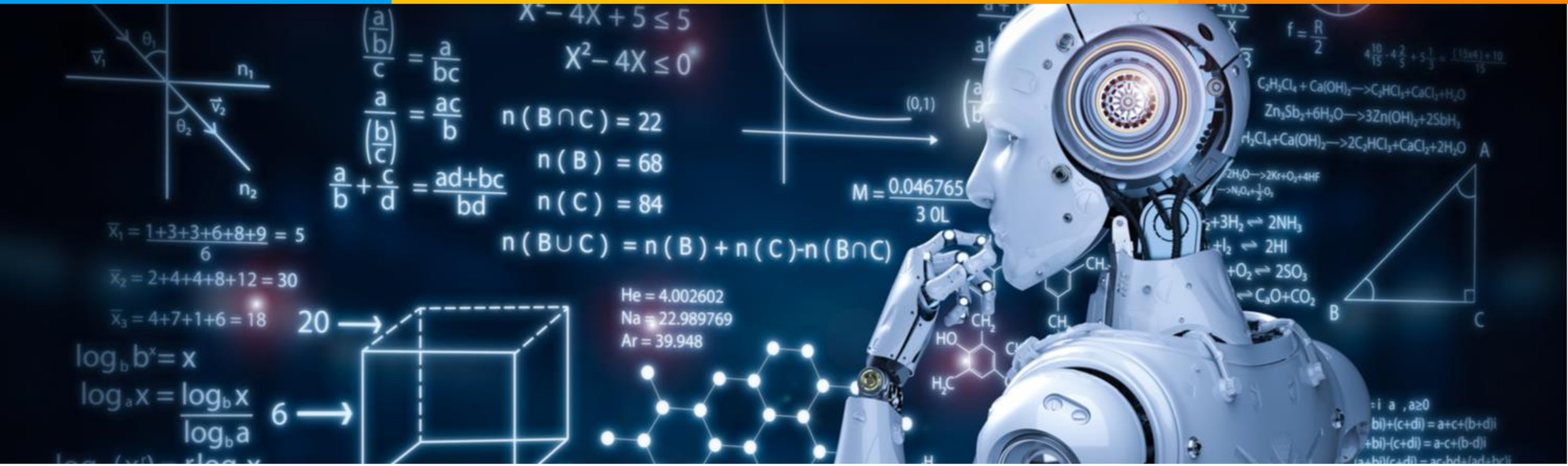


Lecture 24 & 25: Linear Regression Part 3

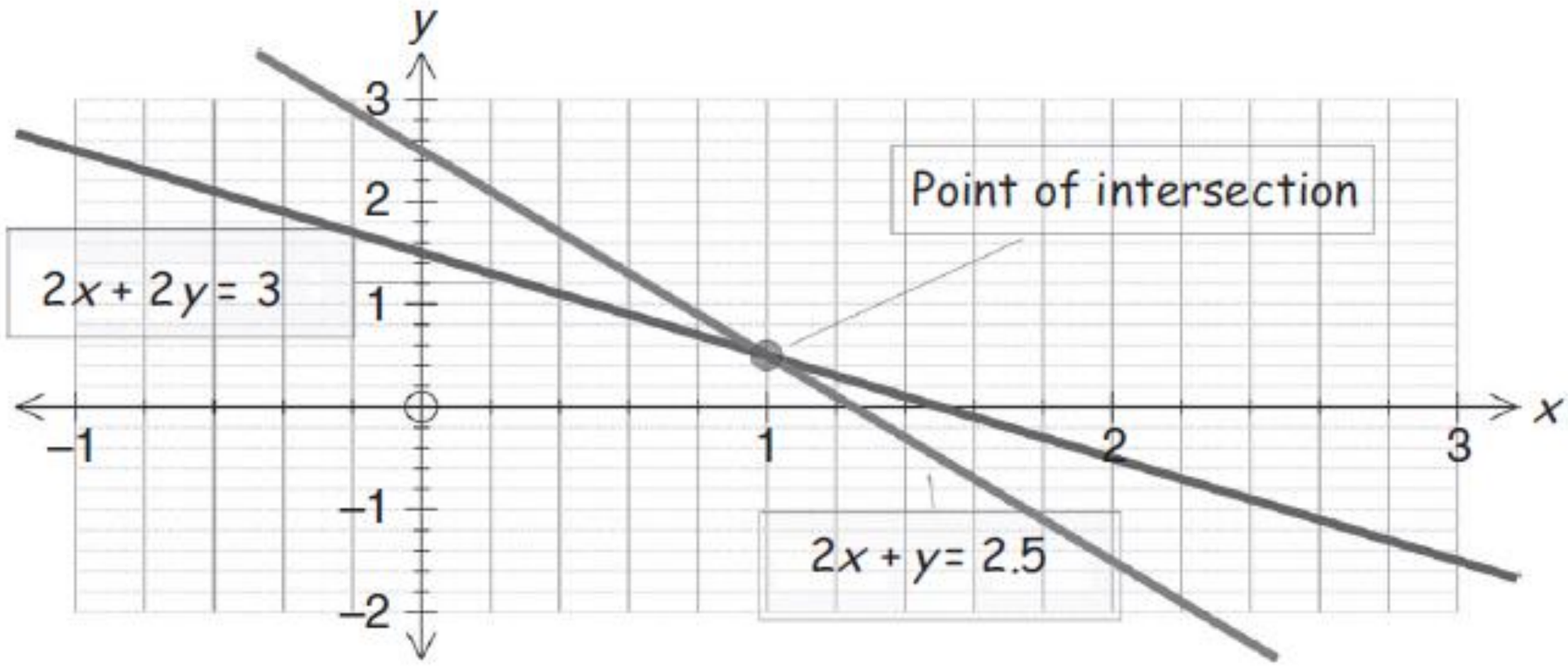
Recap

- Multivariate Calculus refresher
- Vectorized form of Linear Regression Objective function, Gradient Descent

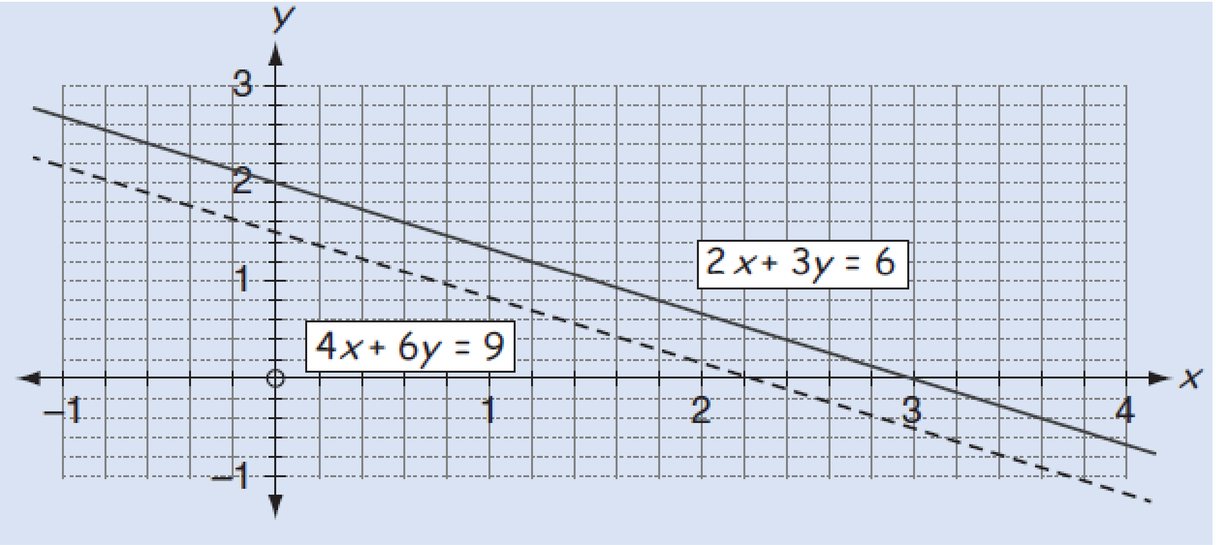
System of Equations

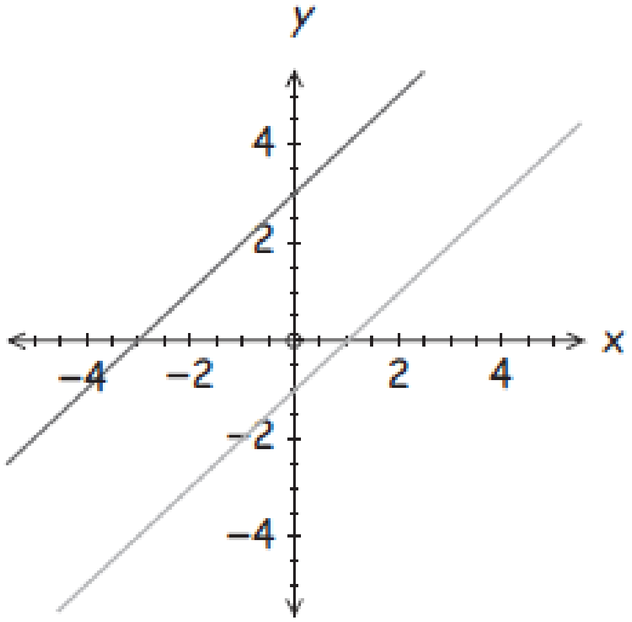
$$2x + 2y = 3$$
$$2x + y = 2.5$$

$$x = 1, \quad y = 0.5$$

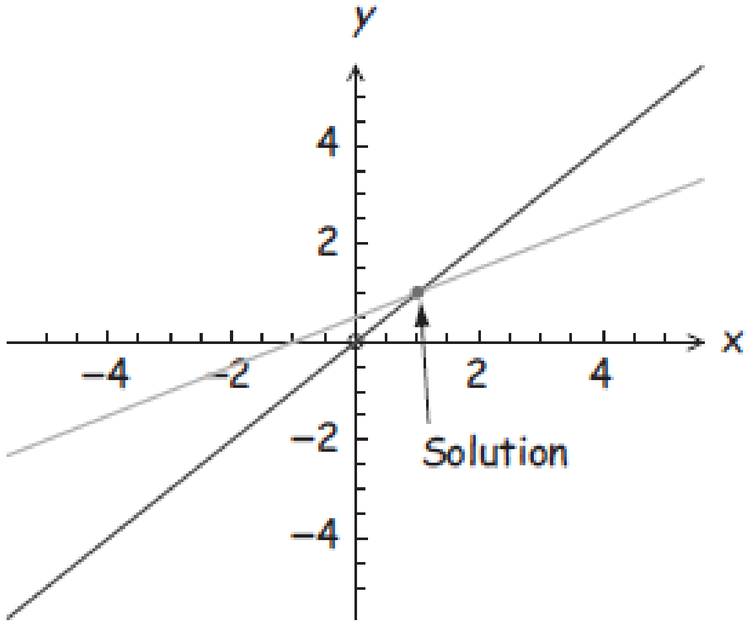


No Solution

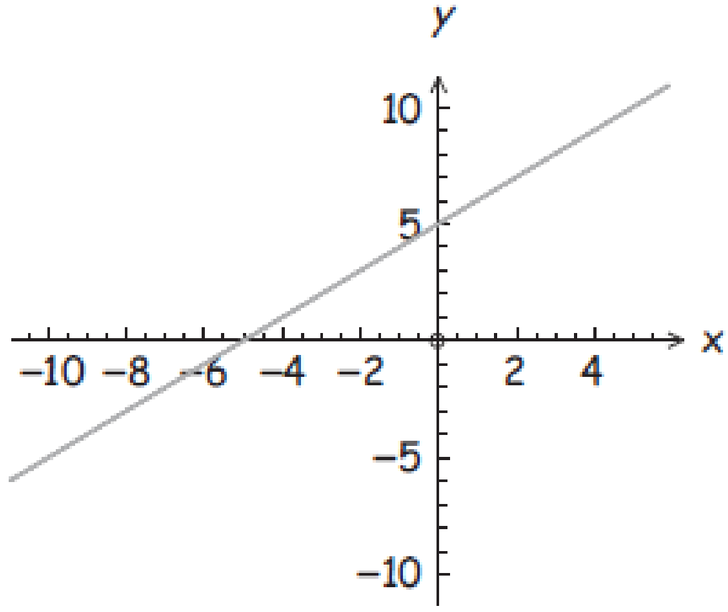




(a) No solution



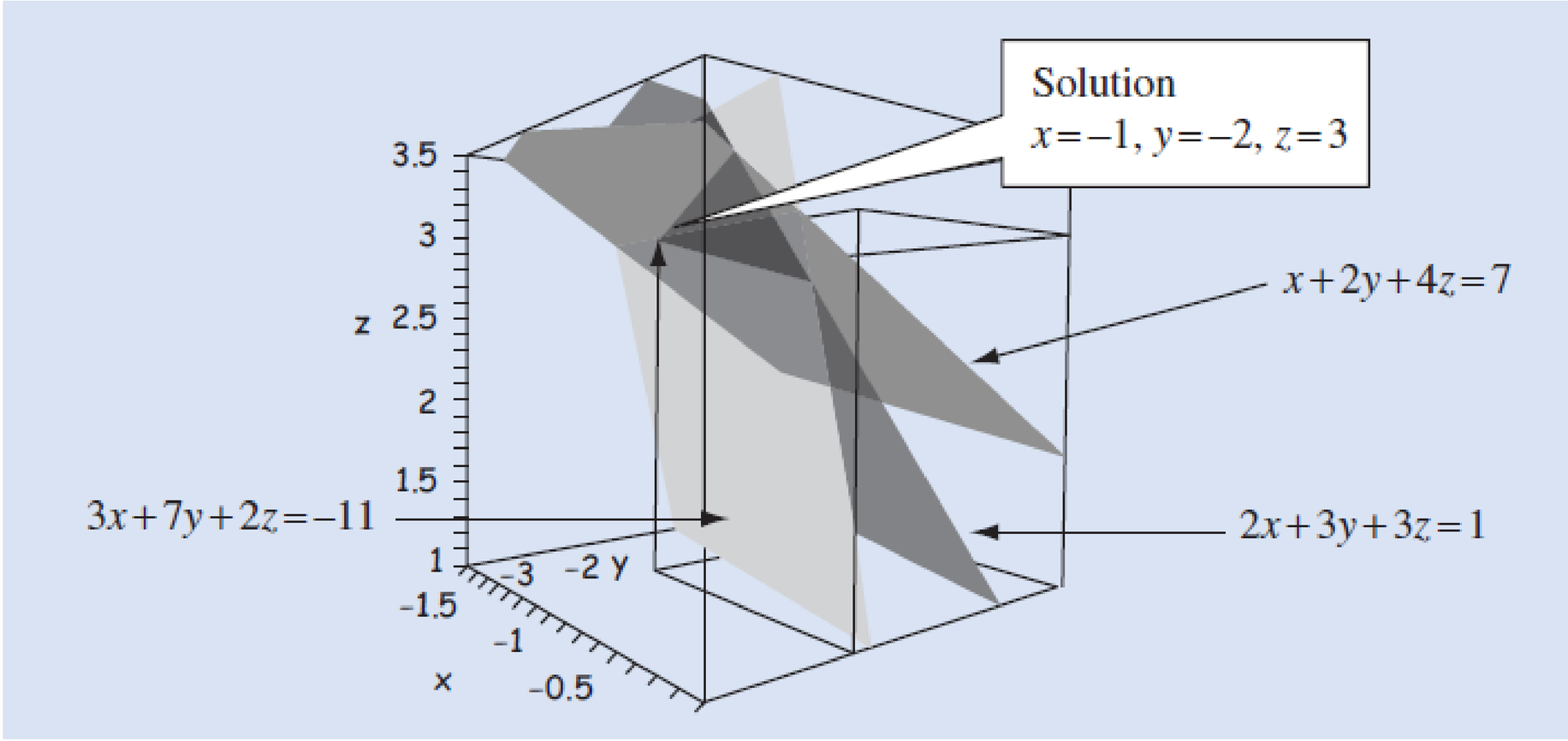
(b) Unique solution



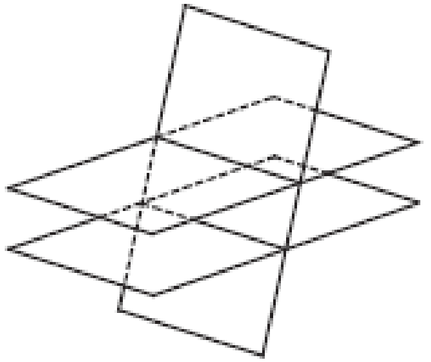
(c) Infinite number of solutions

$$x + 2y + 4z = 7$$
$$3x + 7y + 2z = -11$$
$$2x + 3y + 3z = 1$$

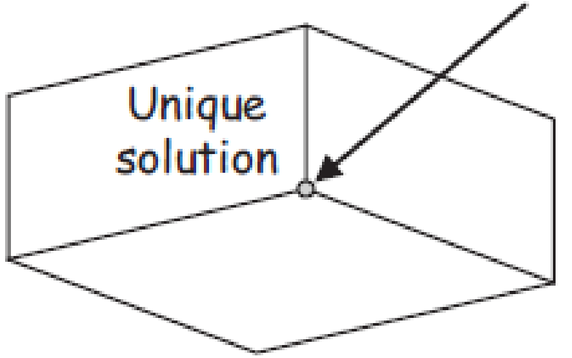
- https://www.geogebra.org/calculator/xuqtru6e



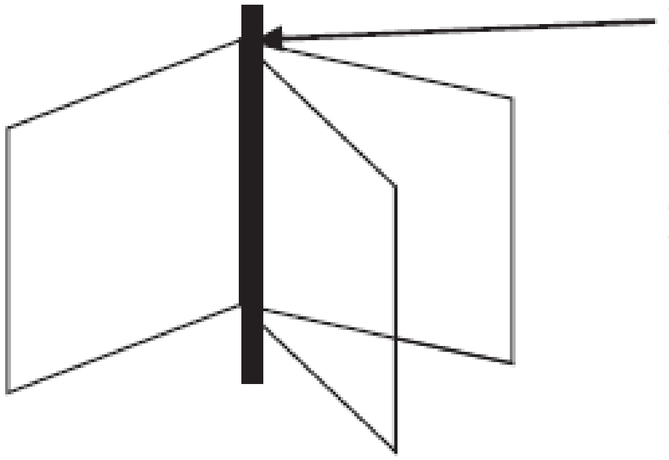
(a) No solution



(b) Unique solution



(c) Infinite number of solutions



Every point on this line is a solution because all three planes meet here.

