Machine Learning from Data

Lecture 13: Spring 2021

Today’s Lecture

- Validation and Model Selection
  - Validation Set
  - Model Selection
  - Cross validation

Regularization (Recap)

Regularization combats the effects of noise by putting a leash on the algorithm.

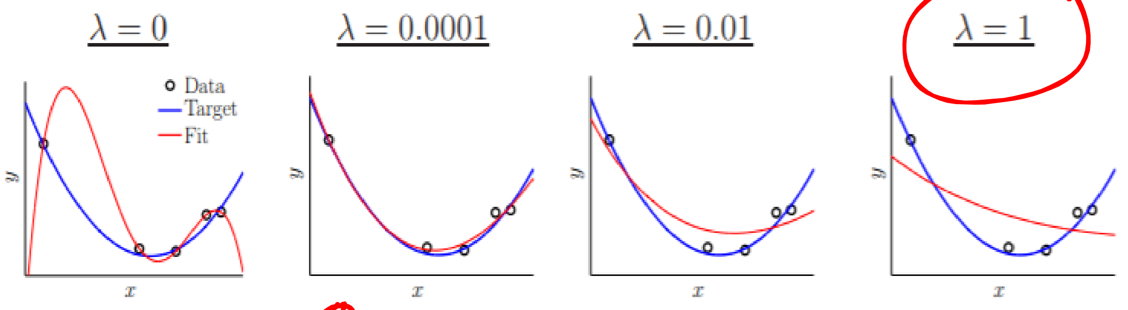
$$E_{\text{aug}}(h) = E_{\text{in}}(h) + \frac{\lambda}{N}\Omega(h)$$

$\Omega(h) \rightarrow$ smooth, simple h

noise is rough, complex.

Different regularizers give different results

can choose λ, the **amount** of regularization.



Overfitting $\rightarrow$ $\rightarrow$ **Underfitting**

Optimal λ balances approximation and generalization, bias and variance.

Validation

$$E_{\text{out}}(g) = E_{\text{in}}(g) + \underbrace{\text{overfit penalty}}_{}$$

VC bounds this using a complexity error bar $\Omega(\mathcal{H})$

regularization estimates this through a heuristic complexity penalty $\Omega(g)$

Validation goes directly for the jugular:

$$\underbrace{E_{\text{out}}(g)}_{\text{validation estimates this directly}} = E_{\text{in}}(g) + \text{overfit penalty}.$$

In-sample estimate of E_{out} is the Holy Grail of learning from data.

