Regularization & Cross Validation

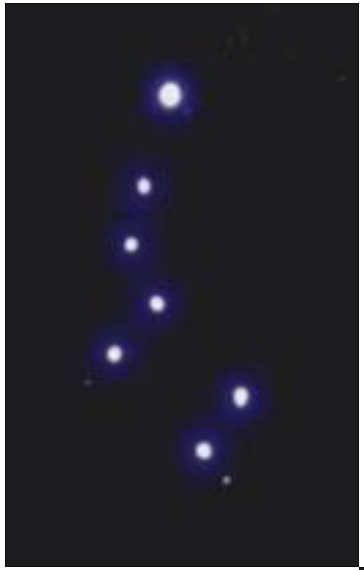
Professor: Srikanth Krishnamurthy
Presented by: Chenlian Xu
Qianli Ma
Mar 3, 2018

Overfitting

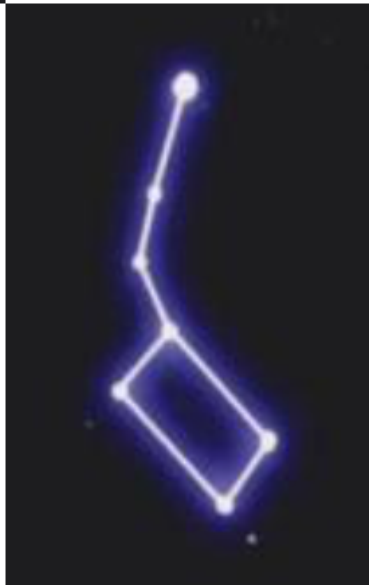
What is it?

How to prevent it?