When do System of equations have solution

$$2x + 2y = 3$$
$$2x + y = 2.5$$

2 variables, 2 equations

$$x + 2y + 4z = 7$$
$$3x + 7y + 2z = -11$$
$$2x + 3y + 3z = 1$$

3 variables, 3 equations

- Num Variables = Num equations
  - Equation MAY have exact solution
  - Need not necessarily be the case

$$2x + 2y = 3$$
$$2x + 2y = 9$$

$$2x + 2y = 3$$
$$4x + 4y = 6$$

Under & over determined system of equations

- Less equations more variables – infinite solutions

Under determined

$$x + 2y + 4z = 7$$

$$3x + 7y + 2z = -11$$

Over determined

- More equations less variables – has no EXACT solution

$$230x_1 + 37.8x_2 = 22.1$$

$$44.5x_1 + 39.3x_2 = 10.4$$

$$17.2x_1 + 45.9x_2 = 9.3$$

$$151.5x_1 + 41.3x_2 = 18.5$$

$$180.8x_1 + 10.8x_2 = 12.9$$

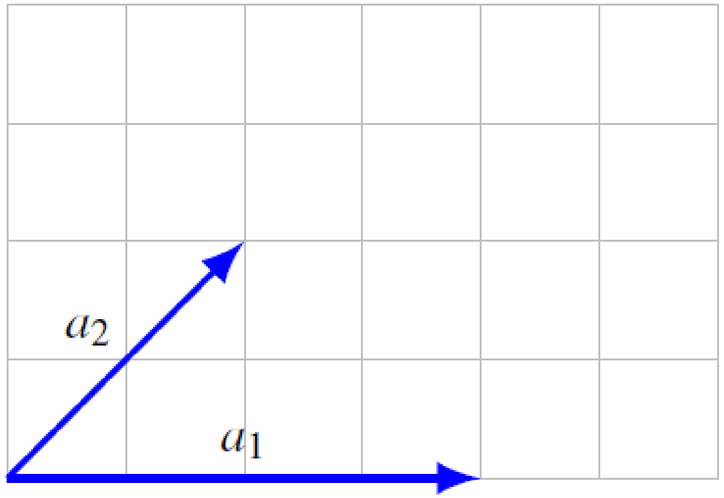
TV	Radio	Sales
230.1	37.8	22.1
44.5	39.3	10.4
17.2	45.9	9.3
151.5	41.3	18.5
180.8	10.8	12.9

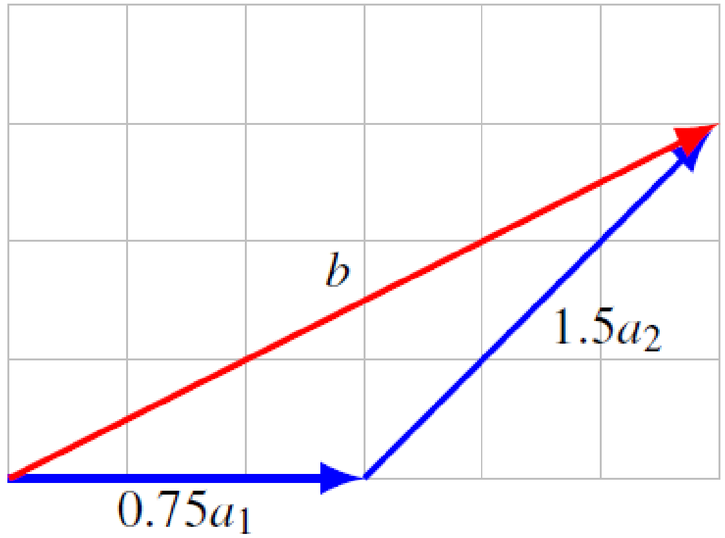
Dataset with m records & n features

m >> n

Linear Combinations

- Definition: $\beta_1 \mathbf{x_1} + \beta_2 \mathbf{x_2} + \ldots + \beta_n \mathbf{x_n}$
  - $\beta_1, \beta_2, \ldots, \beta_n$ are scalars
  - $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}$ are vectors
- Simply put: Scale and add vectors





- https://www.geogebra.org/calculator/eedbpvb9

System of Equations as linear combination

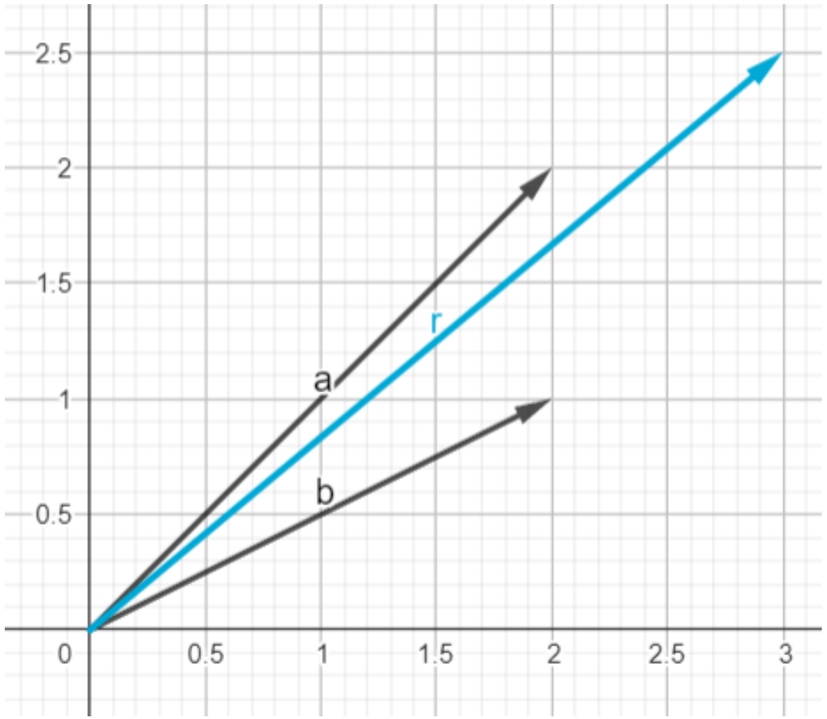
$$2x + 2y = 3$$

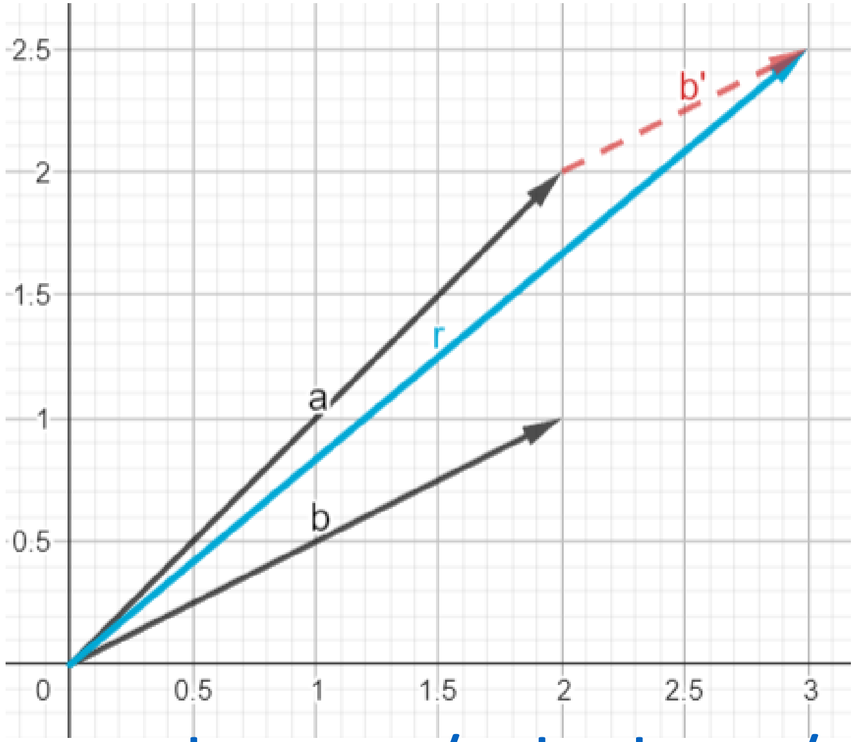
$$2x + y = 2.5$$

$$x = 1, \quad y = 0.5$$

$$x \begin{bmatrix} 2 \\ 2 \end{bmatrix} + y \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 2.5 \end{bmatrix}$$

$$a = \begin{bmatrix} 2 \\ 2 \end{bmatrix} \quad b = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \quad r = \begin{bmatrix} 3 \\ 2.5 \end{bmatrix}$$





- 2D system equation https://www.geogebra.org/calculator/nwueaj

System of Equations as linear combination

$$x + 2y + 4z = 7$$
$$3x + 7y + 2z = -11$$
$$2x + 3y + 3z = 1$$

$$x = -1, \quad y = -2, \quad z = 3$$

$$x \begin{bmatrix} 1 \\ 3 \\ 2 \end{bmatrix} + y \begin{bmatrix} 2 \\ 7 \\ 3 \end{bmatrix} + z \begin{bmatrix} 4 \\ 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 7 \\ -11 \\ 1 \end{bmatrix}$$

$$a = \begin{bmatrix} 1 \\ 3 \\ 2 \end{bmatrix} \quad b = \begin{bmatrix} 2 \\ 7 \\ 3 \end{bmatrix} \quad c = \begin{bmatrix} 4 \\ 2 \\ 3 \end{bmatrix} \quad r = \begin{bmatrix} 7 \\ -11 \\ 1 \end{bmatrix}$$

- 3D system of equation as vectors
  - https://www.geogebra.org/calculator/t6p49dhk

System of equations (SOE) representation

- SOE can be written in traditional format

$$2x + 2y = 3$$
$$2x + y = 2.5$$

- SOE can be written as linear combination

$$x \begin{bmatrix} 2 \\ 2 \end{bmatrix} + y \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 2.5 \end{bmatrix}$$

- Linear combination = Matrix Vector product
  - Therefore SOE can be written as Matrix Vector product

$$\begin{bmatrix} 2 & 2 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 3 \\ 2.5 \end{bmatrix}$$

Ax = b

Xw = y

- 3Blue1Brown Essence of Linear Algebra on youtube
  - https://www.youtube.com/playlist?list=PLZHQObOWTQDPD3MizzM2xVFitgF8hE_ab

Matrix equation solution & notations

$$\begin{bmatrix} 2 & 2 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 3 \\ 2.5 \end{bmatrix} \qquad \begin{bmatrix} x \\ y \end{bmatrix} = \left(\begin{bmatrix} 2 & 2 \\ 2 & 1 \end{bmatrix} \right)^{-1} \begin{bmatrix} 3 \\ 2.5 \end{bmatrix}$$

Inverse of Matrix

Linear Algebra Convention

$$Ax = b \qquad x = A^{-1}y$$

Statistics Convention

$$X\beta = y \qquad \beta = X^{-1}y$$

Andrew Ng Convention

$$X\theta = y \qquad \theta = X^{-1}y$$

Our uniform Convention

$$Xw = y \qquad w = X^{-1}y$$

System of equations as Matrix Vector product

$$2x + 2y = 3$$
$$2x + y = 2.5$$

$$\begin{bmatrix} 2 & 2 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 3 \\ 2.5 \end{bmatrix}$$

2 variables, 2 equations

2 x 2 2 x 1 2 x 1

$$x + 2y + 4z = 7$$
$$3x + 7y + 2z = -11$$
$$2x + 3y + 3z = 1$$

$$\begin{bmatrix} 1 & 2 & 4 \\ 3 & 7 & 2 \\ 2 & 3 & 3 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 7 \\ -11 \\ 1 \end{bmatrix}$$

3 variables, 3 equations

3 x 3 3 x 1 3 x 1

- Num Variables = Num equations
  - Equation MAY have exact solution

Over determined system of equations

- More equations less variables – has no EXACT solution

TV	Radio	Sales
230.1	37.8	22.1
44.5	39.3	10.4
17.2	45.9	9.3
151.5	41.3	18.5
180.8	10.8	12.9

$$230x_1 + 37.8x_2 = 22.1$$
$$44.5x_1 + 39.3x_2 = 10.4$$
$$17.2x_1 + 45.9x_2 = 9.3$$
$$151.5x_1 + 41.3x_2 = 18.5$$
$$180.8x_1 + 10.8x_2 = 12.9$$

$$\begin{bmatrix} 230 & 37.8 \\ 44.5 & 39.3 \\ 17.2 & 45.9 \\ 151.5 & 41.3 \\ 180.8 & 10.8 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 22.1 \\ 10.4 \\ 9.3 \\ 18.5 \\ 12.9 \end{bmatrix}$$

5 x 2 2 x 1 5 x 1

What linear combinations possible if no exact solution?

How about approximate solution?

Unreachable combinations

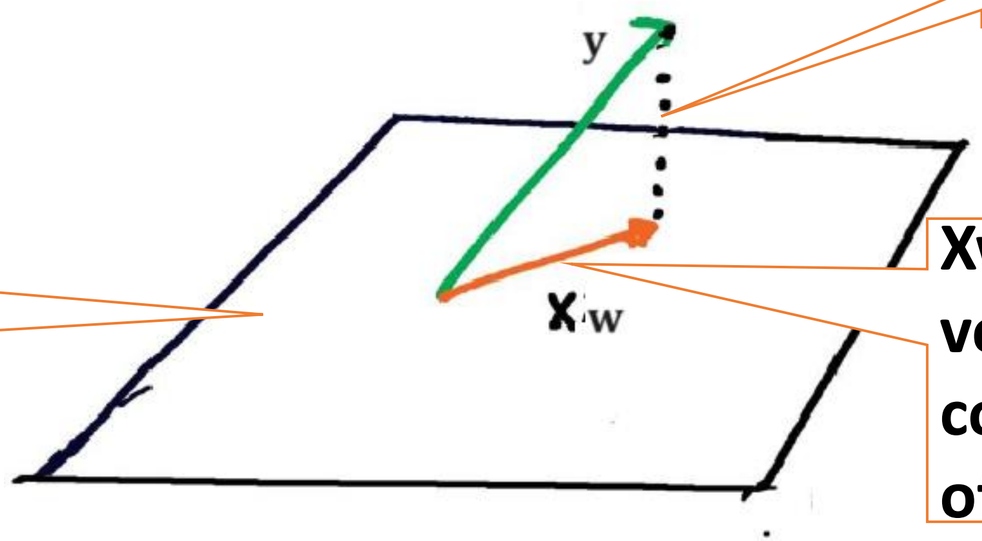
- Recall: Exact solutions to system of equations are vectors reachable by linear combination

- How do unreachable vectors & approximate solutions look like?
  - https://www.geogebra.org/calculator/xansjfxj

- Ambient vector space
- Actual vector sub space (column space)

Linear Algebra Interpretation

TV	Radio	Sales
230.1	37.8	22.1
44.5	39.3	10.4
17.2	45.9	9.3
151.5	41.3	18.5
180.8	10.8	12.9



