REGULARIZATION

The learning problem

- Given a set of examples (the training set) : $\{(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}\}_{i=1}^{n}$
- Find f such that: $f(x) \sim y$
- We need a function with good ***generalization***
  - A function that performs well on new unseen data
- To evaluate model generalization, we need to measure errors
- Cost function **measures the performance of a Machine Learning model** for given data.

Cost functions

- Cost function quantifies the error between predicted values and expected values and **presents it in the form of a single real number**.
- We have various measures of model error
- Mean Absolute Error (MAE)

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}|$$

- Mean Square Error (MSE)

$$MSE = \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y})^2$$

- Root Mean Square Error (RMSE)

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y})^2}$$

An optimization problem

- Building a model is an optimization problem
- Machine learning goal
  - learn $f(x)$ such that the cost function $J(yi, f(x_i))$ is minimized
- For parametric models:
  - find the optimal configuration of model parameters w_i that minimizes the cost function J
- Example of linear regression ($f_w(x) = w^T x + w_0$)

$$\min_{w,w_0} J(w, w_0) = \frac{1}{N} \sum_{i=1}^{N} (f_w(x_i) - y_i)^2$$

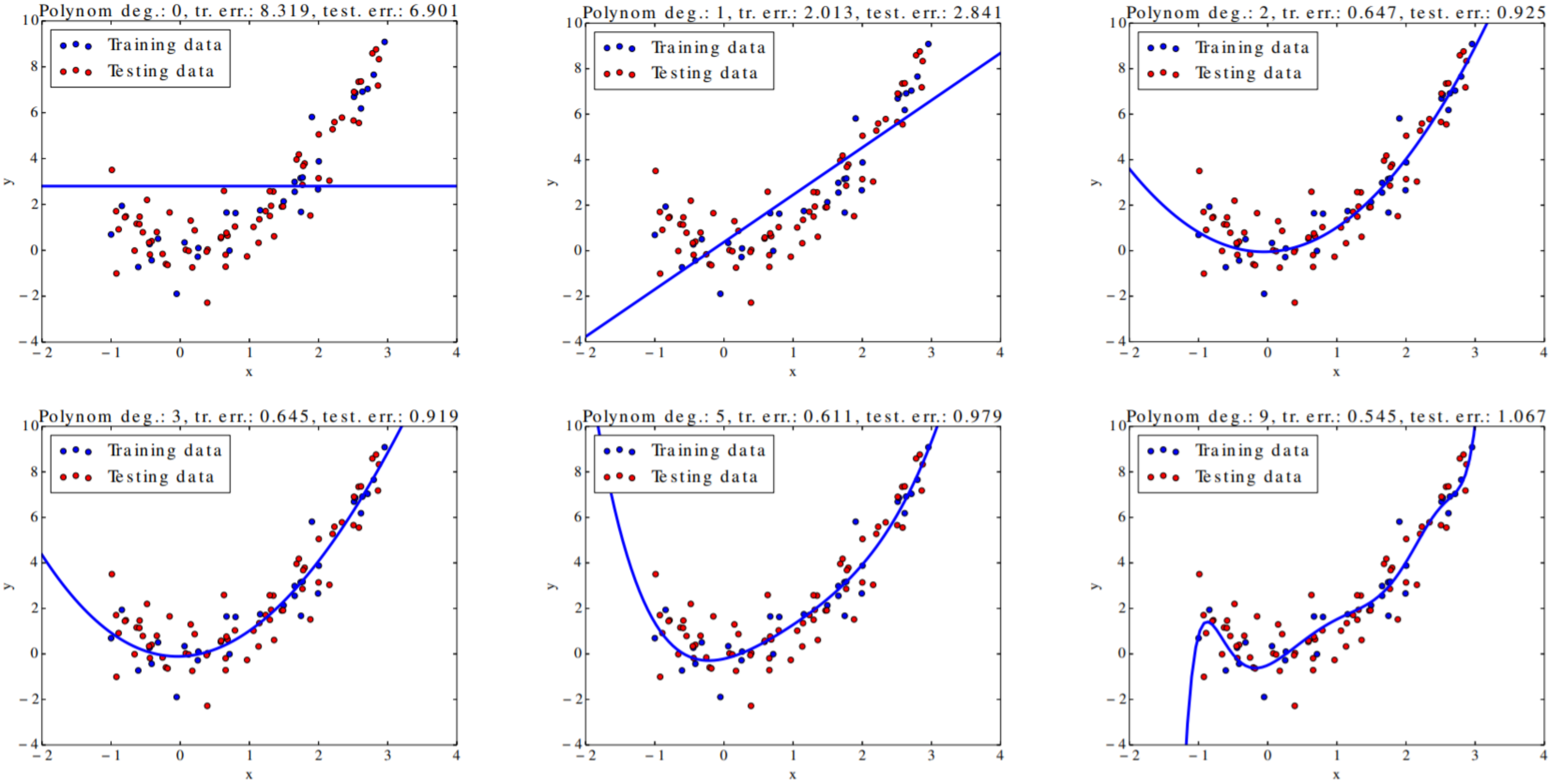
Training and testing error

- Example:

  Polynomial regression with varying degree

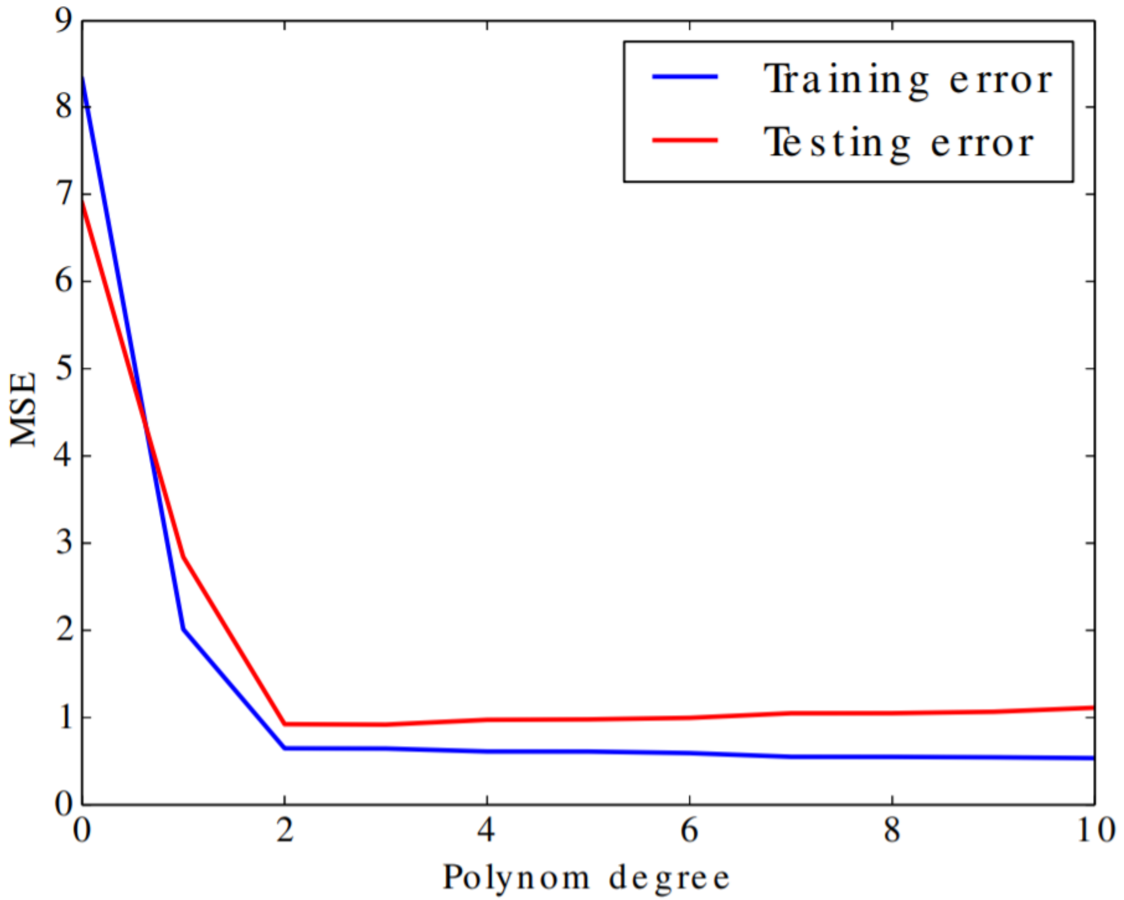
$$X \sim U(-1, 3)$$
$$Y \sim X^2 + N(0, 1)$$

