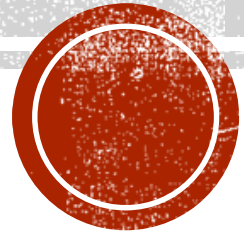
OVERFITTING

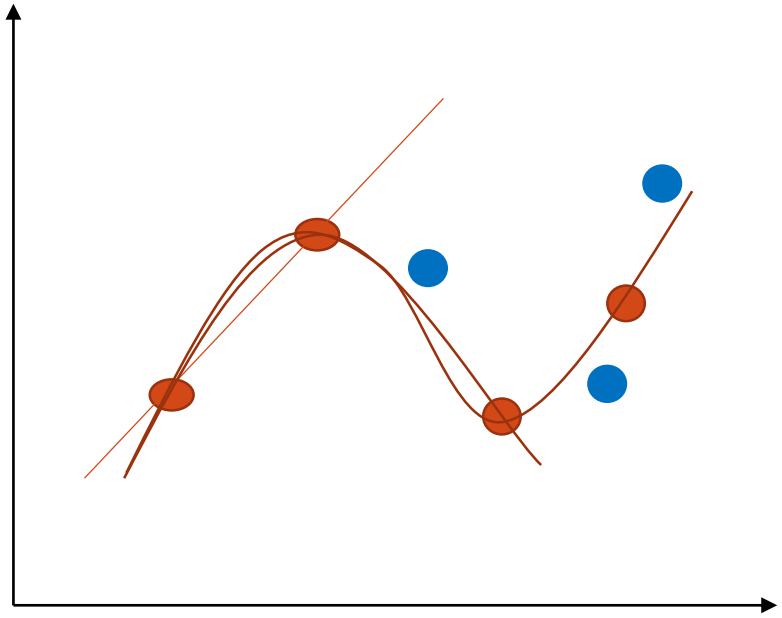
Sonpvh



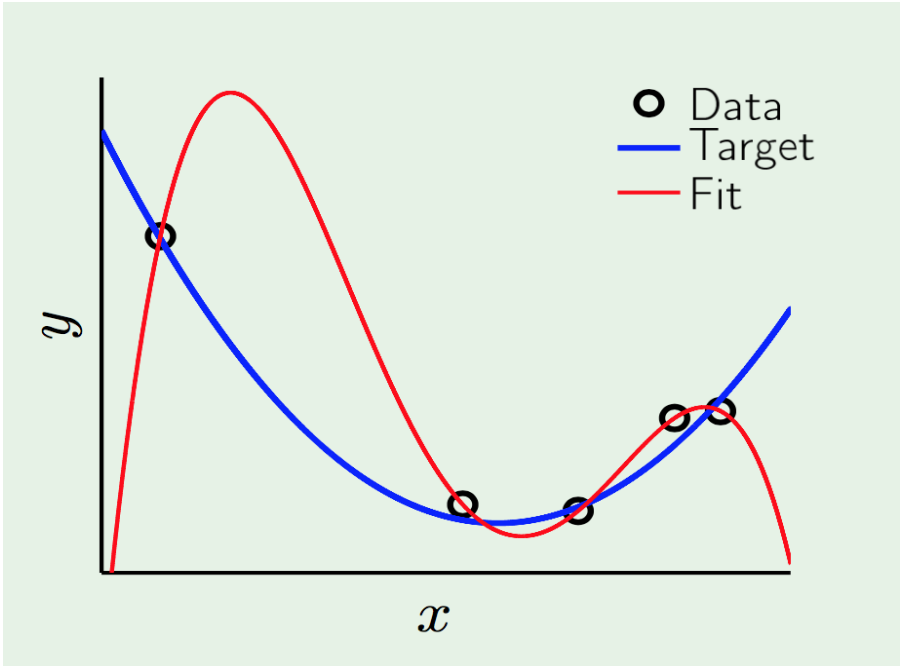
OUTLIER

1. Overfitting
2. Regularization
3. Validation
4. Model selection

1. OVERFITTING [1]



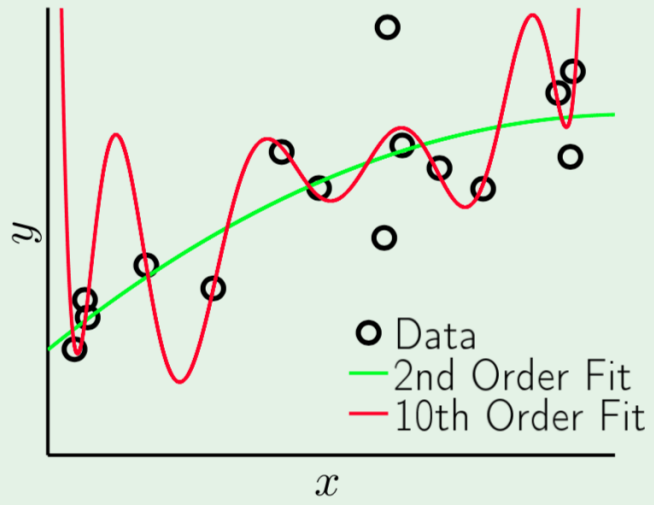
We can fit any function …



But noise …
& not function

1. OVERFITTING [1]

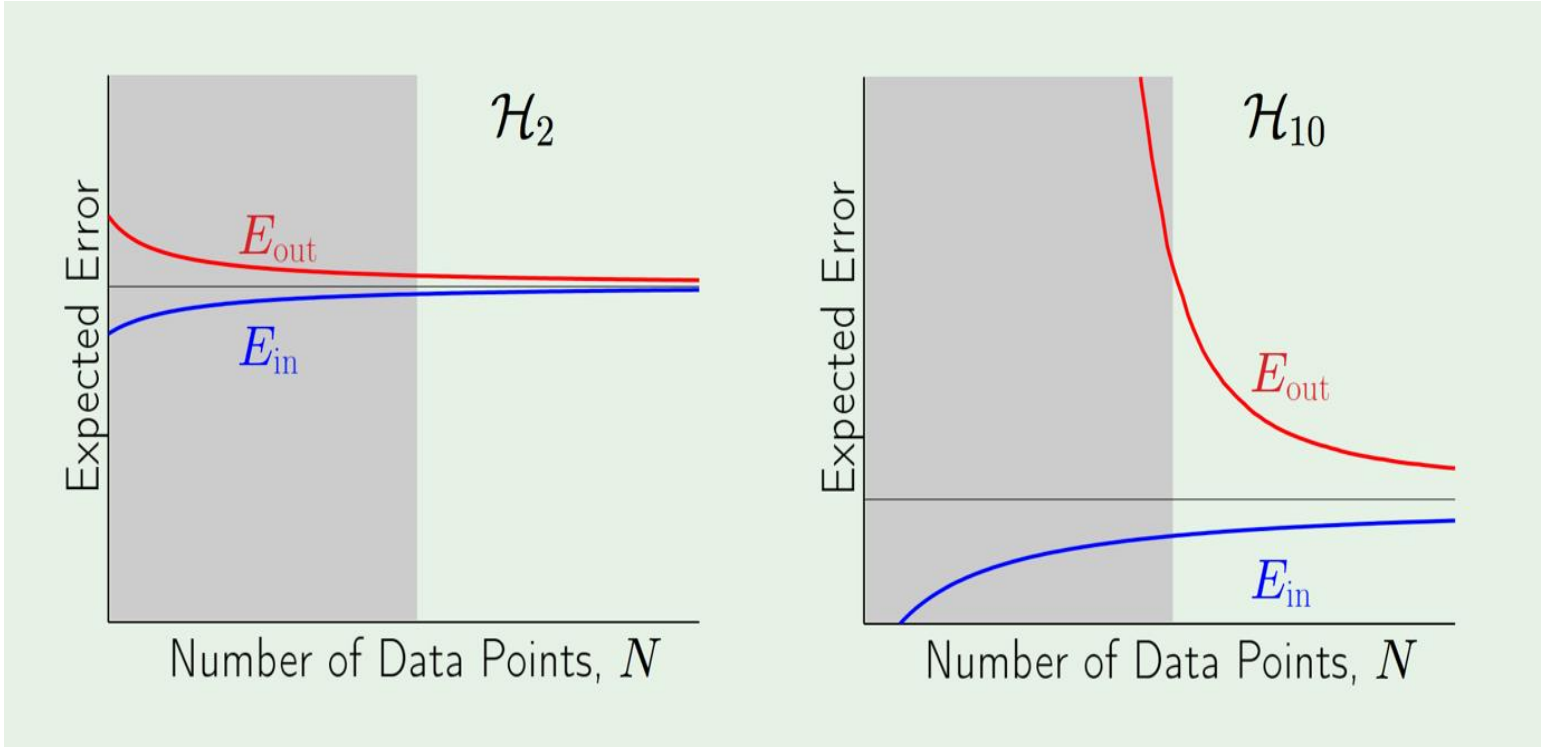
- **Overfitting: “fitting the data more than is warranted” [1]**
- **Fitting the noise**