Xw-y

Column space of X

Xw is special vector in column space of X

$$\begin{bmatrix} 230 & 37.8 \\ 44.5 & 39.3 \\ 17.2 & 45.9 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} 22.1 \\ 10.4 \\ 9.3 \end{bmatrix}$$

$$Xw = \hat{y}$$

$$\arg\min_w \|\hat{y} - y\|^2$$

$$\arg\min_w \mathcal{J}(w) = \frac{1}{m}(Xw - y)^T(Xw - y)$$

Vectorized closed form solution

$$X = \begin{bmatrix} 1 & x_1^{(1)} \\ 1 & x_1^{(2)} \\ 1 & .. \\ 1 & x_1^{(m)} \end{bmatrix} \qquad \mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} \qquad \mathcal{J}(w) = \frac{1}{m}(Xw - y)^T(Xw - y)$$

$$\nabla_w J = \begin{bmatrix} \frac{\partial J}{\partial w_0} \\ \\ \frac{\partial J}{\partial w_1} \end{bmatrix} = \frac{2}{m} X^T(Xw - y) = 0$$

$$X^T(Xw - y) = 0 \implies X^T X w = X^T y$$

$$w = (X^T X)^{-1} X^T y$$

Normal equation

Left Inverse (Moore Penrose Pseudo Inverse)

Inverse of Matrix

$$w = X^{-1} y$$

Compare with non vectorized solution

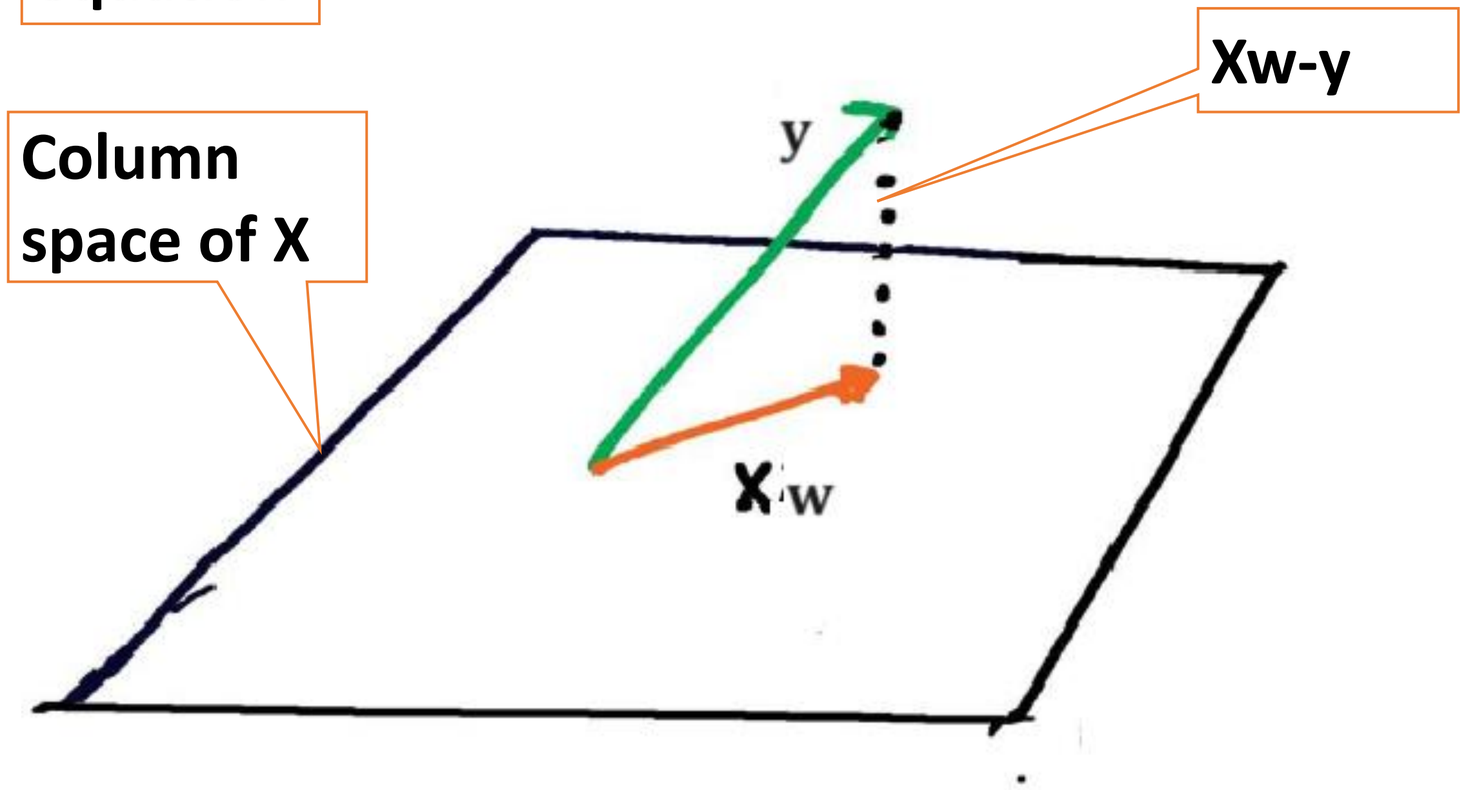
$$w_1 = \frac{n \sum_{i=1}^m x^{(i)} y^{(i)} - \sum_{i=1}^m x^{(i)} \sum_{i=1}^m y^{(i)}}{n \sum_{i=1}^m {x^{(i)}}^2 - (\sum_{i=1}^m x^{(i)})^2}$$

$$w_0 = \frac{\sum_{i=1}^m y^{(i)} - w_1 \sum_{i=1}^m x^{(i)}}{n}$$

Normal equation from Linear Algebra

$$w = (X^T X)^{-1} X^T y$$

Normal equation



$$X^T(Xw - y) = 0$$

$$X^T(Xw) - X^T y = 0$$

$$(X^T X)w = X^T y$$

$$w = (X^T X)^{-1} X^T y$$

Interpreting Linear Regression Coefficients

Interpreting Linear Regression Coefficients

$$\hat{y} = w_0 + w_1 x_1 + w_2 x_2$$

$$\frac{\partial \hat{y}}{\partial x_1} = w_1$$

Example with Boston house prices

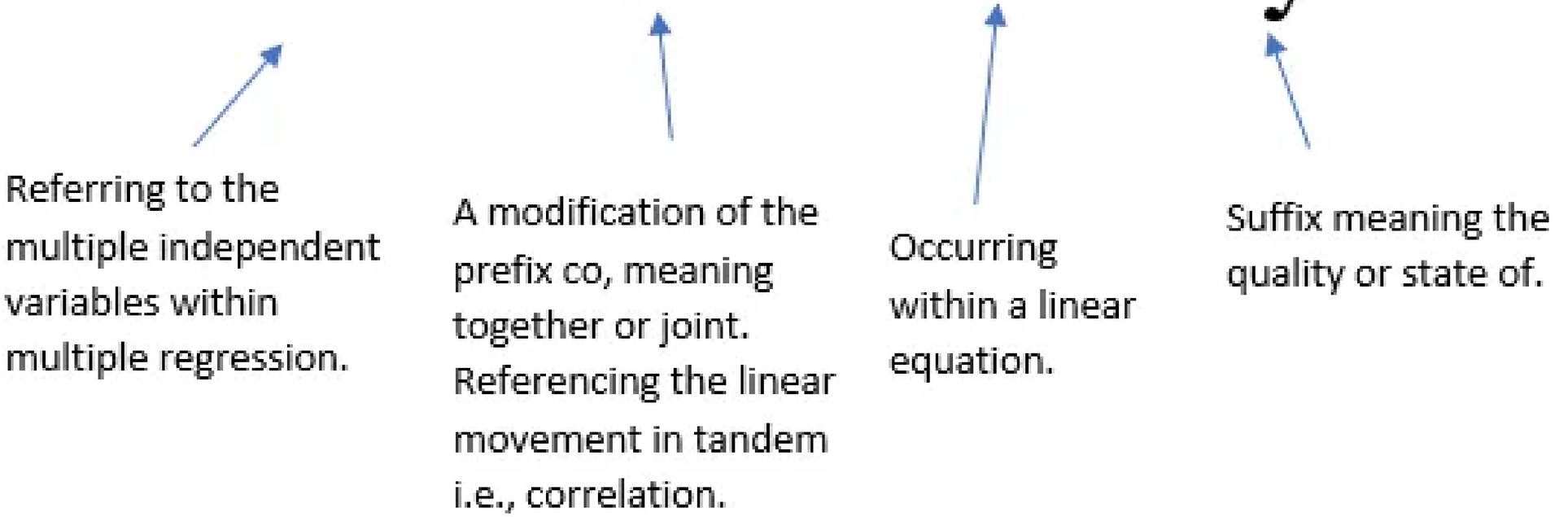
- For a unit change in x, what is the change in y-hat
  - Assumes all others stay constant
  - By definition of partial derivative
  - Not true for correlated features
- What about categorical variables?

True for numeric variables

Multicollinearity

Word Anatomy of Multicollinearity

Multi-col-linear-ity



What is it? Why is it bad?

- Significant correlation among features
- Change in one impacts another. Model is not robust
- Take a look at normal equation will make it obvious
  - Covariance matrix has no inverse when features are fully correlated

$$\hat{y} = w_0 + w_1 x_1 + w_2 x_2$$

$$\frac{\partial \hat{y}}{\partial x_1} = w_1$$

$$w = (X^T X)^{-1} X^T y$$

$$X^T X = \Sigma = \begin{bmatrix} \sigma_1^2 & Cov_{12} & \dots & Cov_{1n} \\ Cov_{21} & \sigma_2^2 & \dots & Cov_{2n} \\ .. & .. & .. & .. \\ .. & .. & \sigma_{n-1}^2 & .. \\ Cov_{n1} & Cov_{n2} & .. & \sigma_n^2 \end{bmatrix}$$

How to find multicollinearity

- Seaborn correlation heatmap
- Eigen decomposition of Covariance matrix $X^T X$
  - Find features with low/zero eigen values

No Inverse

$$\begin{bmatrix} 2 & 33 & 4 \\ 3 & 39 & 6 \\ 1 & 41 & 2 \\ 4 & 46 & 8 \\ 6 & 44 & 12 \end{bmatrix}$$

Correlation=1

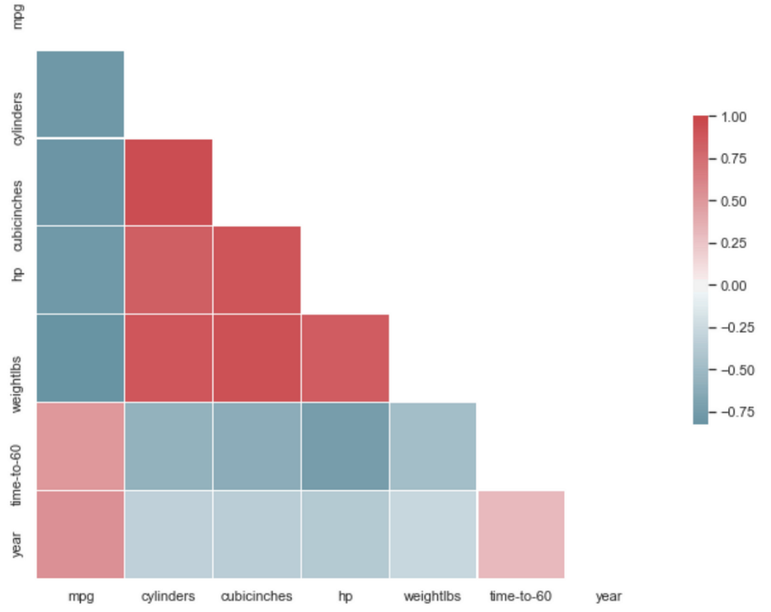
$$\lambda_1 = 3, \lambda_2 = 0.9, \lambda_3 = 0$$

Inverse exists with determinant close to 0

$$\begin{bmatrix} 2 & 33 & 3.6 \\ 3 & 39 & 5 \\ 1 & 41 & 2.6 \\ 4 & 46 & 7.1 \\ 6 & 44 & 13 \end{bmatrix}$$

Correlation=0.85

$$\lambda_1 = 3, \lambda_2 = 0.9, \lambda_3 = 10^{-4}$$

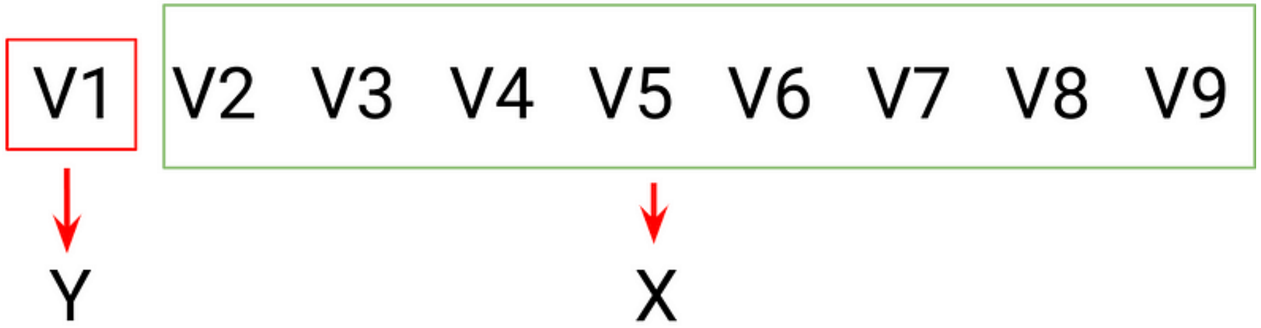


Variance Inflation Factor (VIF)

- Given features (variables)