Overfitting



Data

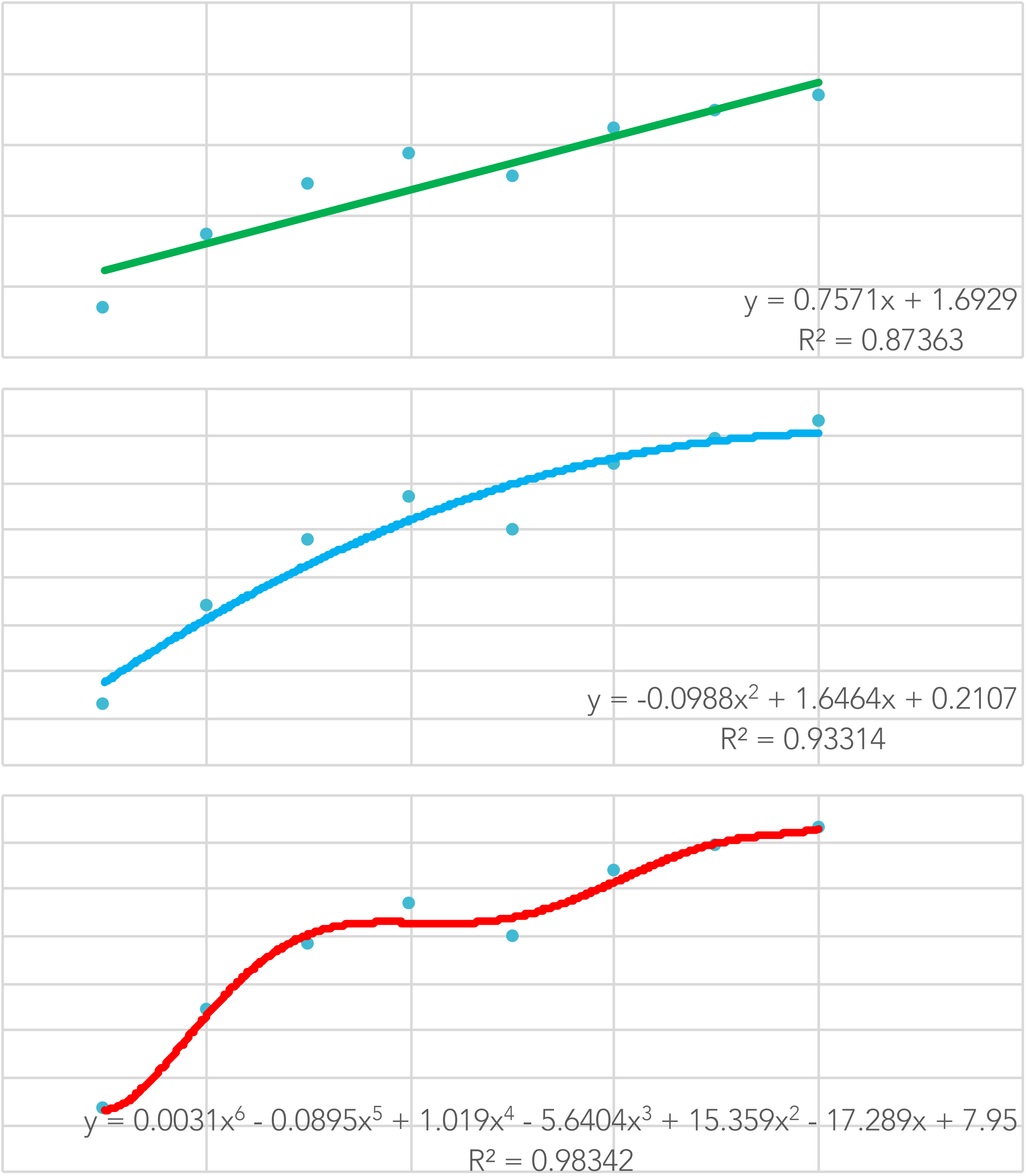


Normal Fitting

Overfitting

Serious Overfitting

Overfitting



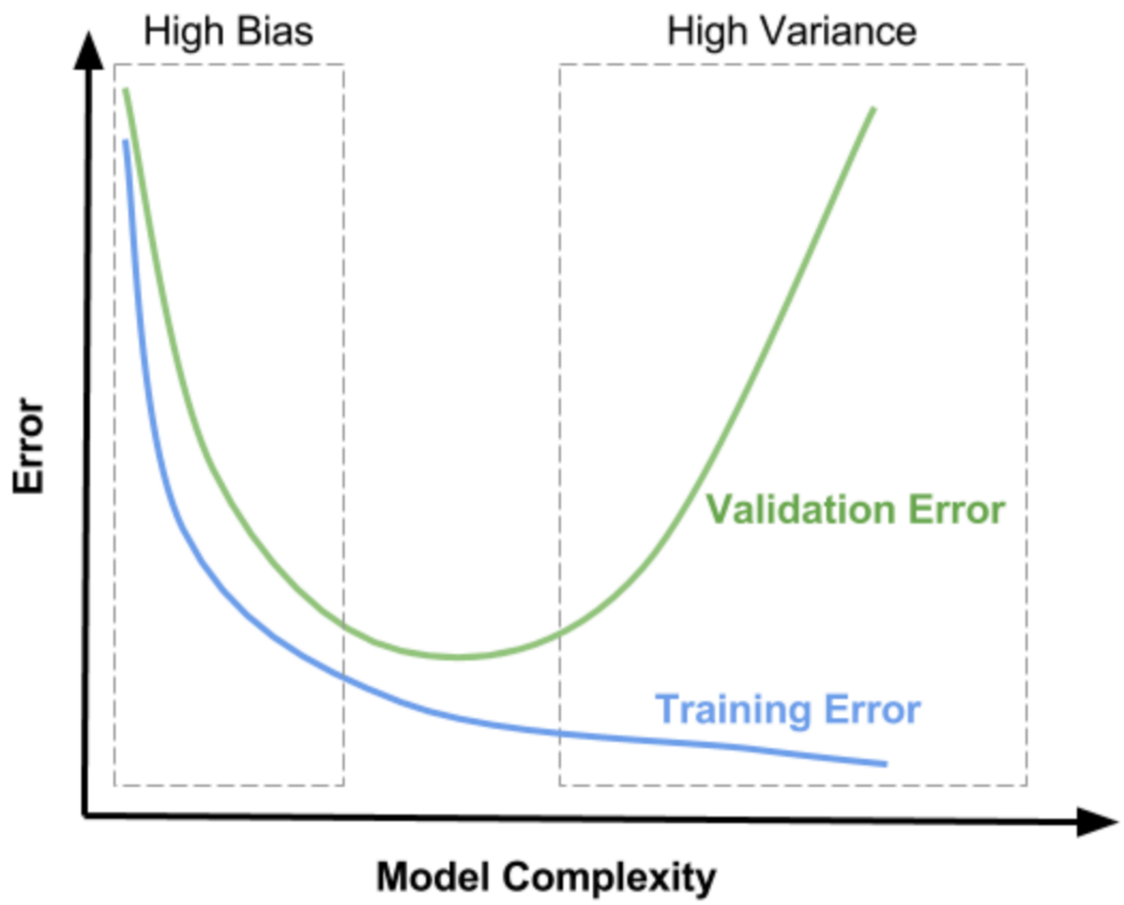
Underfitting
High Bias

Normal Fitting
Bias-Variance Tradeoff