Training and testing error



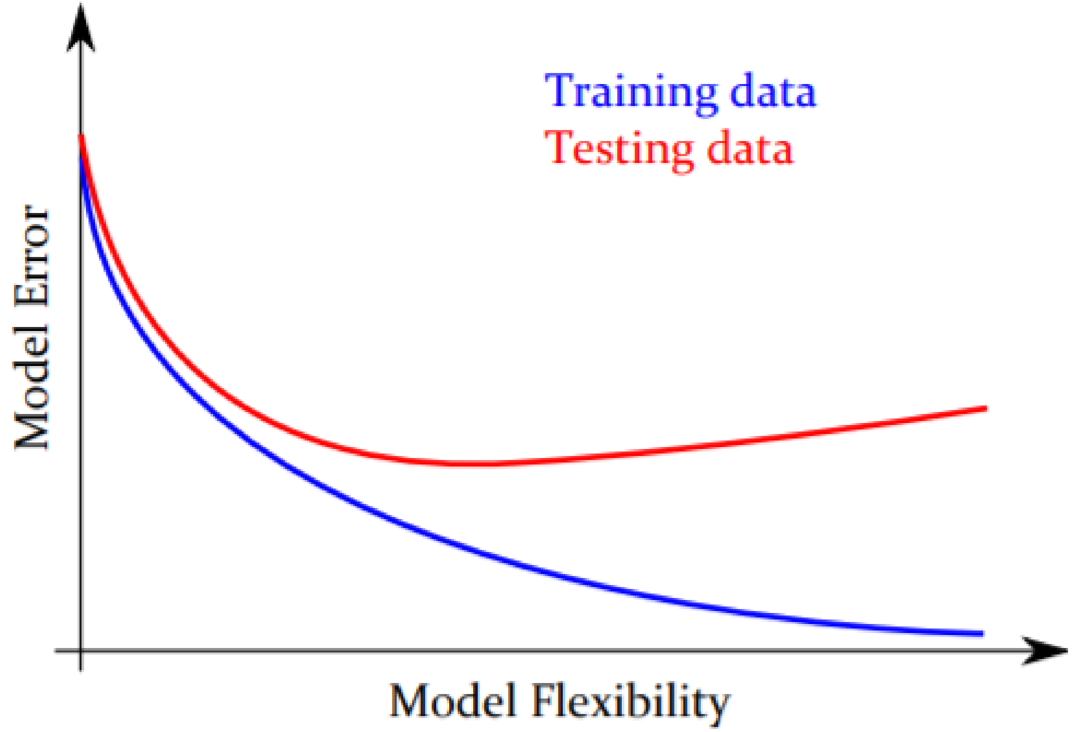
Training and Testing Errors

- The training error decreases with increasing model flexibility (the curviness).
- The testing error is minimal for certain degree of model flexibility.

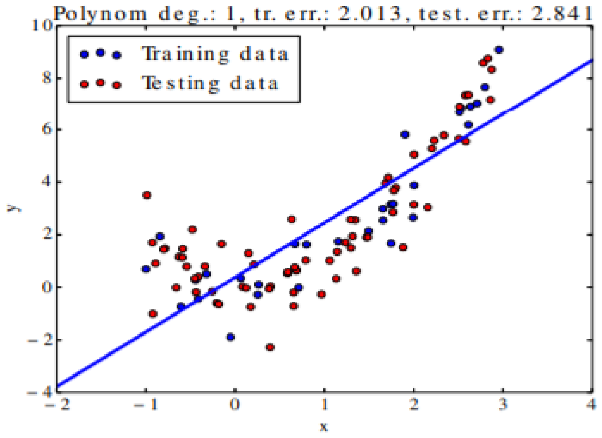


Overfitting

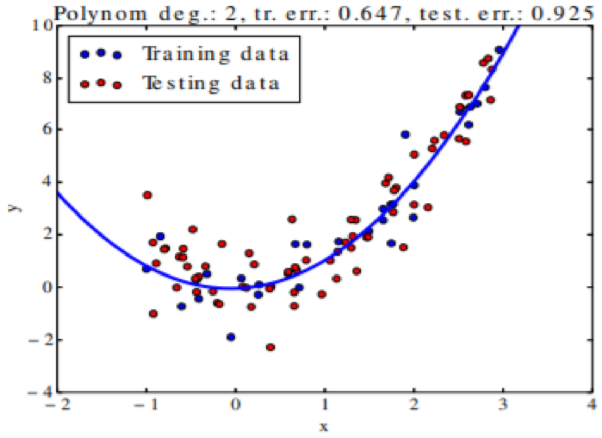
- Overfitting is a general phenomenon affecting all kinds of inductive learning.
- When overfitted, the model works well for the training data, but fails for new (testing) data.