D_N

Hyp + LA

$\downarrow$

g

$\downarrow$

$E_{out}(g)$

x

$e(g(x), y)$

$e(g)$

$E[e(g)] = E_{out}(g)$

$\underline{Var(e(g))}$

Test data set

$x_1 \; x_2 \; \cdots \; x_K$

$\downarrow \quad \downarrow \quad \downarrow$

$e_1 \quad e_2 \quad e_K$

$$E_{test} = \frac{1}{K} \sum_{k=1}^{K} e_k$$

$$\underline{E[E_{test}]} = E\left[\frac{1}{K} \sum_{k=1}^{K} e_k\right]$$

$$= \frac{1}{K} \sum_{k=1}^{K} E(e_k)$$

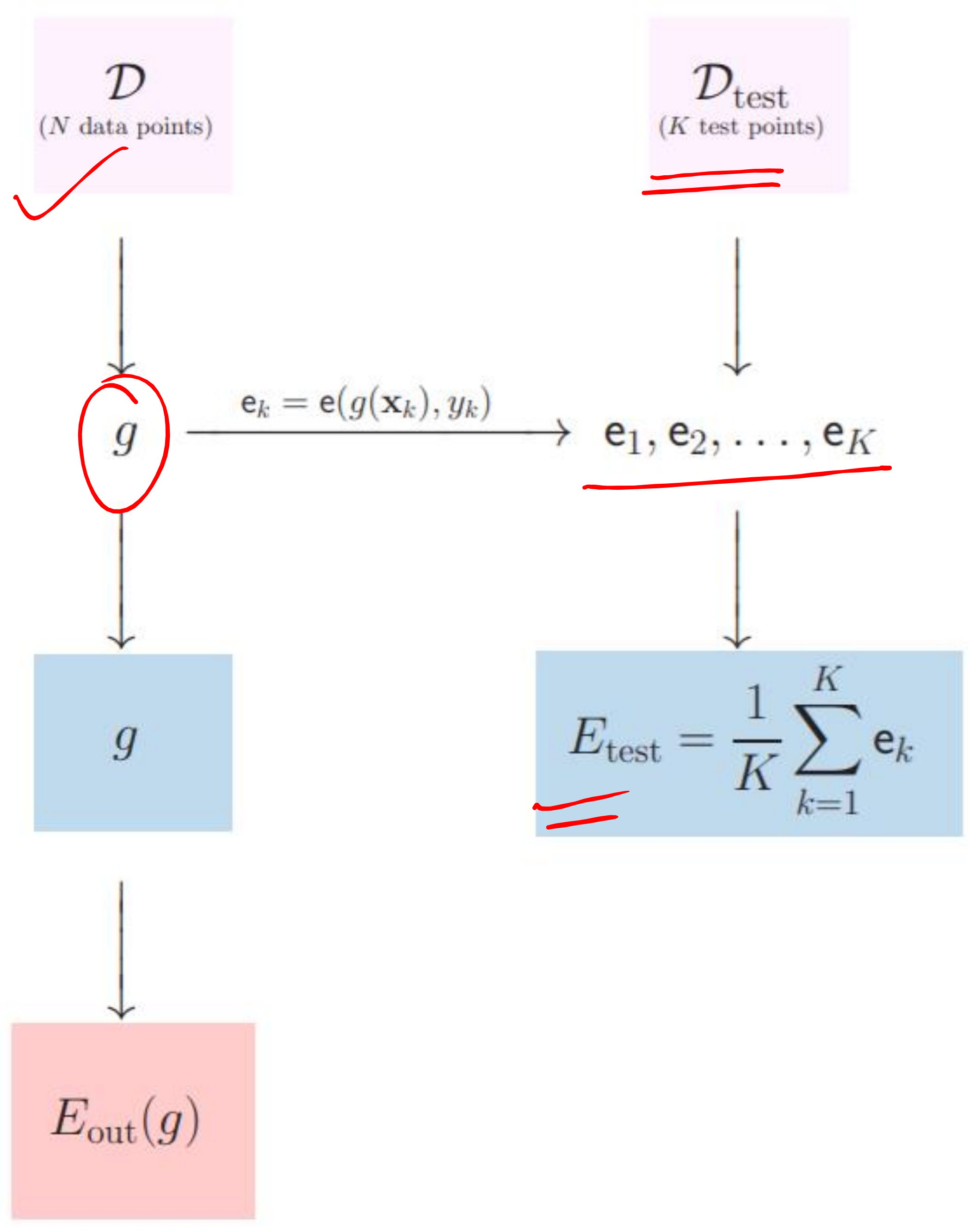
$$= \frac{1}{K} \cancel{K} \; E_{out}(g)$$

$$\mathrm{Var}(E_{test}) = \mathrm{Var}\left[\frac{1}{K}\sum_{k=1}^{K} e_k\right]$$

$$= \frac{1}{K^2}\mathrm{Var}\left[\sum_{k=1}^{K} e_k\right]$$

$$\mathrm{Var}(E_{test}) = \frac{1}{K^2}\cdot K\cdot\mathrm{Var}(e_k) = \frac{1}{K}\mathrm{Var}(e_k)$$

$$E_{out} \simeq E_{test} \pm \frac{1}{\sqrt{K}}\sqrt{\mathrm{Var}(e_k)}$$



E_{test} is an estimate for $E_{\text{out}}(g)$

$$\mathbb{E}_{\mathcal{D}_{\text{test}}}[\mathrm{e}_k] = E_{\text{out}}(g)$$

$$\begin{aligned}
\mathbb{E}[E_{\text{test}}] &= \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}[\mathrm{e}_k] \\
&= \frac{1}{K}\sum_{k=1}^{K} E_{\text{out}}(g) = E_{\text{out}}(g)
\end{aligned}$$

$\mathrm{e}_1, \ldots, \mathrm{e}_K$ are *independent*

$$\begin{aligned}
\mathrm{Var}[E_{\text{test}}] &= \frac{1}{K^2}\sum_{k=1}^{K} \mathrm{Var}[\mathrm{e}_k] \\
&= \frac{1}{K}\mathrm{Var}[e]
\end{aligned}$$

decreases like $\frac{1}{K}$
bigger $K \implies$ more reliable E_{test}.

