Learning From Data
Lecture 13
Validation and Model Selection

The Validation Set
Model Selection
Cross Validation

M. Magdon-Ismail
CSCI 4100/6100

RECAP: Regularization

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

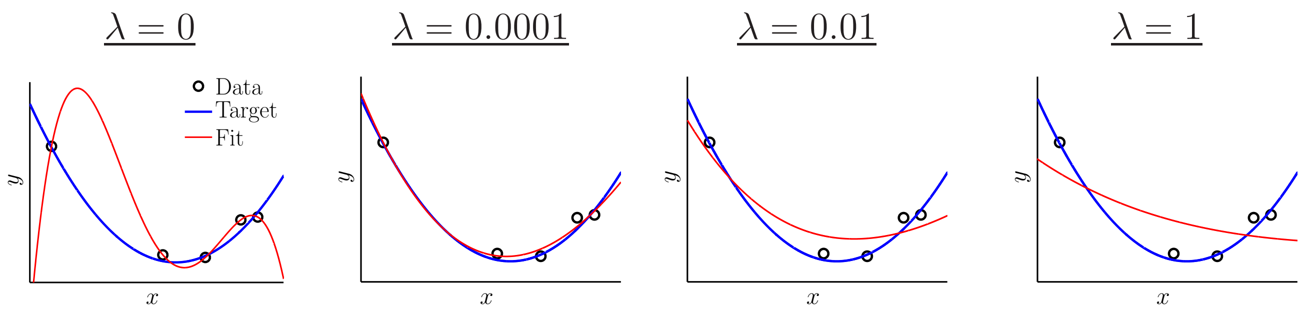
$$E_{\text{aug}}(h) = E_{\text{in}}(h) + \frac{\lambda}{N}\Omega(h)$$

$\Omega(h) \rightarrow$ 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: A Sneak Peek at E_{out}

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

VC bounds this using a complexity error bar for $\mathcal{H}$

regularization estimates this through a heuristic complexity penalty for 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.

The Test Set

$$
\begin{tikzcd}
\boxed{\begin{array}{c}\mathcal{D}\\ (N \text{ data points})\end{array}} \arrow[d] & & \boxed{\begin{array}{c}\mathcal{D}_{\text{test}}\\ (K \text{ test points})\end{array}} \arrow[d] \\
g \arrow[d] \arrow[rr, "\mathsf{e}_k = \mathsf{e}(g(\mathbf{x}_k), y_k)"] & & \mathsf{e}_1, \mathsf{e}_2, \ldots, \mathsf{e}_K \arrow[d] \\
\boxed{g} \arrow[d] & & \boxed{E_{\text{test}} = \frac{1}{K}\sum_{k=1}^{K} \mathsf{e}_k} \\
\boxed{E_{\text{out}}(g)} & &
\end{tikzcd}
$$

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

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

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

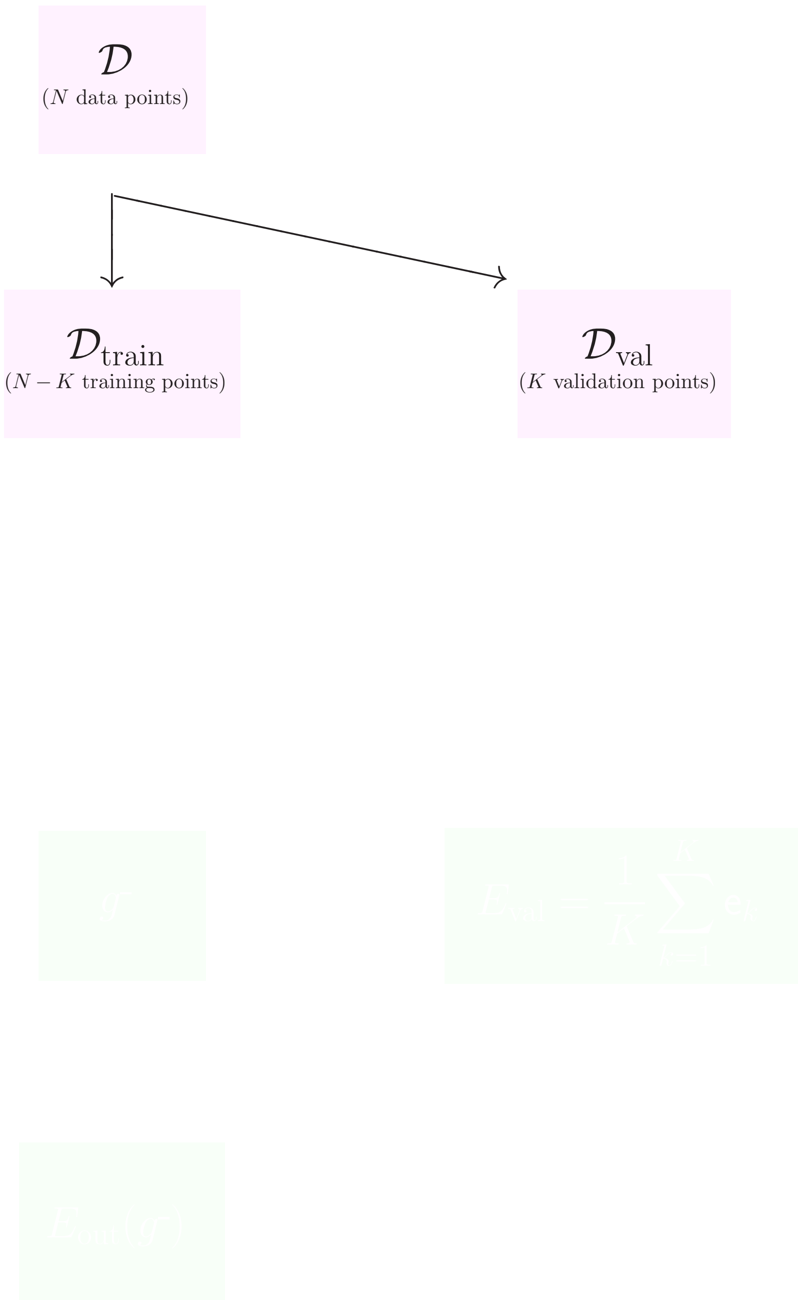
$\mathsf{e}_1, \ldots, \mathsf{e}_K$ are *independent*

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

decreases like $\frac{1}{K}$

bigger $K \implies$ more reliable E_{test}.

