Regularization

Importance of Regularization

- Used by almost all the linear models such as Linear Regression, Logistic regression as well as neural network

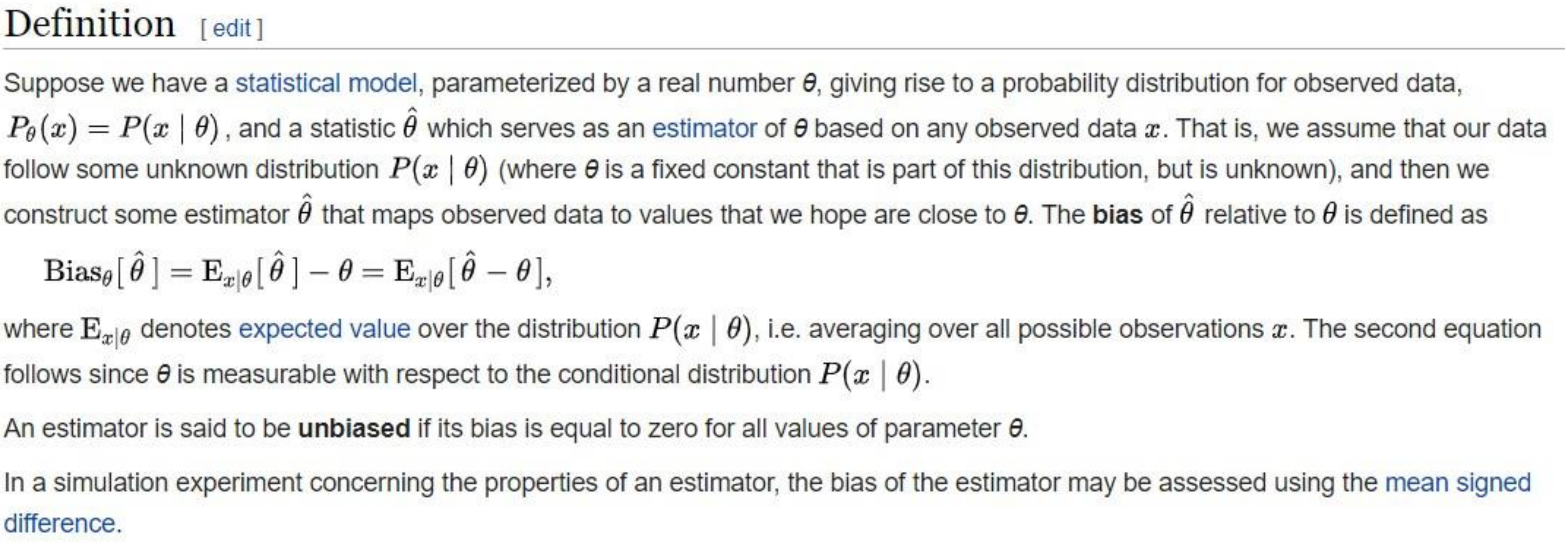
- One of the most important parameters

What we usually hear about regularization?

- Regularization prevents overfitting and improves generalization
- L1 or Lasso and L2 or Ridge regression or L1-L2 regularization
- Adds a penalty to the error term
- One penalizes the absolute term while the other penalizes in squared manner
- Used for the Bias-Variance trade-off
- One makes the coefficients to zero while the other makes them near zero

Bias Variance Trade Off

What is Bias?



Definition [edit]

Suppose we have a statistical model, parameterized by a real number θ, giving rise to a probability distribution for observed data, $P_\theta(x) = P(x \mid \theta)$, and a statistic $\hat{\theta}$ which serves as an estimator of θ based on any observed data x. That is, we assume that our data follow some unknown distribution $P(x \mid \theta)$ (where θ is a fixed constant that is part of this distribution, but is unknown), and then we construct some estimator $\hat{\theta}$ that maps observed data to values that we hope are close to θ. The **bias** of $\hat{\theta}$ relative to θ is defined as

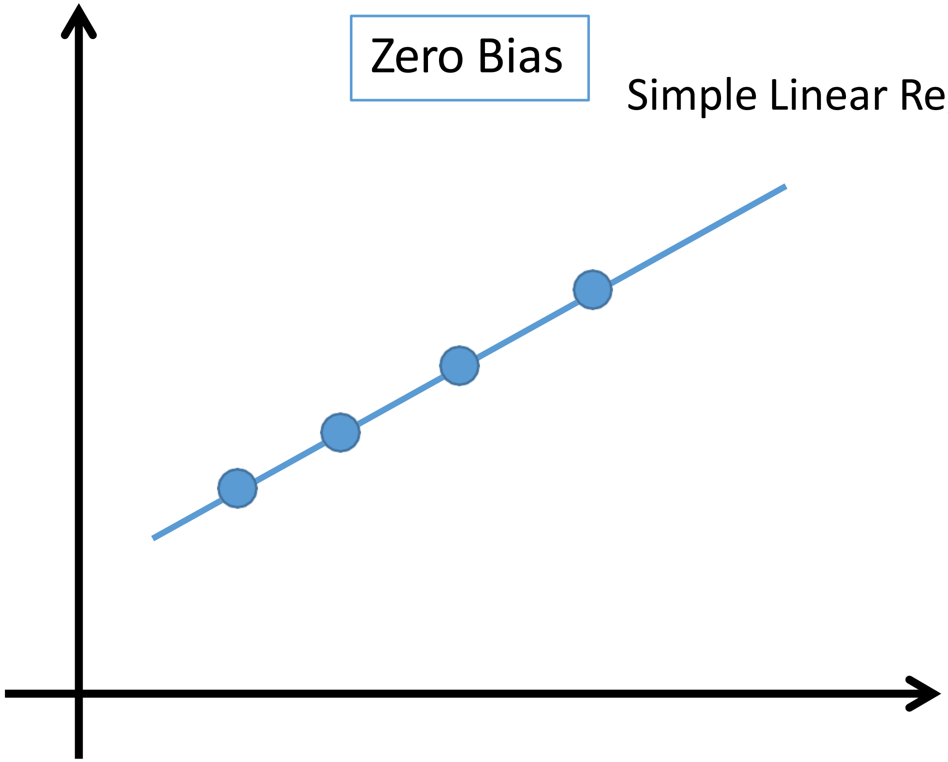
$$\operatorname{Bias}_\theta[\,\hat{\theta}\,] = \operatorname{E}_{x|\theta}[\,\hat{\theta}\,] - \theta = \operatorname{E}_{x|\theta}[\,\hat{\theta} - \theta\,],$$

where $\operatorname{E}_{x|\theta}$ denotes expected value over the distribution $P(x \mid \theta)$, i.e. averaging over all possible observations x. The second equation follows since θ is measurable with respect to the conditional distribution $P(x \mid \theta)$.

An estimator is said to be **unbiased** if its bias is equal to zero for all values of parameter θ.