Validation

$D_N \longrightarrow \boxed{D_K \quad x_1, x_2 \dots x_K}$ validation set

$\downarrow$

Training set $\boxed{D_{N-K}}$

$\downarrow$

g^- (deficient hypothesis)

$\downarrow$

$E_{out}(g^-)$

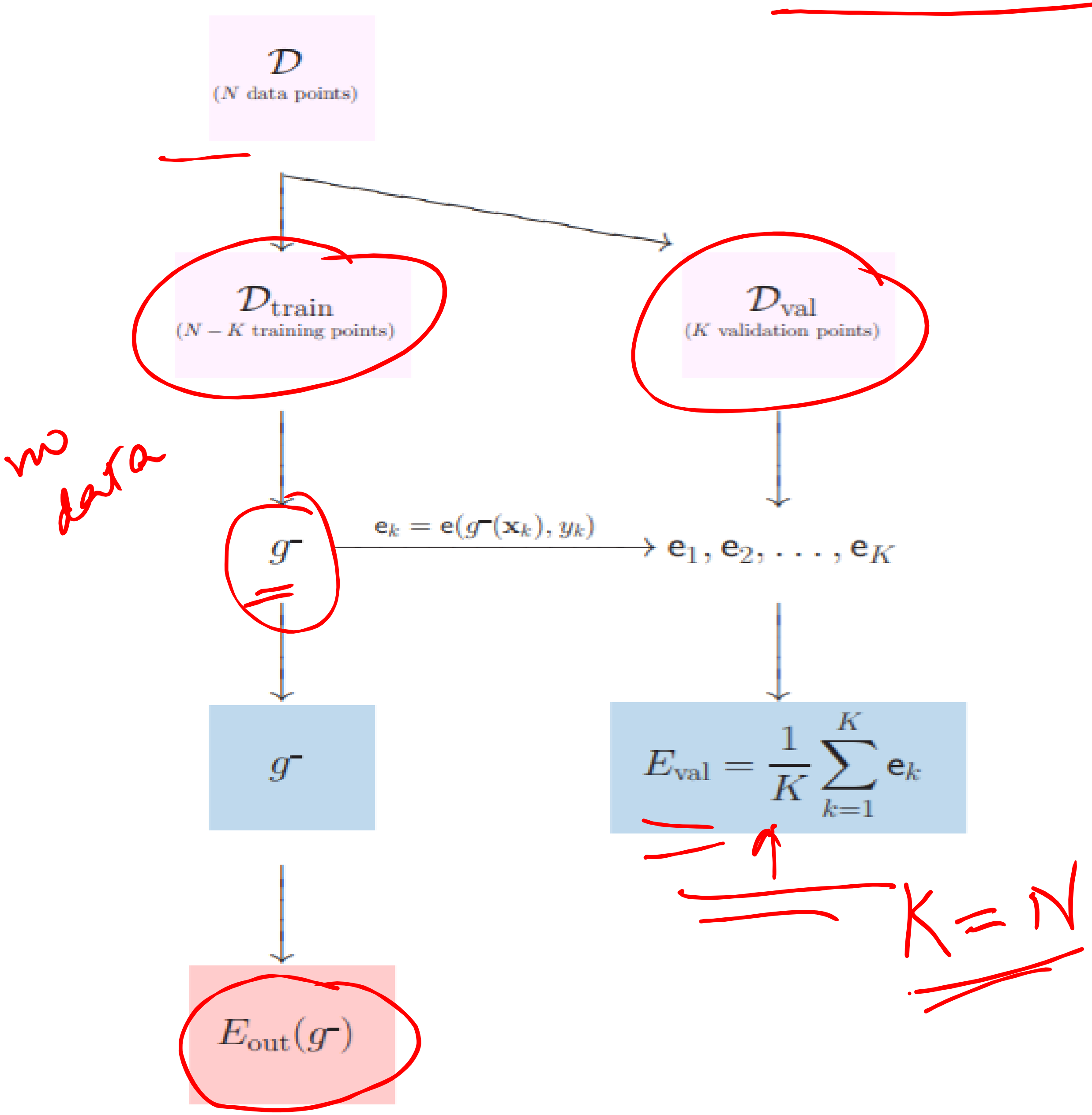
$\boxed{x_1, x_2 \dots x_K} \longrightarrow e_1(g^-), e_2(g^-) \dots e_K(g^-)$