Noisy low-order target

	2nd Order	10th Order
E_{in}	0.050	0.034
E_{out}	0.127	9.00

1. OVERFITTING



1. OVERFITTING [1]

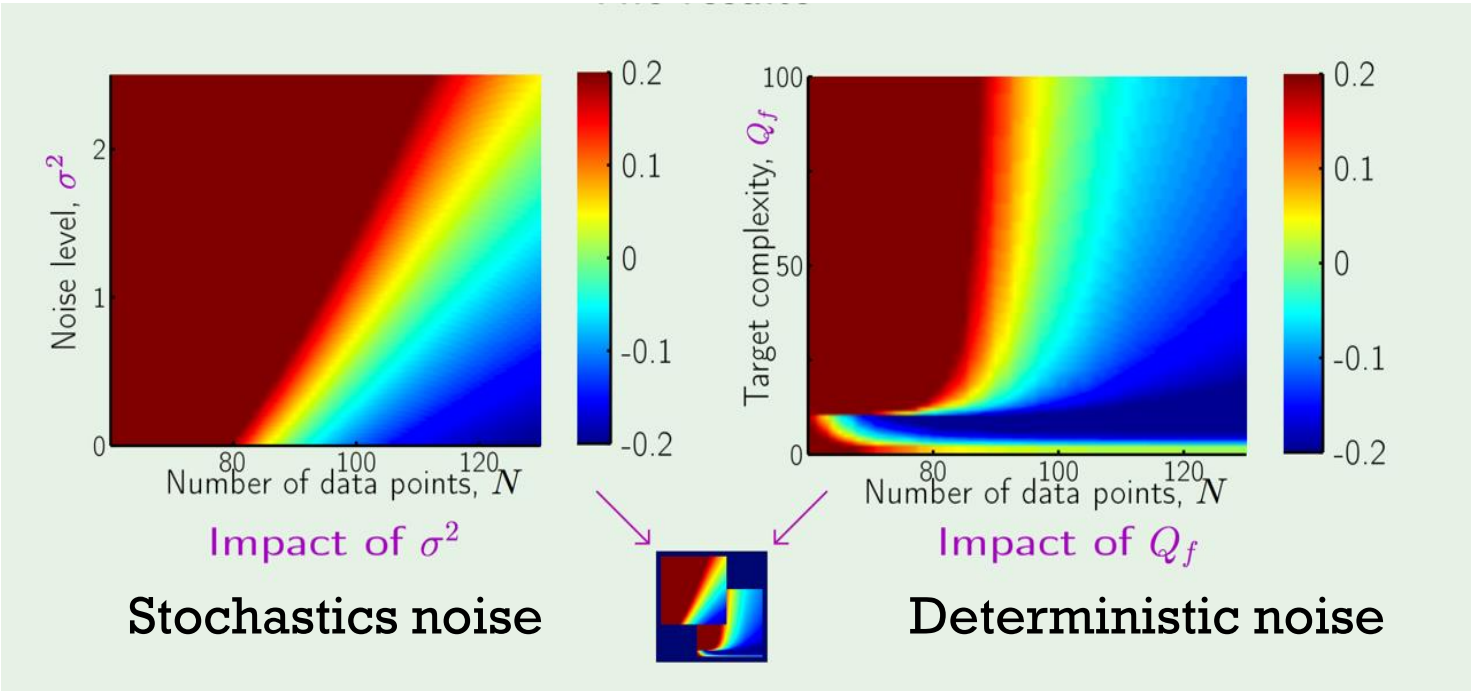
$$\underbrace{y}_{\text{Observation}} = \underbrace{f(x)}_{\text{Target Function}} + \underbrace{\epsilon(x)}_{\text{Noise}} = \underbrace{\sum_{q=0}^{Q} \alpha_i x_i}_{\text{Target complexity}} + \underbrace{\sigma^2}_{\text{Noise}}$$

Q : target complexity ↑ ↑

σ^2: noise level ↑ ↑

N : sample size ↑ ↓

Overfitting



1. OVERFITTING [1]

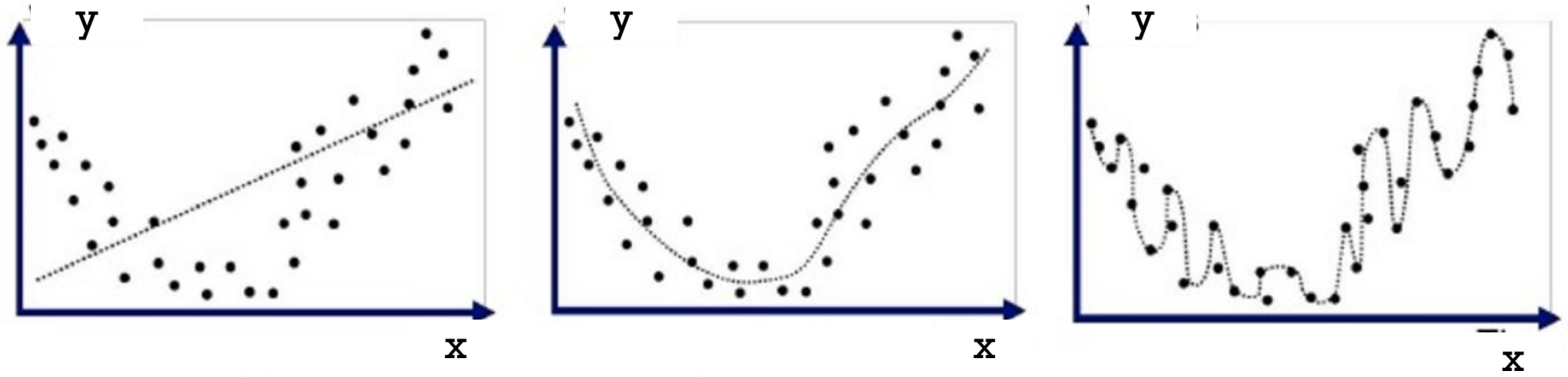
$$E_{out}\left(g^{(D)}\right) = \underbrace{\mathbb{E}_D\left[\left(g^{(D)}(x) - \bar{g}(\mathrm{x})\right)^2\right]}_{\text{Bias}} + \underbrace{\mathbb{E}_D\left[\left(\bar{g}(\mathrm{x}) - f(x)\right)^2\right]}_{\text{Variance}} + \underbrace{\mathbb{E}_x\left[(\epsilon(\mathrm{x}))^2\right]}_{\text{Noise}}$$