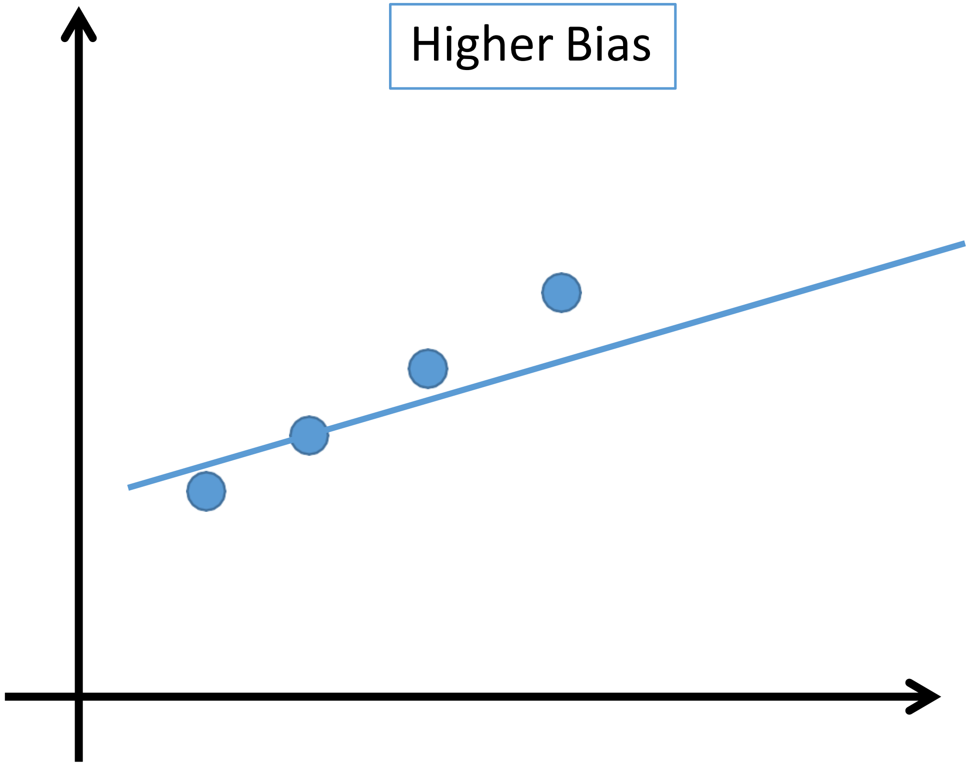
In a simulation experiment concerning the properties of an estimator, the bias of the estimator may be assessed using the mean signed difference.

What is Bias?