$$E_{val} = \frac{1}{K} \sum_{k=1}^{K} e_k(g^-)$$

$$E[E_{val}] = E_{out}(g^-)$$

$$Var[E_{val}] = \left(\frac{1}{K}\right) Var(e(g^-))$$

The Validation Set



E_{val} is an estimate for $E_{\text{out}}(g^-)$

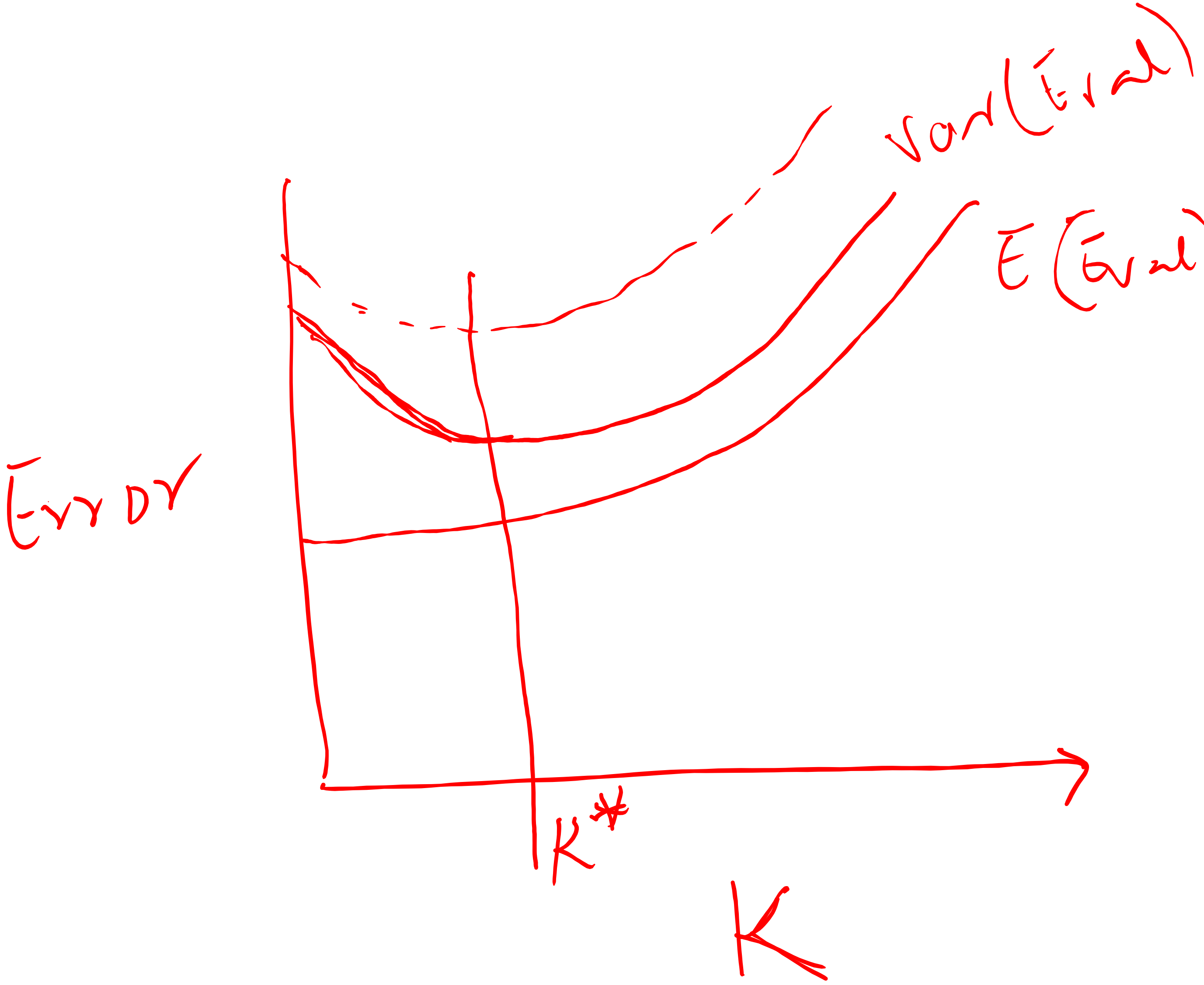
$$\mathbb{E}_{\mathcal{D}_{\text{val}}}[e_k] = E_{\text{out}}(g^-)$$

$$\begin{aligned}\mathbb{E}[E_{\text{test}}] &= \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[e_k] \\ &= \frac{1}{K}\sum_{k=1}^{K}E_{\text{out}}(g^-) = E_{\text{out}}(g^-)\end{aligned}$$

$e_1, \ldots, e_K$ are *independent*

$$\begin{aligned}\text{Var}[E_{\text{val}}] &= \frac{1}{K^2}\sum_{k=1}^{K}\text{Var}[e_k] \\ &= \frac{1}{K}\text{Var}[e(g^-)]\end{aligned}$$

decreases like $\frac{1}{K}$
depends on g^-, not $\mathcal{H}$
bigger $K \implies$ more reliable E_{val}?