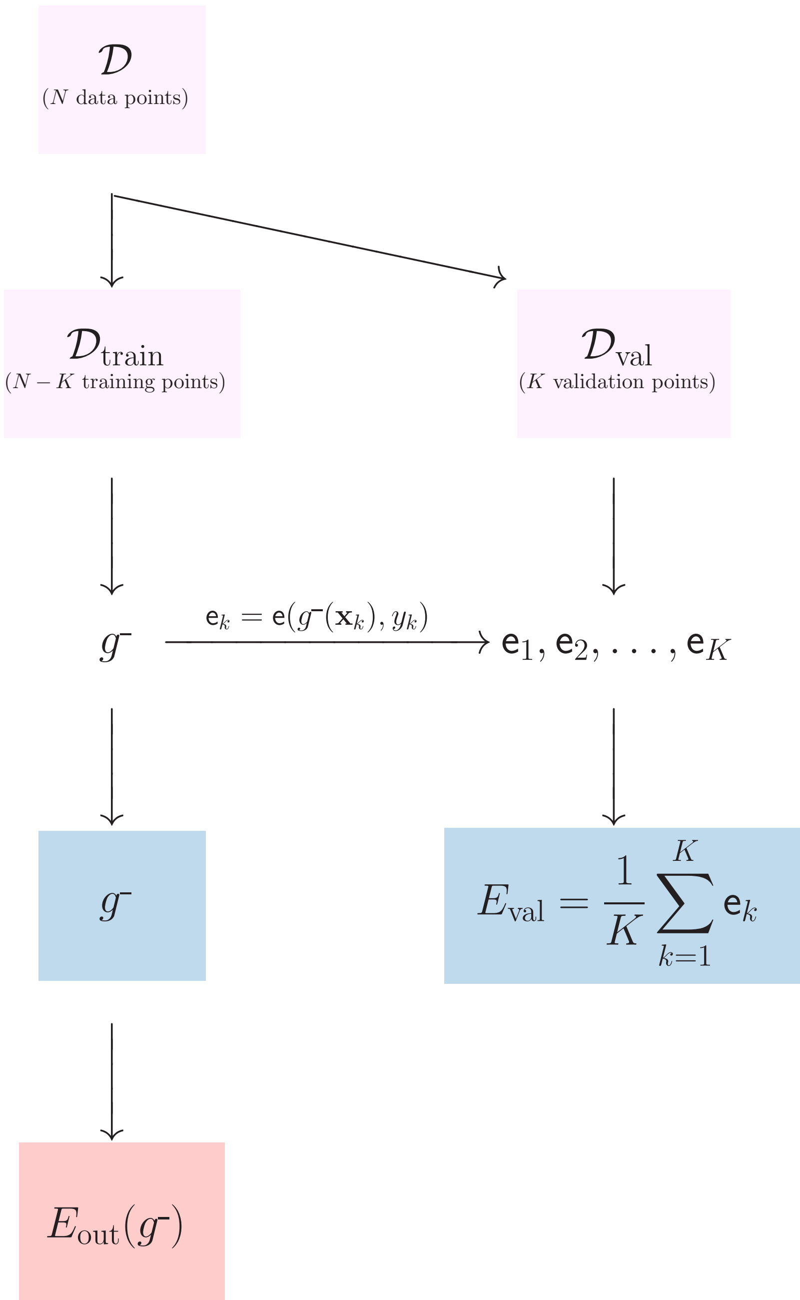
The Validation Set



1. Remove K points from $\mathcal{D}$

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

The Validation Set



1. Remove K points from $\mathcal{D}$

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

2. Learn using $\mathcal{D}_{\text{train}} \longrightarrow g^-$.

3. Test g^- on $\mathcal{D}_{\text{val}} \longrightarrow E_{\text{val}}$.

4. Use error E_{val} to estimate $E_{\text{out}}(g^-)$.

The Validation Set

$$
\begin{tikzcd}
\underset{(N\text{ data points})}{\mathcal{D}} \arrow[d] \arrow[dr] & \\
\underset{(N-K\text{ training points})}{\mathcal{D}_{\text{train}}} \arrow[d] & \underset{(K\text{ validation points})}{\mathcal{D}_{\text{val}}} \arrow[d] \\
g^{-} \arrow[r, "\mathsf{e}_k = \mathsf{e}(g^{-}(\mathbf{x}_k), y_k)"] \arrow[d] & \mathsf{e}_1, \mathsf{e}_2, \ldots, \mathsf{e}_K \arrow[d] \\
\boxed{g^{-}} \arrow[d] & \boxed{E_{\text{val}} = \frac{1}{K}\sum_{k=1}^{K} \mathsf{e}_k} \\
\boxed{E_{\text{out}}(g^{-})} &
\end{tikzcd}
$$