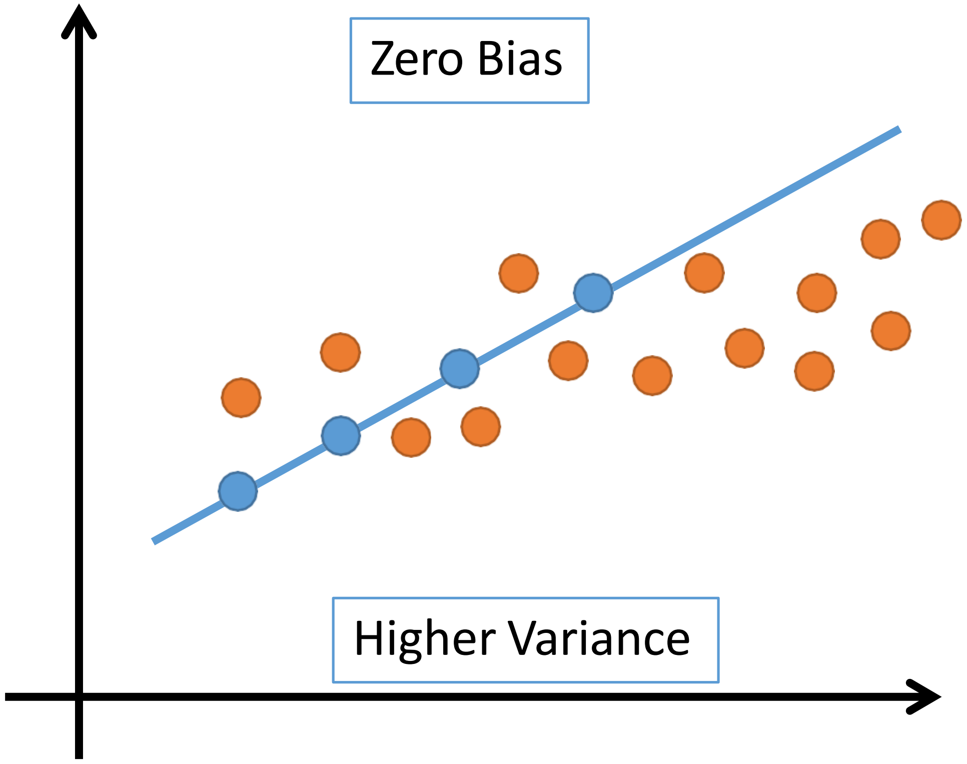
Zero Bias

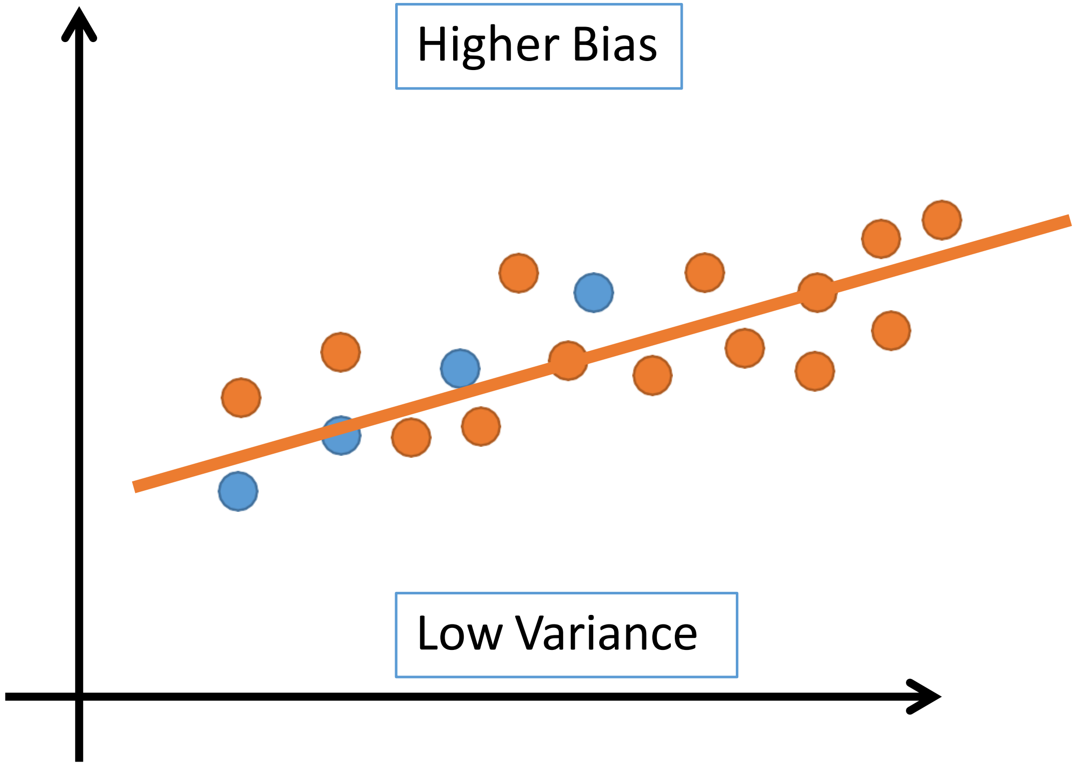
Simple Linear Regression Model



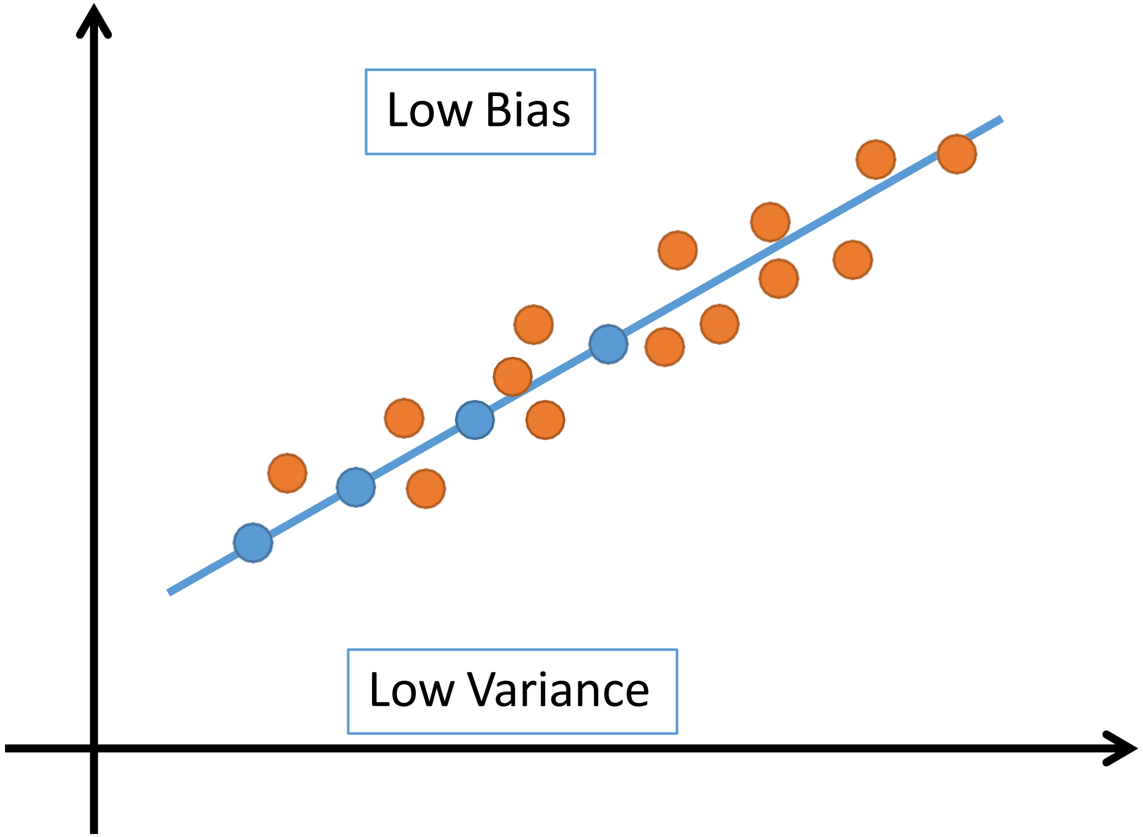
Higher Bias

What is Variance?