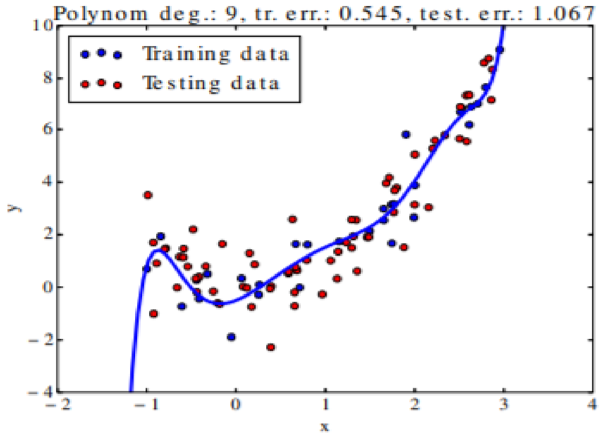
Bias vs Variance



High bias:
model not flexible enough
(Underfit)



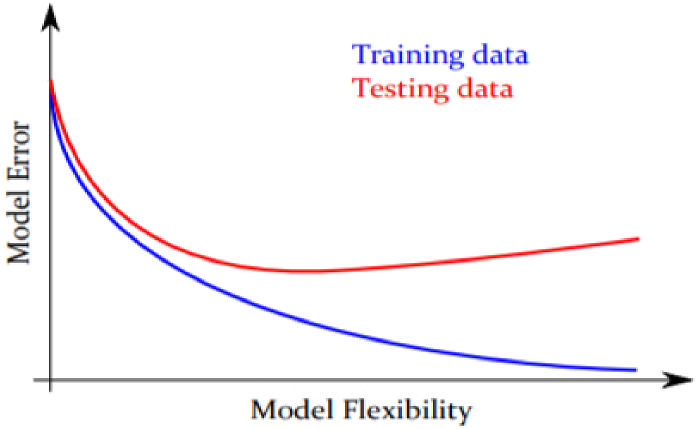
"Just right"
(Good fit)



High variance:
model flexibility too high
(Overfit)

High bias problem:

- Err_{Tr} is high
- $Err_{Tst} \approx Err_{Tr}$



High variance problem:

- Err_{Tr} is low
- $Err_{Tst} >> Err_{Tr}$

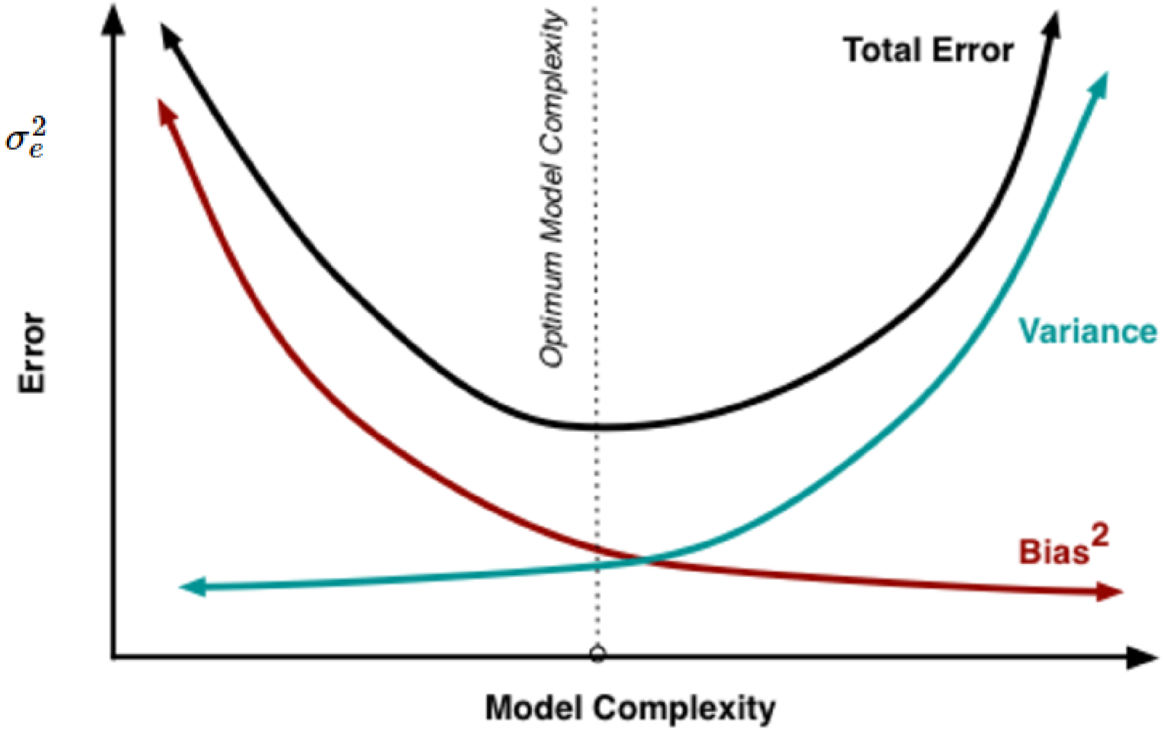
Bias-Variance Tradeoff

- To get good predictions, we need to find a balance of Bias and Variance that minimizes total error

$$Err(x) = \left(E[\hat{f}(x)] - f(x)\right)^2 + E\left[\left(\hat{f}(x) - E[\hat{f}(x)]\right)^2\right] + \sigma_e^2$$

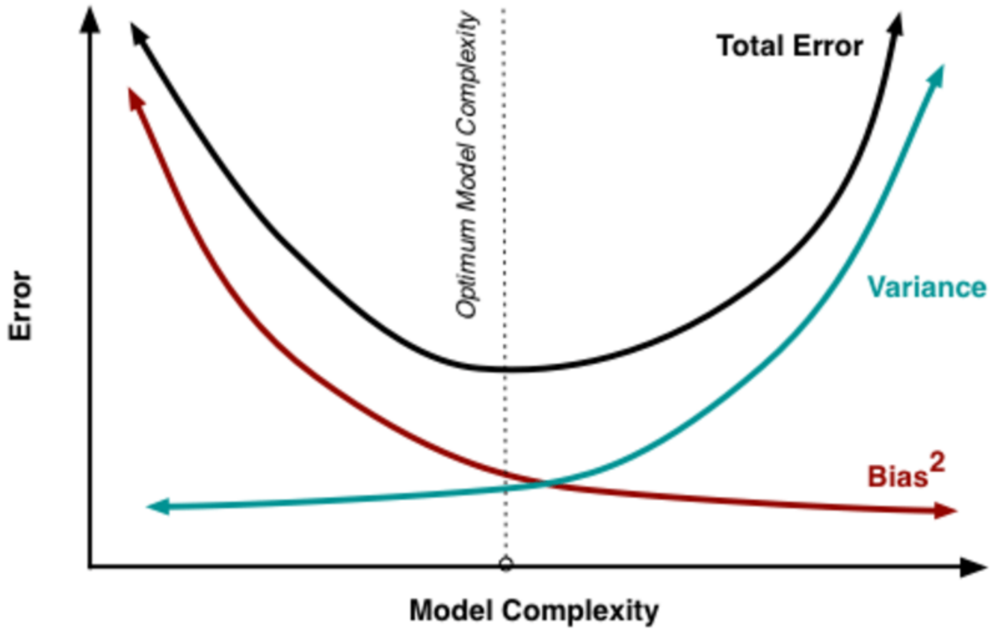
$$Err(x) = \mathrm{Bias}^2 + \mathrm{Variance} + \mathrm{Irreducible\ Error}$$

For error decomposition proof:
https://robjhyndman.com/files/2-biasvardecomp.pdf



Regularization

- Reduce model variance at the cost of introducing some bias.
- A regularization technique is a penalty mechanism which applies shrinkage (driving them closer to zero) of coefficient to build a more robust and parsimonious model.
  - Reduce model complexity