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

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

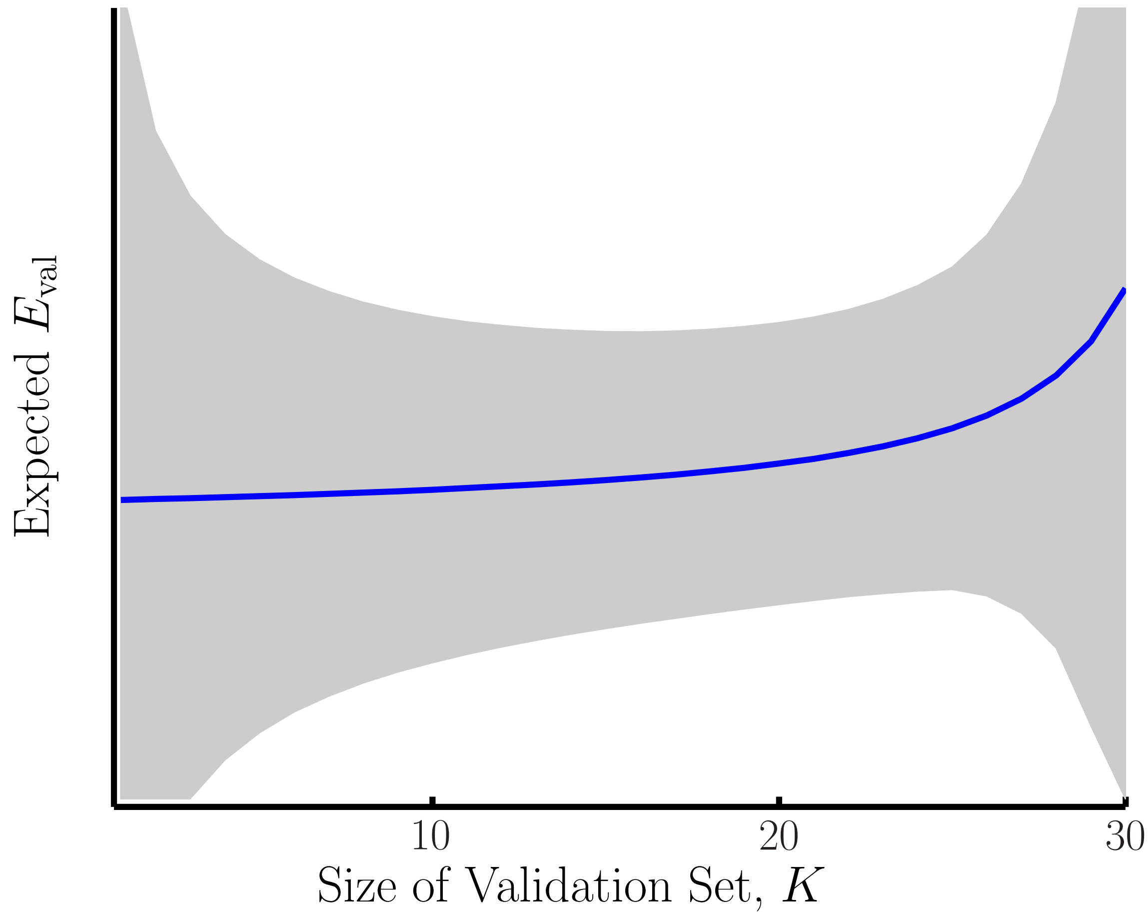
$$
\begin{aligned}
\mathbb{E}[E_{\text{test}}] &= \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}[\mathsf{e}_k] \\
&= \frac{1}{K}\sum_{k=1}^{K} E_{\text{out}}(g^{-}) = E_{\text{out}}(g^{-})
\end{aligned}
$$

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

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

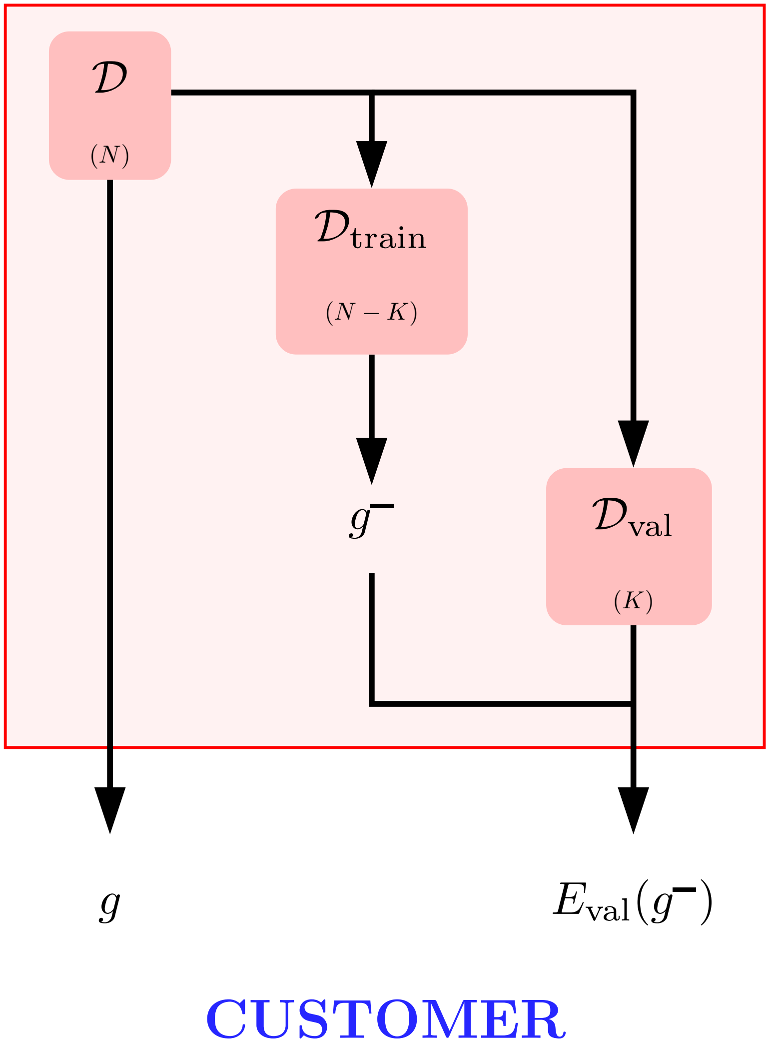
decreases like $\frac{1}{K}$
depends on g^{-}, not $\mathcal{H}$
bigger $K \implies$ more reliable E_{val}?

Choosing K



Rule of thumb: $K^* = \frac{N}{5}$.

Restoring $\mathcal{D}$



Primary goal: output best hypothesis.

g was trained on *all* the data.

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

g^- is behind closed doors.

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

E_{val} Versus E_{in}

Biased error bar depends on $\mathcal{H}$.

$$E_{\text{out}}(g) \;\leq\; E_{\text{in}}(g) + O\left(\sqrt{\frac{d_{\text{VC}}}{N}\log N}\right)$$

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

learning curve is decreasing
(a practical truth, not a theorem)

Unbiased error bar depends on g^{-}.

$E_{\text{val}}(g)$ usually wins as an estimate for $E_{\text{out}}(g)$, especially when the learning curve is not steep.