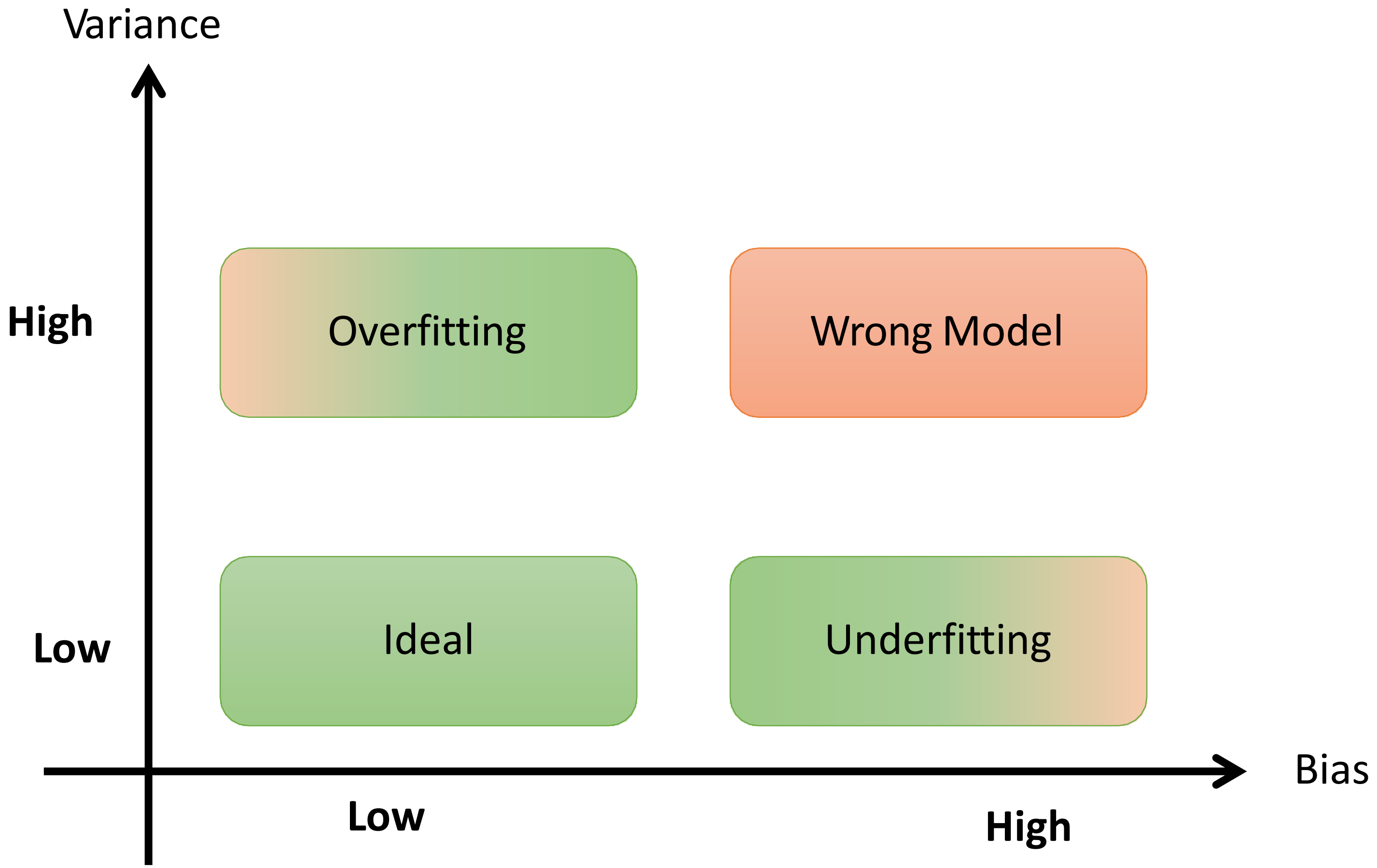
Ideal Scenario



Low Bias

Low Variance

Bias-Variance Tradeoff



Variance

High

Overfitting

Wrong Model

Low

Ideal

Underfitting

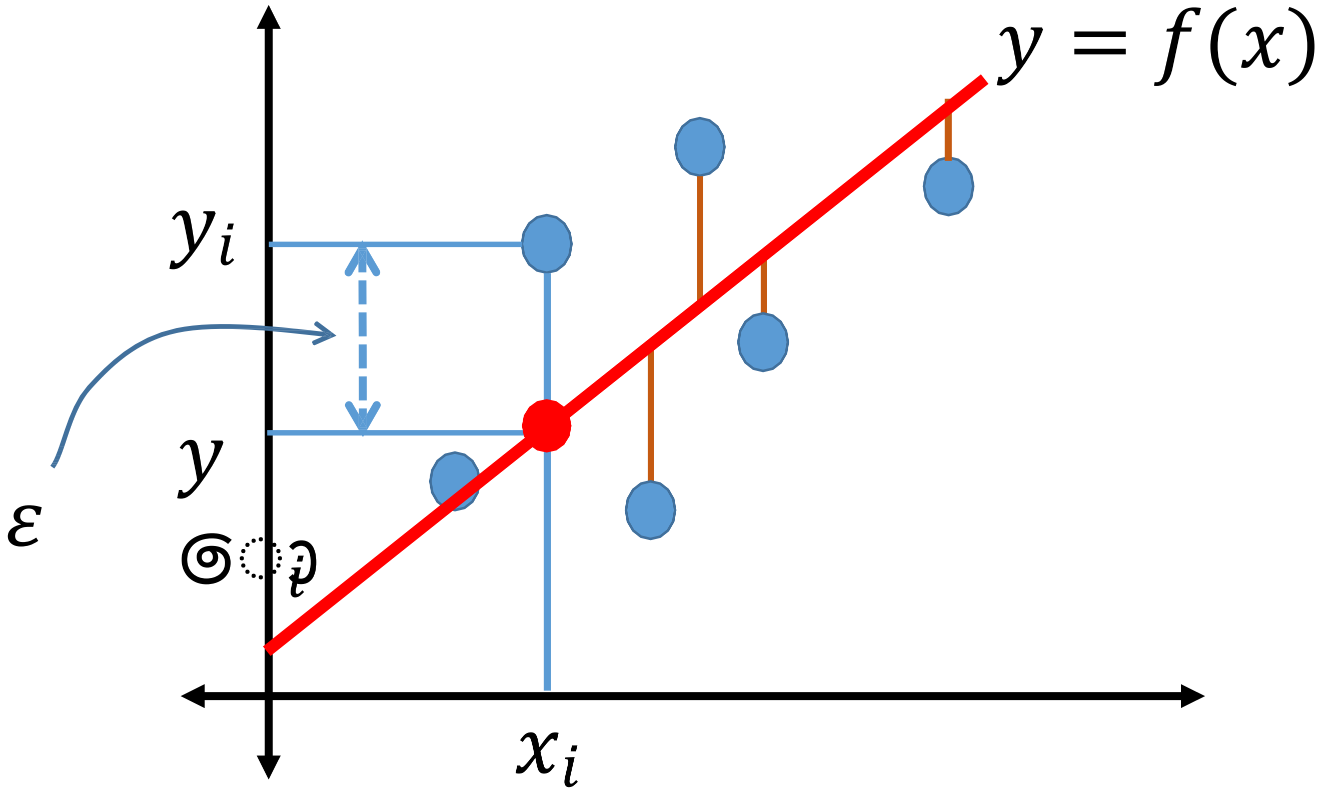
Bias

Low

High

Ridge Regression
or
L2 Regression

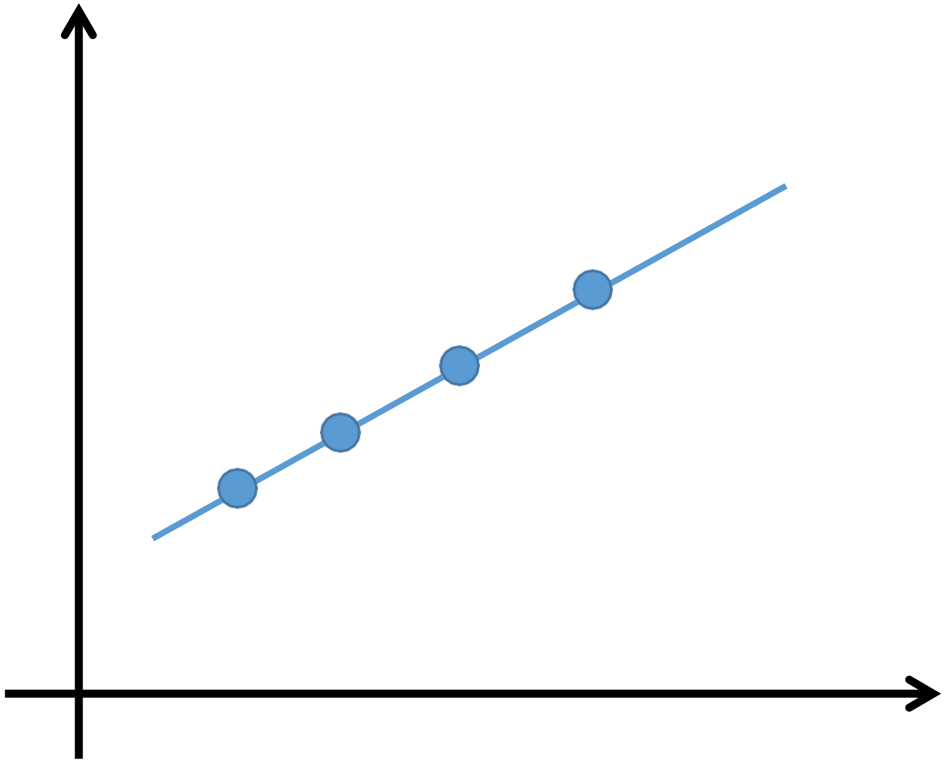
Ordinary Least Square Revisited



Minimum

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