$g^- \rightarrow$ worse

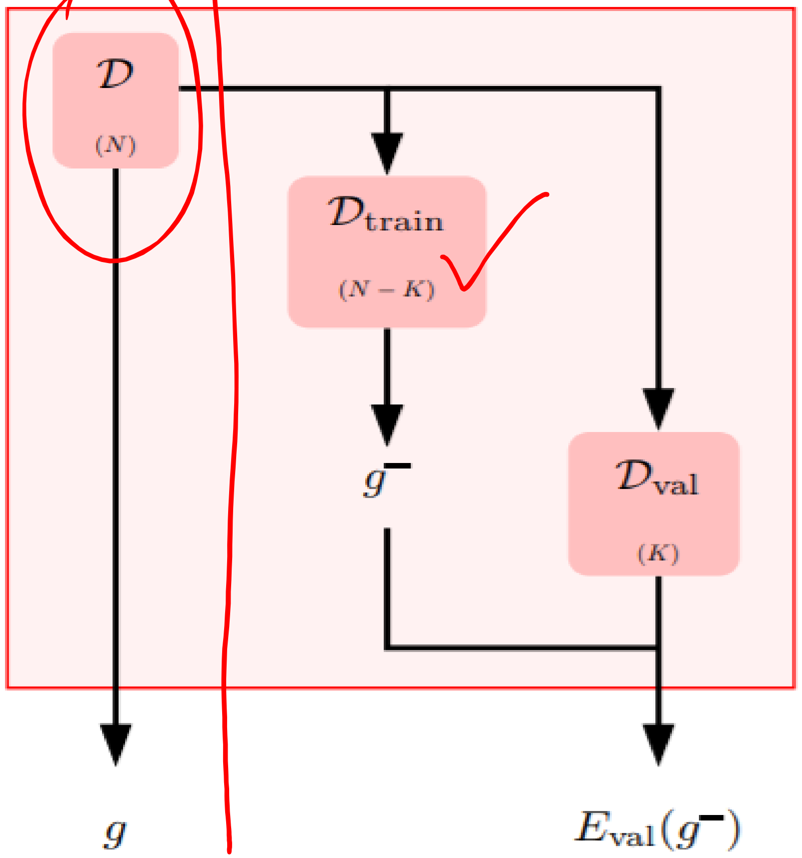
$e(g^-)$

In practice,

$K^* \simeq \frac{N}{5}$ or 20%

Behind closed doors

Restoring $\mathcal{D}$



CUSTOMER

g^-

Primary goal: output best hypothesis.

g was trained on *all* the data. $\rightarrow$ N

Secondary goal: estimate $E_{\text{out}}(g)$.

g^- is behind closed doors.

$$\underbrace{\begin{array}{cc} E_{\text{out}}(g) & E_{\text{out}}(g^-) \\ \downarrow & \downarrow \\ E_{\text{in}}(g) & E_{\text{val}}(g^-) \end{array}}_{\text{which should we use?}}$$

E

E_{val} vs E_{in}

$$E_{out}(g) \le E_{in}(g) + O\left(\sqrt{\frac{d_{vc}\ln N}{N}}\right)$$

H-set

$$E_{out}(g) \le E_{out}(g^{-}) \le E_{val} + O\left(\frac{1}{\sqrt{K}}\right)$$

single hypothesis

Model Selection

- $D_N \rightarrow D_{N-K}$ (train) D_{train}

  $D_N \rightarrow D_K$ (validation) D_{val}

H_1 $\quad$ H_2

$\boxed{M_1}$ $\quad$ $\boxed{M_2}$ $\quad$ $\boxed{M_3}$ $\quad$ - - - $\quad$ $\boxed{M_M}$

D_{train}

g_1 $\quad$ g_2 $\quad$ g_3 $\quad$ g_M

$E_{val} \rightarrow E_{out}(g_1)$

$$
\begin{array}{lccccc}
 & \boxed{\mathcal{H}_1} & \boxed{\mathcal{H}_2} & \boxed{\mathcal{H}_3} & \cdots & \boxed{\mathcal{H}_M} \\
 & \downarrow & \downarrow & \downarrow & & \downarrow \\
\underline{D_{train}} & g_1^- & g_2^- & \underline{\underline{g_3^-}} & & g_M^- \; \Big\} \text{ finalists} \\
 & \downarrow & \downarrow & \downarrow & & \downarrow \\
\underline{\underline{D_{val}}} & E_1 & E_2 & g_*^- & \cdots & E_M
\end{array}
$$

$$\approx E_{out}(g_1^-) \quad \approx E_{out}(g_2^-)$$

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

finite

finite hyp. sets

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

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

λ

$\lambda_1 = 0$ $\boxed{H_1}$ $\quad$ $\lambda_2 = 0.001$ $\boxed{H_2}$ $\quad \cdots\cdots$ $\quad$ $\lambda_M = 0.1$ $\boxed{H_M}$