Overfitting
High Variance

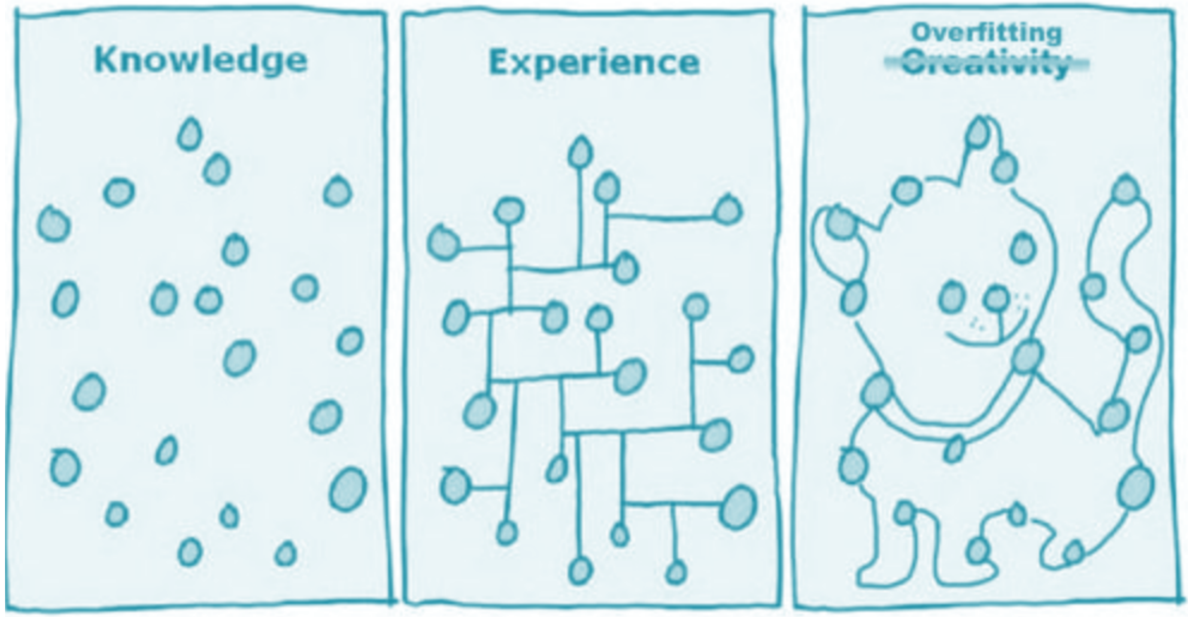
Overfitting

- Bias-Variance Tradeoff



Overfitting

- Too many features
- Fit the train set very well
- Fail to generalize to new examples



How to prevent overfitting?