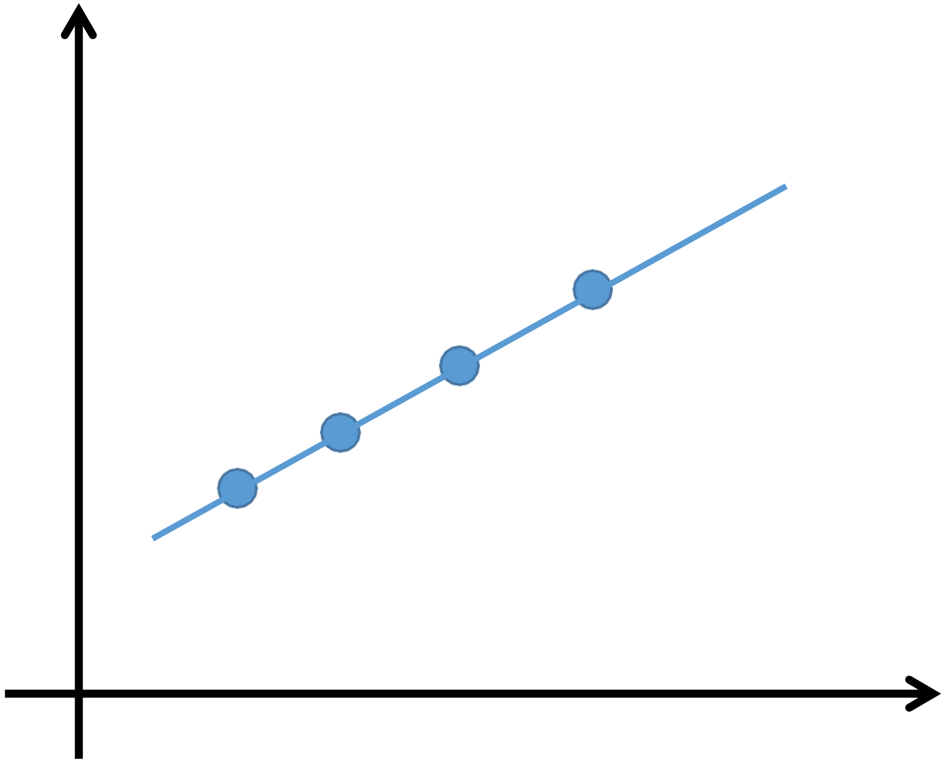
Ridge Regression



Minimum

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

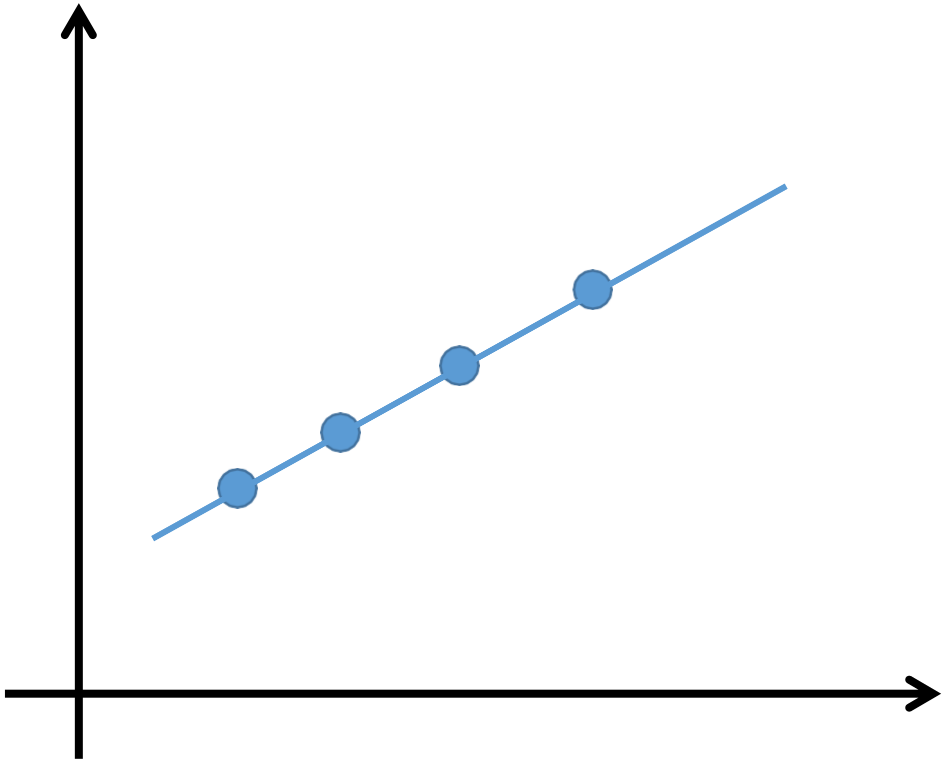
Ridge Regression



Minimum

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \color{red}{Penalty}$$

Ridge Regression

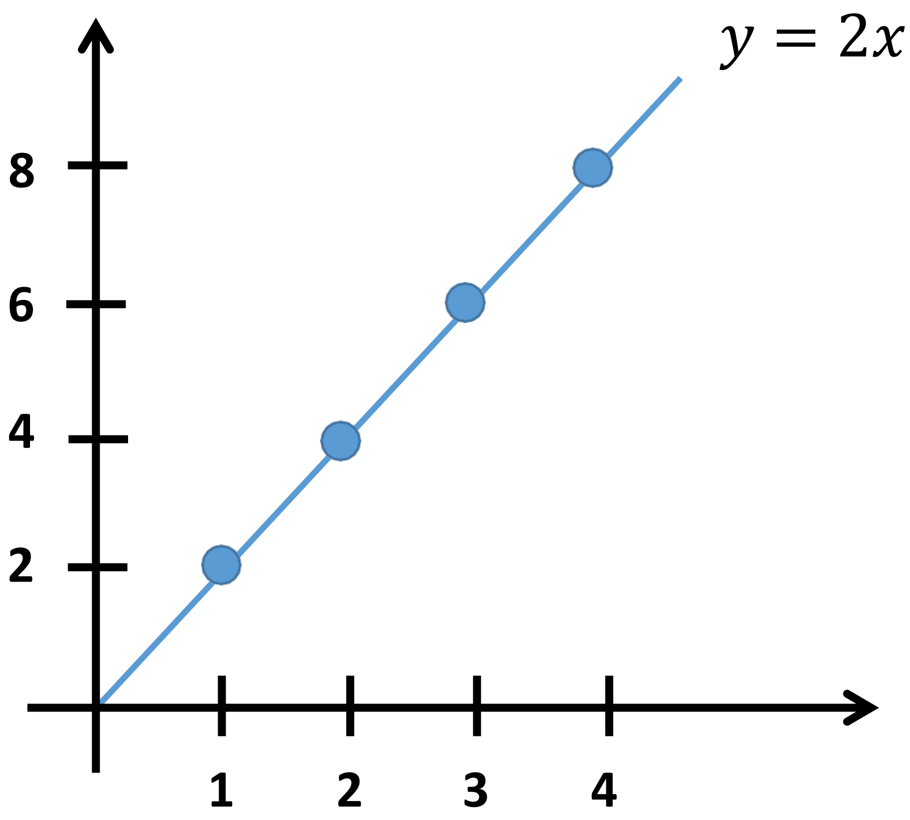


Minimum

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda * Slope^2$$

Understand using an Example !!