Deterministic Noise	Variance	Stochastic Noise

Depend on Hypothesis H

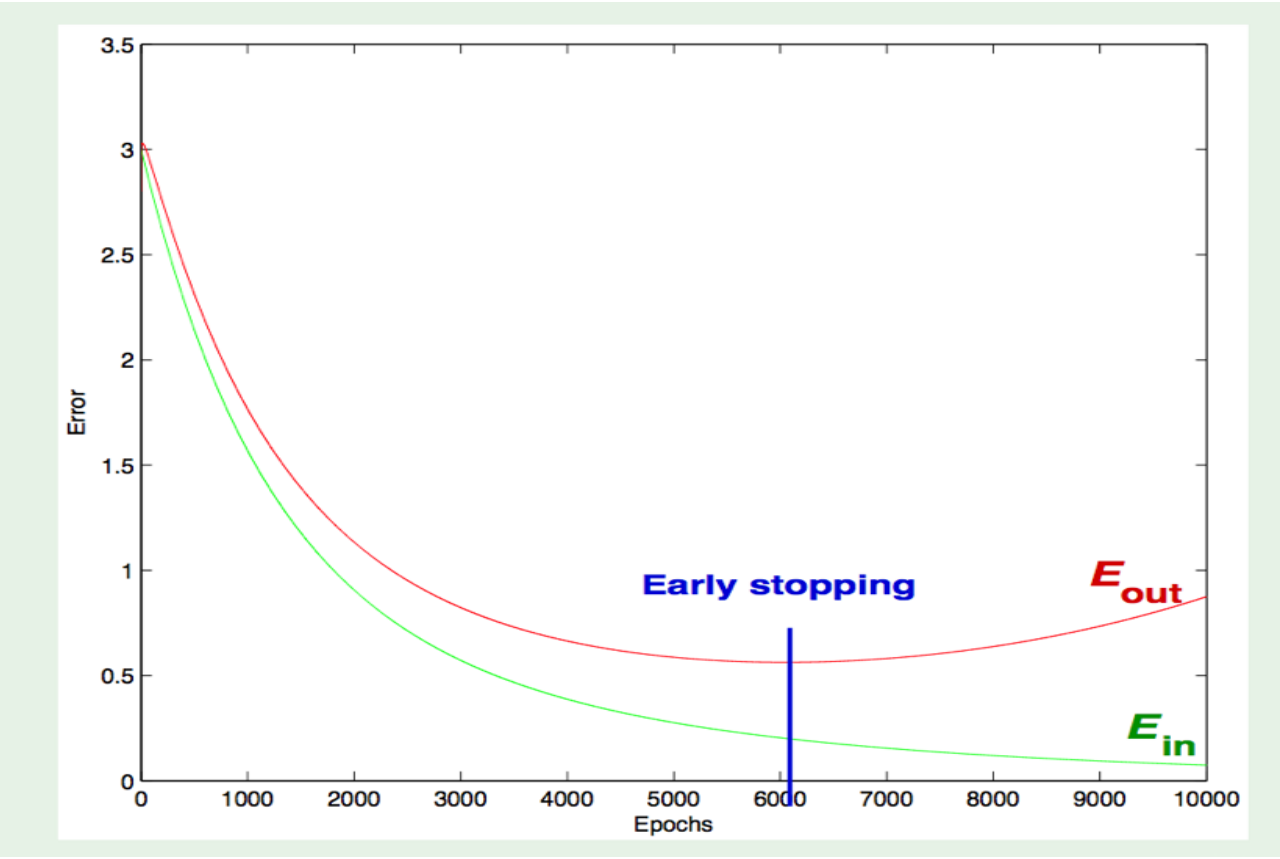
Overfitting

1. OVERFITTING [2]



Underfitted | Good Fit/Robust | Overfitted

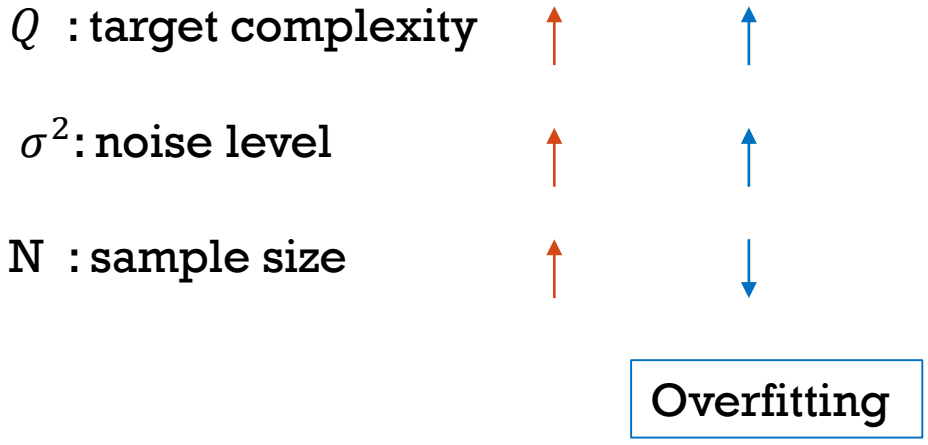
1. OVERFITTING [1]



Use Learning Curve to detect Overfitting

2. REGULARIZATION [3]

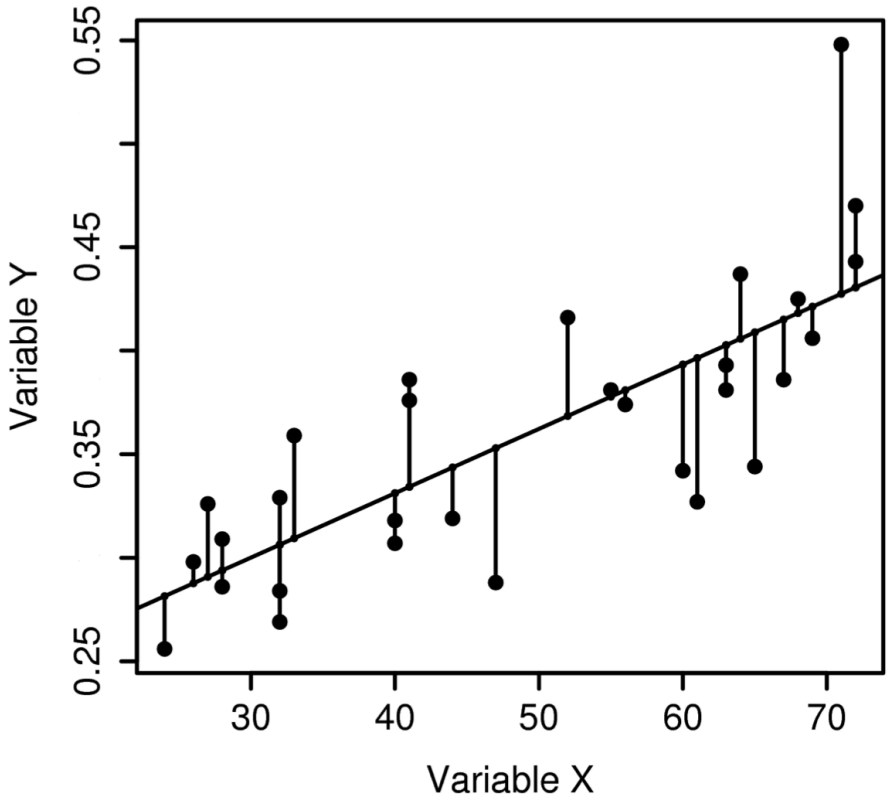
- Definition: "any modification we make to a learning algorithm that is intended to reduce its generalization error but not its training error" [3]



2. REGULARIZATION [3]

1. Parameter Norm Penalties
2. Norm Penalties as Constrained Optimization
3. Regularization and Under-Constrained Problems
4. Dataset Augmentation
5. Noise Robustness
6. Semi-supervised learning
7. Multi-task Learning
8. Early Stopping
9. Parameter Typing and Parameter Sharing
10. Sparse Representation
11. Bagging and Other Ensemble methods
12. Dropout
13. Adversarial Training
14. Tangent Distance, Tangent prop …

2. REGULARIZATION – PARAMETER NORM PENALTIES [3]



Cost function:

$$\sum (y - \hat{y})^2 = \sum \left((y - \sum (w_i x_i))^2 \right)$$

Constrain:

$$\sum |w_i| < C \quad \text{or} \quad \sum (w_i)^2 < C$$

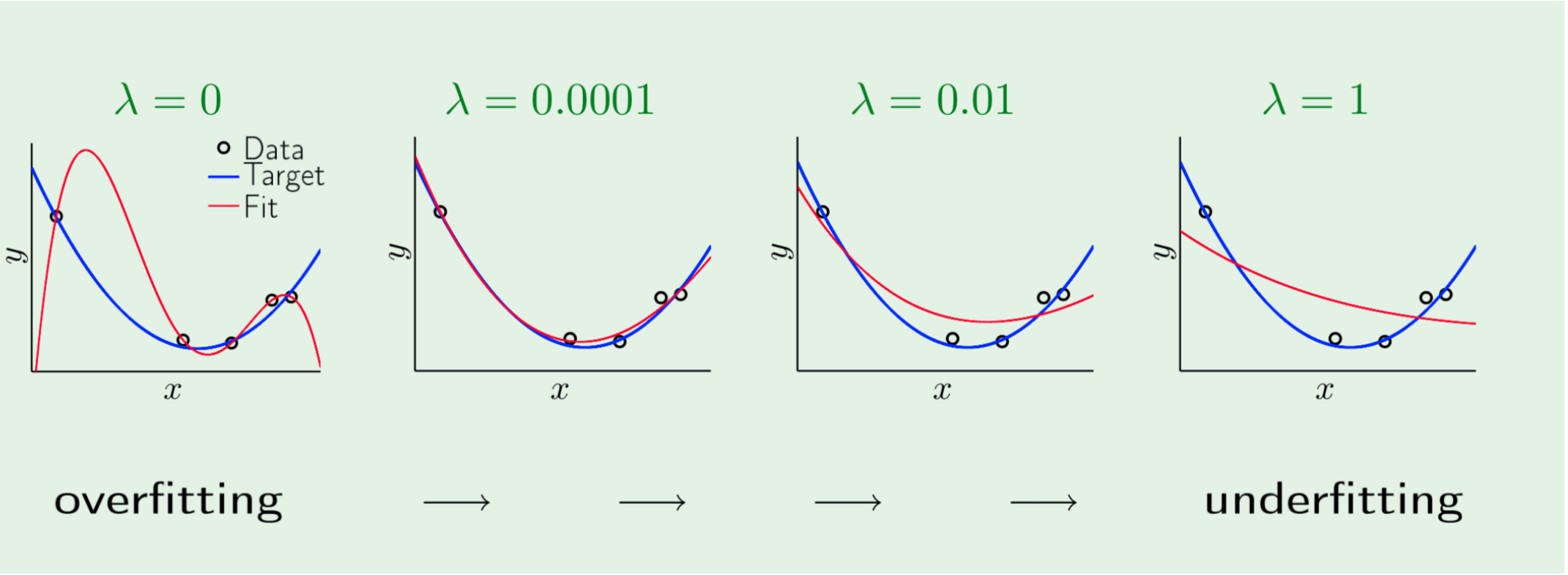
L1 – Lasso Reg L2 – Ridge Reg

Cost:

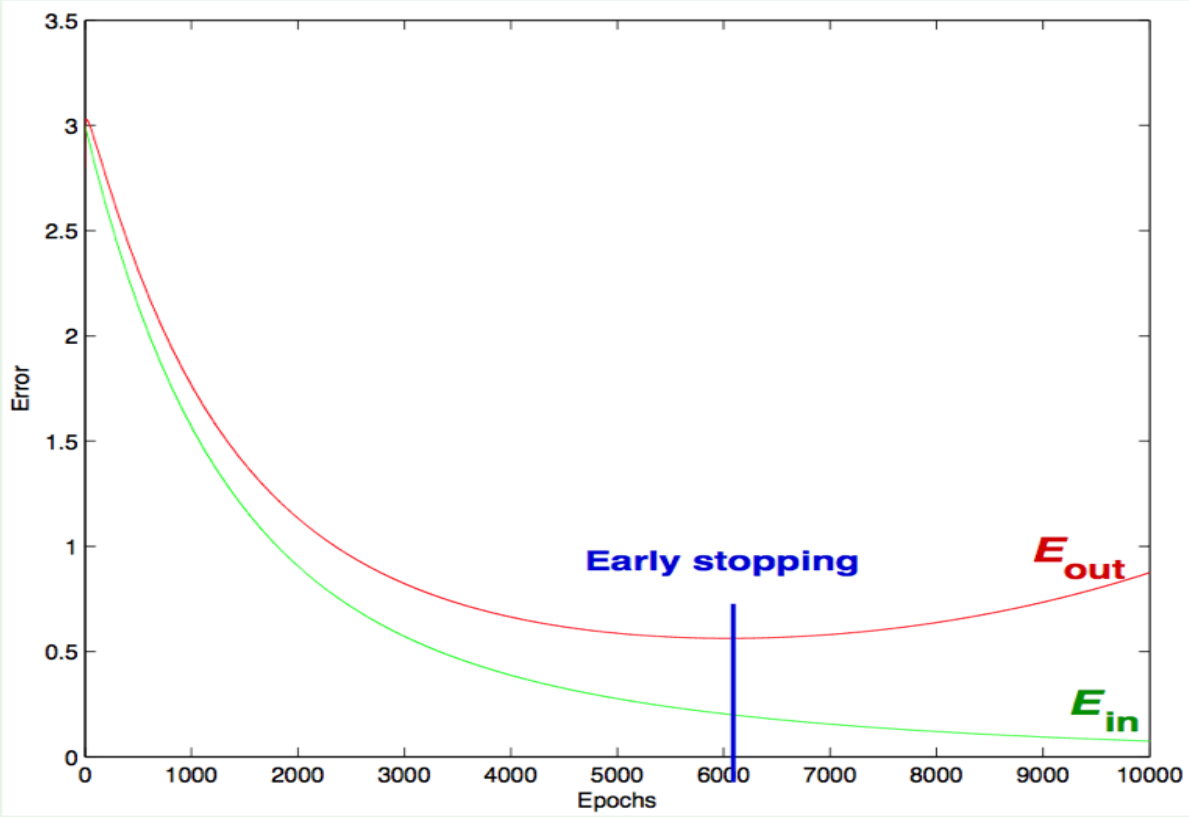
$$\sum \left((y - \sum (w_i x_i))^2 \right) + \lambda \sum |w_i|$$

2. REGULARIZATION – PARAMETER NORM PENALTIES [1]

$$\sum\left((y - \sum(w_i x_i))^2\right) + \lambda \sum |w_i| \quad \text{or} \quad \sum\left((y - \sum(w_i x_i))^2\right) + \lambda \sum (w_i)^2$$