V1 V2 V3 V4 V5 V6 V7 V8 V9

- Pick one as target, rest as predictors. Fit OLS



- Calculate R2. Good R2 = good fit. Hence bad predictor

$$R^2 = 1 - \frac{\text{MSE(model)}}{\text{MSE(baseline)}}$$

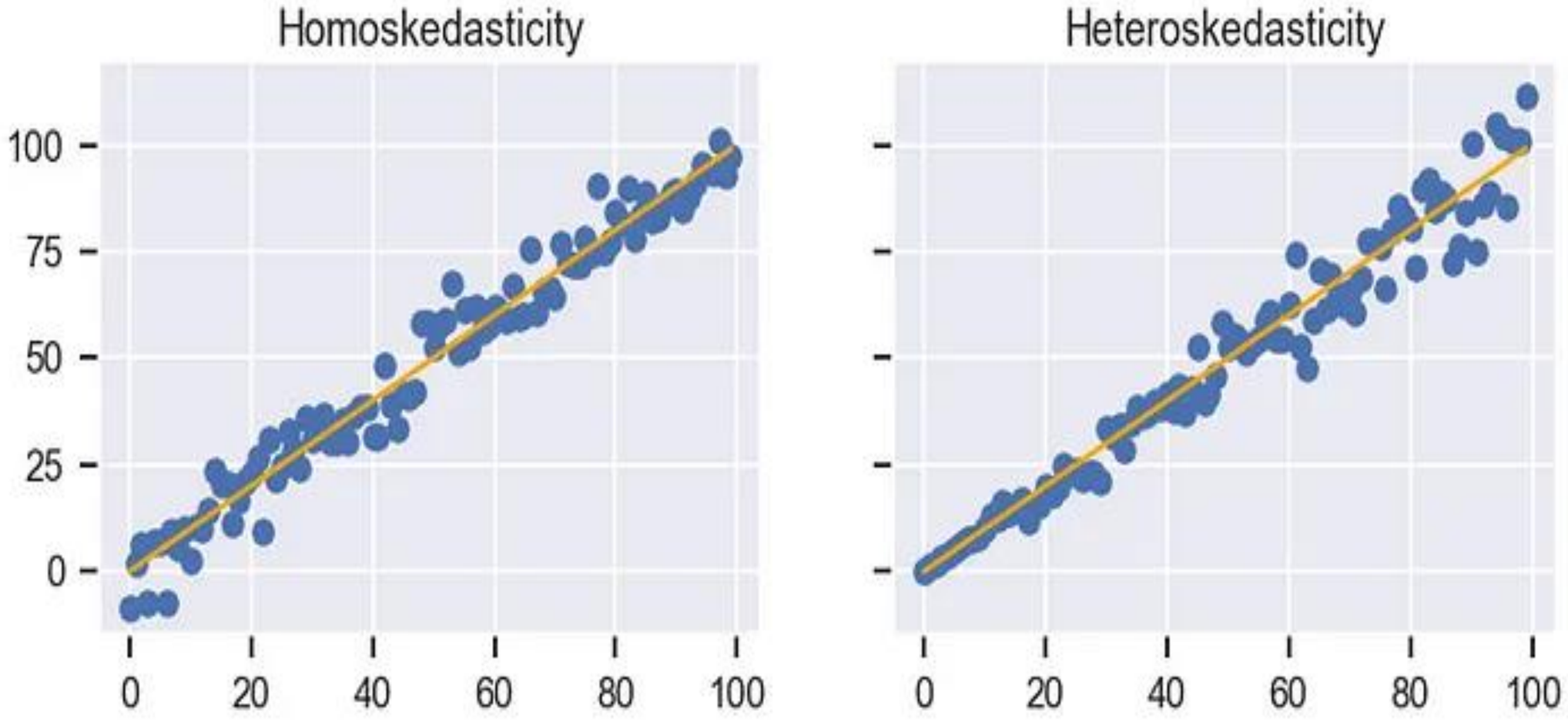
$$\text{VIF} = \frac{1}{1 - R^2}$$

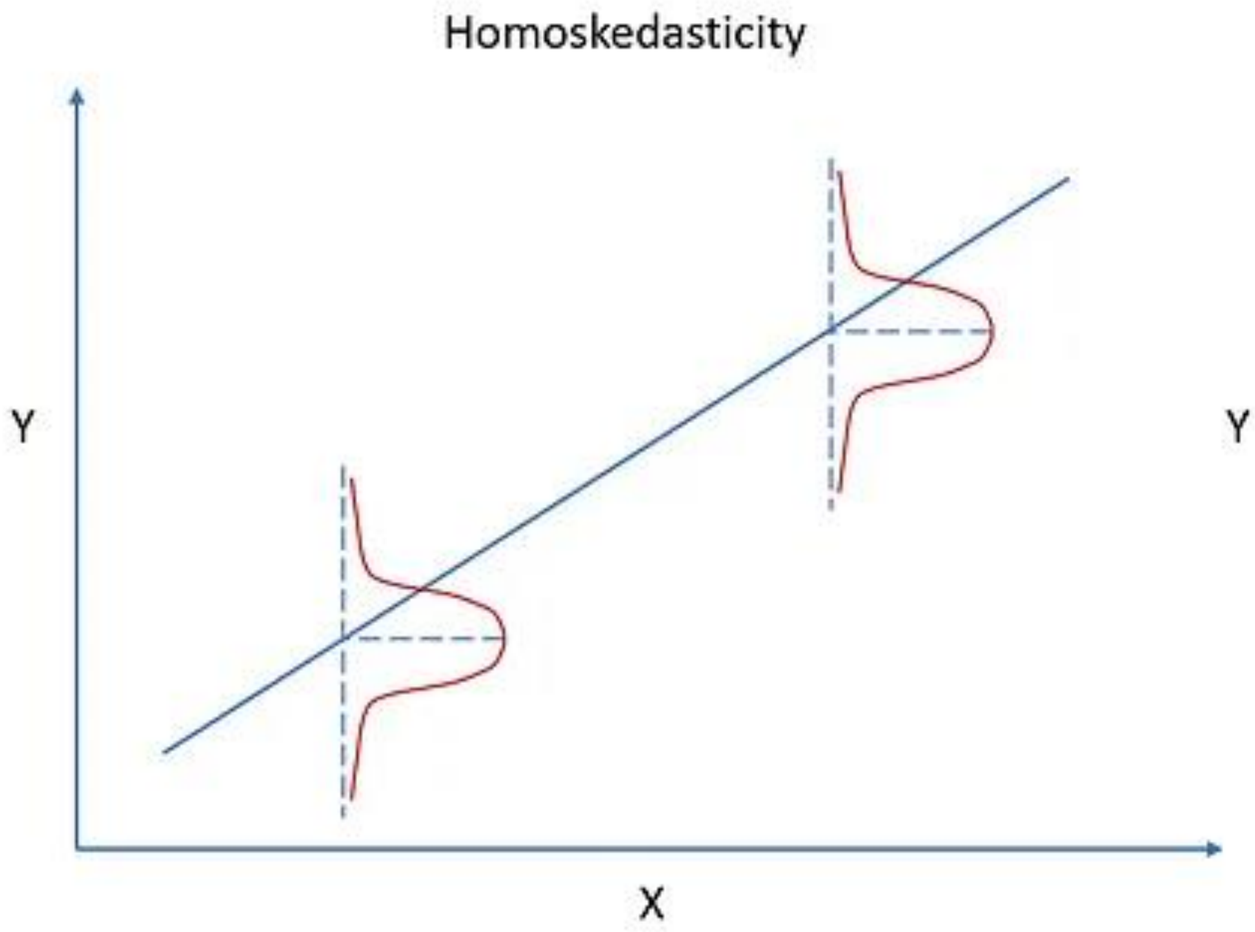
VIF > 5 is bad feature

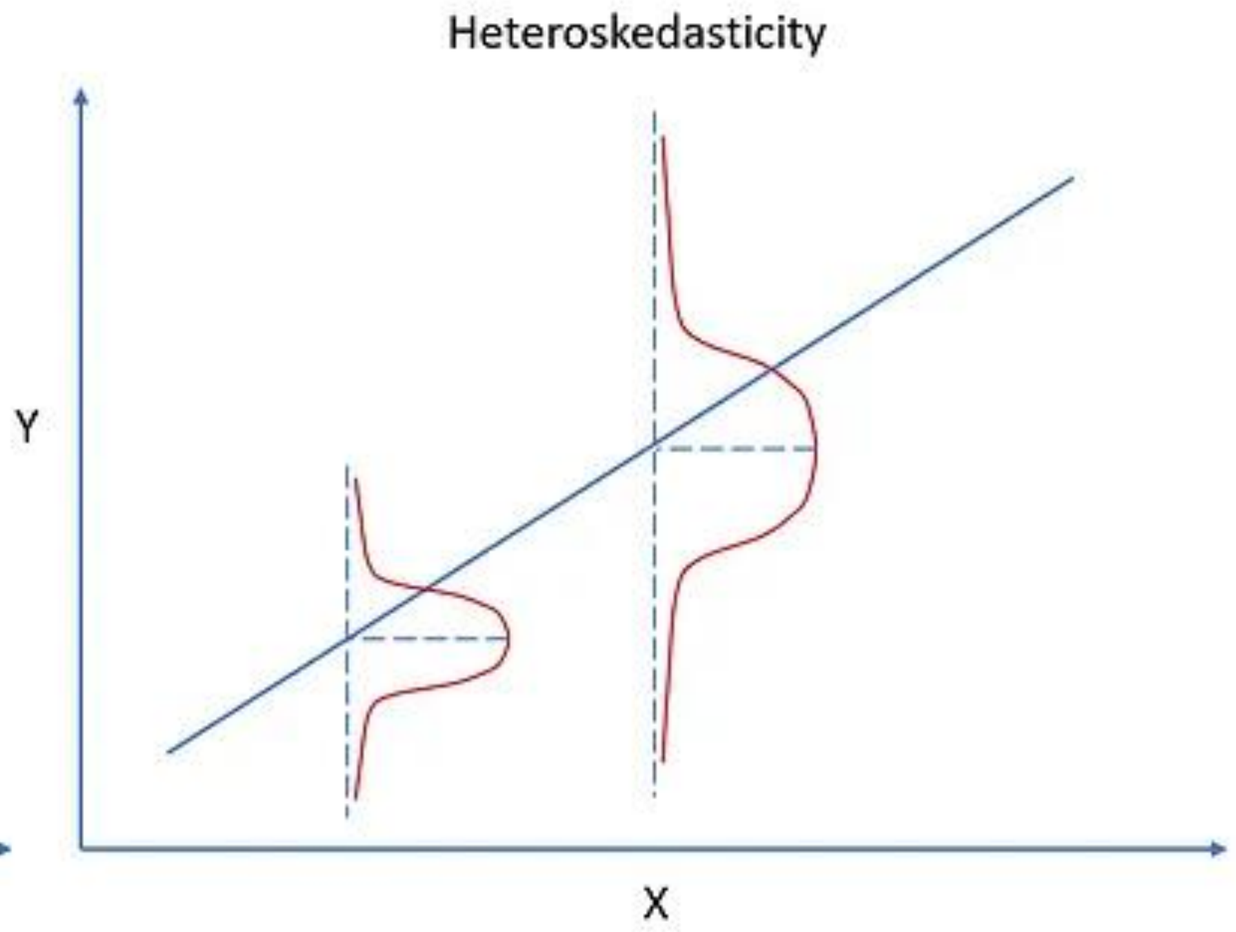
Variance Inflation Factor (VIF)

1	Low to no correlation with 1 or more of the other variables.
1-5	Moderate correlation to other variables.
5 or more	High correlation, this variable should be considered for removal.

Heteroskedasticity

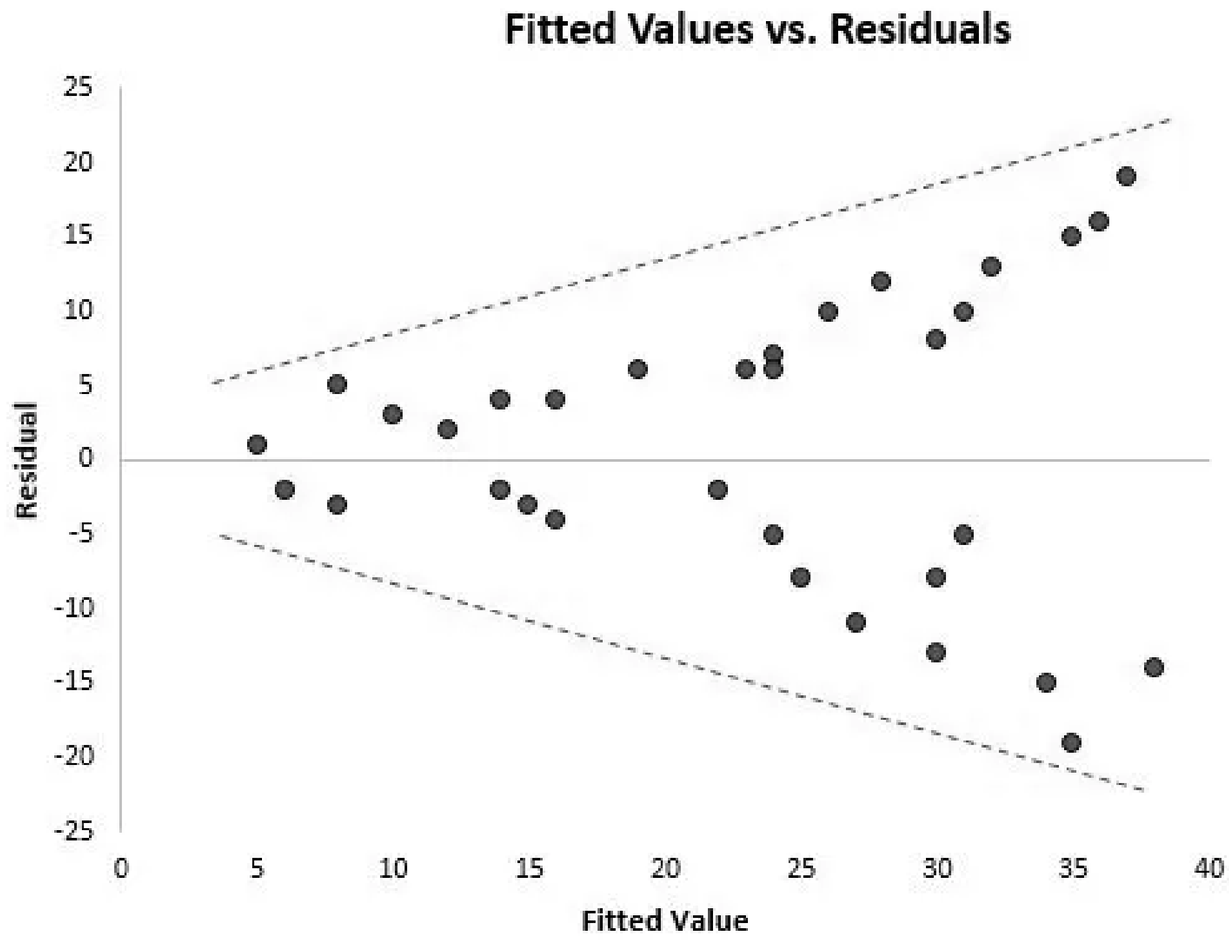

Homoskedasticity
Heteroskedasticity
100
75
50
25
0
0
20
40
60
80
100


Homoskedasticity
Y
X


Heteroskedasticity
Y
X

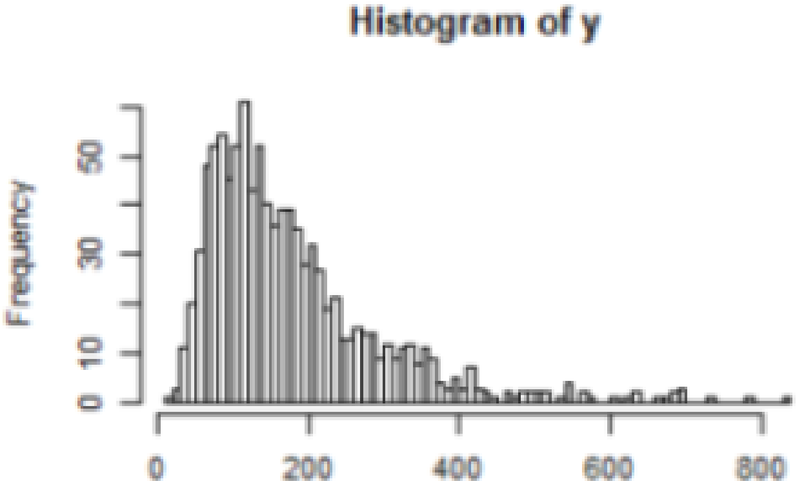
How to find

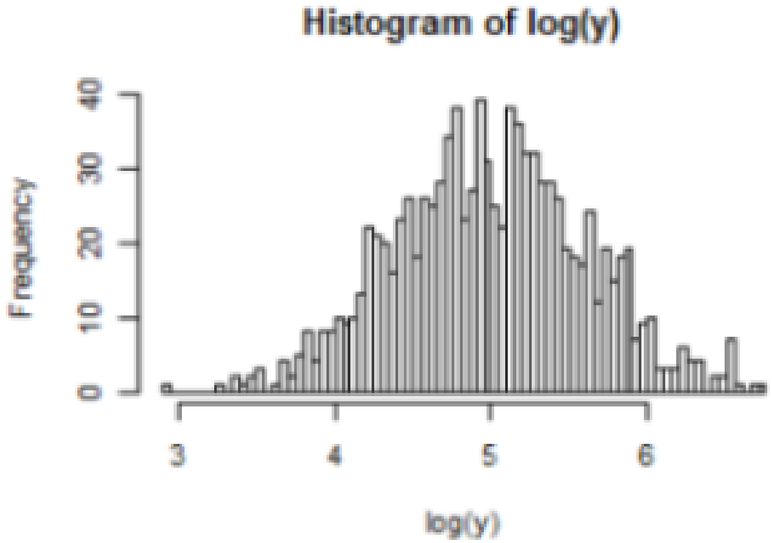
- Visually via Residual plot
- Statistical tests
  - Breusch Pagan Test
  - White Test