Model Selection

The most important use of validation

$$
\begin{array}{ccccc}
\boxed{\mathcal{H}_1} & \boxed{\mathcal{H}_2} & \boxed{\mathcal{H}_3} & \cdots & \boxed{\mathcal{H}_M} \\
\downarrow & \downarrow & \downarrow & & \downarrow \\
g_1 & g_2 & g_3 & \cdots & g_M
\end{array}
$$

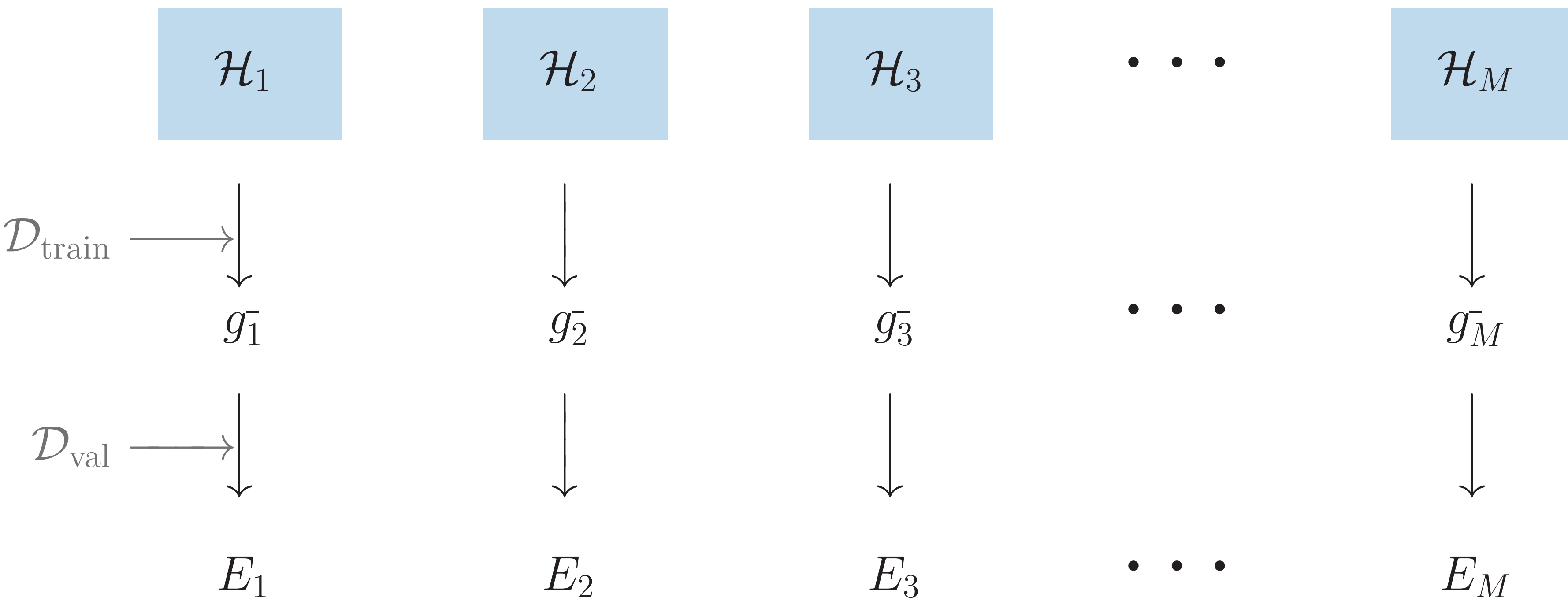
The most important use of validation

Validation Estimate for $(\mathcal{H}_1, g_1)$

The most important use of validation

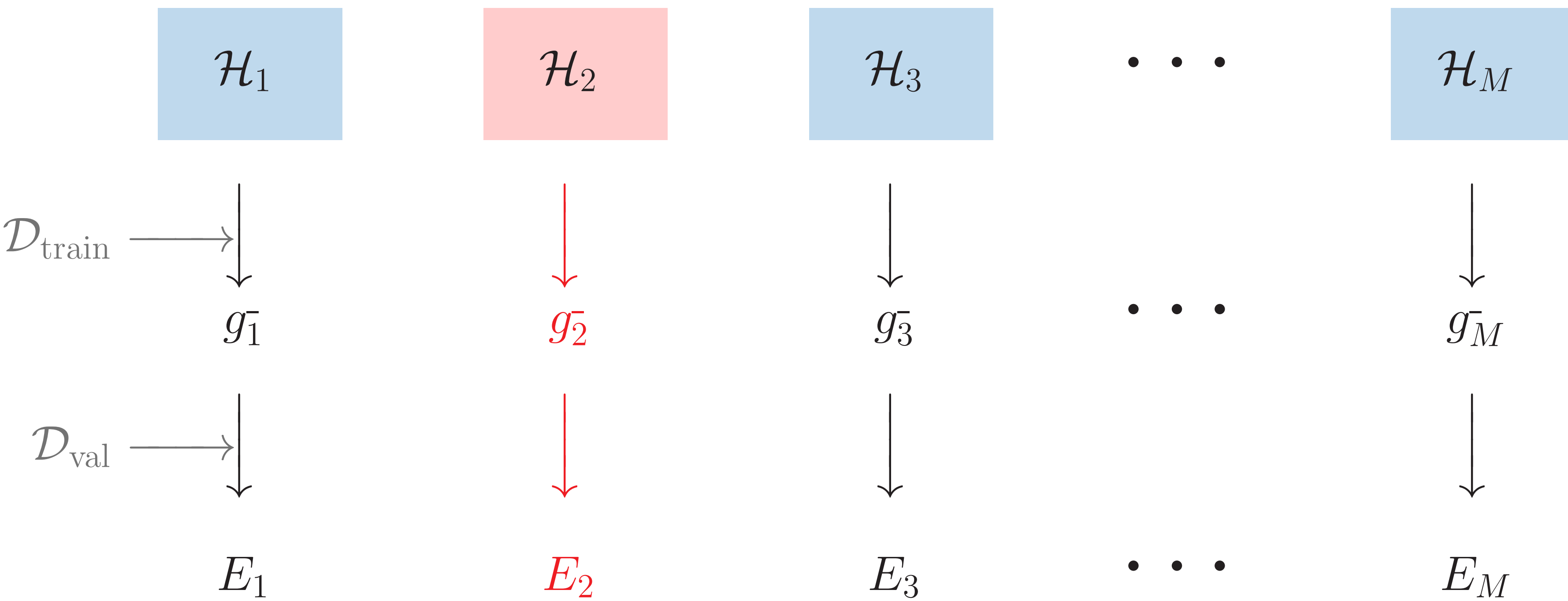
Compute Validation Estimates for All Models