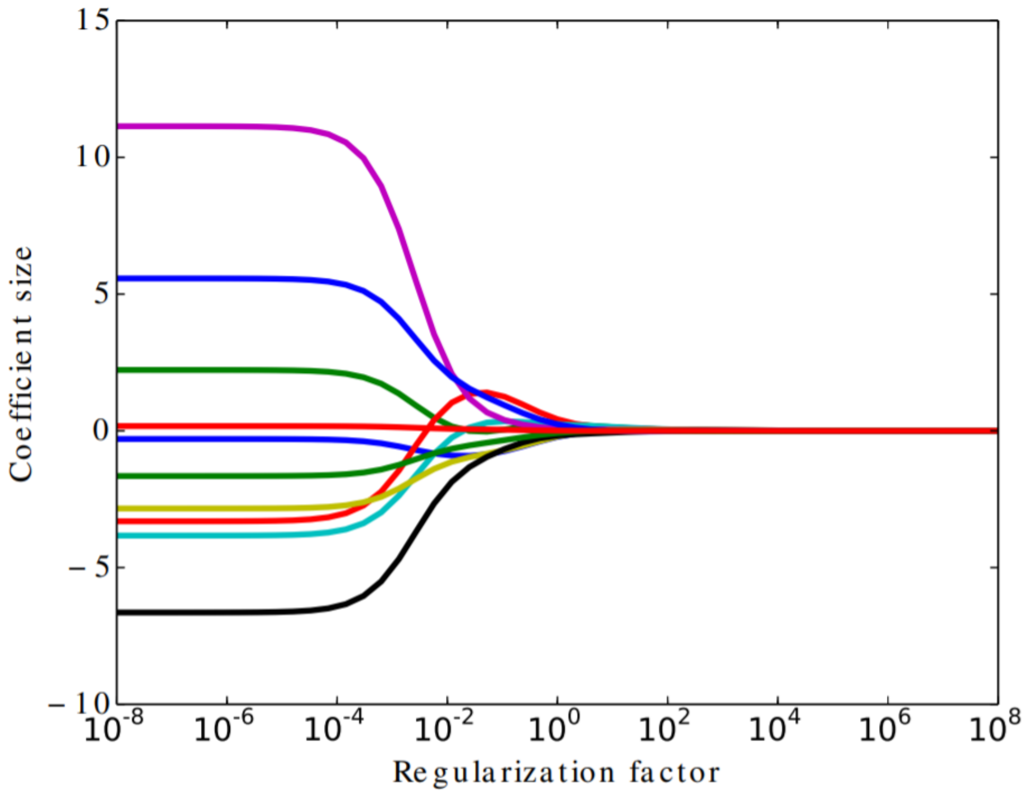
Ridge regularization (L2 regularization)

- Ridge regularization imposes a penalty on the size of the model coefficients.

- The penalty is equal to the sum of the squared value of the coefficients w_i

$$Error_{Ridge} = Error + \alpha \sum_{i=1}^{d} {w_i}^2$$

- α is the regularization parameter
- Ridge forces **the parameters to be relatively small**
- The bigger the penalization, the smaller (and the more robust) the coefficients are.



- $\alpha \to 0, \quad w^{ridge} \to w^{Error}$
- $\alpha \to \infty, \quad w^{ridge} \to 0$

Lasso regularization (L1 regularization)