- 1 Reduce number of features
  - Manually select
  - Model selection algorithm
- 2 Regularization
  - sparsity
  - Reduce values of parameters

Regularization

L1, L2, Elastic Net

Regularization

- Reduces overfitting by adding a complexity penalty to the loss function

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}\left[\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2 + penalty\right]$$

L1 & L2 Regularization

- L1 LASSO (Least absolute shrinkage and selection operator)

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}\left[\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2 + \boxed{\lambda\sum_{j=1}^{n}|\theta_j|}\right]$$

- L2 Ridge

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}\left[\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2 + \boxed{\lambda\sum_{j=1}^{n}\theta_j^2}\right]$$

Elastic Net

- Elastic Net
  - linearly combines the L1 and L2 penalties of the lasso and ridge methods

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}\left[\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2 + \underbrace{\lambda\sum_{j=1}^{n}|\theta_j|}_{\text{L1}} + \underbrace{\lambda\sum_{j=1}^{n}\theta_j^2}_{\text{L2}}\right]$$

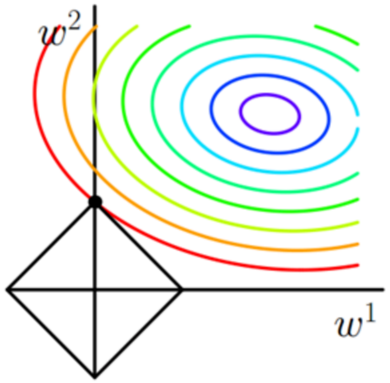
Effect of L1 Regularization

- L1 LASSO (Least absolute shrinkage and selection operator)

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}\left[\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2 + \lambda\sum_{j=1}^{n}|\theta_j|\right]$$

L1 Regularization encourages sparsity

$$\theta \dashrightarrow \theta' = \theta - \eta\lambda\cdot sgn(\theta) - \eta\frac{\partial C0}{\partial\theta}$$



Effect of L2 Regularization

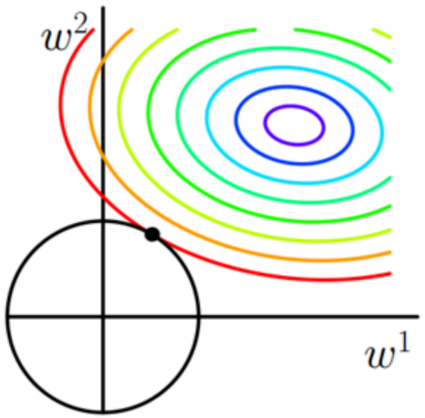
- L2 Ridge

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}\left[\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2 + \boxed{\lambda}\boxed{\sum_{j=1}^{n}\theta_j^2}\right]$$

As λ increases, sum of squares decreases

Weight decay

$$\theta \dashrightarrow \theta' = (1-\eta\lambda)\theta - \eta\frac{\partial C0}{\partial\theta}$$