The most important use of validation

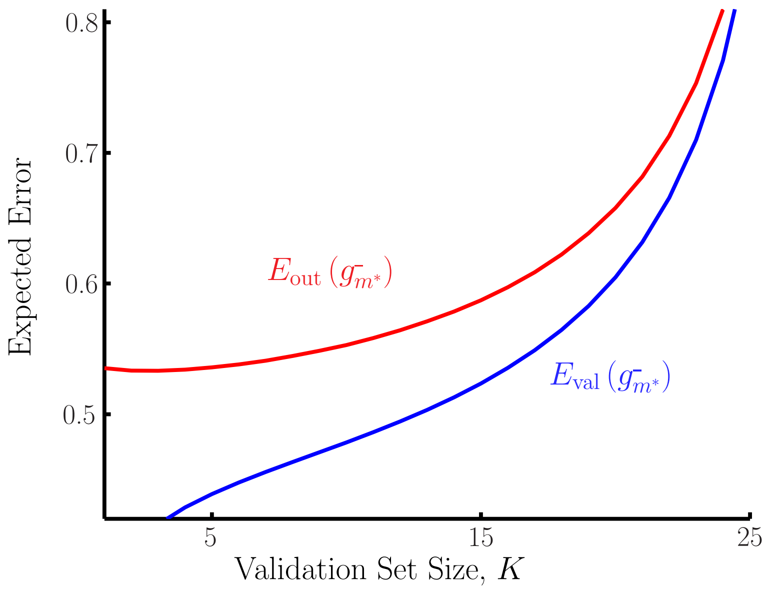


Pick The Best Model According to Validation Error

The most important use of validation



$E_{\text{val}}(g^-_{m^*})$ is not Unbiased For $E_{\text{out}}(g^-_{m^*})$



. . . because we *choose* one of the M finalists.

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

$\uparrow$

VC error bar for selecting a hypothesis from M using a data set of size K.

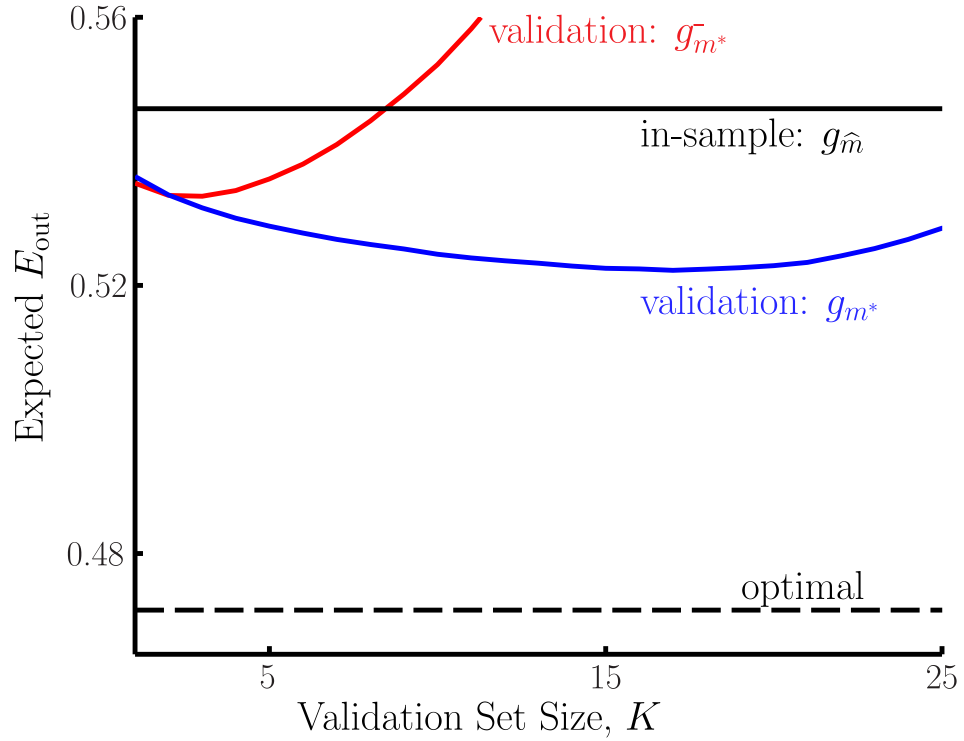
Restoring $\mathcal{D}$

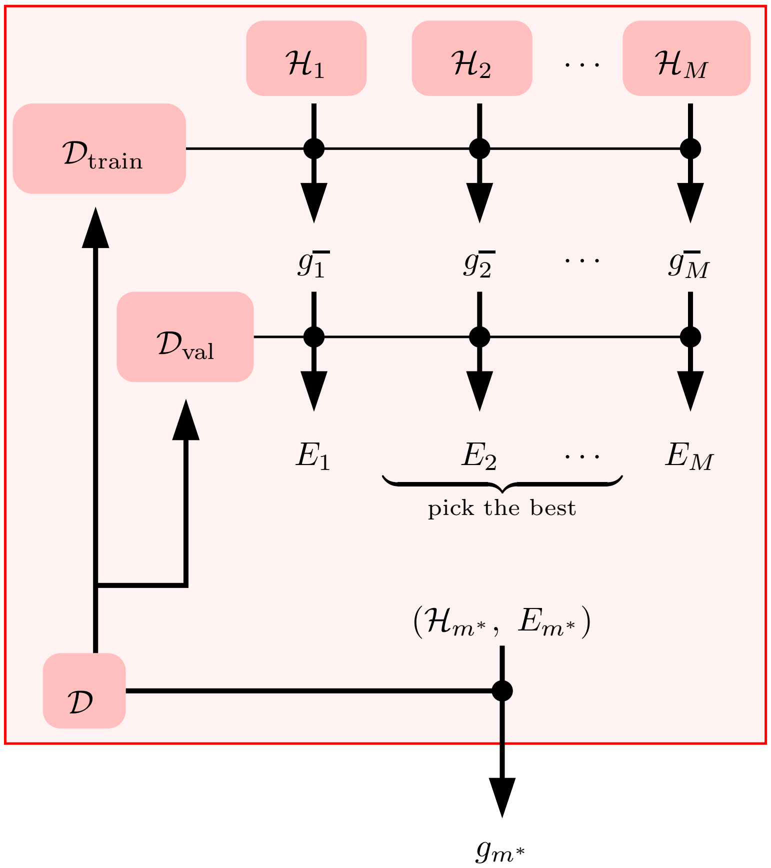
$$
\begin{tikzcd}
\mathcal{H}_1 \arrow[d] & \mathcal{H}_2 \arrow[d] & \mathcal{H}_3 \arrow[d] & \cdots & \mathcal{H}_M \arrow[d] \\
g_1 & g_2 & g_3 & \cdots & g_M
\end{tikzcd}
$$

Model with best g also has best g^- ← leap of faith

We can find model with best g^- using validation ← true modulo E_{val} error bar

Comparing E_{in} and E_{val} for Model Selection





Application to Selecting λ

Which regularization parameter to use?