How to fix

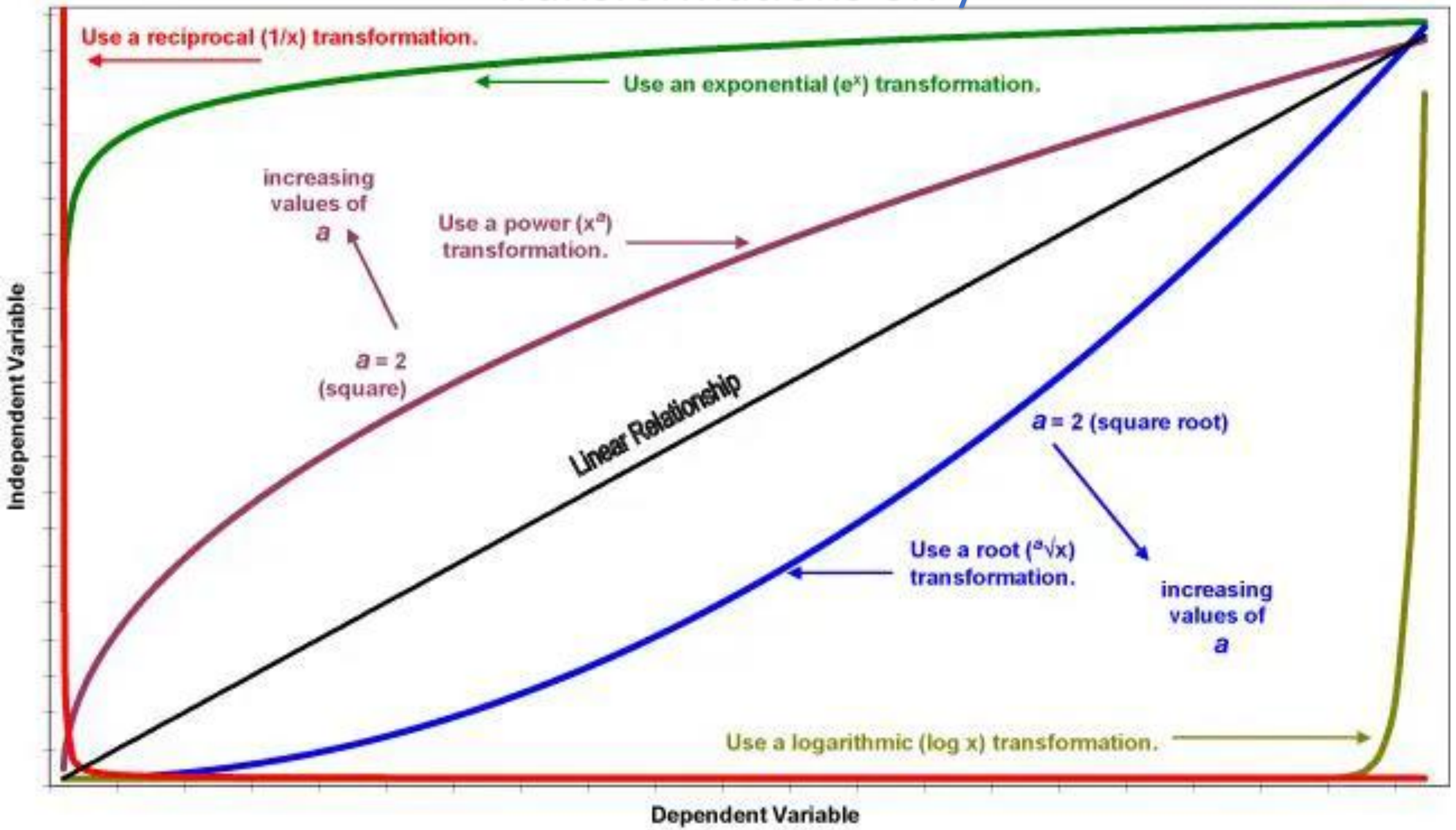
- Fit to log y instead of y





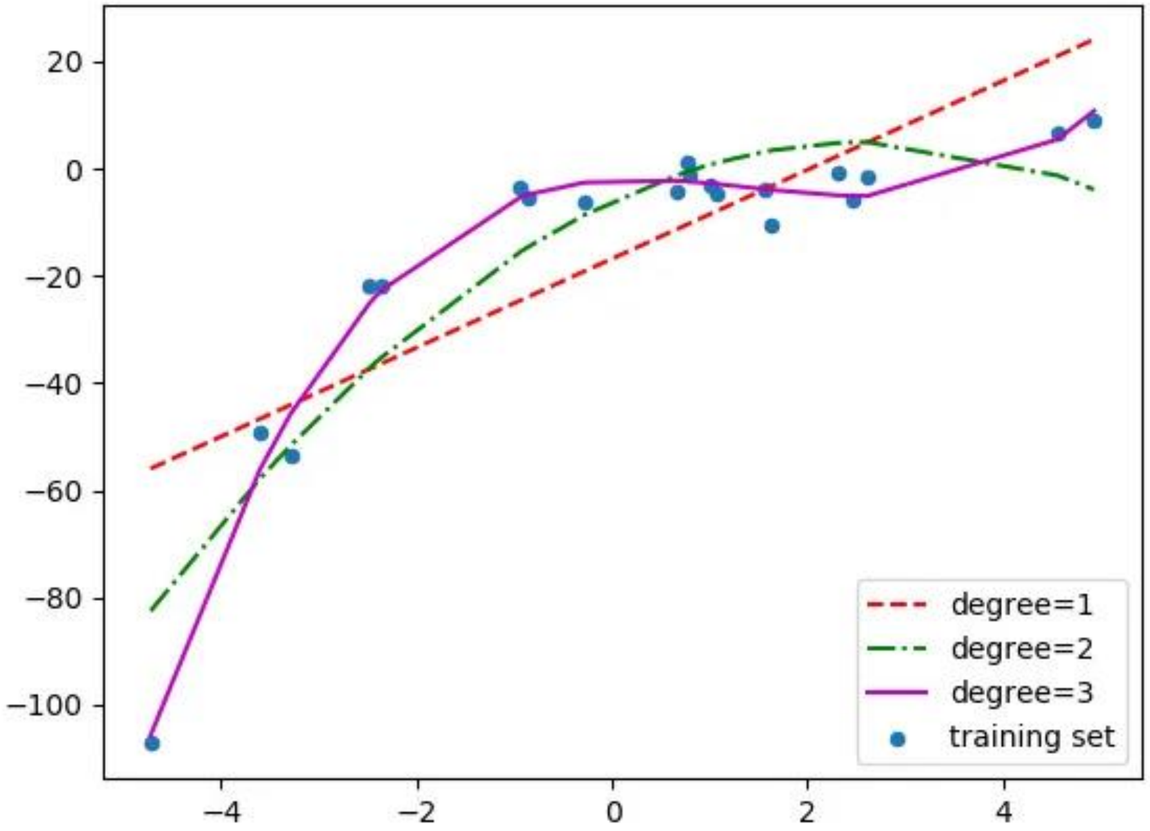
- Box-Cox transformations on X and y
- Weighted Least Squares

Transformations on y



Polynomial Linear Regression

Polynomial Regression



$$y = \mathbf{w}^T \mathbf{x} + b$$

$$y = w_1 x + w_2 x^2 + w_3 x^3 + b$$

$$y = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix}^T \underbrace{\begin{bmatrix} x \\ x^2 \\ x^3 \end{bmatrix}}_{Features} + b$$

Non-linear relationship

$$\hat{y} = w_0 + w_1 x + w_2 x^2$$

$$\begin{bmatrix} 1 & x^{(1)} & {x^{(1)}}^2 \\ 1 & x^{(2)} & {x^{(2)}}^2 \\ 1 & x^{(3)} & {x^{(3)}}^2 \\ 1 & x^{(4)} & {x^{(4)}}^2 \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} \hat{y}^{(1)} \\ \hat{y}^{(2)} \\ \hat{y}^{(3)} \\ \hat{y}^{(4)} \end{bmatrix}$$

Feature Cross

$$\hat{y} = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1^2 + w_4 x_1 x_2$$

$$\begin{bmatrix} 1 & x_1^{(1)} & x_2^{(1)} & {x_1^{(1)}}^2 & x_1^{(1)} x_2^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & {x_1^{(2)}}^2 & x_1^{(2)} x_2^{(12)} \\ 1 & x_1^{(3)} & x_2^{(3)} & {x_1^{(3)}}^2 & x_1^{(3)} x_2^{(3)} \\ 1 & x_1^{(4)} & x_2^{(4)} & {x_1^{(4)}}^2 & x_1^{(4)} x_2^{(4)} \\ 1 & x_1^{(5)} & x_2^{(5)} & {x_1^{(5)}}^2 & x_1^{(5)} x_2^{(5)} \\ 1 & x_1^{(6)} & x_2^{(6)} & {x_1^{(6)}}^2 & x_1^{(6)} x_2^{(6)} \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ w_3 \\ w_4 \end{bmatrix} = \begin{bmatrix} \hat{y}^{(1)} \\ \hat{y}^{(2)} \\ \hat{y}^{(3)} \\ \hat{y}^{(4)} \\ \hat{y}^{(5)} \end{bmatrix}$$

Polynomial Regression sklearn

- Feature Cross grows exponentially
- sklearn.preprocessing.PolynomialFeatures

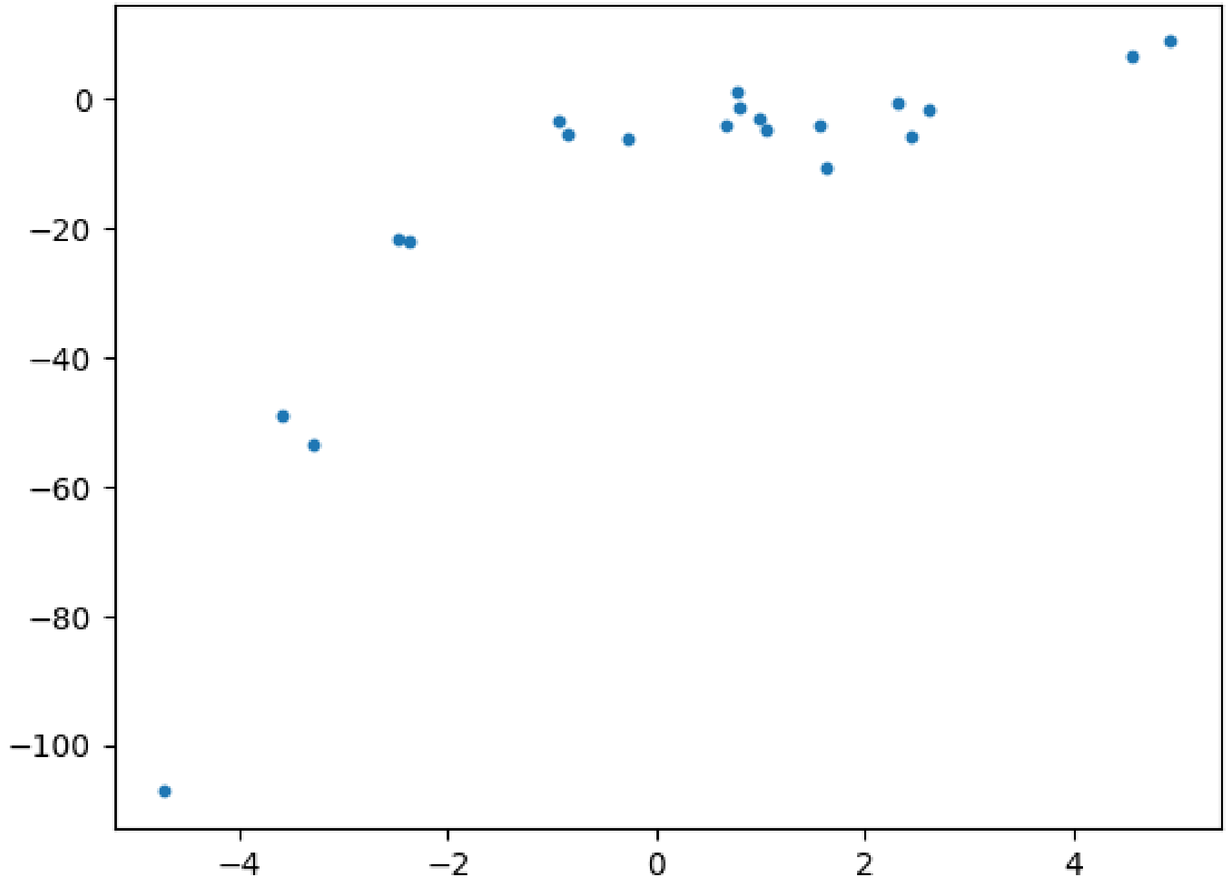
```
from sklearn.preprocessing import PolynomialFeatures

poly = PolynomialFeatures(2)
```

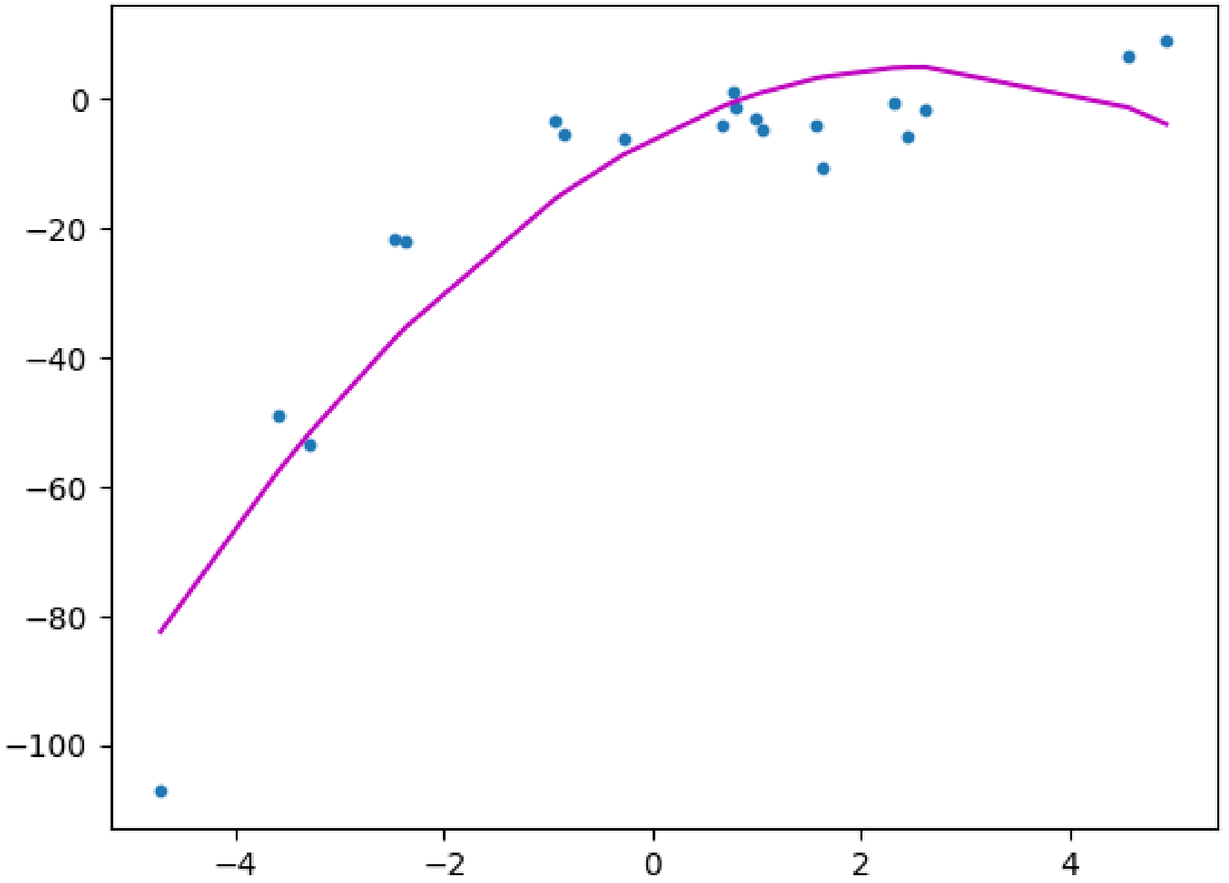
```
poly.fit_transform(X)
```

```
array([[ 1.,  0.,  1.,  0.,  0.,  1.],
       [ 1.,  2.,  3.,  4.,  6.,  9.],
       [ 1.,  4.,  5., 16., 20., 25.]])
```

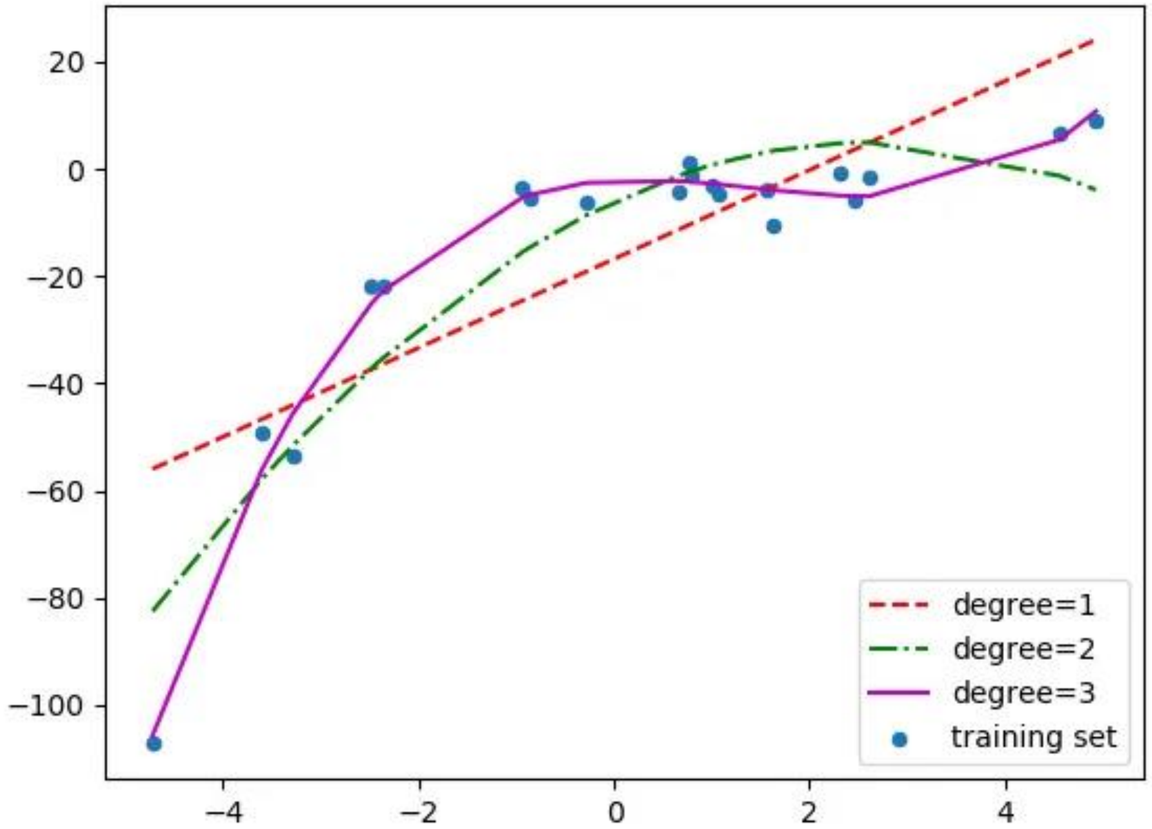
Underfitting, Overfitting, Bias Variance trade off