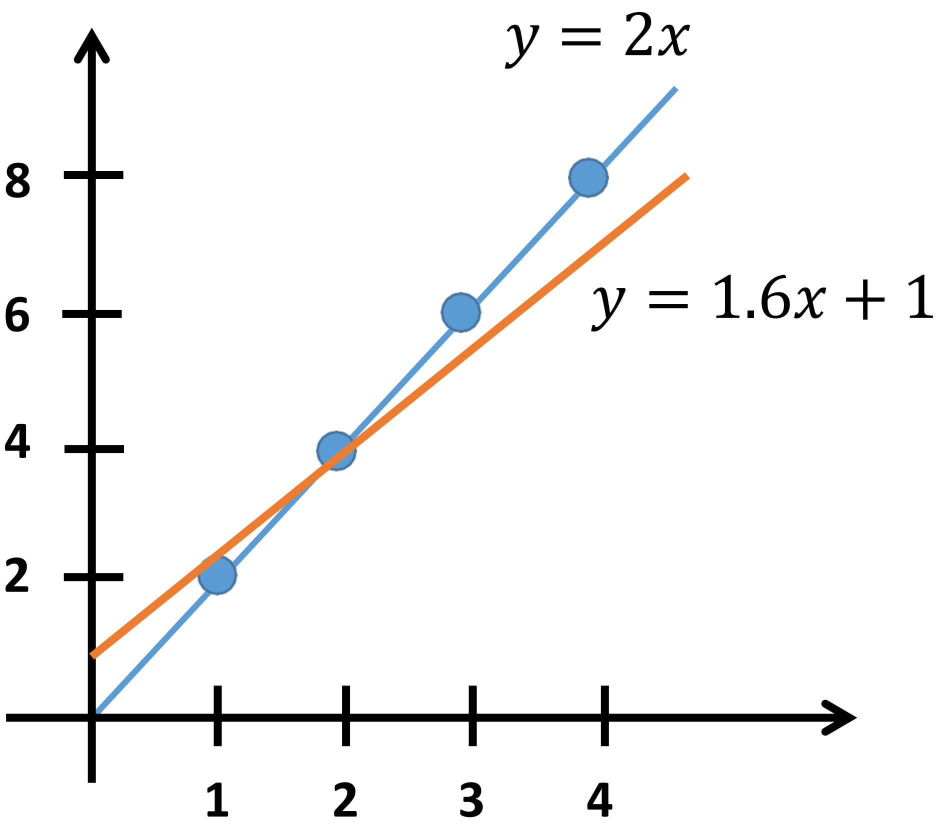
Ridge Regression



$Slope = 2 \qquad \lambda = 1$

OLS	Ridge
$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda * Slope^2$
	$0 + 1 * 2^2$
0	4

Ridge Regression



$Slope = 1.6$ $\lambda = 1$

$Intercept = 1$

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda * Slope^2$$

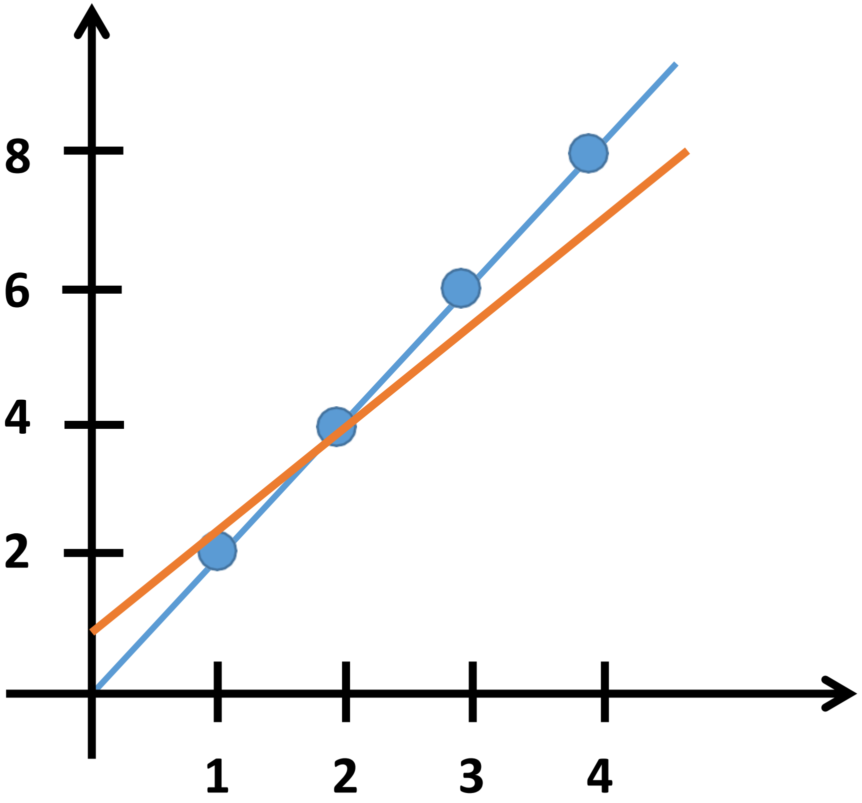
x	y	$\hat{y}$	$(y - \hat{y})^2$
1	2	2.6	0.36
2	4	4.2	0.04
3	6	5.8	0.04
4	8	7.4	0.36
Sum of Squared Differences			**0.80**

$0.8 + 2.56$

$\mathbf{3.36 < 4}$

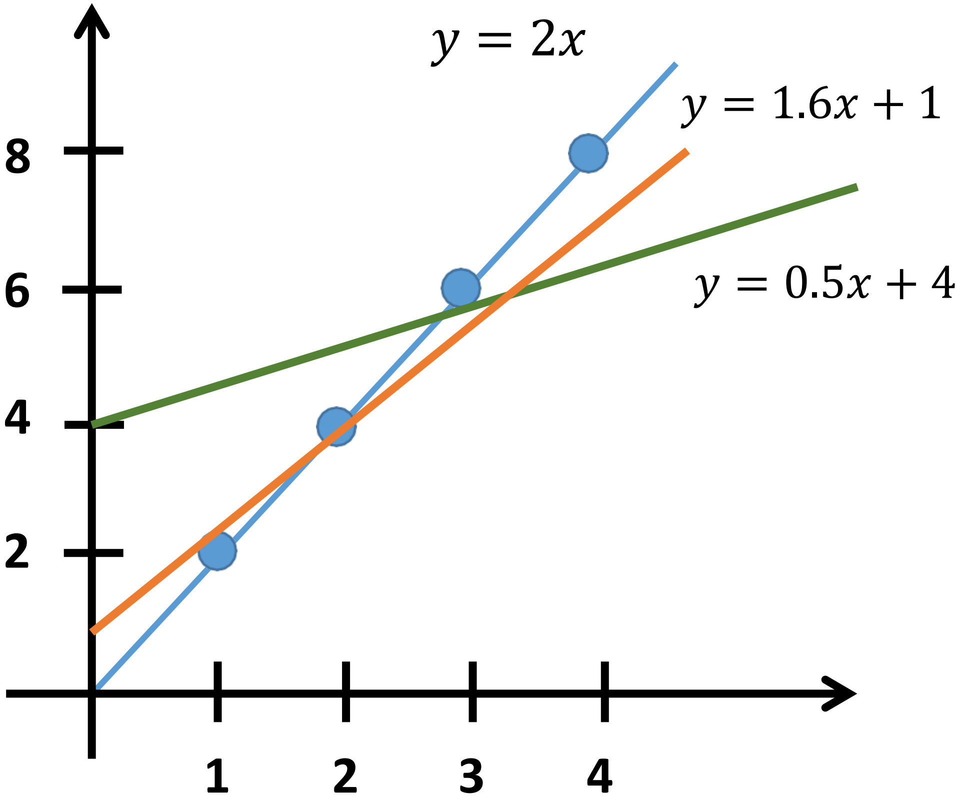
$Penalty = \lambda * Slope^2 = 1 * 1.6^2 = \mathbf{2.56}$