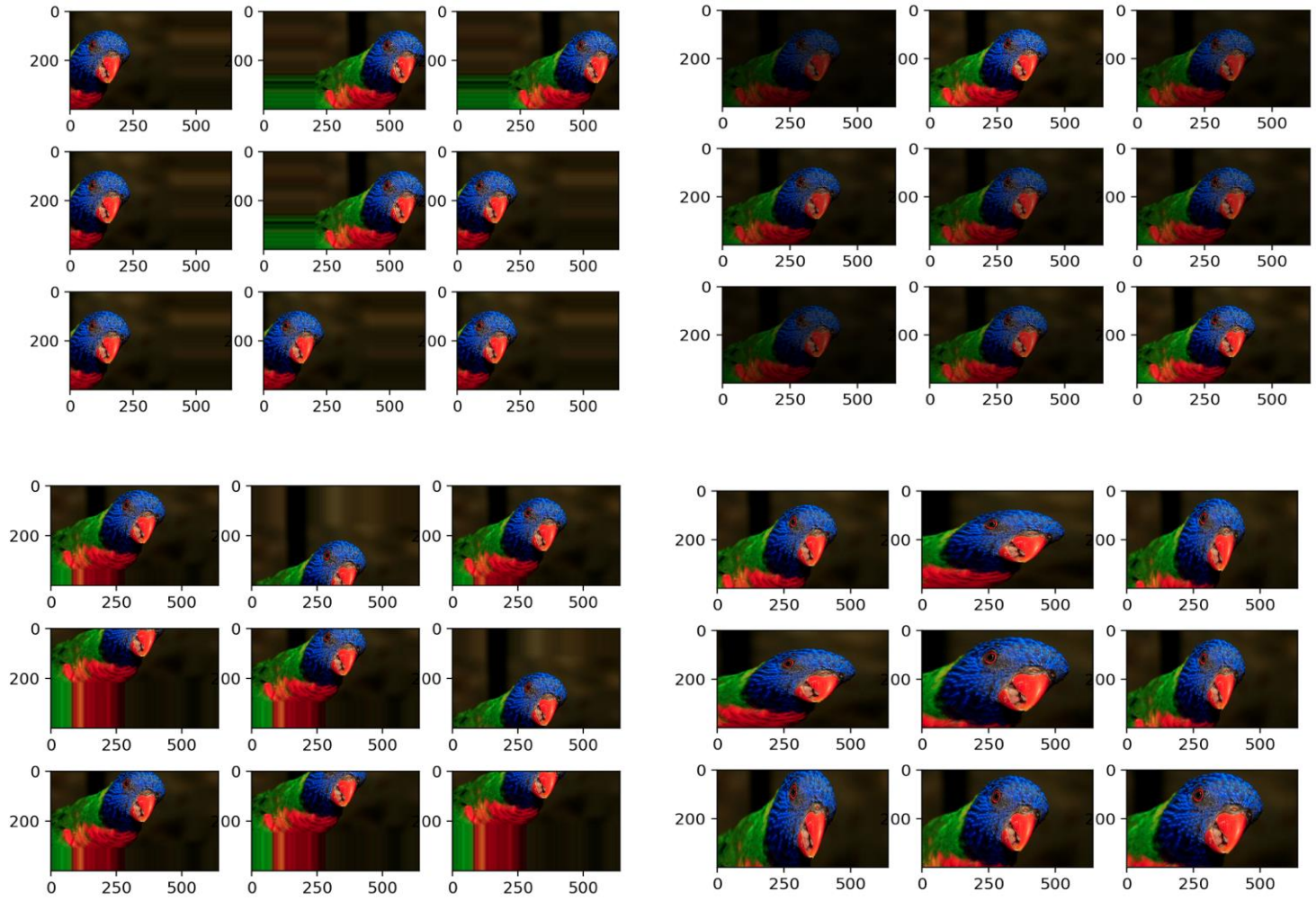


2. REGULARIZATION – EARLY STOPPING [1,3]

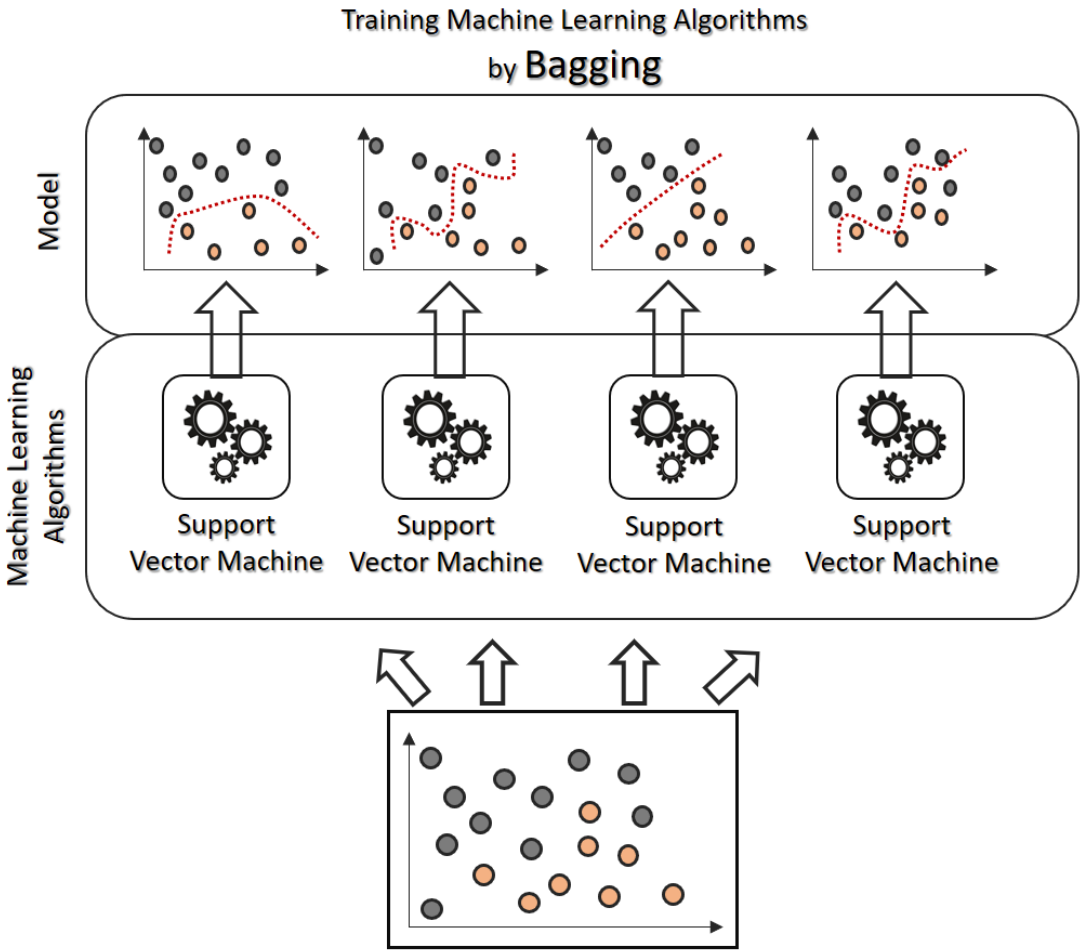


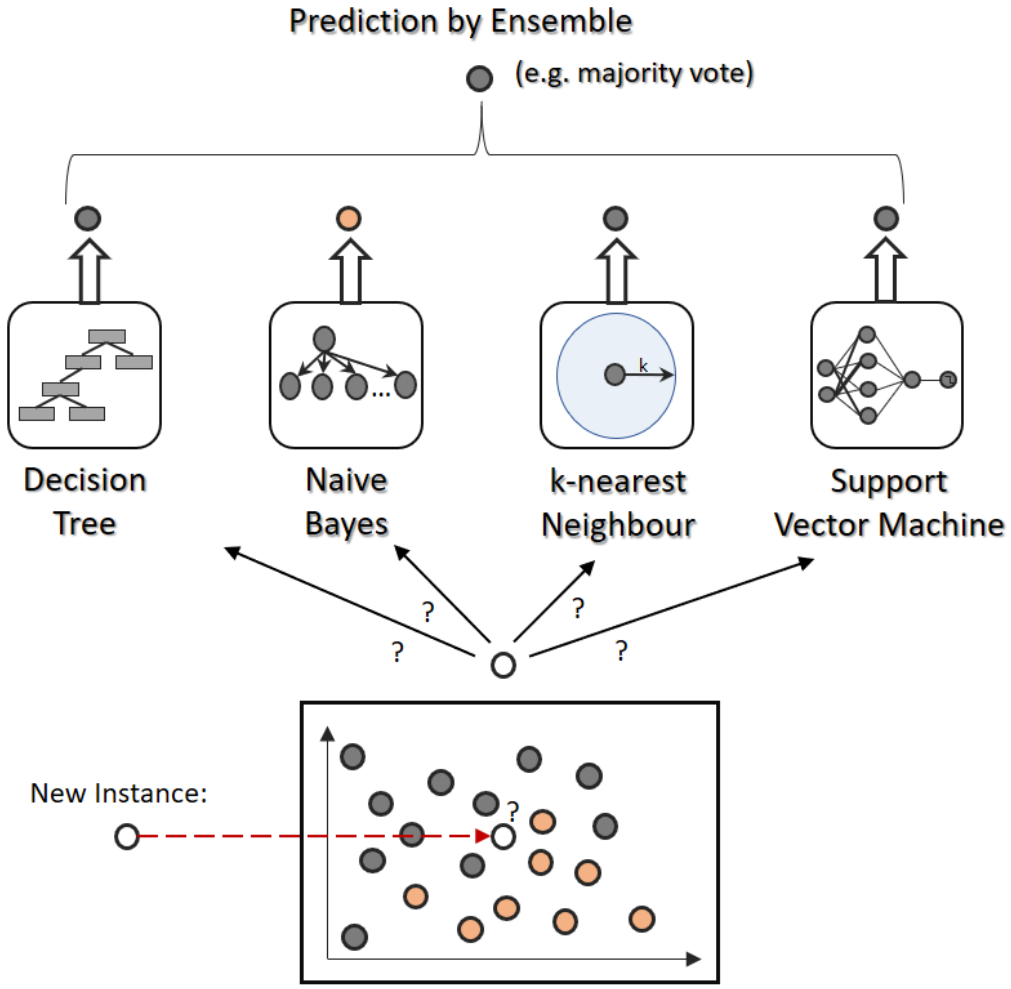
Noise Fitting ….

2. REGULARIZATION – EARLY STOPPING [3,4]



2. REGULARIZATION – BAGGING & ENSEMBLE METHODS

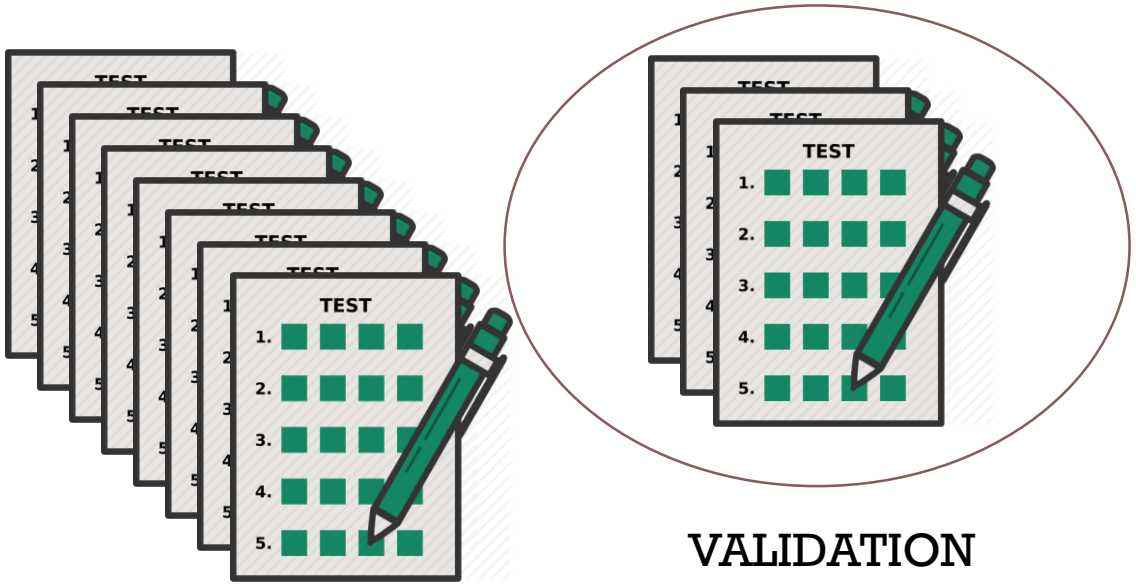




[3,5]

3. VALIDATION [1]

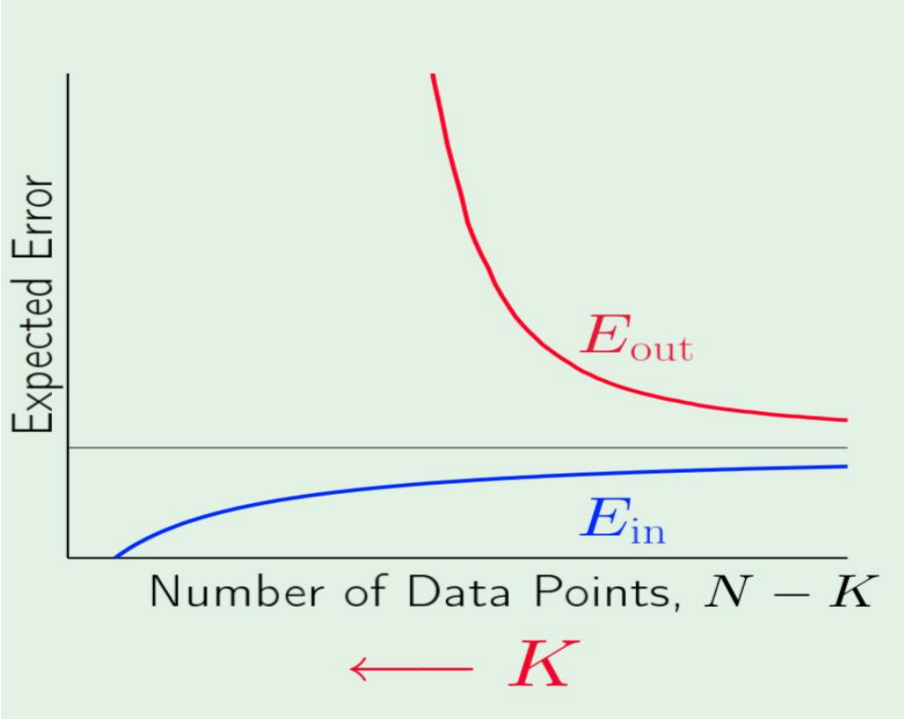
$$\underbrace{E_{out}(h)}_{\text{Validation estimate this quantity}} = E_{in}(h) + \underbrace{\text{Overfitting penalty}}_{\text{Regularization estimate this quantity}}$$



VALIDATION

Target complexity
Noise level
Sample size

3. VALIDATION [1]

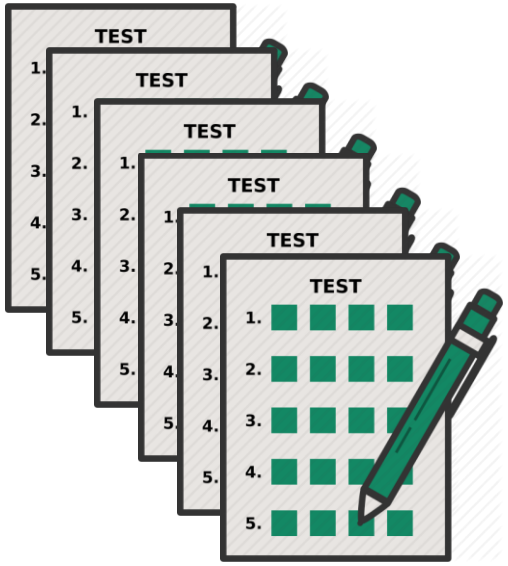


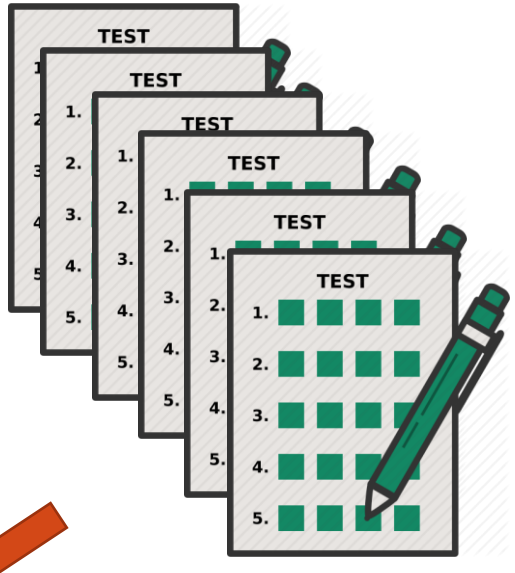
How much K?