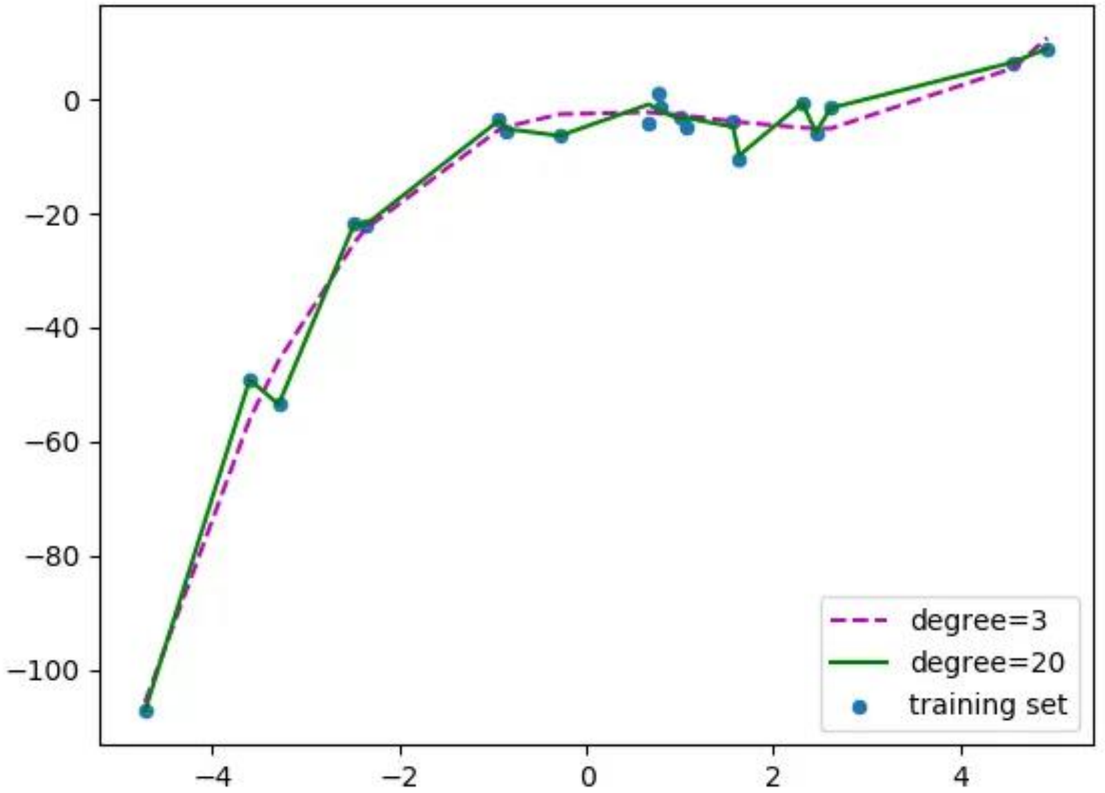
Fitting a curve



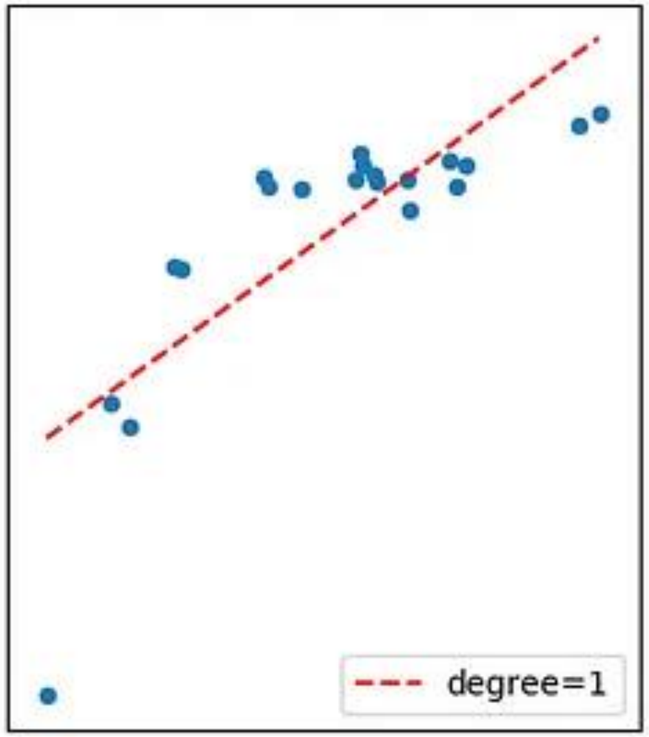


How many degrees is good?

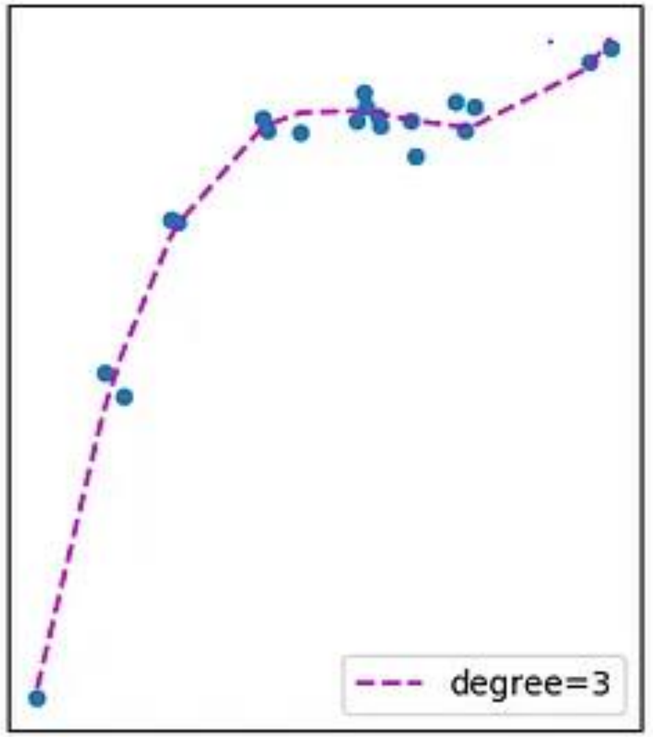




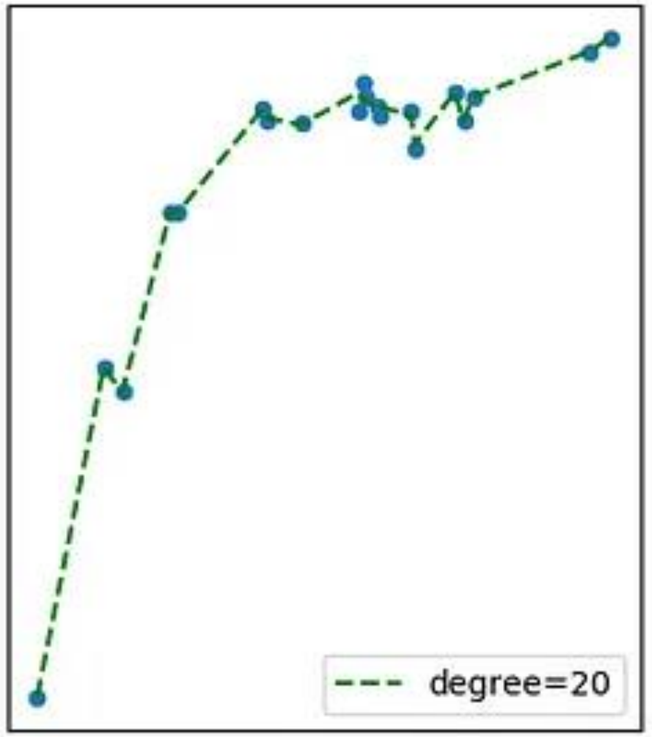
Underfitting & overfitting



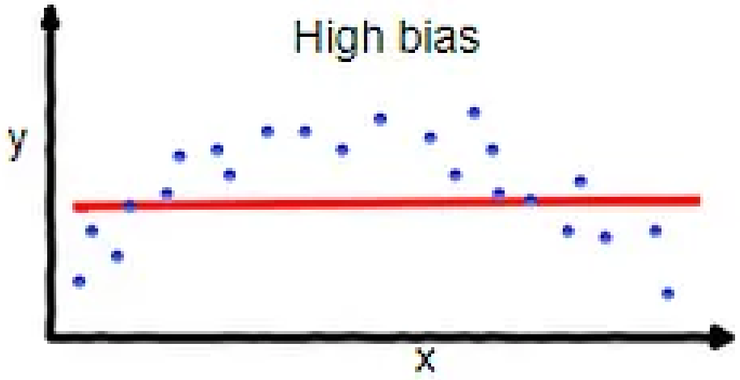
Underfit
High Bias
Low Variance



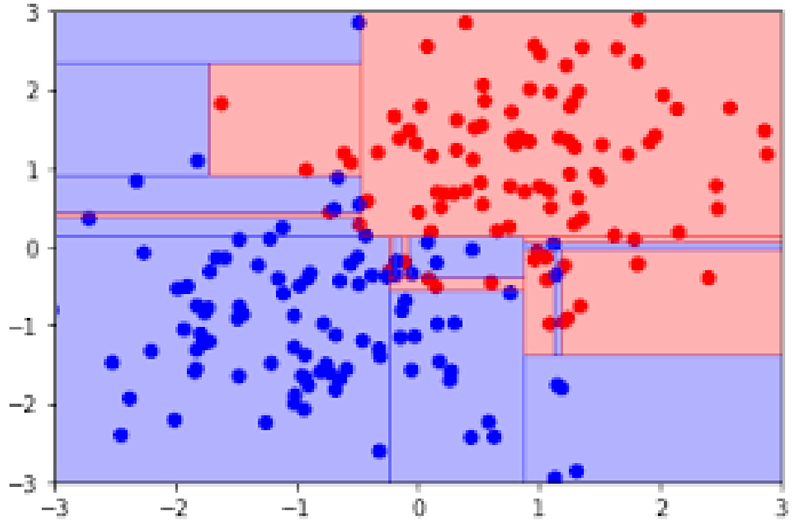
Correct Fit
Low Bias
Low Variance



Overfit
Low Bias
High Variance



Add some new data points, remove some. What happens in each case?



Think Linear Regression

Think Decision Tree

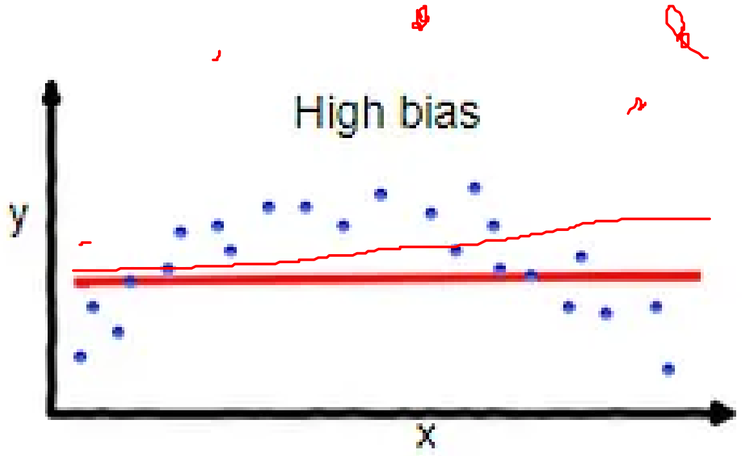
Bias occurs when an algo has *limited flexibility* to learn the true signal from a dataset.

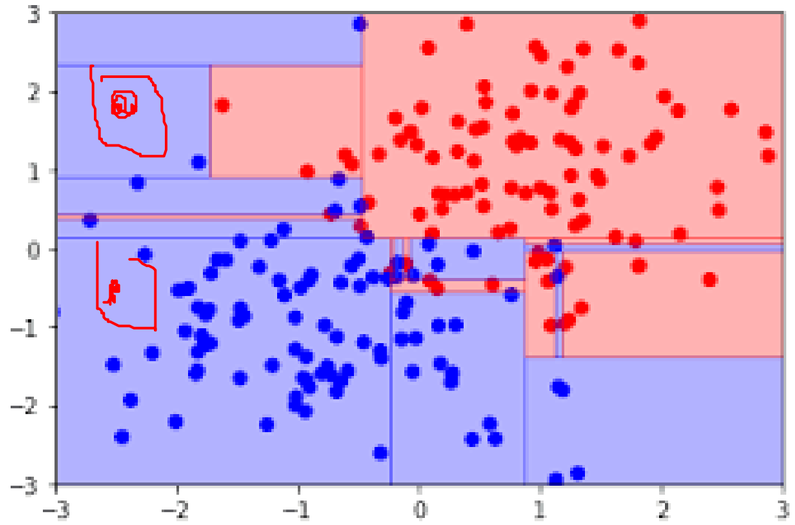
Variance refers to an algo's *sensitivity* to specific sets of training data.

- 5 different training sets (imagine bootstrapping)
- Same algorithm trained on 5 data sets

High bias, low variance algorithms train models that are consistent, but inaccurate *on average*.

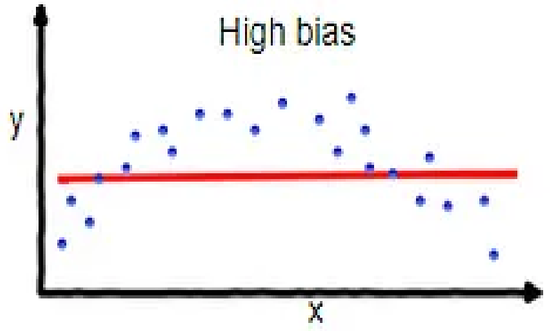
High variance, low bias algorithms train models that are accurate *on average*, but inconsistent.



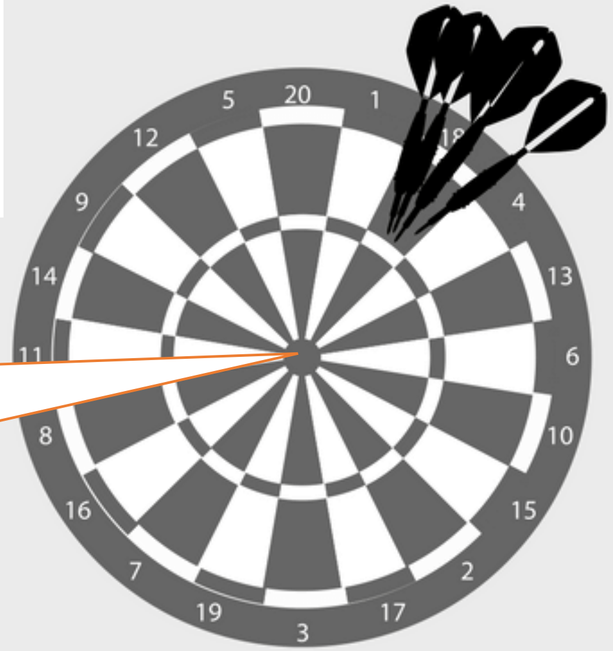


High bias, low variance algorithms train models that are consistent, but inaccurate *on average*.

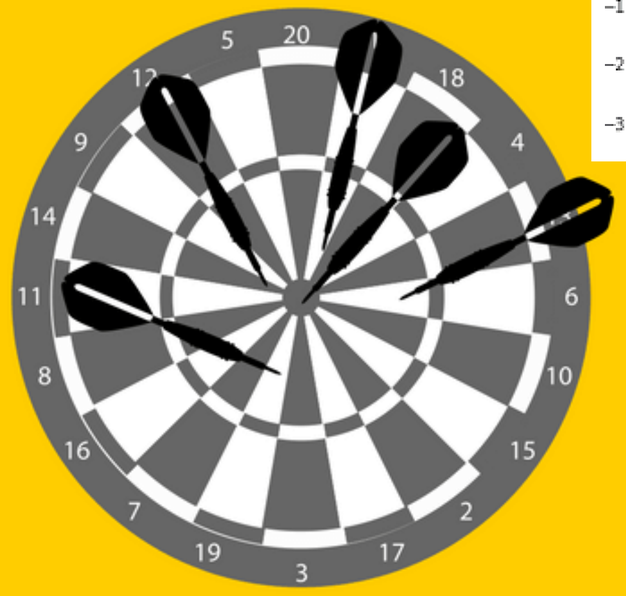
High variance, low bias algorithms train models that are accurate *on average*, but inconsistent.

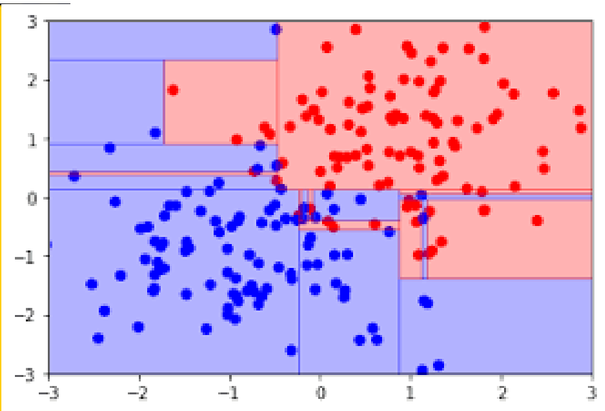


High Bias
Low Variance



High Variance
Low Bias





Bulls eye means best model & consistent model

But why is there a tradeoff?