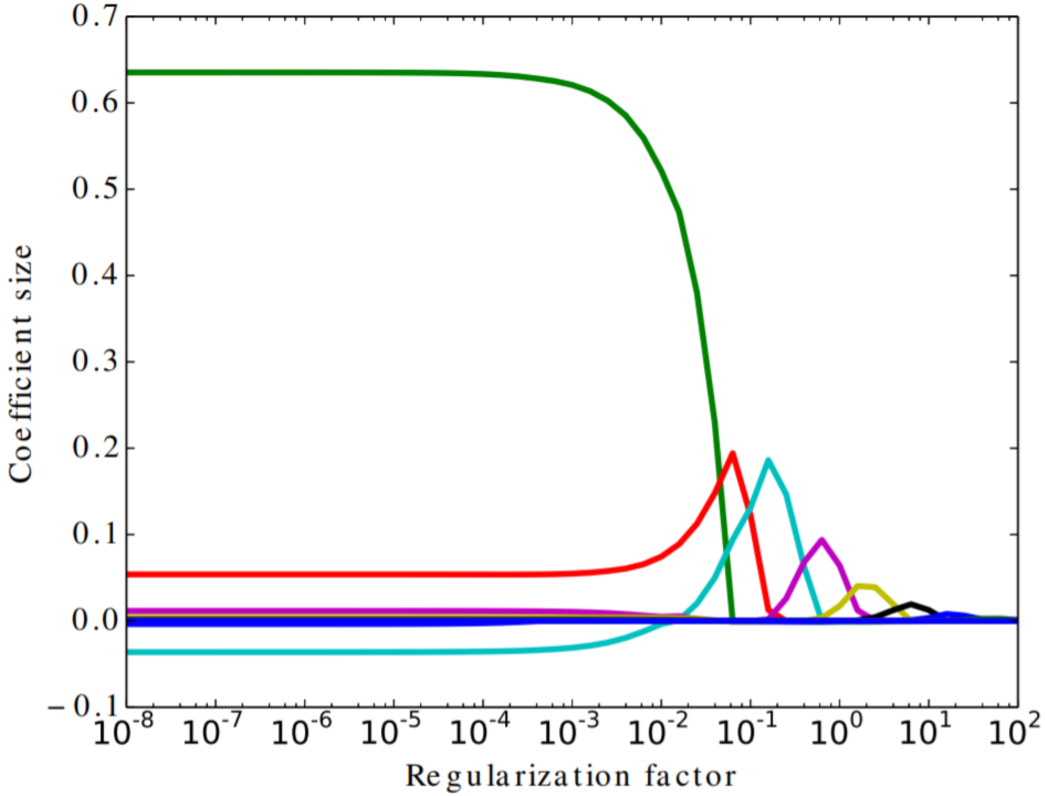
- Like ridge, Lasso regularization penalizes the size of the model coefficients.

- The penalty is equal to the sum of the absolute value of the coefficients w_i

$$Error_{Lasso} = Error + \alpha \sum_{i=1}^{d} |w_i|$$

- Lasso regularization will **shrink some parameters to zero**.

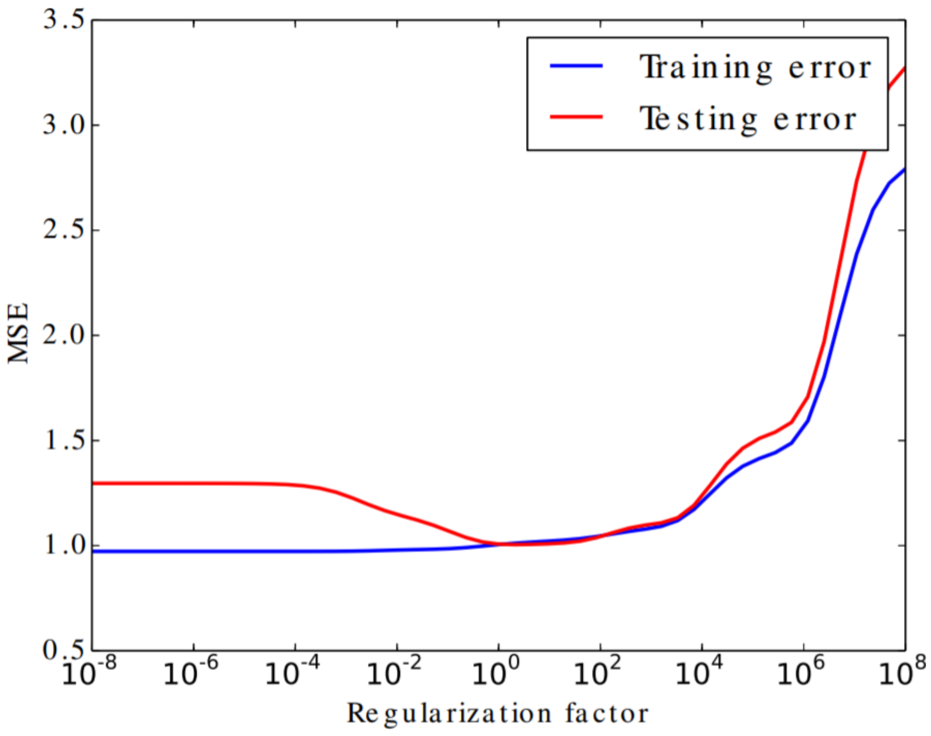
- It can be seen as a way to select features in a model.