Rule of thumb: 20% train data

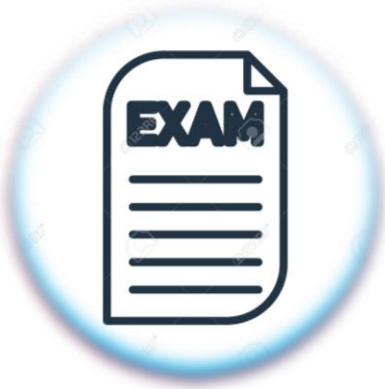
TRAIN

VALIDATE





BUT ….

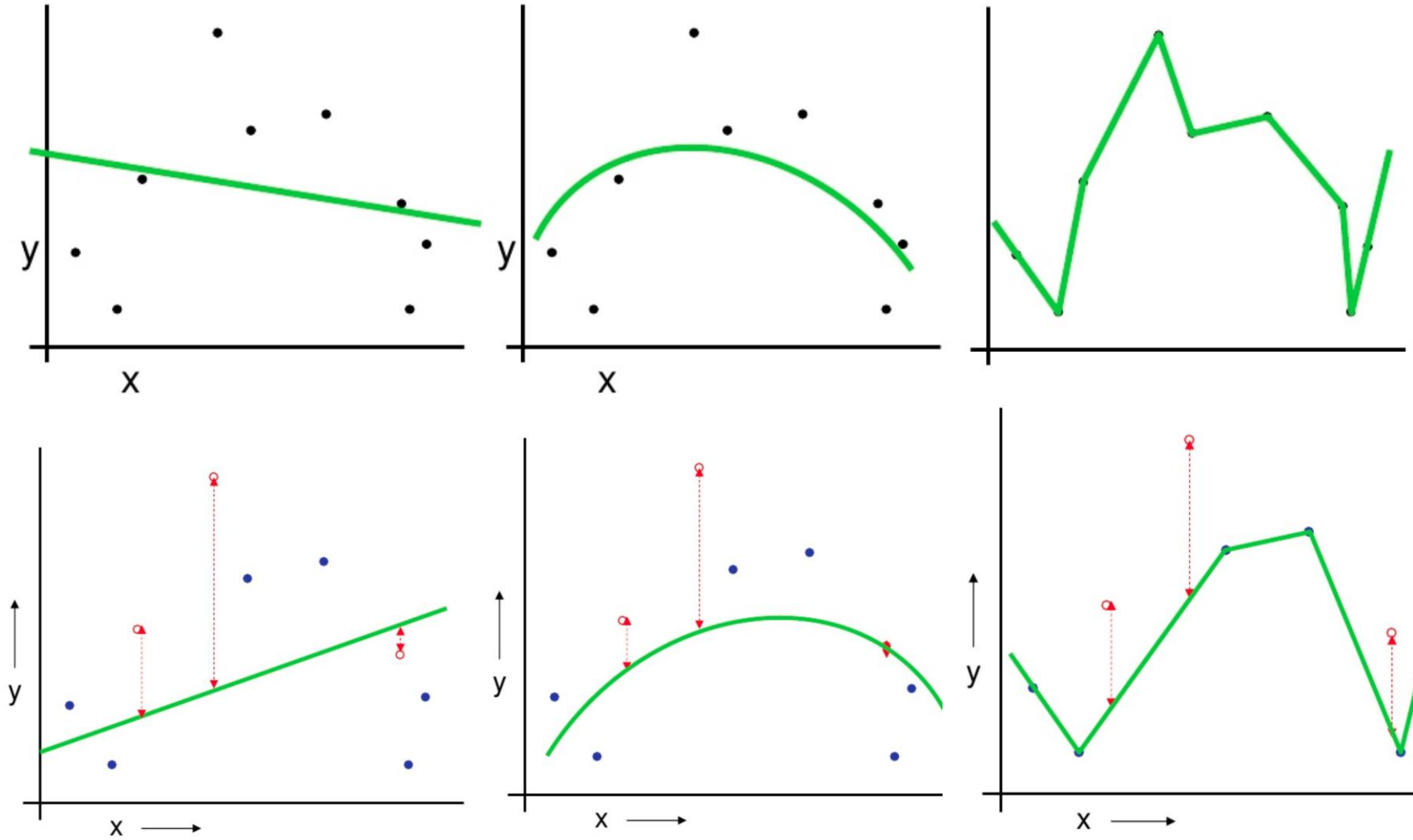


PUBLIC

3. VALIDATION

1. K-fold Cross validation
2. Holdout or Train/Test split
3. Stratified K-Fold Cross Validation
4. Repeated Cross validation
5. Leave-one-out cross validation - LOOCV
6. …