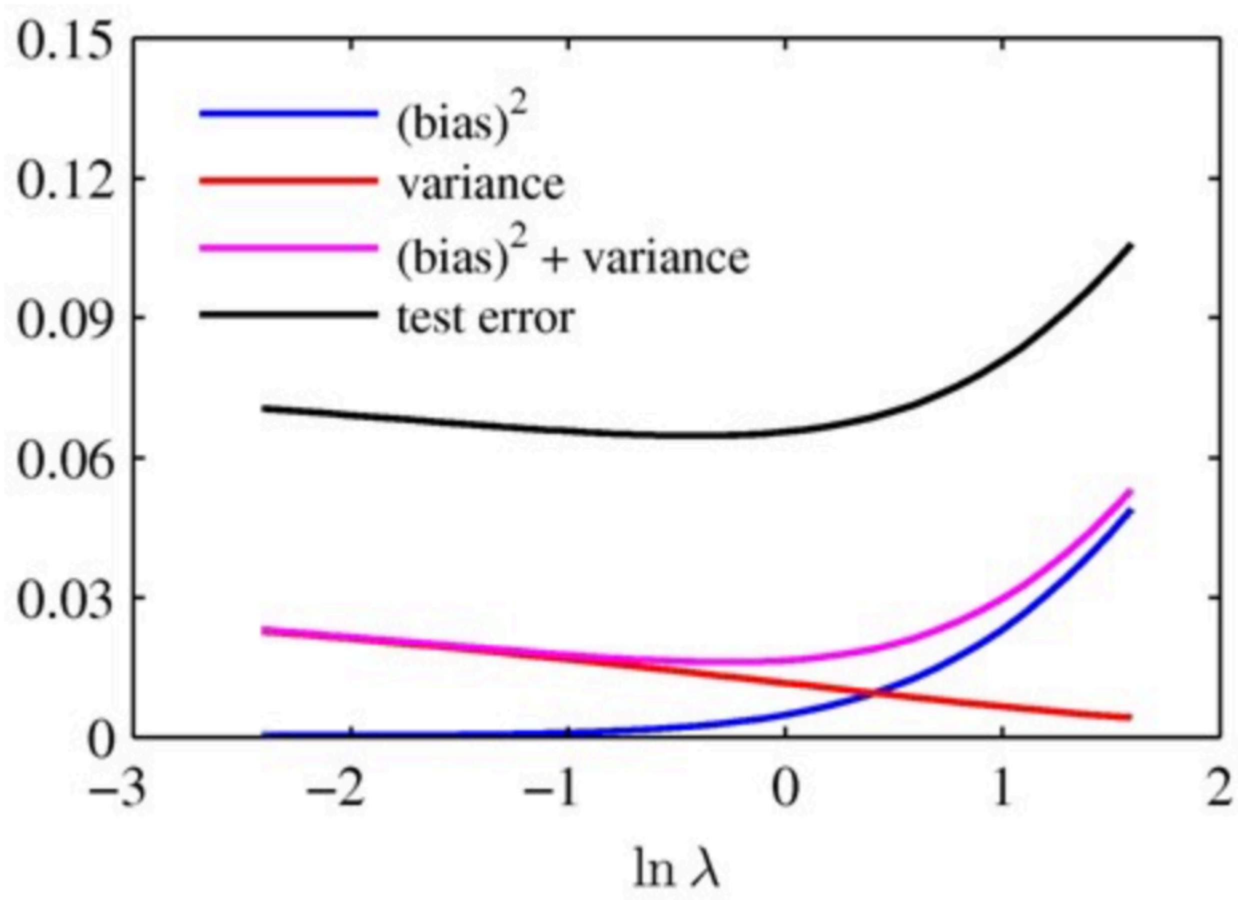


Regularization

- Reduces overfitting
- Reduces variance
- Minimizes the test-set error

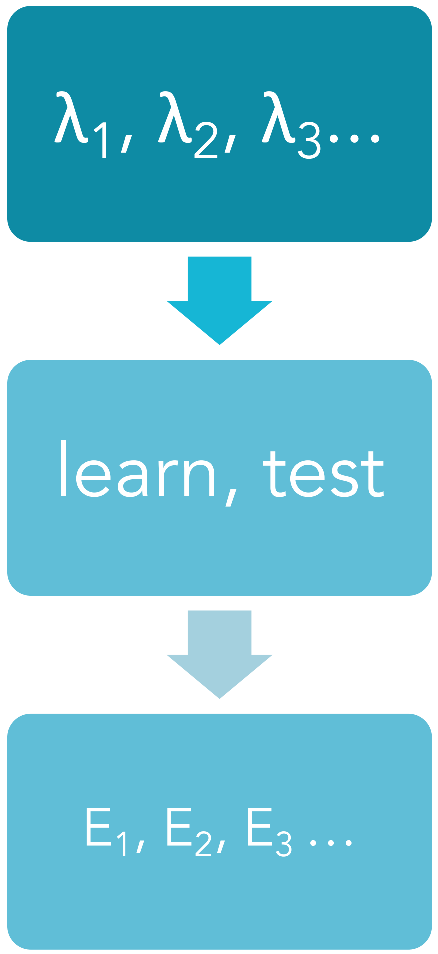
- Minimizes the R_{in}

Regularization

