FoDA L13 : Cross Validation

Data (x, y) $X \in \mathbb{R}^{n \times d}$ $(d=1)$ $y \in \mathbb{R}^n$

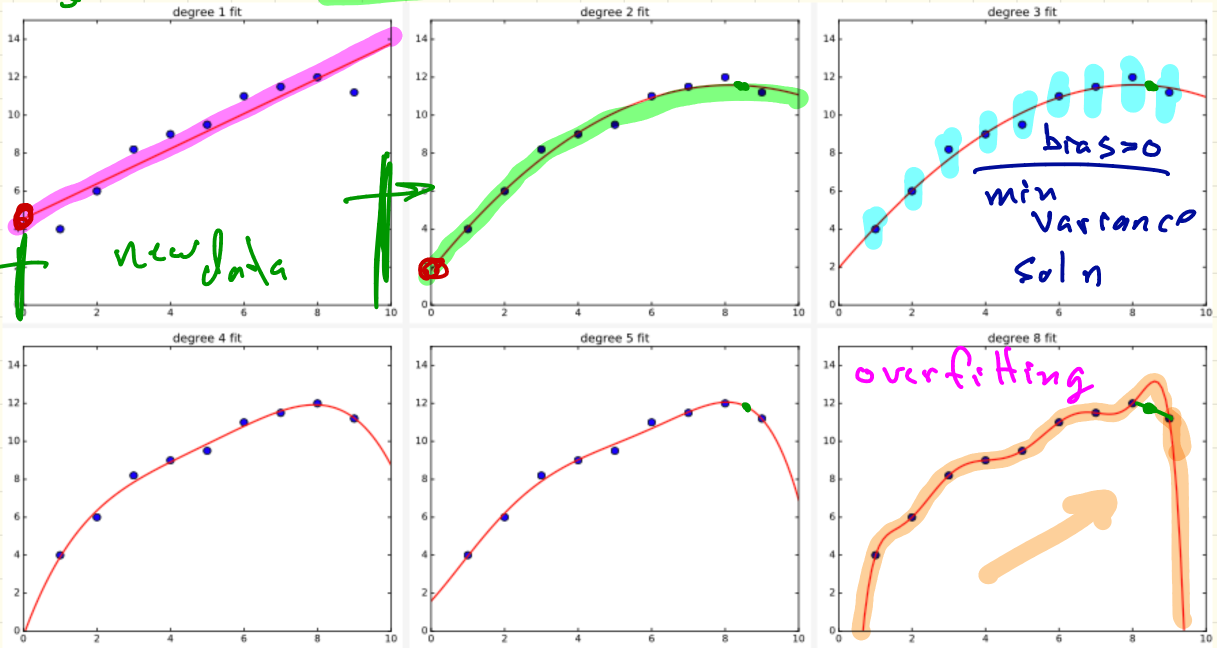
$\hat{y}_i = M_\alpha(x_i) \longrightarrow y_i$

polynominal model

residual $\hat{y} - y_i$

$$M^{(p)}_\alpha(x) = \sum_{j=0}^{p} \alpha_j x^P = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \dots$$

$$SSE = \sum_i (r_i)^2$$



Goal: Make good predictions on new unseen data.

modeling as residual $r_i = |M_\alpha(x_i) - y_i|$

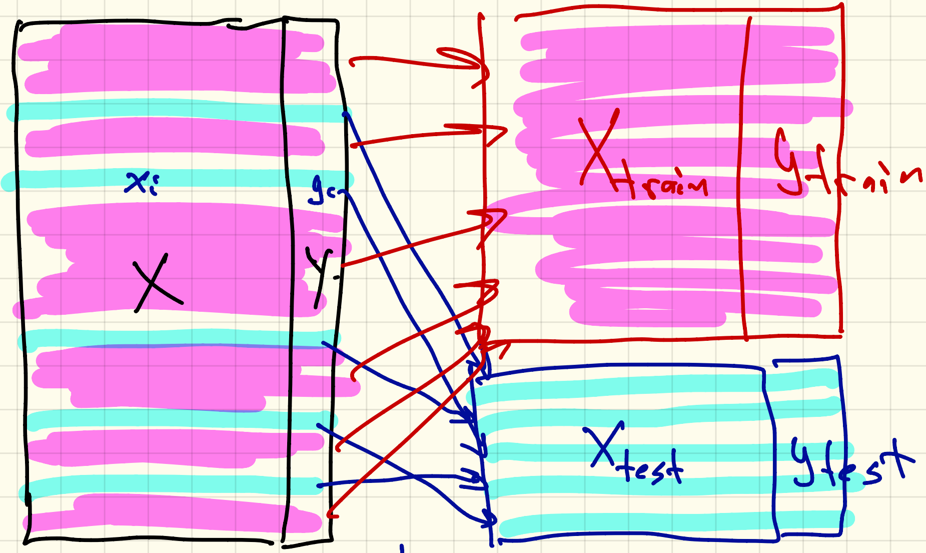
"generalization"