Low variance algos tend to be **less complex**, with simple or rigid underlying structure.

- e.g. Regression
- e.g. Naive Bayes
- *Linear algos*
- *Parametric algos*

Low bias algos tend to be **more complex**, with flexible underlying structure.

- e.g. Decision trees
- e.g. Nearest neighbors
- *Non-linear algos*
- *Non-parametric algos*

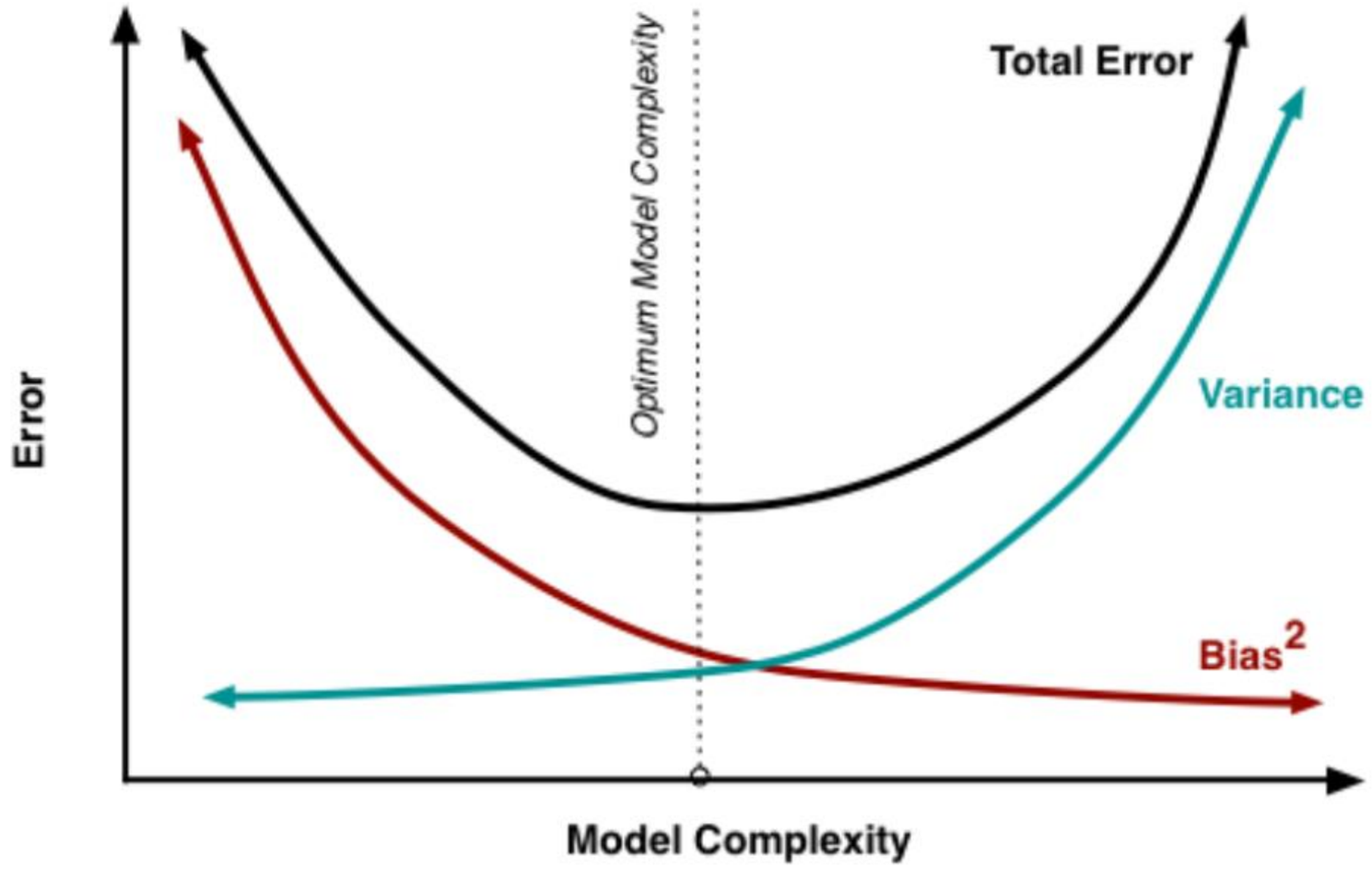
Within each algo family, there's a tradeoff too...

For example, regression can be **regularized** to further reduce complexity.

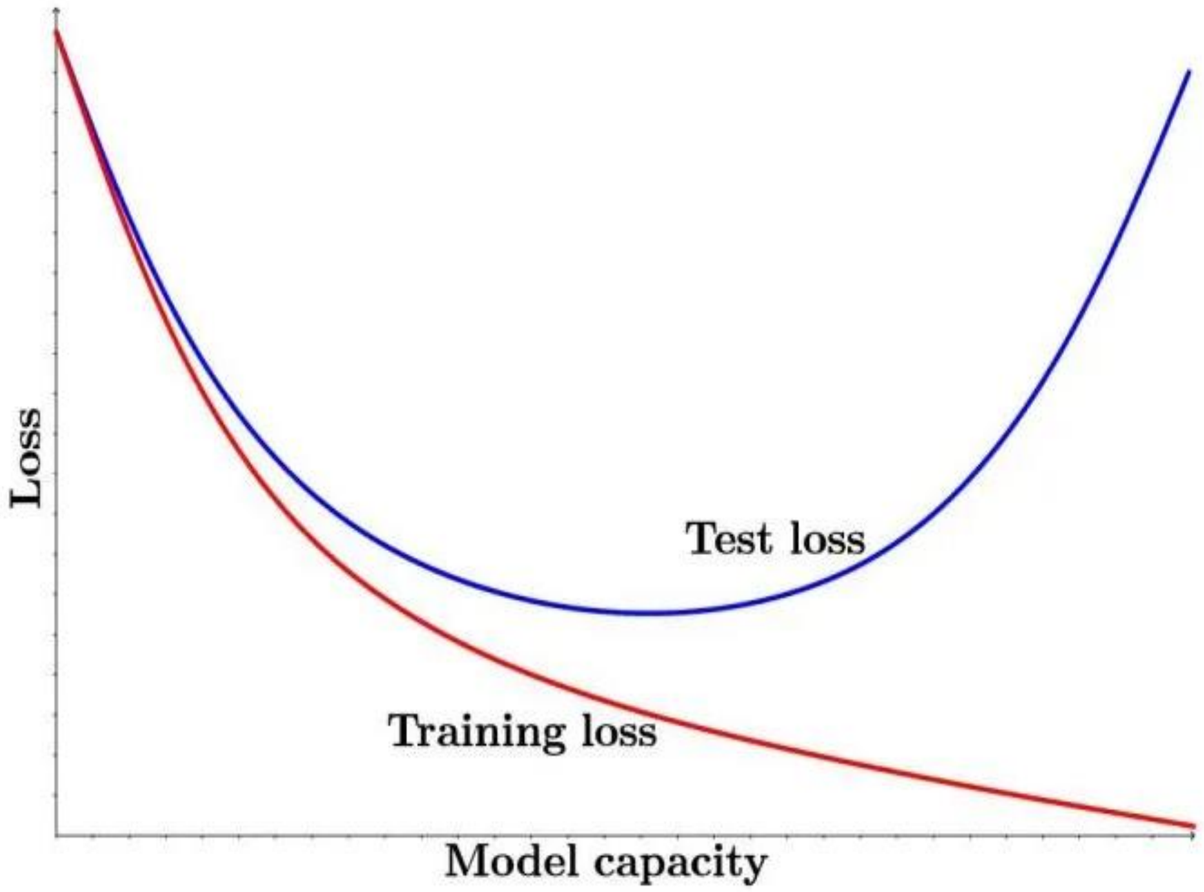
For example, decision trees can be **pruned** to reduce complexity.

Bias Variance Plot

Total Error = Bias^2 + Variance + Irreducible Error



Regularization



- As model capacity increases
  - Training loss goes down
  - Model remembers more learns less
  - Cannot generalize

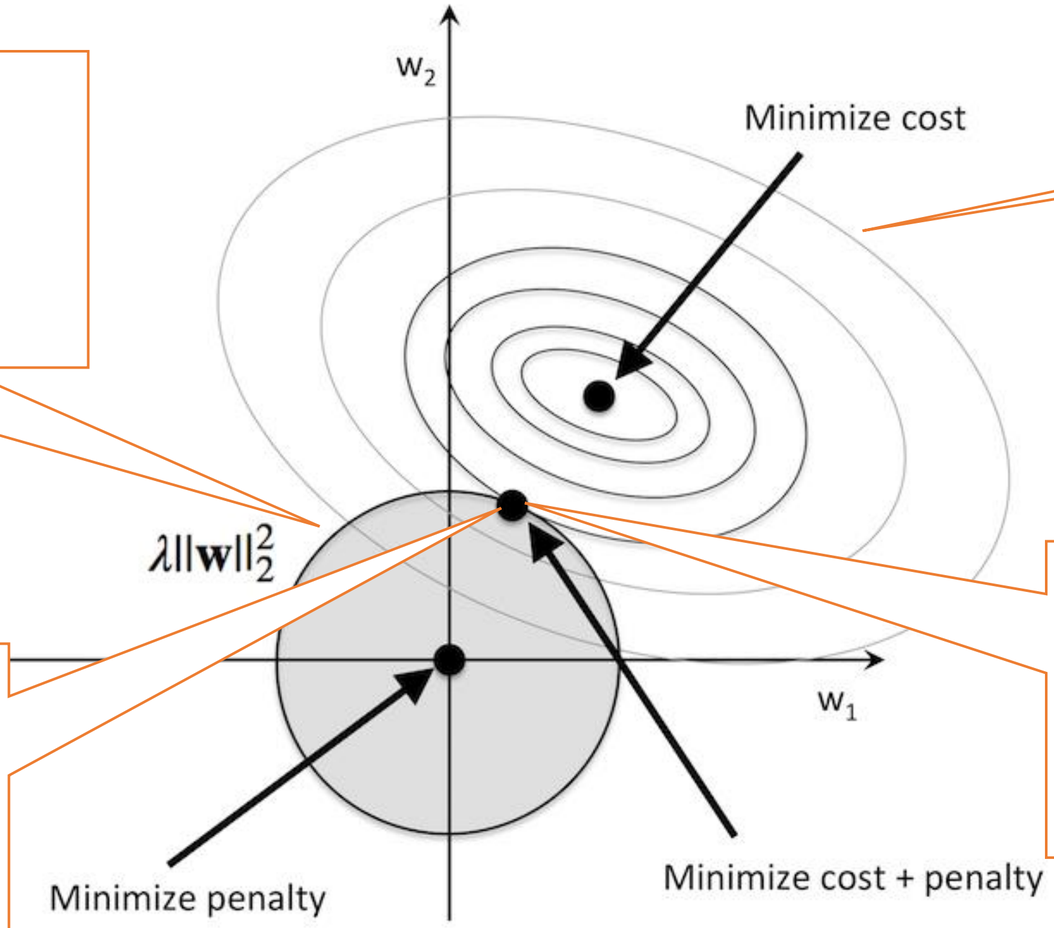
L2 Regularization

- Stop (Adjust) model before it overfits
- https://www.geogebra.org/calculator/c8tvsbrn

All points on circle are equidistant from origin when using Euclidean distance

Contour plot

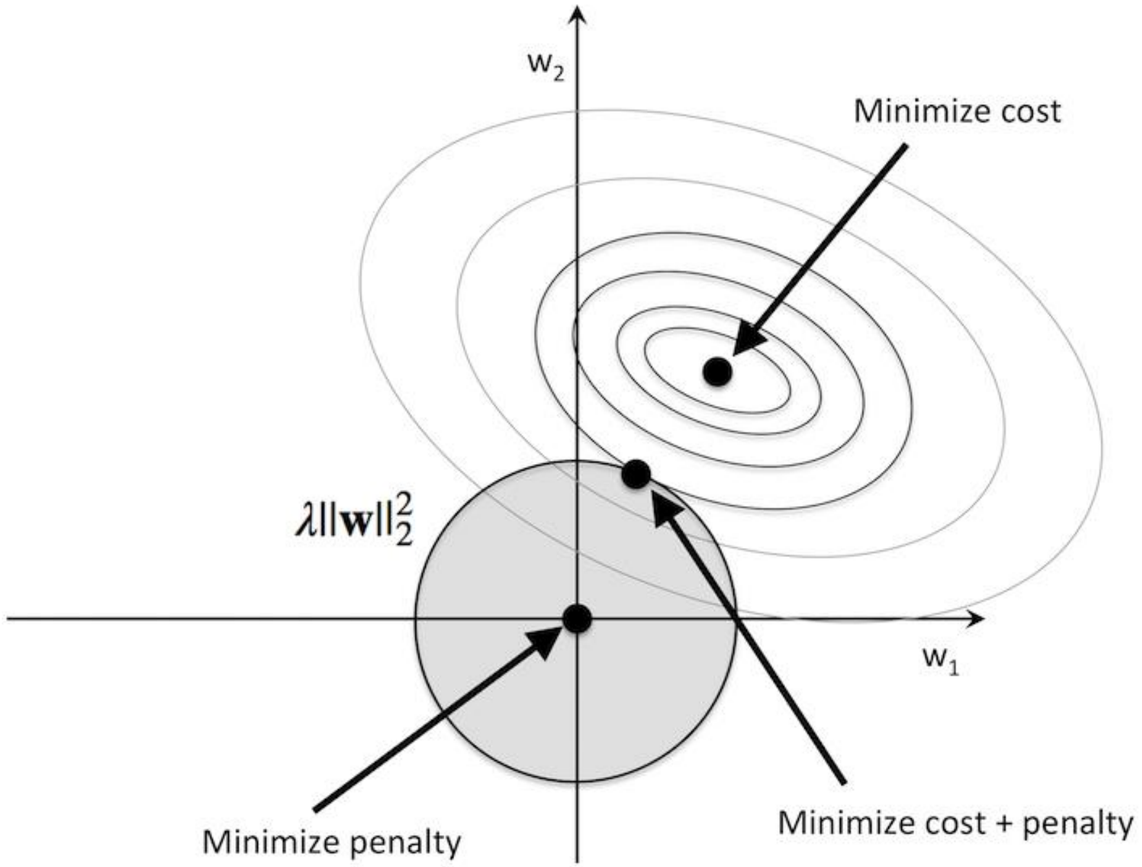
$$\nabla_w \mathcal{J} = \lambda \nabla_w \|w\|^2$$



Common tangent implies gradient vector of one is multiple of another

Both circle and one contour curve share a common tangent at the point of "kiss"

Cost function adjusted for L2 Regularization



$$\mathcal{J}(\mathbf{w}) = \frac{1}{m}\sum_{i=1}^{m}\left(\mathbf{w}^T x^{(i)} - y^{(i)}\right)^2$$

Lagrange Multiplier

$$\nabla_w \mathcal{J} = \lambda \nabla_w \|w\|^2$$

Objective function in Lagrangian notation

$$\nabla_w \mathcal{J} + \lambda \nabla_w \|w\|^2 = 0$$

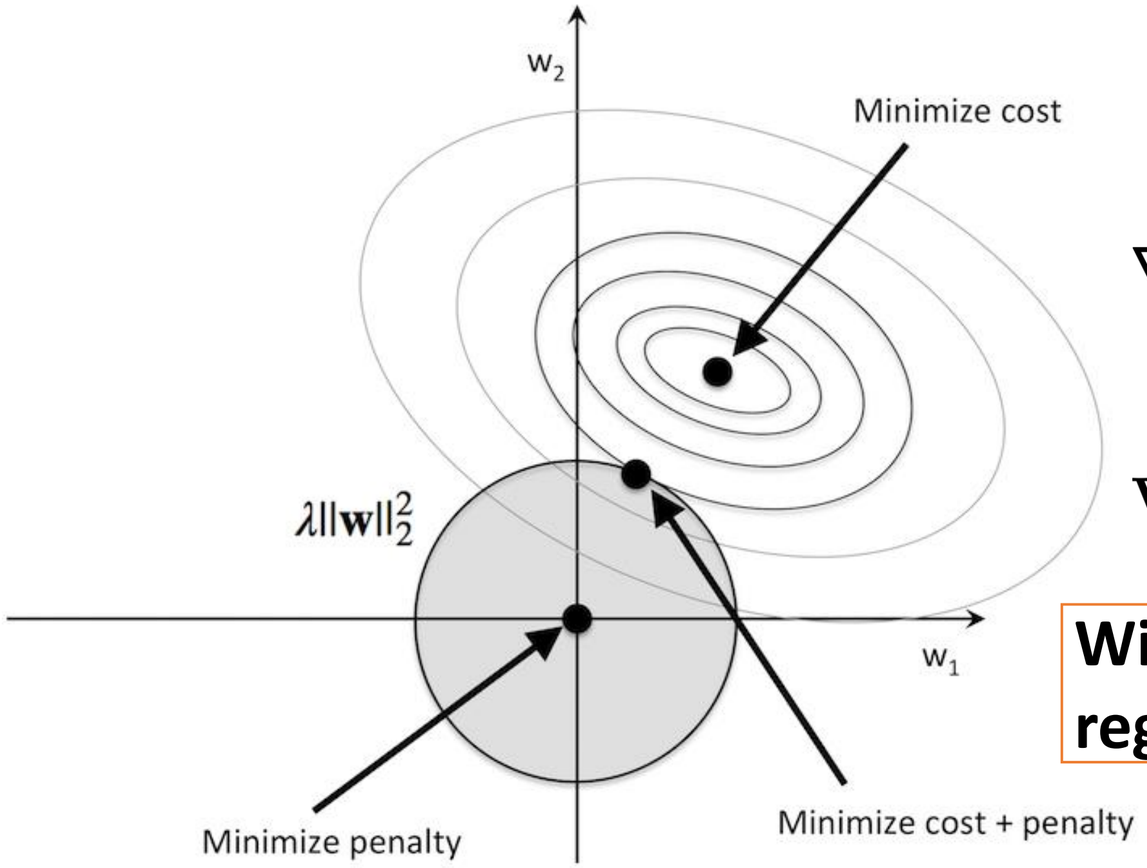
$$\arg\min_w \nabla_w \mathcal{J} + \lambda \nabla_w \|w\|^2$$