3. VALIDATION [7]



Train-Test split

+ Simple, cheap

- Waste data …

MSE: 2.4

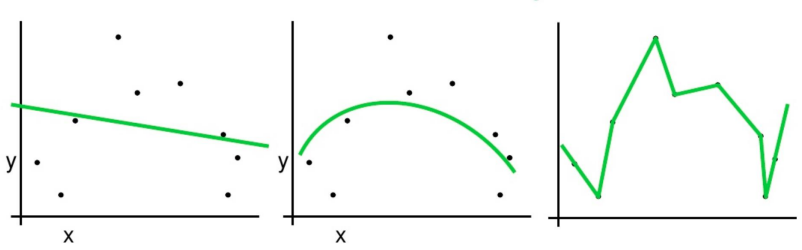
MSE: 0.9

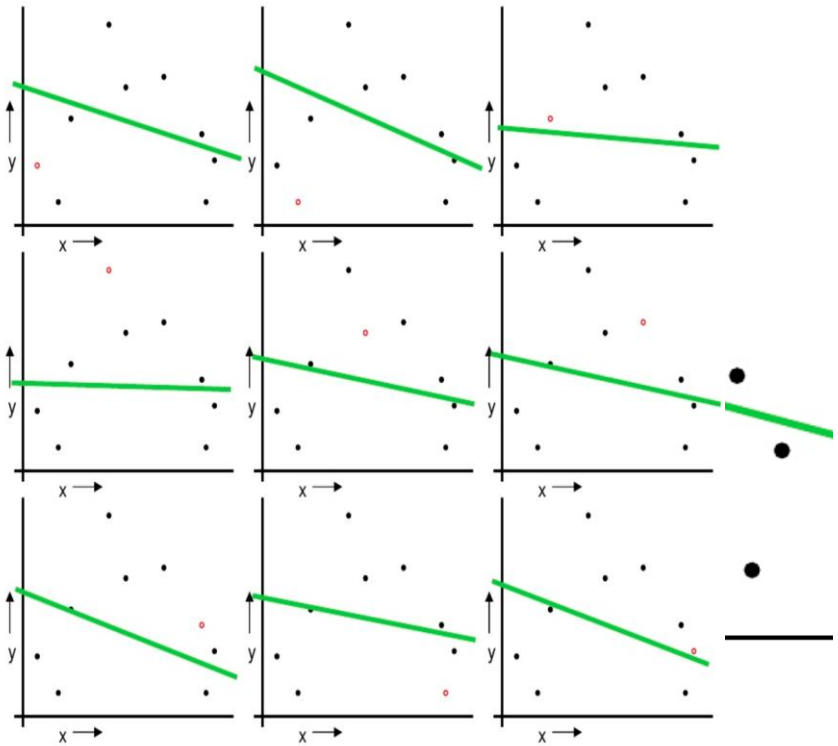
MSE: 2.2

3. VALIDATION [7]

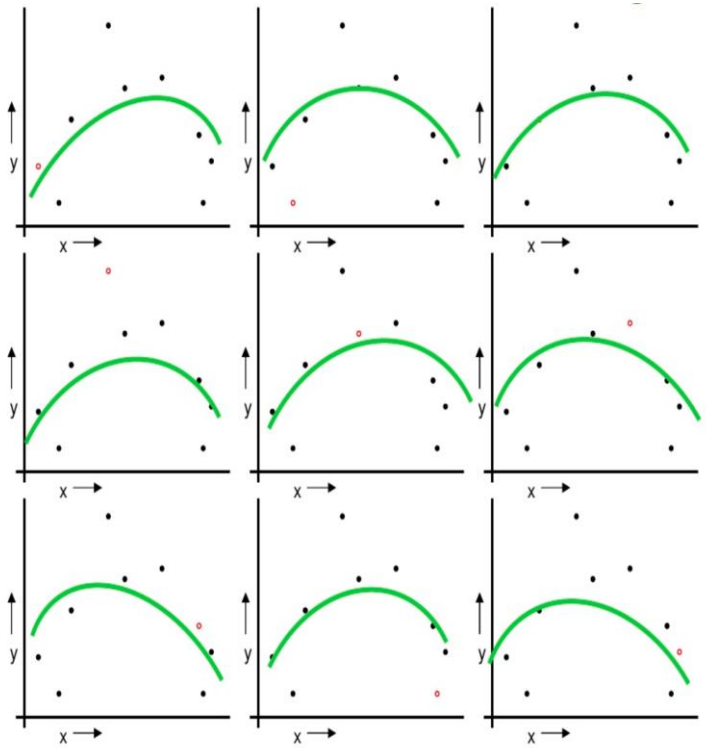
LOOCV

+ Doesn't waste data
- Expensive
- Weird behaviors

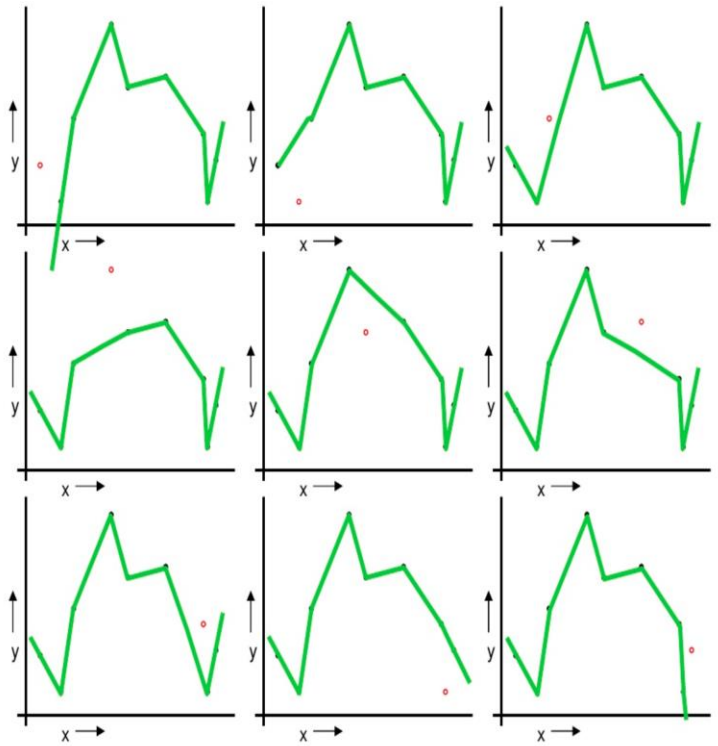




MSE: 2.12



MSE: 0.962



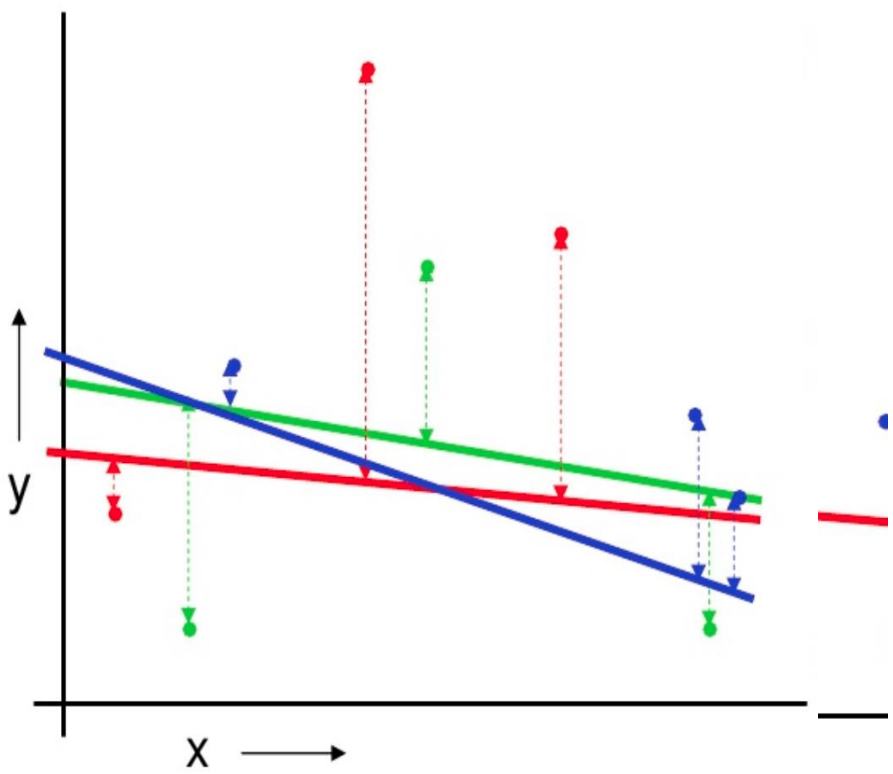
MSE: 3.33

3. VALIDATION [7]

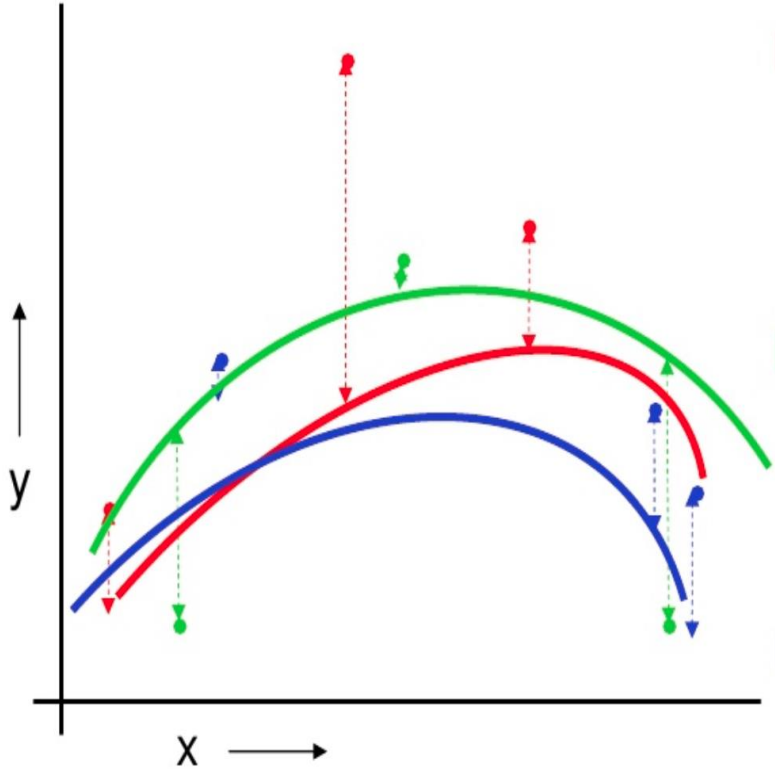
K-fold CV

+ Only waste $\frac{N}{K}$ data

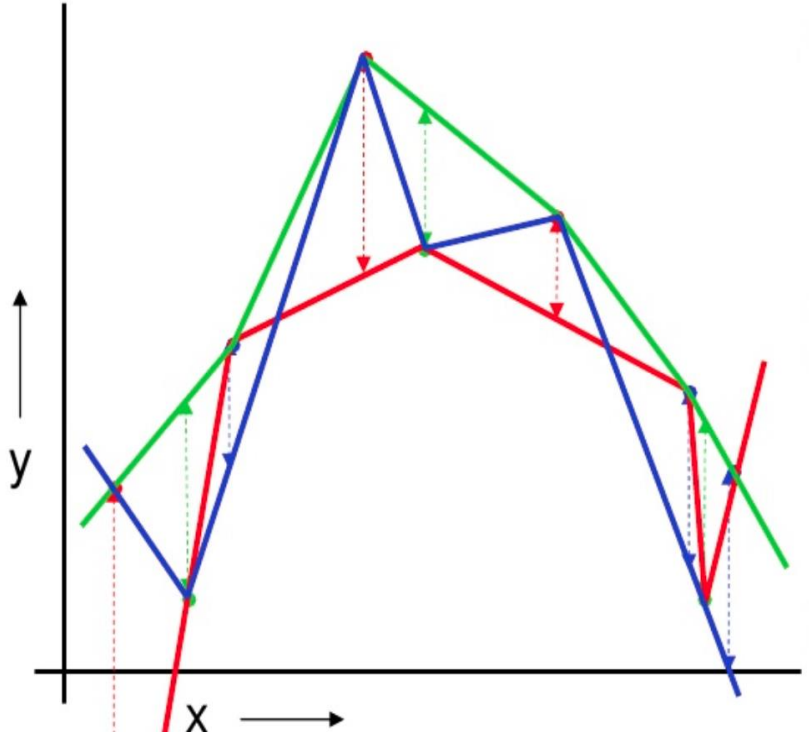
+ Only K times more expensive than train-test split