- Collect new data
  ↳ try on this new data

- "Save" some data for testing

Data

assumption each (x_i, y_i) iid f



random hold out

$$\alpha = (X_{tr}^T X_{tr})^{-1} X_{tr}^T y_{tr}$$

$\downarrow$

train eval

$$SSE[(x_{test}, y_{test}), \alpha_{train}] = \sum_{(x_i, y_i) \in (x_{test}, y_{test})} \left(M_{\alpha_{train}}(x_i) - y_i \right)^2$$

How well will work on new data?

$$\sqrt{\frac{1}{|X_{test}|} SSE((X_{test}, y_{test}), \alpha_{train})}$$

RMSE

How large should the test set be?

<u>Common</u> test size

- 10%
- 33%

Evaluate expected value of error by averaging n_{test} observations.

↳ CLT

→ unbiased

→ variance $\frac{Var}{n_{test}}$

more data
less test percentage

---

more complex model
more test size

What is cross-validation used for?

- see how model generalizes to new data.

or

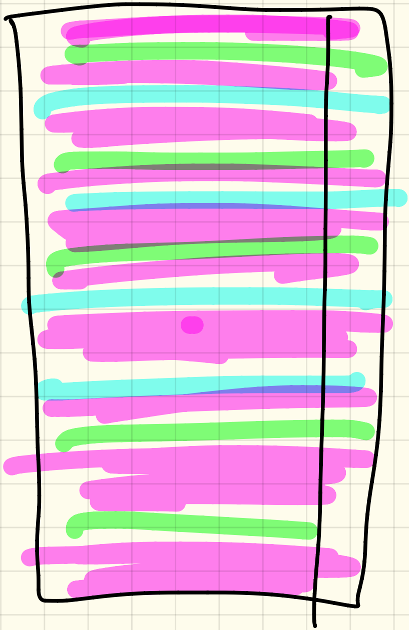
- to select a parameter in model (ex. p in $M_\alpha^{(p)}$)

$$p^* = \underset{p\in[1...8]}{\arg\min}\ SSE\left((X_{test}, y_{test}), M^{(p)}_{\alpha_{train}}\right)$$

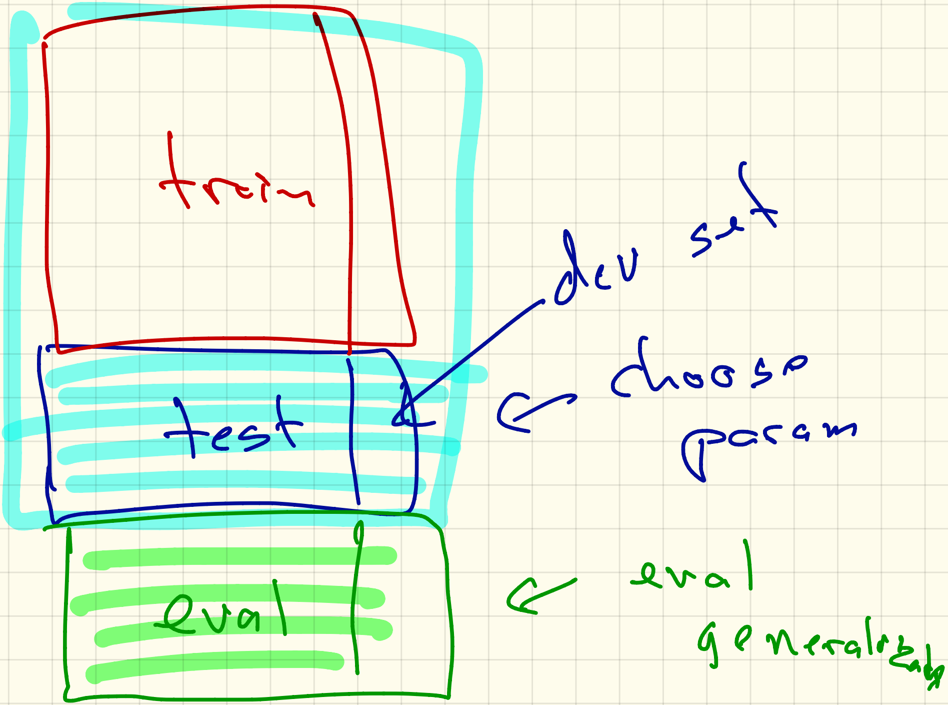
choose best

train & model

not both



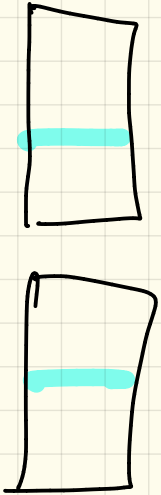
Ok!