D_{train}

g_1^- $\quad$ g^-

D_{val}

E_1 $\quad$ E_2 ✓ $\quad$ E_M $\quad$ g

Validation (general philosophy)

$$\downarrow \qquad g \simeq g^-$$

$$g^- \; (g) \qquad K \;,\; N-K$$

$$\downarrow$$

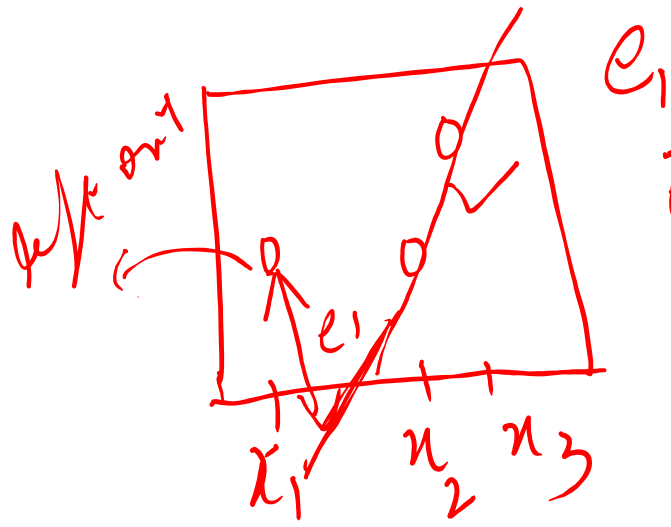
$$E_{out} \qquad E_{out}(g) \simeq E_{out}(g^-) \simeq E_{val}(g^-) + O\left(\frac{1}{\sqrt{K}}\right)$$

K small — tight

K big — tight

Can we have $K=1$?

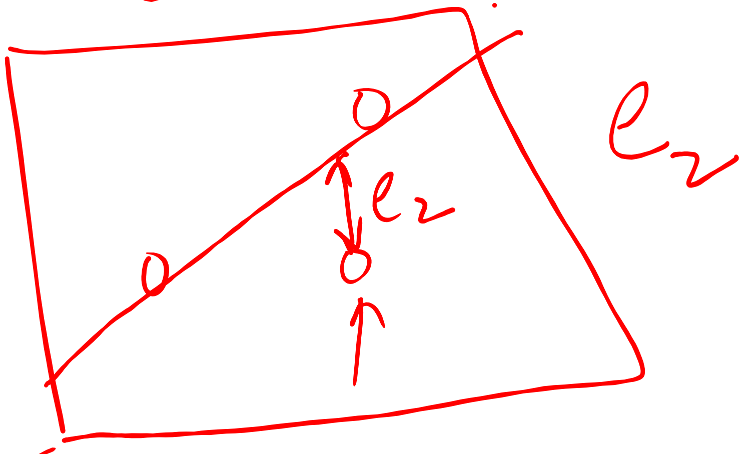
CROSS VALIDATION. (K=1?)



