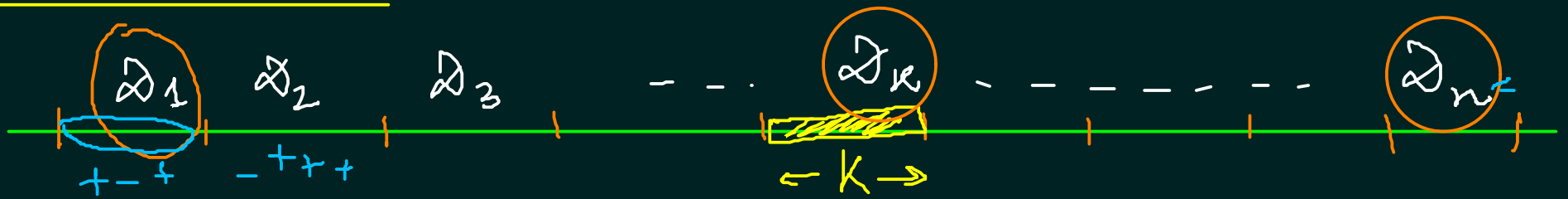


w/o val: $E_{in} = 0\%$ $E_{out} = 2.5\%$

$(1, x_1, x_2) \rightarrow (1, x_1, x_2, x_1^2, x_1x_2, x_2^2, \cdots)$ 40%

w/ val: $E_{in} = 0.8\%$ $E_{val} = 1.5\%$

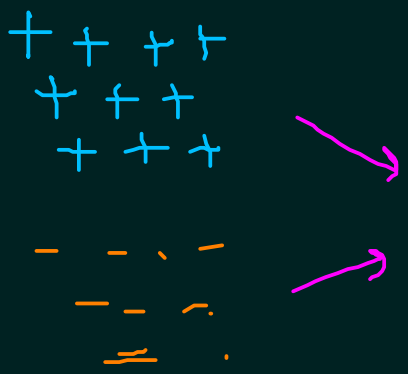
N can be large



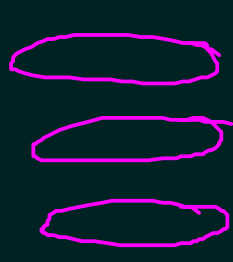
Train on $\{\mathcal{D}_i\} - \mathcal{D}_K$
Validate on only $\mathcal{D}_K$
} Repeat this n time

$\frac{N}{K}$ Training Sessions // $(N-K)$ points to train

(K)-fold Cross-Validation



Strata →



→ Stratified Approach ✓