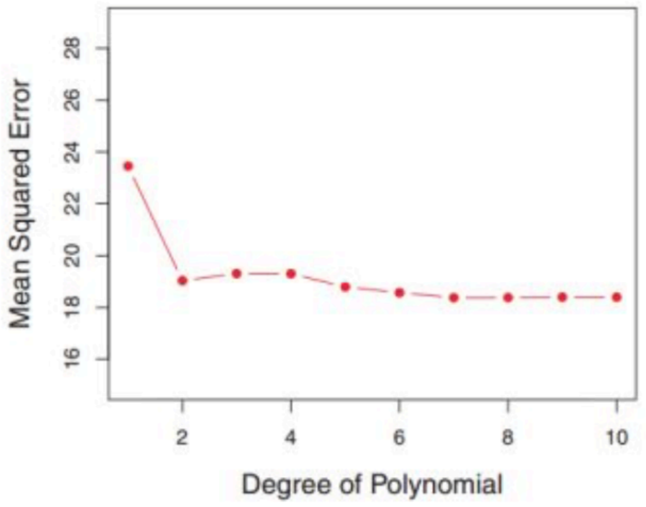


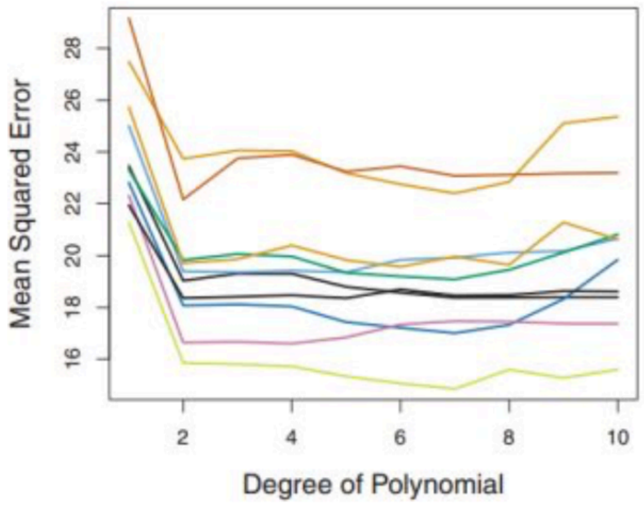
Only Regularization?

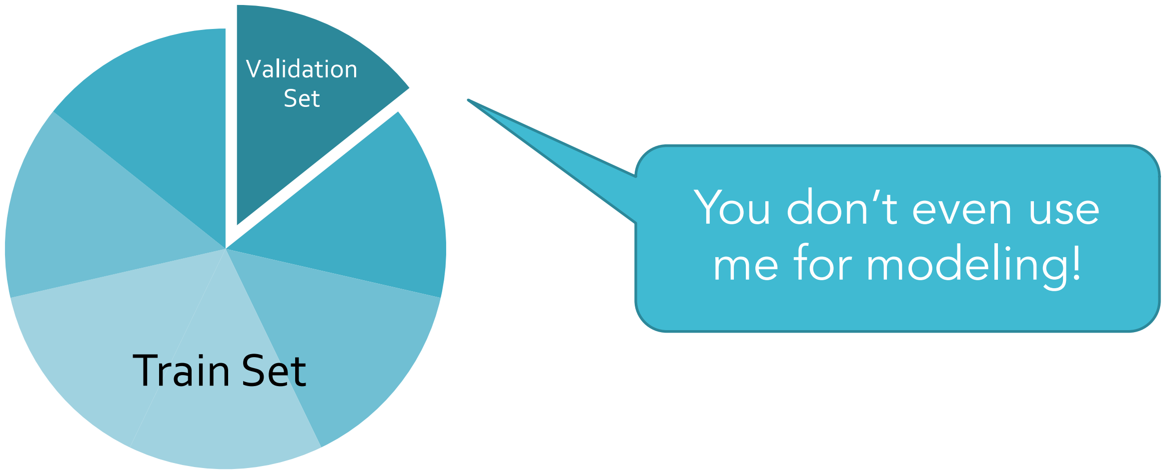
- "Optimal" λ ?
- We need a validation set



Why do we need Cross Validation?