choose param
+
evaluate generalization

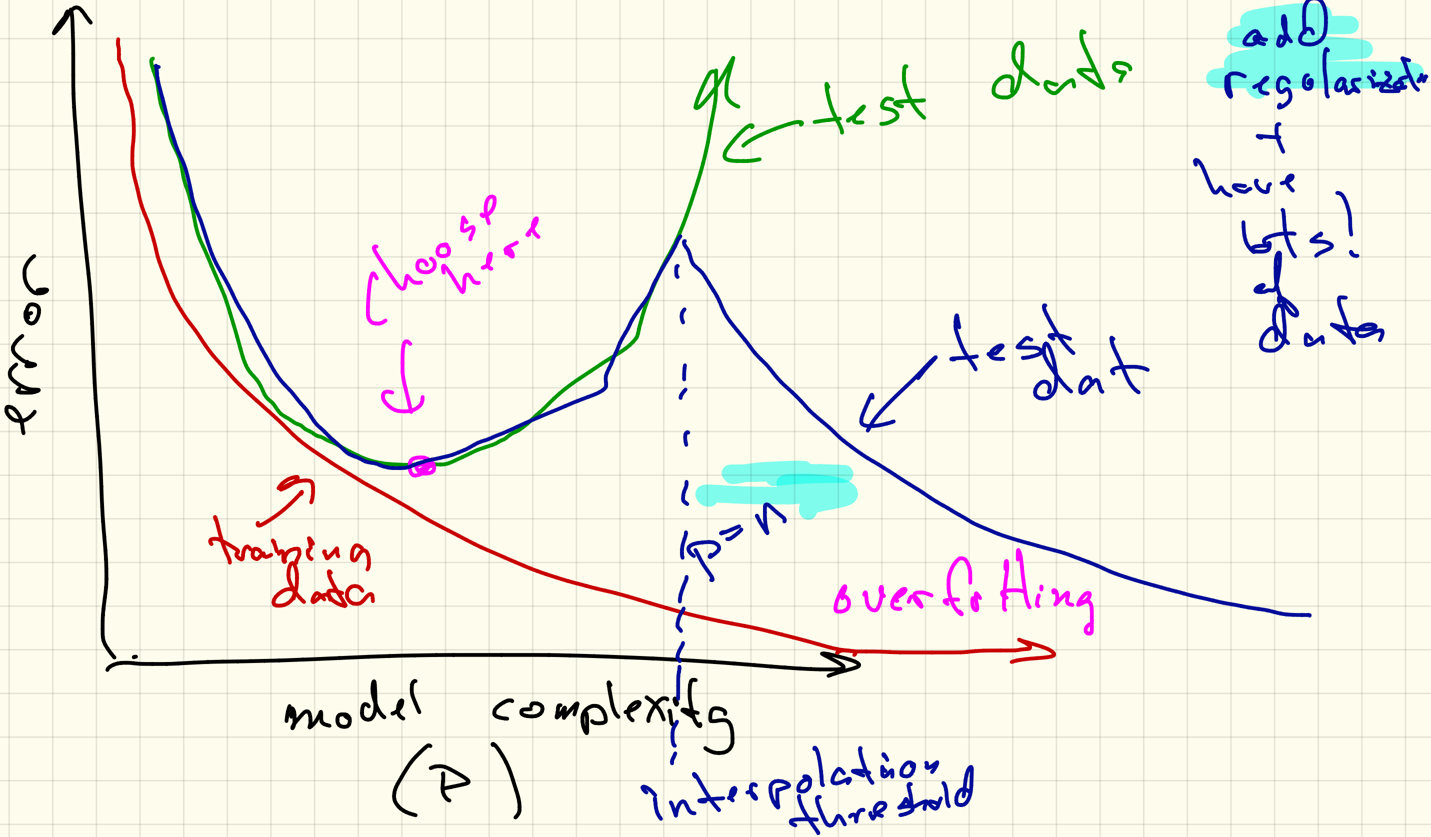
If your data is <u>small</u>.
don't want to waste data on test.

Leave-one-out CV.



test set size 1
but try all

Double Descent



if you add regularization + have lots of data