Ridge Regression



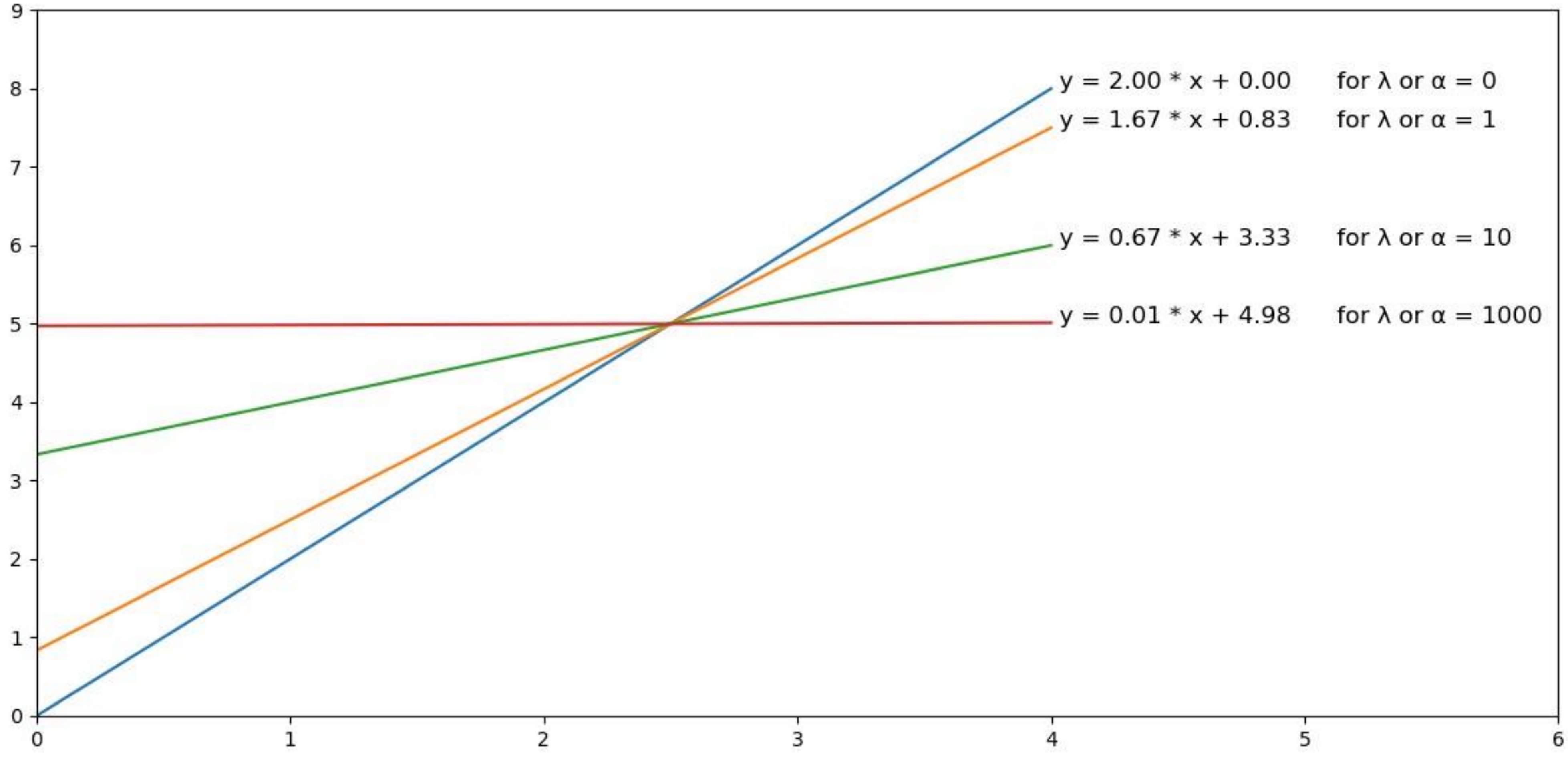
OLS	Ridge
$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda * Slope^2$
0	3.36 < 4
$y = 2x$	$y = 1.6x + 1$
Higher Dependency on X	**Lesser Dependency on X**

Effect of Lambda Values



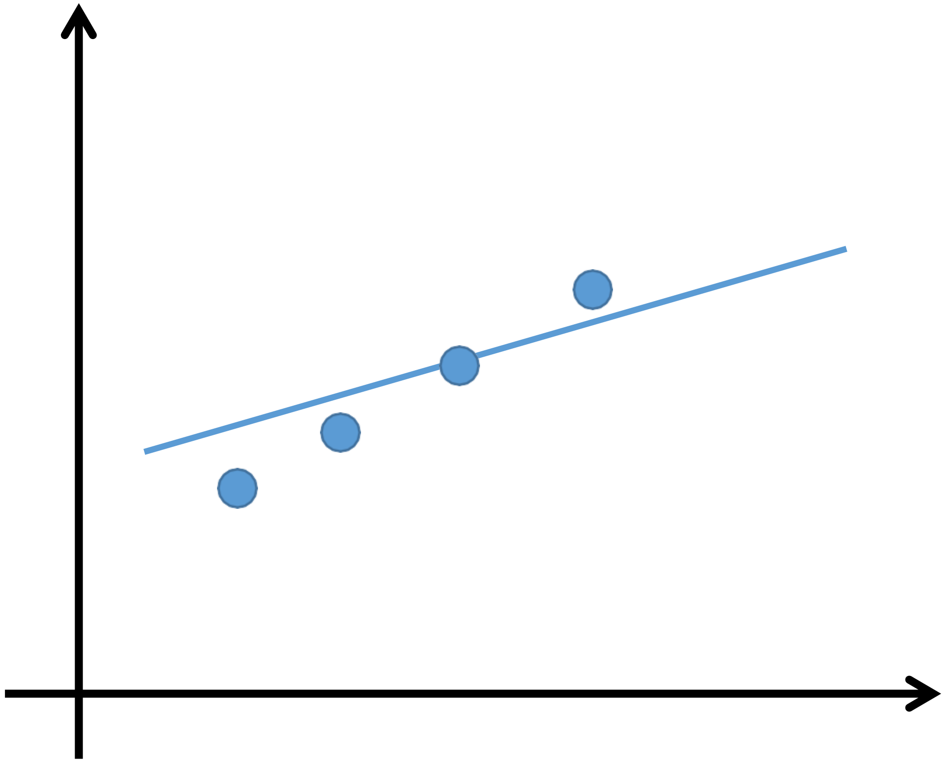
$\lambda = 1$	$\lambda = 10$
$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda * Slope^2$	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda * Slope^2$
3.36 < 4	14 < 40
$y = 1.6x + 1$	$y = 0.5x + 4$
Lesser Dependency on X	**Reduced Dependency on X**

Effect of Penalty Parameter



Lasso
or
L1 Regularization

Lasso Regression



Minimum

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda * |Slope|$$

Ridge

$\lambda * Slope^2$

Shrinks some of the coefficient **to near zero**.