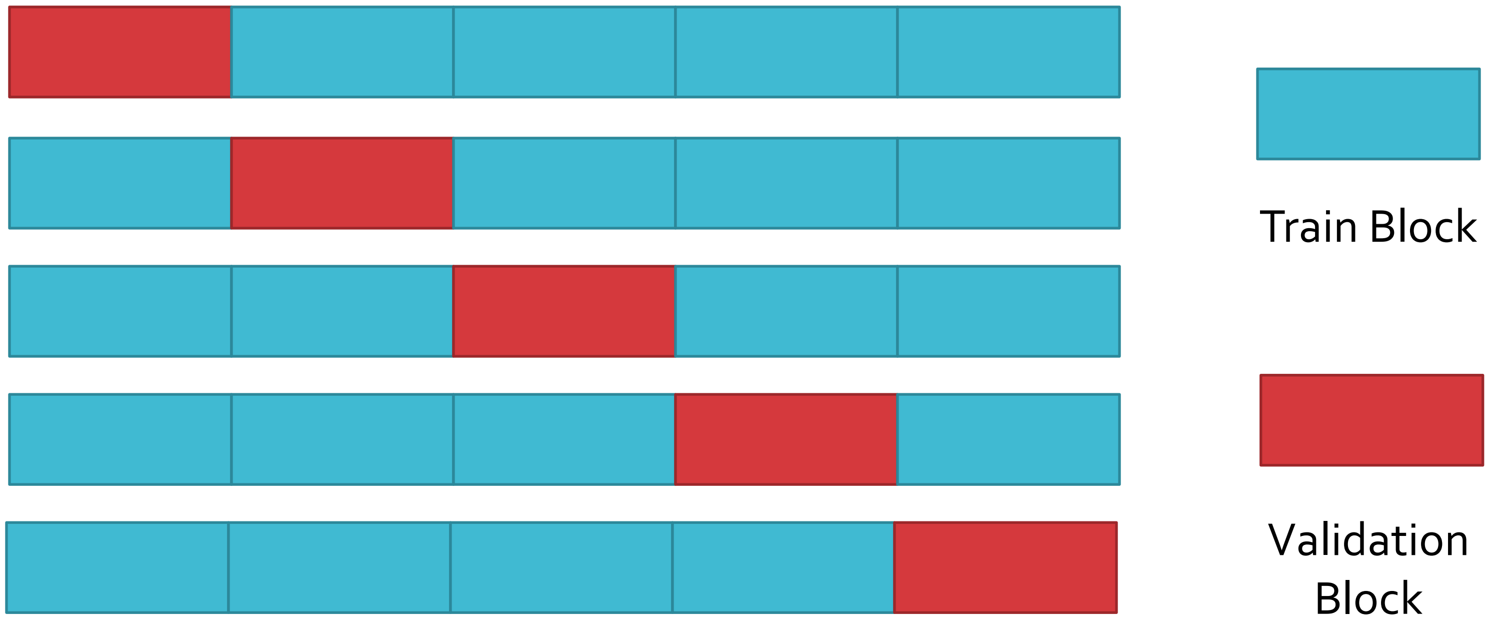
small validation set ⇒ large error in estimated loss
large validation set ⇒ small training set ⇒ bad model