$E(e_1) = E_{out}(g_1^-)$

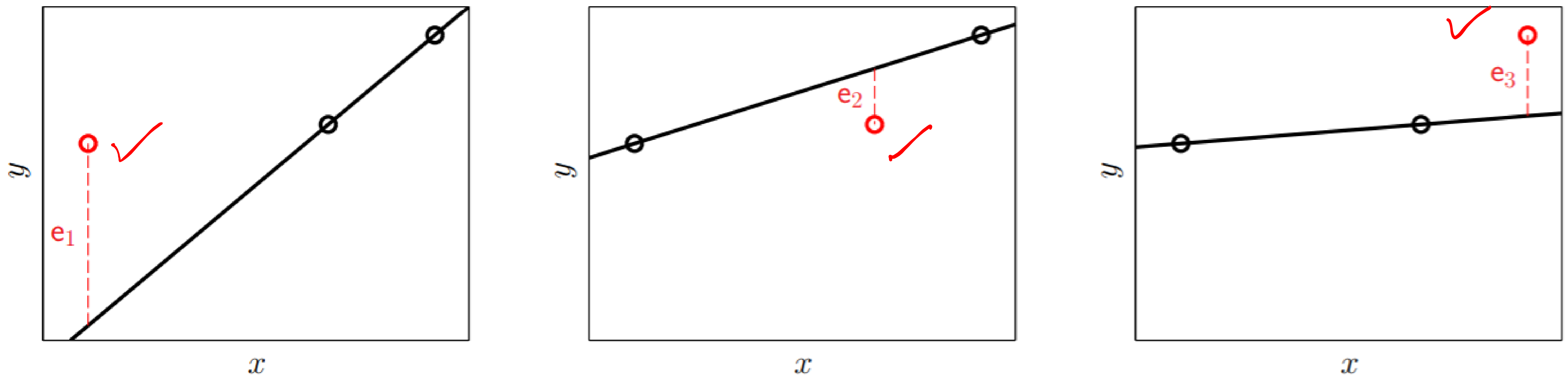
$E(e_2) = E_{out}(g_2^-)$

$E(e_3) = E_{out}(g_3^-)$



Pick all of them $\longrightarrow$ An average leave one out

$$E_{CV} = \frac{1}{N}\sum_{n=1}^{N} e_n \longrightarrow (x_n, y_n)$$



$$E_{\text{cv}} = \frac{1}{N} \sum_{n=1}^{N} \mathrm{e}_n$$

Theorem $\quad \underbrace{E_{D_N}[E_{cv}]}_{\Downarrow} = E_{out}(N-1)$

$\underbrace{E_{out}(g) \le E_{cv}}$ &

$E_{out} \le E_{cv} + O\left(\frac{1}{\sqrt{N}}\right) \rightarrow$ average

$e_1 \longrightarrow (x_1, y_1) \rightarrow g_1^-$

$e_2 \longrightarrow (x_2, y_2) \rightarrow g_2^-$

subtle way

N's $\longrightarrow e_1 \; e_2 \; e_3 \; \cdots \; e_N$

Assume almost independent.

K-fold cross validation

N-data points $\longrightarrow$ N regression problems

$\therefore$ N+1 training sets.

10 fold CV $\rightarrow$ 10% of your data

Analytically $\longrightarrow$ Regression

$E_{cv}(\lambda_1)$ $E_{cv}(\lambda_2)$ ....

→ Digits data

1. Get features : Symmetry & Intensity.
2. PLA
3. $\begin{bmatrix} 1 \\ x_1 \\ x_2 \end{bmatrix} \xrightarrow{\Phi} \left.\begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_1^5 \\ x_2^5 \end{bmatrix}\right\}$ 20 dim in Z space ✓

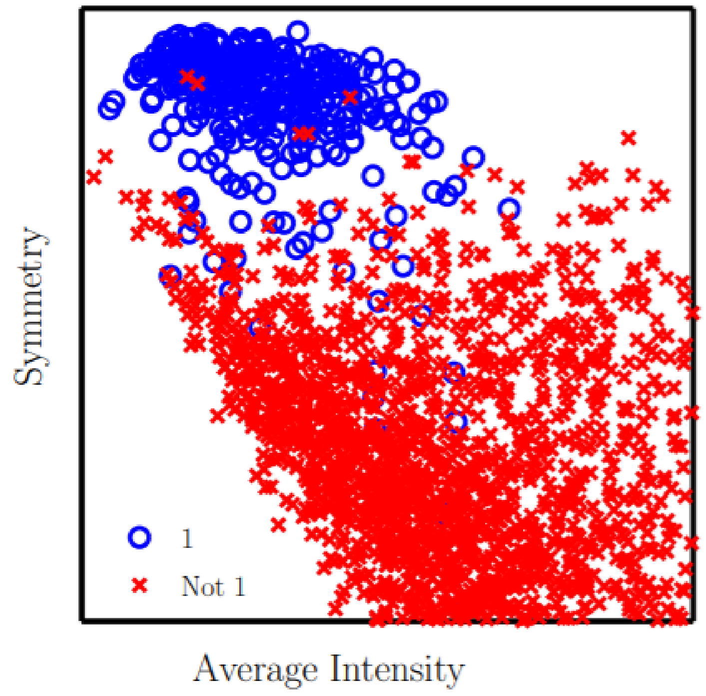
4. Model selection

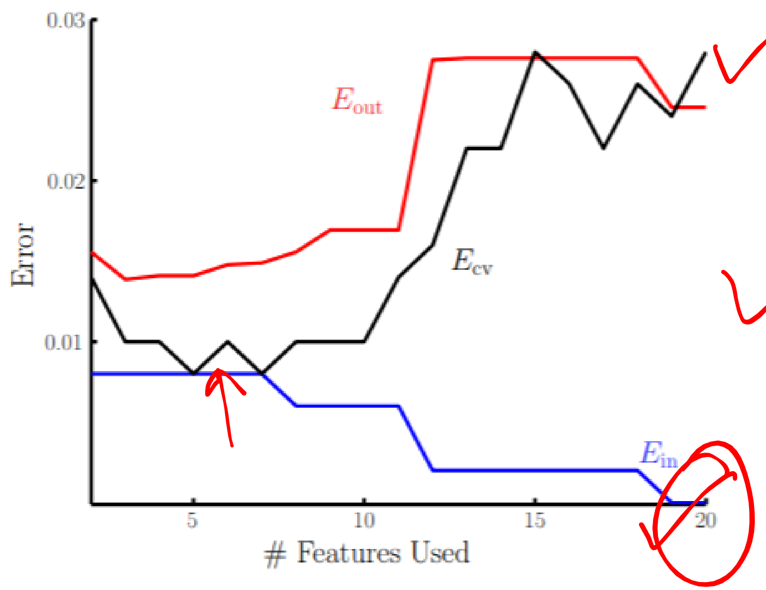
$M_1 \to \begin{bmatrix} 1 \\ z_1 \end{bmatrix}$ $M_2 \to \begin{bmatrix} 1 \\ z_1 \\ z_2 \end{bmatrix}$ - - - M_{20}