$$\lambda_1, \lambda_2, \ldots, \lambda_M.$$

This is a special case of *model selection* over M models,

$$\begin{array}{ccccc} (\mathcal{H}, \lambda_1) & (\mathcal{H}, \lambda_2) & (\mathcal{H}, \lambda_3) & \cdots & (\mathcal{H}, \lambda_M) \\ \downarrow & \downarrow & \downarrow & & \downarrow \\ g_1 & g_2 & g_3 & \cdots & g_M \end{array}$$

Picking a model amounts to chosing the optimal λ

The Dilemma When Choosing K

Validation relies on the following chain of reasoning,

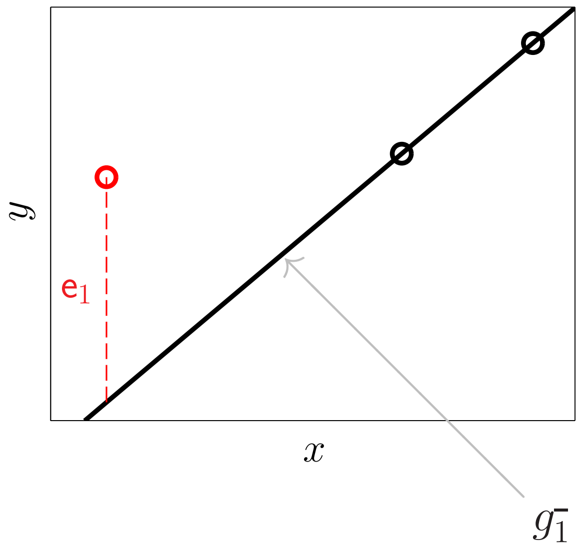
$$E_{\text{out}}(g) \approx E_{\text{out}}(g^-) \approx E_{\text{val}}(g^-)$$

(small K) (large K)

Can we get away with $K = 1$?

Yes, almost!

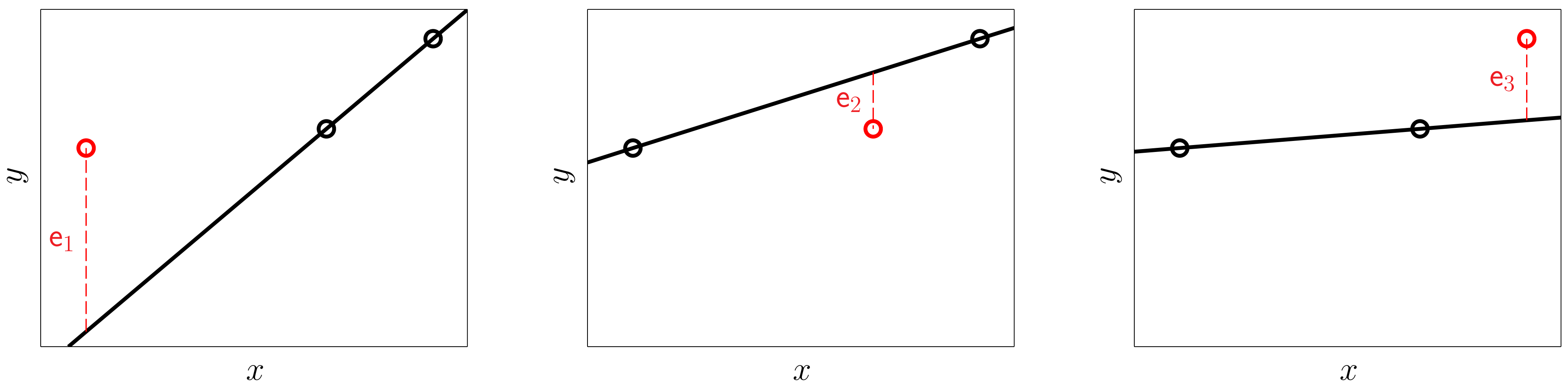
The Leave One Out Error $(K = 1)$



$$\mathbb{E}[e_1] = E_{\text{out}}(g_1^-)$$

. . . but it is a **wild** estimate

The Leave One Our Errors



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

Cross Validation is Unbiased

Theorem. E_{cv} is an unbiased estimate of $\bar{E}_{\text{out}}(N-1)$.

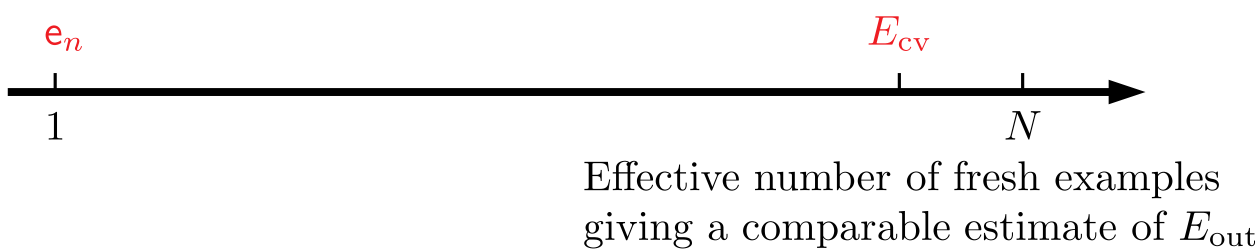
Expected E_{out} when learning with $N-1$ points.

Reliability of E_{cv}

e_n and e_m are not independent.

e_n depends on g_n^- which was trained on $(\mathbf{x}_m, y_m)$.

e_m is evaluated on $(\mathbf{x}_m, y_m)$.



Cross Validation is Computationally Intensive

N epochs of learning each on a data set of size $N-1$.