MSE: 2.05



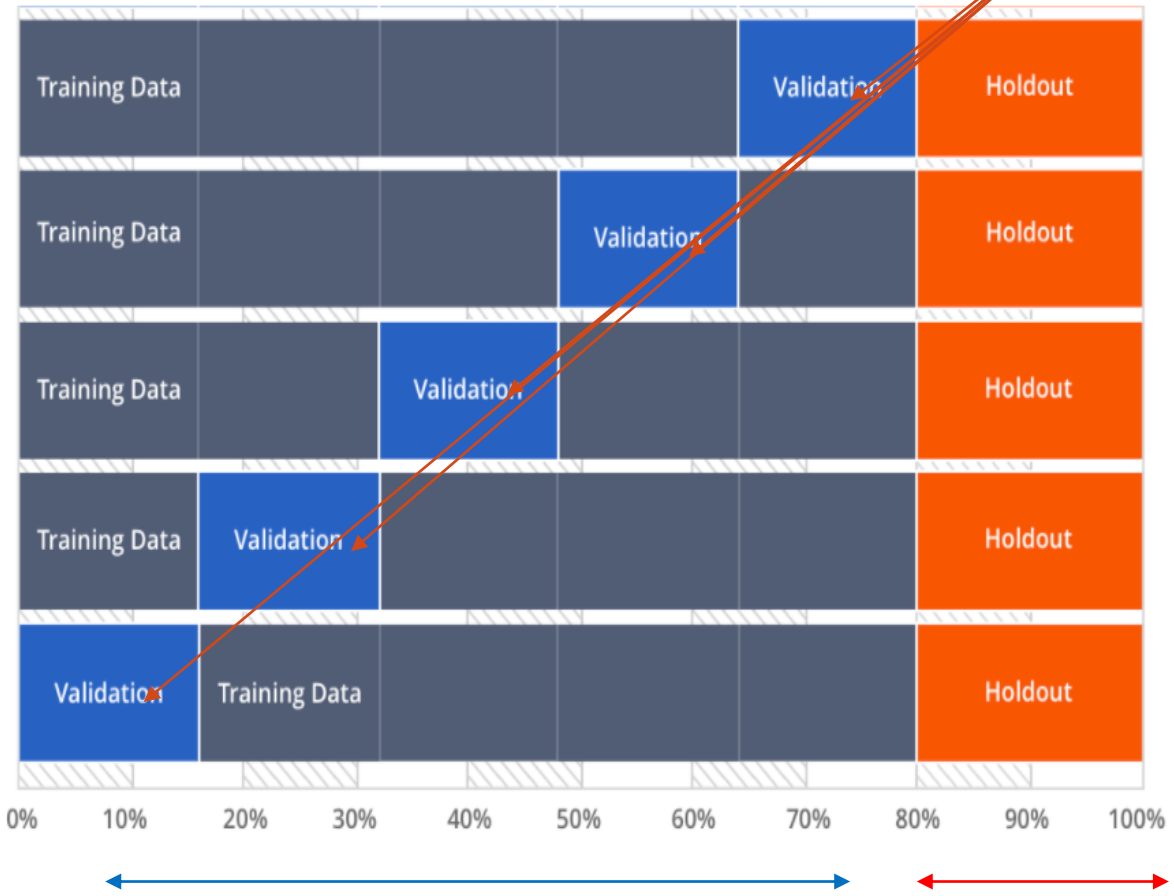
MSE: 1.11

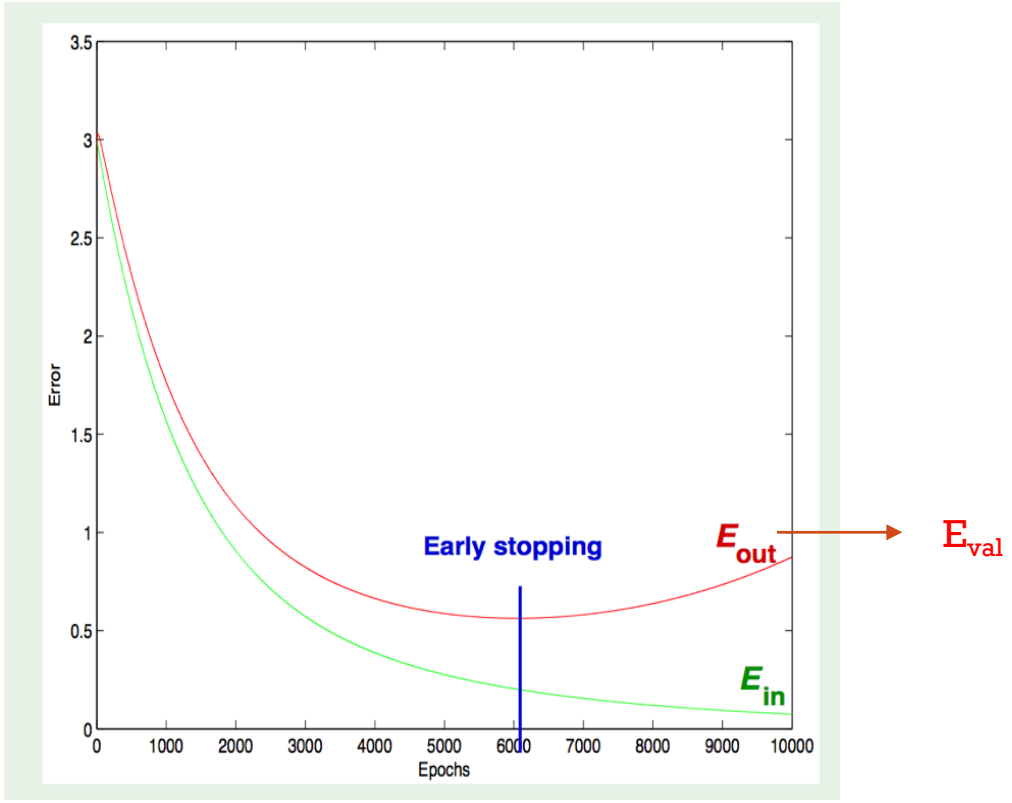


MSE: 2.93

3. VALIDATION

K-fold (K ==5)



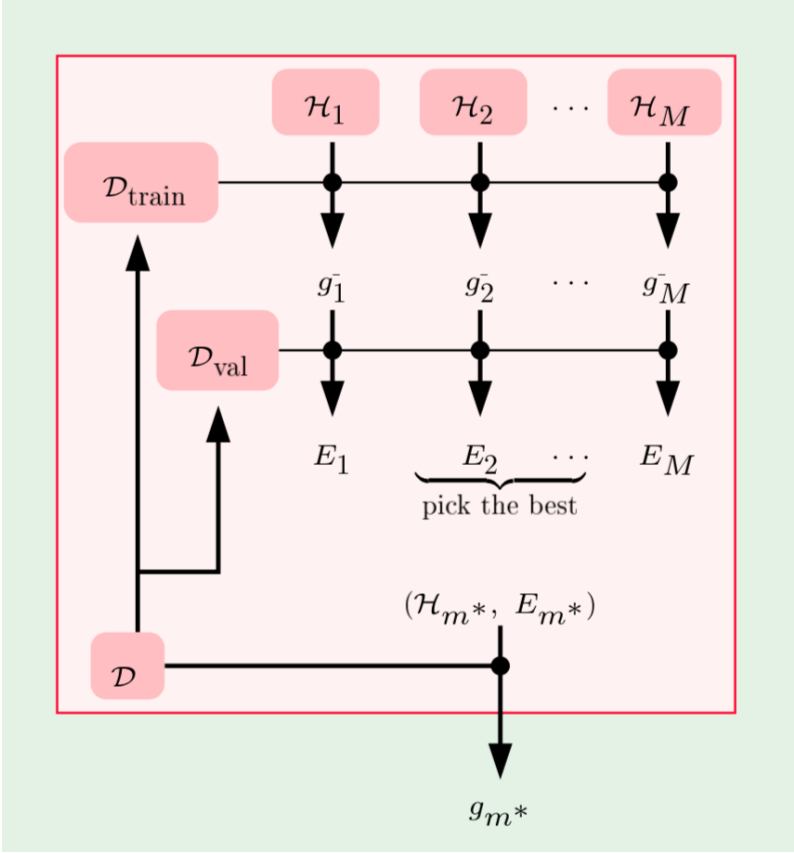


[6]

Holdout or Train – Test split

Early stopping

4. MODEL SELECTION



[3]

Algorithm	TRAINERR	10-FOLD-CV-ERR	Choice
0 hidden units			
1 hidden units			
2 hidden units			⊠
3 hidden units			
4 hidden units			
5 hidden units			

[7]

REFERENCES:

1. Yaser S. Abu-Mostafa, Malik Magdon-Ismail, Hsuan-Tien Lin-Learning From Data. A short course-AMLBook (2012)
2. https://medium.com/greyatom/what-is-underfitting-and-overfitting-in-machine-learning-and-how-to-deal-with-it-6803a989c76
3. Ian GoodFellow, Yoshua Bengio, Aaron Courville – Deep learning – Chapter 7. Regularization for Deep Learning
4. https://machinelearningmastery.com/how-to-configure-image-data-augmentation-when-training-deep-learning-neural-networks/
5. https://data-science-blog.com/blog/2017/12/03/ensemble-learning/
6. https://nagornyy.me/courses/data-science/fair_models/
7. https://www.slideshare.net/guestfee8698/crossvalidation