All features are important.

Lasso

$\lambda * |Slope|$

Shrinks some of the coefficients **to zero.**

Some features can be eliminated.

Effect of Lasso and Ridge

Feature Selection

Multicollinearity

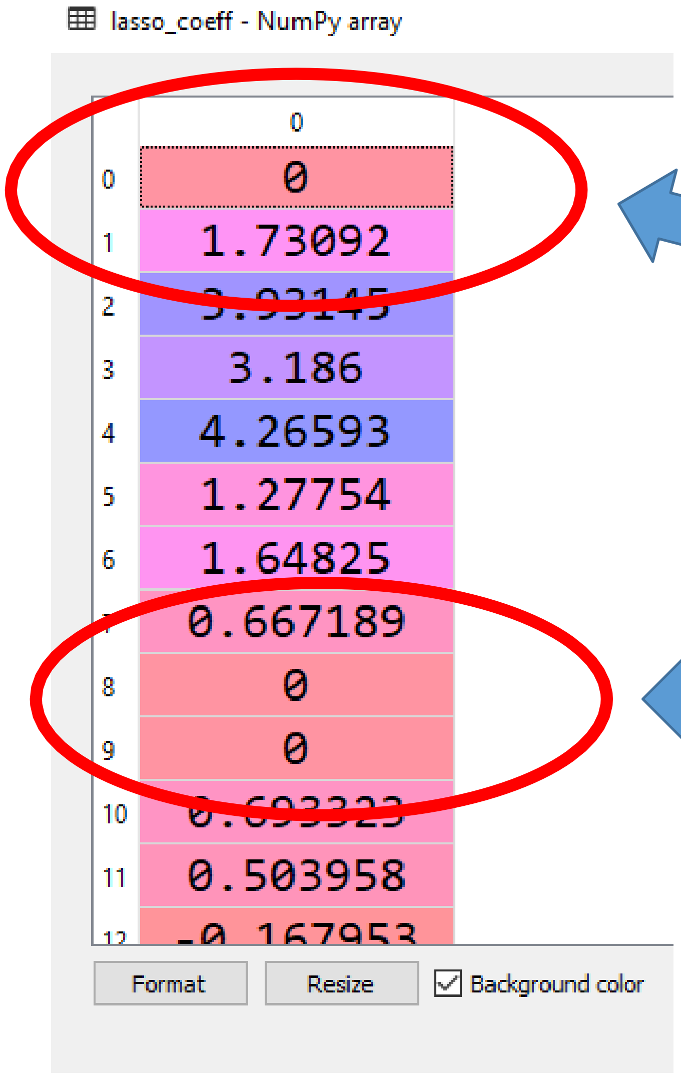
Dataset

$$x_2 = 1.8 * x_1$$

X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	Y
7	12.6	2	16	12	19	19	2	5	11	13	12	6	20	10	294.958
4	7.2	13	14	12	16	11	18	20	1	9	6	17	17	19	344.721
10	18	20	9	2	10	14	7	3	9	15	19	2	14	14	343.366
15	27	1	20	2	18	18	15	8	14	11	4	19	5	6	280.772
6	10.8	20	2	17	16	15	11	4	13	20	2	19	20	19	374.397
16	28.8	2	7	15	1	8	20	5	14	11	1	6	18	2	296.258
5	9	14	9	3	8	20	10	7	10	3	15	1	5	14	304.648

Decreasing Coefficients

Multicollinearity

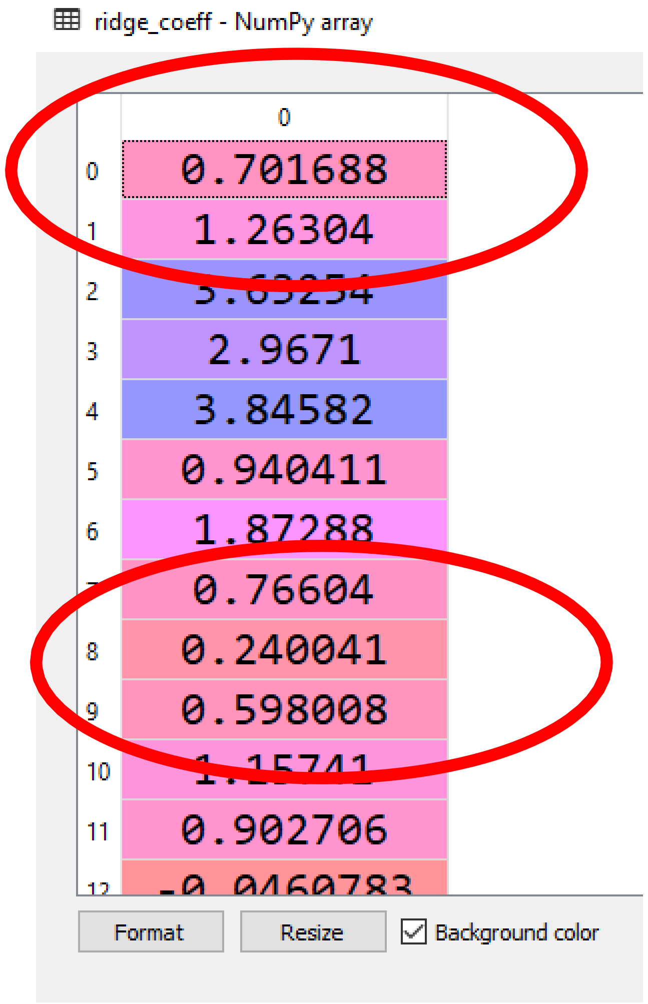


lasso_coeff - NumPy array

	0
0	0
1	1.73092
2	3.93145
3	3.186
4	4.26593
5	1.27754
6	1.64825
7	0.667189
8	0
9	0
10	0.693323
11	0.503958
12	-0.167953

Lasso Reduces features to zero.

Built-in or Embedded Feature Selection



ridge_coeff - NumPy array

	0
0	0.701688
1	1.26304
2	3.69254
3	2.9671
4	3.84582
5	0.940411
6	1.87288
7	0.76604
8	0.240041
9	0.598008
10	1.15741
11	0.902706
12	-0.0460783

Ridge

$\lambda * Slope^2$

Shrinks some of the coefficient **to near zero**.

Can not be used for feature selection.

Makes correlated features coefficients smaller.

Makes sense when all features are important.

Lasso

$\lambda * |Slope|$

Shrinks some of the coefficients **to zero.**

Performs Embedded feature Selection

Makes some of the correlated features irrelevant.

Can be used when some features can be eliminated.