Cross Validation

Estimate the "optimal" λ by using it

K-fold cross validation

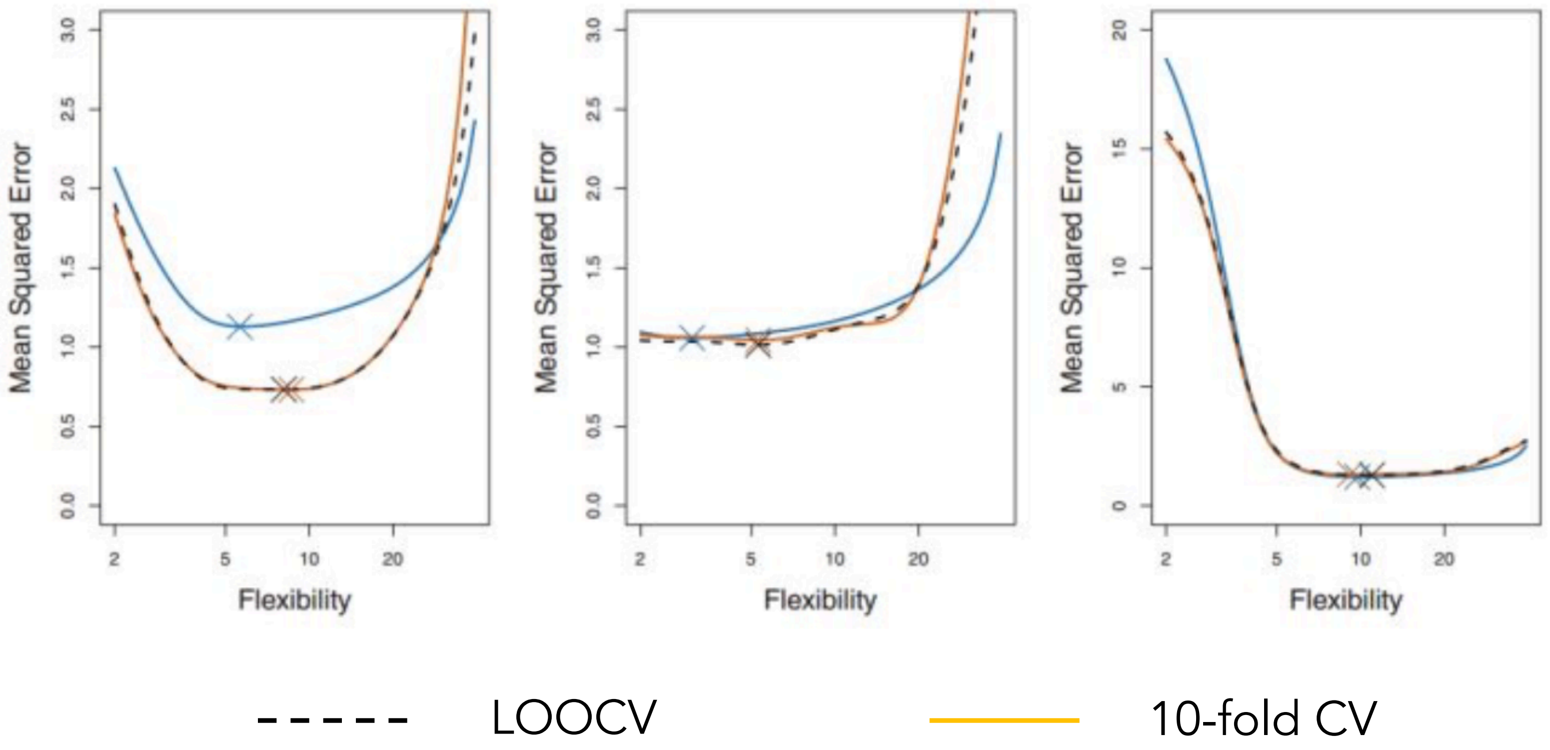
Divide the dataset into k blocks
for k = 1 to k
train on blocks except kth block, test on kth block
average the results, choose best λ.



K-fold cross validation

Common cases: K = 5, 10 or K = N (LOOCV)

High computation cost: K folds × many choices of model or λ



Summary

Regularization

&

Cross Validation

Regularization & Cross Validation

- Trading off bias and variance is hard.
  - Degree of Polynomial ↑ Bias ↓ Variance ↑
- Regularization penalizes hypothesis complexity
  - L2 regularization leads to small weights
  - L1 regularization leads to many zero weights (sparsity)
- Cross-validation enables selection of regularization penalties by estimating test-set error on parts of the training set

Demo

Regularization

&

Cross Validation