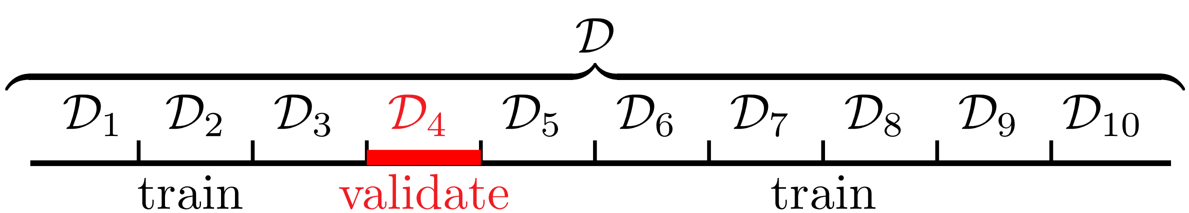
- Analytic approaches, for example linear regression with weight decay

$$\mathbf{w}_{\text{reg}} = (\mathrm{Z}^{\mathrm{T}}\mathrm{Z} + \lambda\mathrm{I})^{-1}\mathrm{Z}^{\mathrm{T}}\mathbf{y}$$

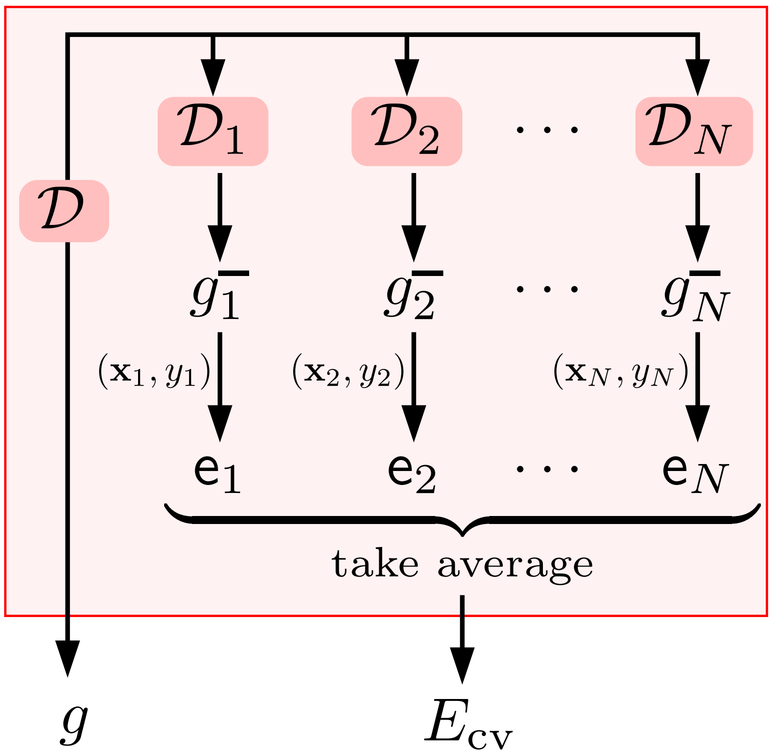
$$E_{\text{cv}} = \frac{1}{N}\sum_{n=1}^{N}\left(\frac{\hat{y}_n - y_n}{1 - \mathrm{H}_{nn}(\lambda)}\right)^2$$

$\mathrm{H}(\lambda) = \mathrm{Z}(\mathrm{Z}^{\mathrm{T}}\mathrm{Z} + \lambda\mathrm{I})^{-1}\mathrm{Z}^{\mathrm{T}}$.

- 10-fold cross validation



Restoring $\mathcal{D}$

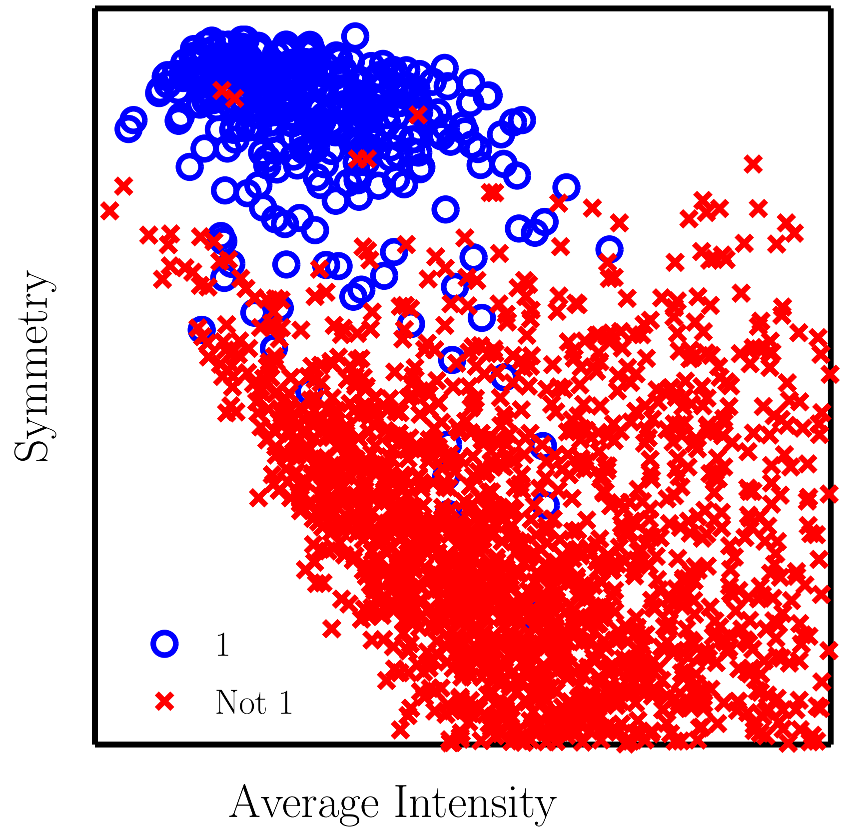


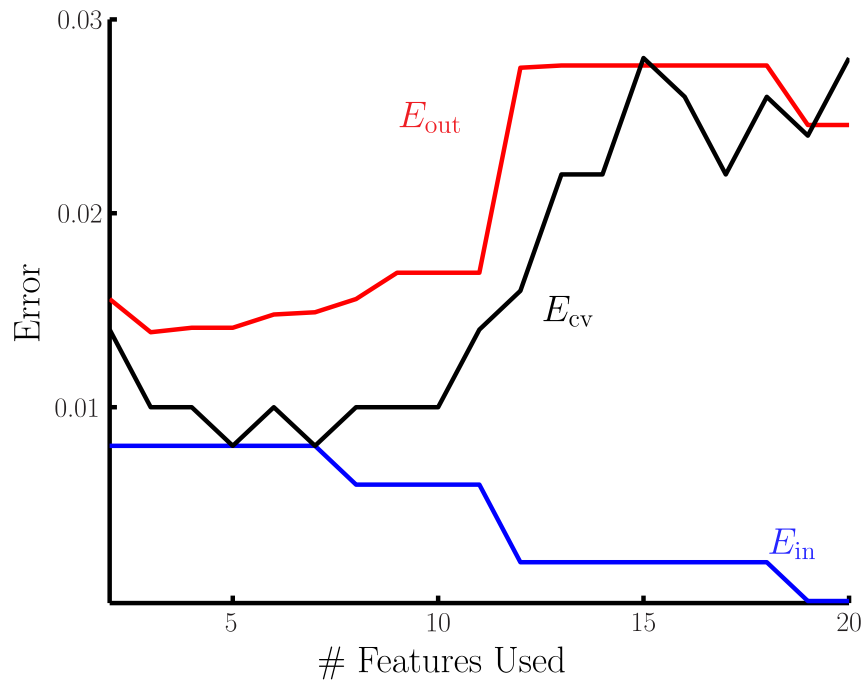
CUSTOMER

$$E_{\text{out}}(g^{(N)}) \underset{\substack{\uparrow \\ \text{learning curve}}}{\leq} \bar{E}_{\text{out}}(N-1) \leq E_{\text{cv}} + \underset{\substack{\uparrow \\ \text{nearly independent } \mathsf{e}_n}}{O\left(\frac{1}{\sqrt{N}}\right)}.$$

E_{cv} can be used for model selection just as E_{val}, for example to choose λ.