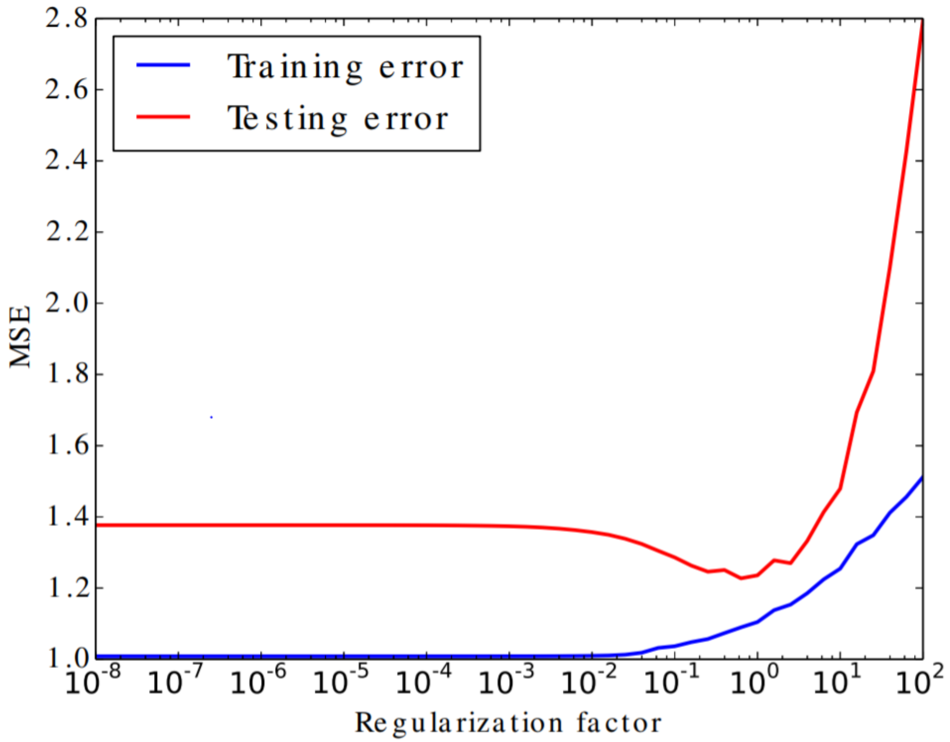
- $\alpha \to 0, \quad w^{Lasso} \to w^{Error}$
- $\alpha \to \infty, \quad w^{Lasso} = 0$

Ridge vs Lasso

- Evolution of error vs regularization factor



Ridge (L2 regularization)



Lasso (L1 regularization)

When the LASSO fails?

- In the $p > n$ case, the lasso can select at most n variables before it saturates
- If there is a group of variables with very high pairwise correlations, the lasso tends to select only one variable from the group, not caring which one
- For usual $n > p$ situations, if there are high correlations between predictors, the prediction performance of lasso is poor with respect to ridge.

Elastic-net

- The lasso sometimes does not perform well with highly correlated variables, and often performs worse than ridge in prediction
- Elastic-net is a mix of **both Ridge and Lasso regularizations**.

$$Error_{Elastic} = Error + \alpha \left(\rho \sum_{i=1}^{d} |w_i| + \left(\frac{1-\rho}{2}\right) \sum_{i=1}^{d} {w_i}^2\right)$$

- α is a shared penalization parameter
- ρ is a mixing parameter, it sets the ratio between ridge (L2) and lasso (L1) regularization
  - $\rho = 0$: Ridge regularization
  - $\rho = 1$: Lasso regularization

Summary

- **Ridge**: regularization
  - Shrink coefficients to zero but can not produce a parsimonious model
  - Similar estimated coefficient for highly correlated predictors
  - Exactly identical variables will have same coefficients
- **LASSO**: regularization + variable selection
  - Unlike Ridge, can set variables exactly to zero
  - If p>>n, select n variables only
  - Can not do grouped selection; select one variable if highly correlated variables
- **Elastic-net**: regularization + variable selection
  - Overcomes LASSO limitations by borrowing strength from Ridge
  - Allow selecting more than n variables
  - Allow selection of groups of correlated variables

Summary

Criteria to choose regularization method:

- Ridge Regression
  - When all the features you have are important to your model
  - When you don't want to do feature selection as well as feature removing
- Lasso Regression
  - When you have too many features
  - And you know some of them don't have any significance to your model
  - When you want to remove the features with less importance
- Elastic Net Regression
  - When you don't know whether all the features have significance or not
  - when there are strong correlations between features