5. CV to pick M_i

Digits Problem: '1' Versus 'Not 1'

2 cases

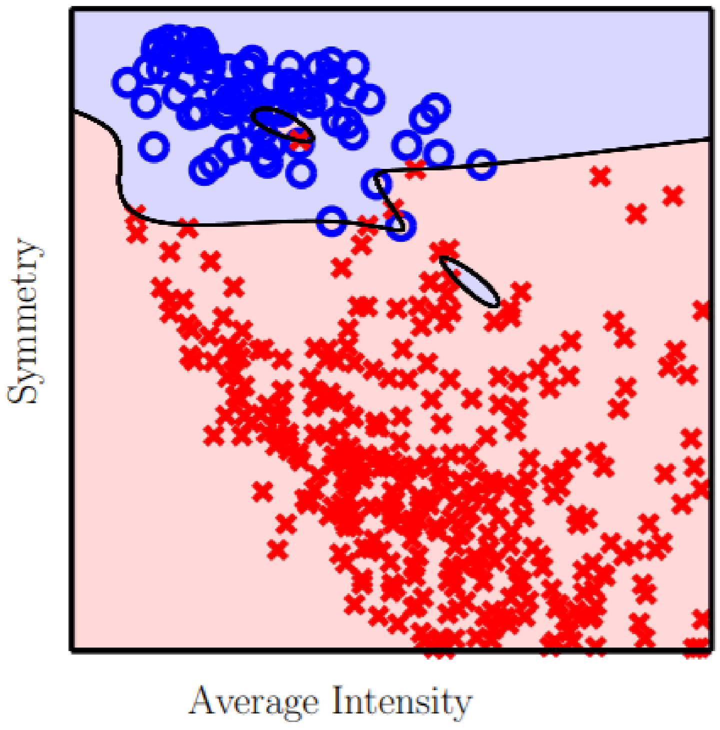




$$\mathbf{x} = (1, x_1, x_2)$$

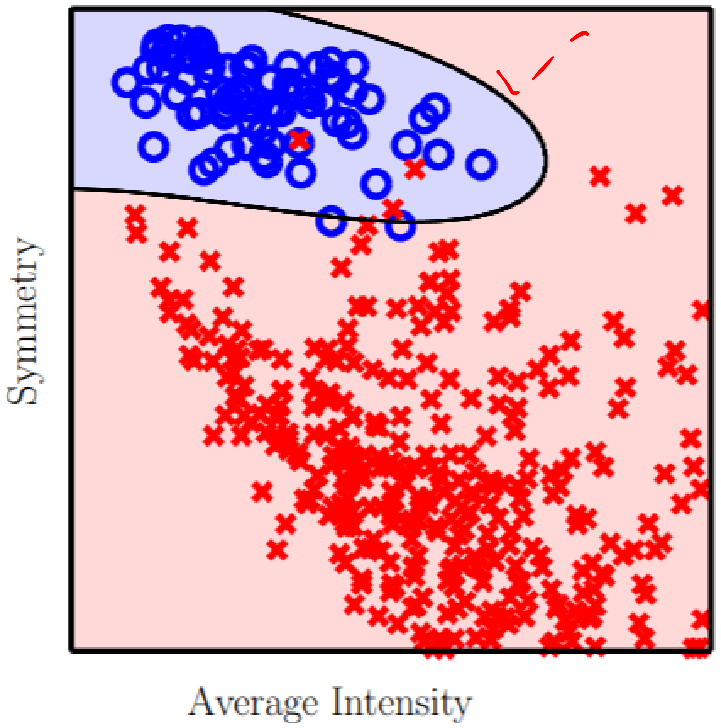
$$\mathbf{z} = \underbrace{(1, x_1, x_2, x_1^2, x_1x_2, x_2^2, x_1^3, x_1^2x_2, x_1x_2^2, x_2^3, \ldots, x_1^5, x_1^4x_2, x_1^3x_2^2, x_1^2x_2^3, x_1x_2^4, x_2^5)}_{\text{5th order polynomial transform} \longrightarrow \text{20 dimensional non linear feature space}}$$

Validation Wins In the Real World



no validation (20 features)

$E_{in} = 0\%$
$E_{out} = 2.5\%$



cross validation (6 features)

$E_{in} = 0.8\%$
$E_{out} = \mathbf{1.5\%}$

Thanks!