Digits Problem: '1' Versus 'Not 1'

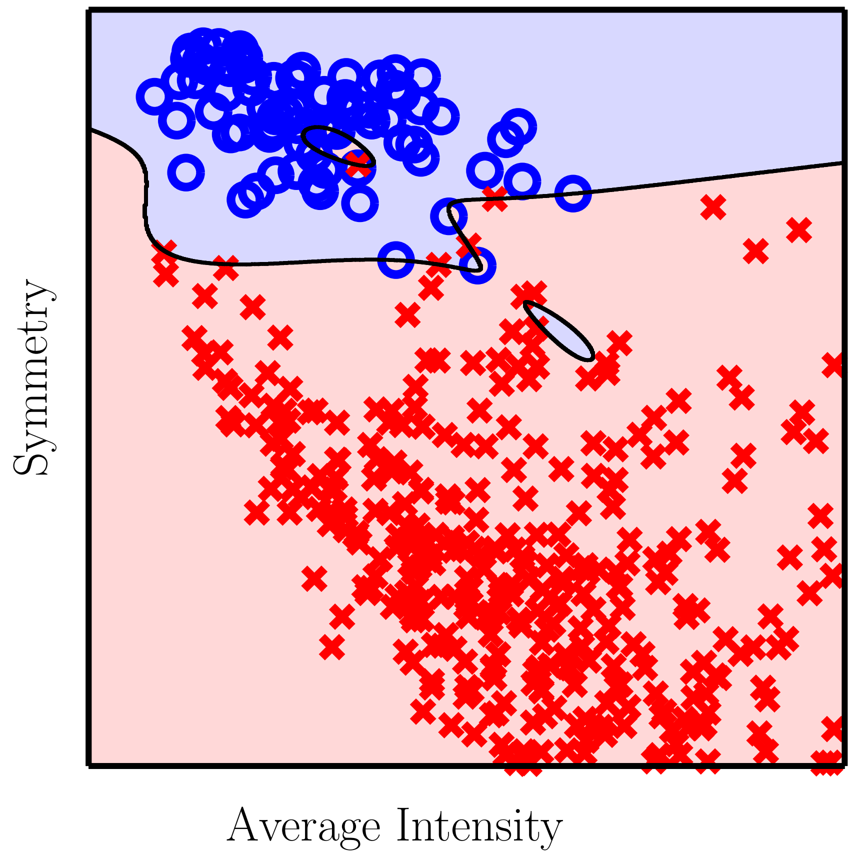




$$\mathbf{x} = (1, x_1, x_2)$$

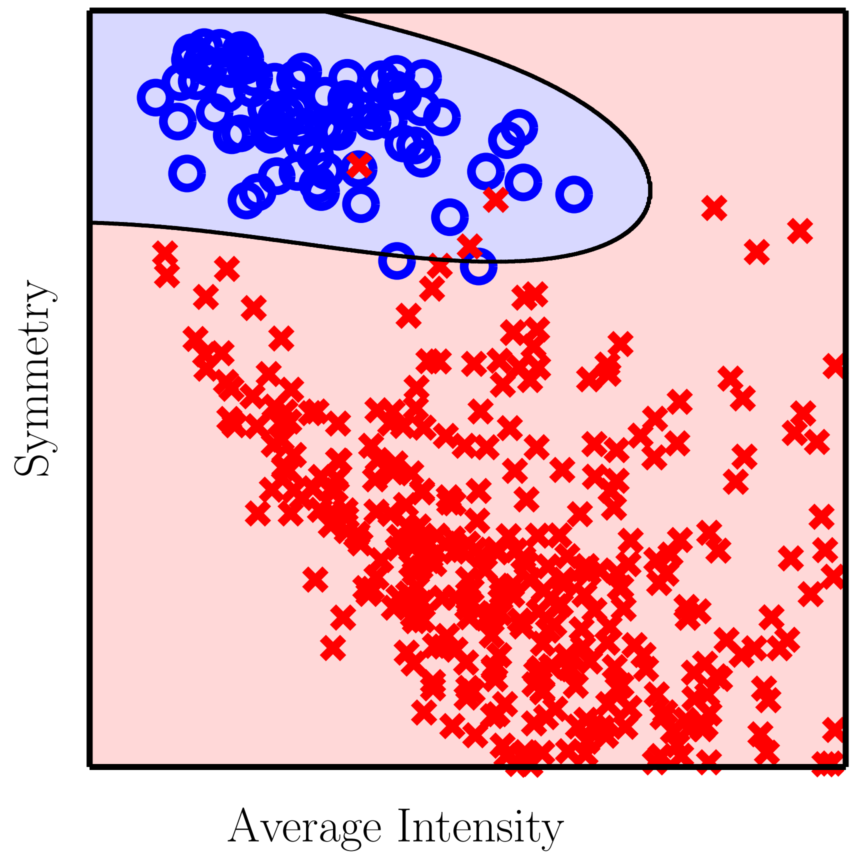
$$\mathbf{z} = \underbrace{(1, x_1, x_2, x_1^2, x_1 x_2, x_2^2, x_1^3, x_1^2 x_2, x_1 x_2^2, x_2^3, \ldots, x_1^5, x_1^4 x_2, x_1^3 x_2^2, x_1^2 x_2^3, x_1 x_2^4, x_2^5)}_{\text{5th order polynomial transform } \longrightarrow \text{ 20 dimensional non linear feature space}}$$

Validation Wins In the Real World



no validation (20 features)

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



cross validation (6 features)

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