Without w0

$$\nabla_w \|w\|^2 = 2w$$

$$\mathcal{J}(w) = \frac{1}{m}(Xw - y)^T(Xw - y)$$

Cost function adjusted for L2 Regularization



$$\arg\min_{w} \nabla_w \mathcal{J} + \lambda \nabla_w \|w\|^2$$

$$\nabla_w \mathcal{J} = \frac{2}{m} X^T (Xw - y) \qquad \nabla_w \|w\|^2 = 2w$$

$$\mathbf{w} = \mathbf{w} - \eta \nabla_w \mathcal{J} \qquad \mathbf{w} = \mathbf{w} - \eta \nabla_w \mathcal{J} - 2\eta\lambda w$$

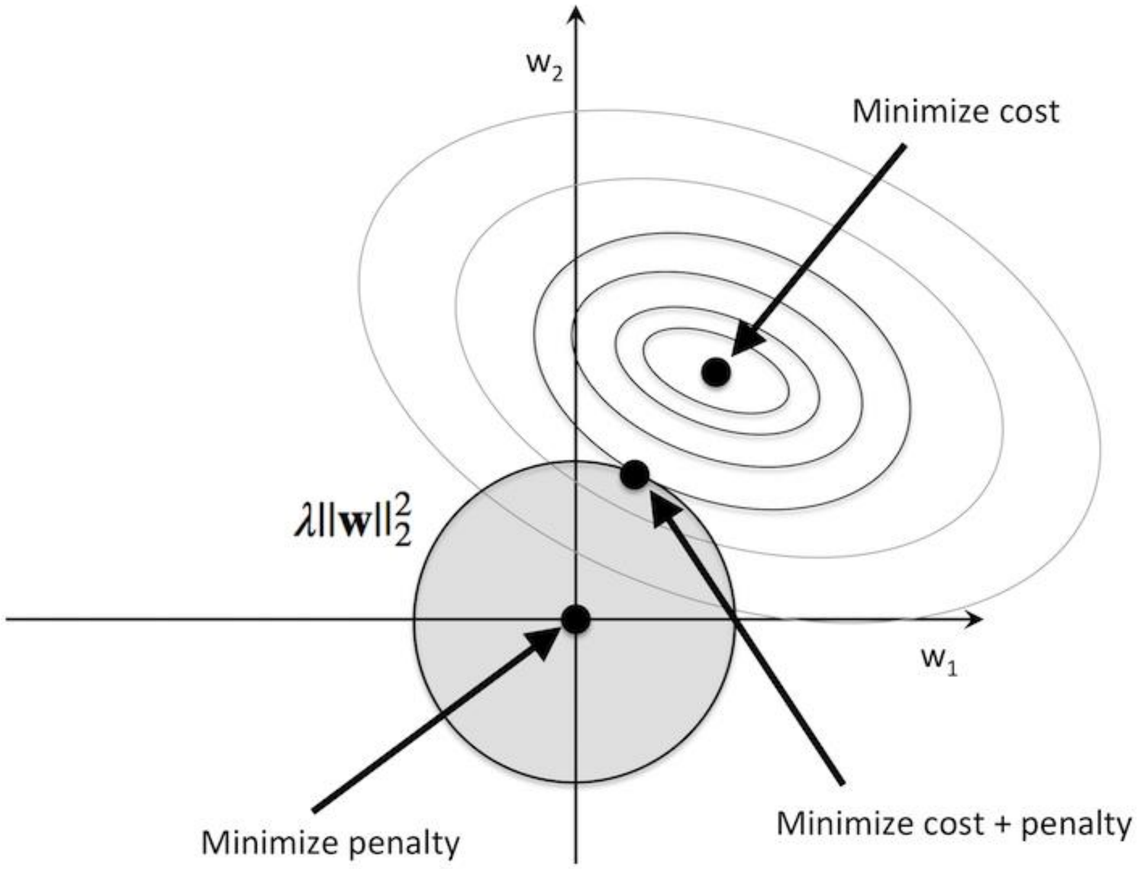
$$\mathbf{w} = (\mathbf{w} - 2\eta\lambda w) - \eta \nabla_w \mathcal{J}$$

$$\mathbf{w} = \mathbf{w}(1 - 2\eta\lambda) - \eta \nabla_w \mathcal{J}$$

Without regularization

A progressively small number keeps getting subtracting from a small w. Net effect w tends to 0, but does not become 0

L2 Regularization - Hyperparam Lambda



$$\arg\min_{w} \nabla_w \mathcal{J} + \lambda \nabla_w \|w\|^2$$

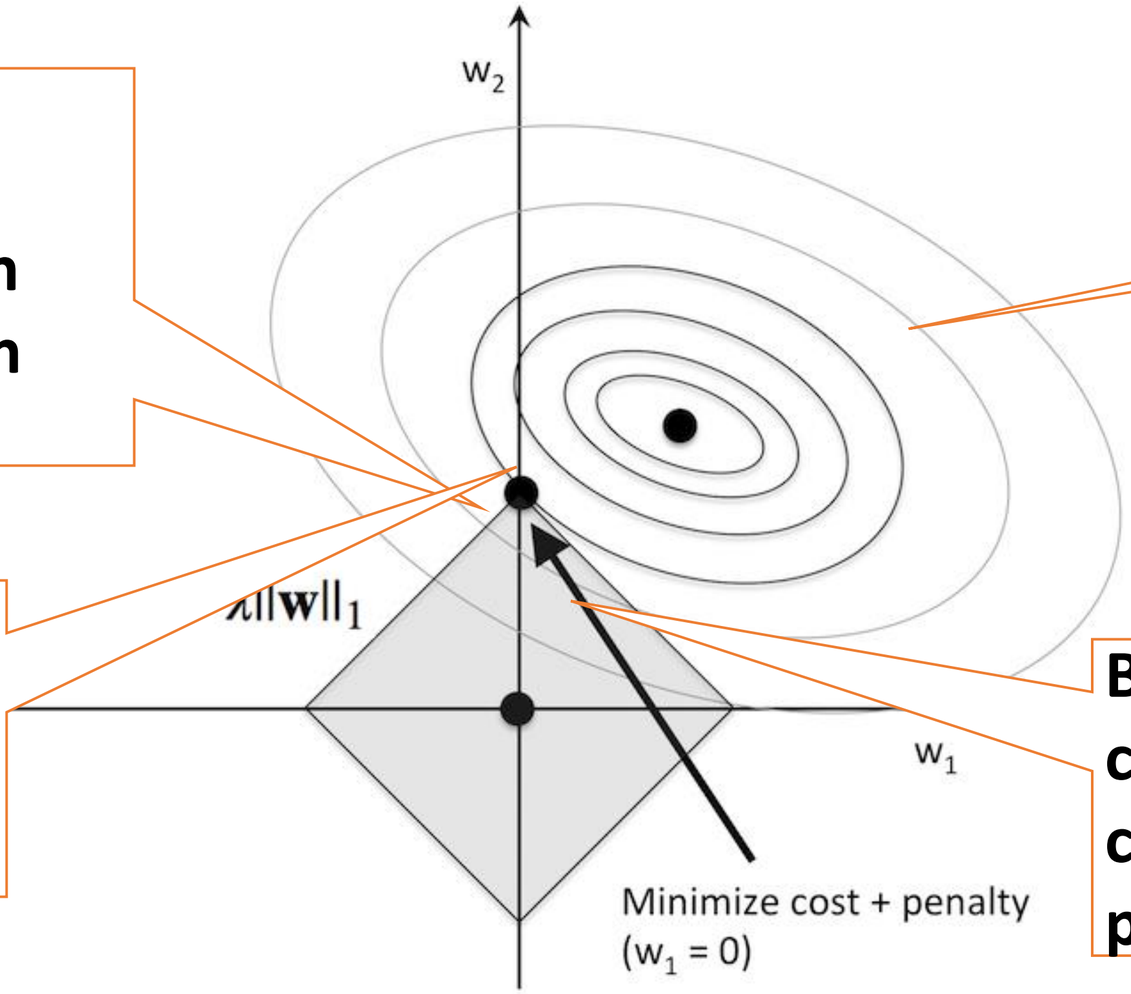
- What happens when
  - Lambda = 0
  - Lambda is very high
- Lambda=0 is equivalent to no regularization
- High lambda overrides objective function

L1 Regularization

- Stop (Adjust) model before it overfits

All points on rhombus circumference are equidistant from origin when using Manhattan distance

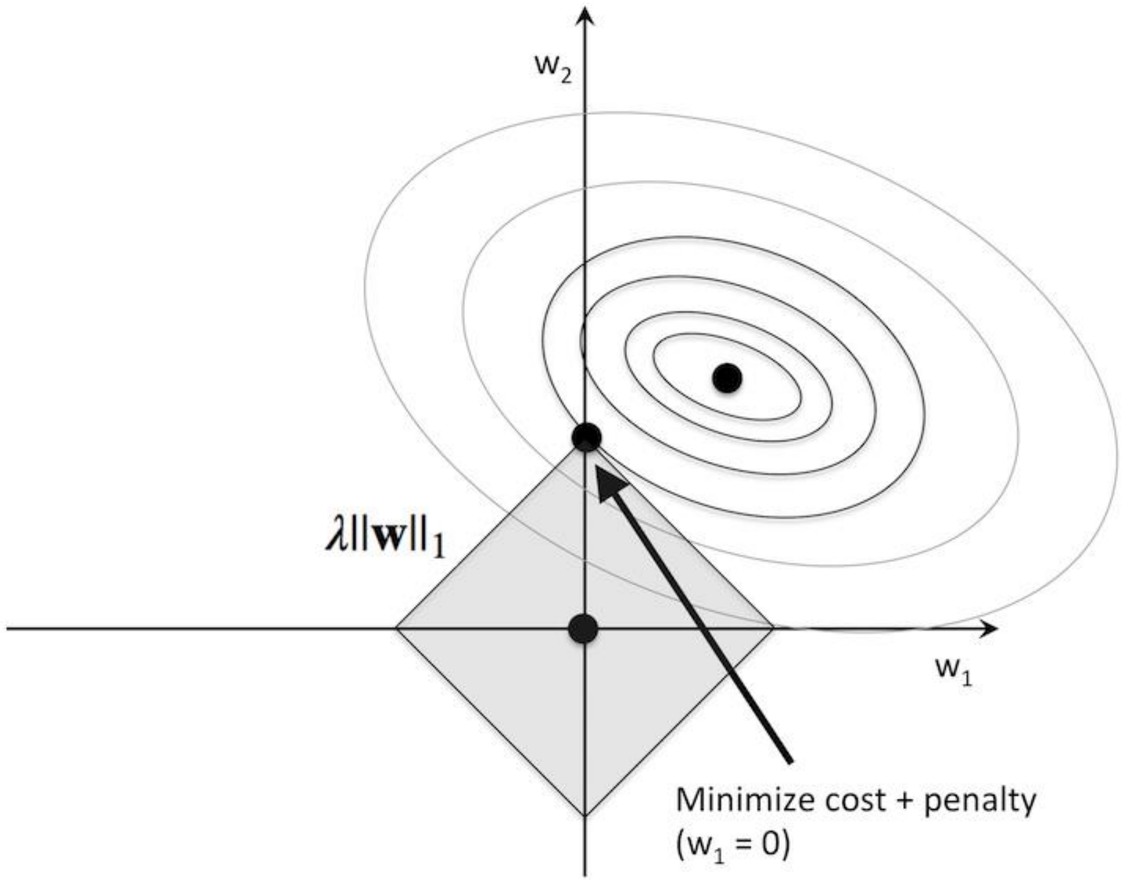
Contour plot



Common tangent implies gradient vector of one is multiple of another

Both rhombus and one contour curve share a common tangent at the point of "kiss"

Cost function adjusted for L1 Regularization



$$\mathcal{J}(\mathbf{w}) = \frac{1}{m}\sum_{i=1}^{m}(\mathbf{w}^T x^{(i)} - y^{(i)})^2$$

Lagrange Multiplier

$$\nabla_w \mathcal{J} = \lambda \nabla_w \|w\|_1$$

Objective function in Lagrangian notation

$$\nabla_w \mathcal{J} + \lambda \nabla_w \|w\|_1 = 0$$

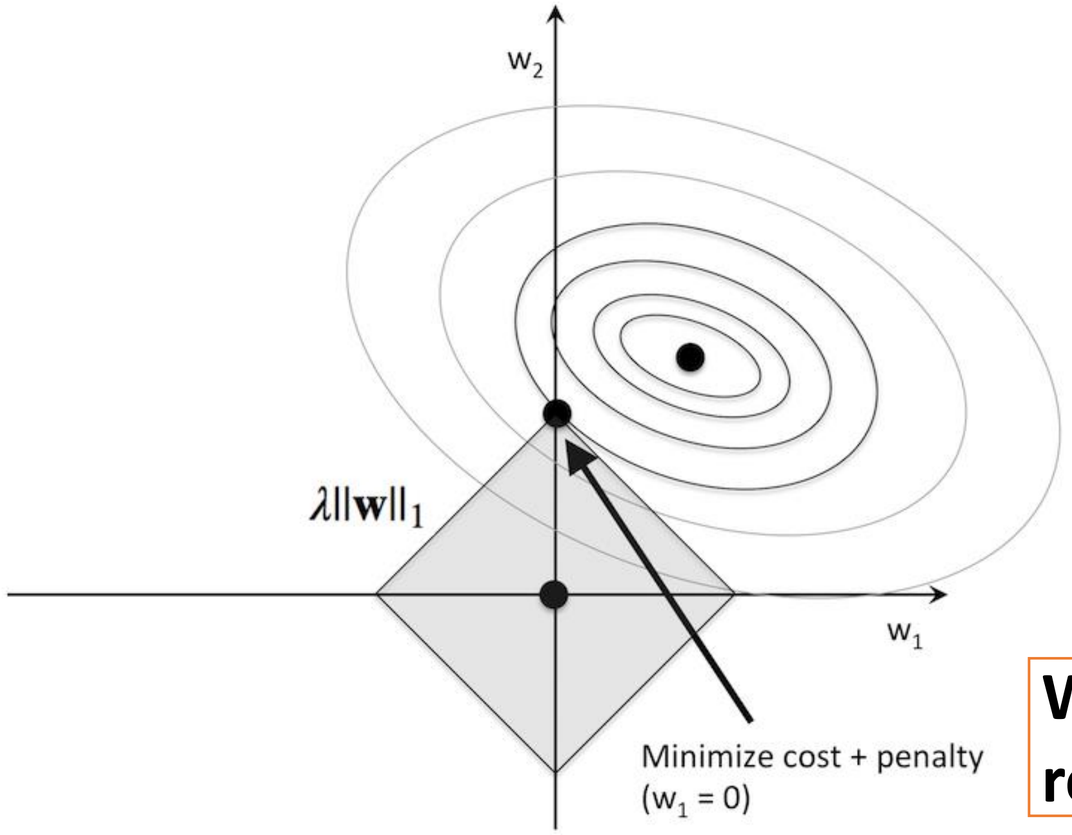
$$\arg\min_w \nabla_w \mathcal{J} + \lambda \nabla_w \|w\|_1$$

Without w0

$$\nabla_w \|w\|^2 = \mathbf{1}$$

$$\mathcal{J}(w) = \frac{1}{m}(Xw - y)^T(Xw - y)$$

Cost function adjusted for L1 Regularization



$$\arg\min_w \nabla_w \mathcal{J} + \lambda \nabla_w \|w\|_1$$

$$\nabla_w \mathcal{J} = \frac{2}{m} X^T (Xw - y) \qquad \nabla_w \|w\|_1 = \mathbf{1}$$

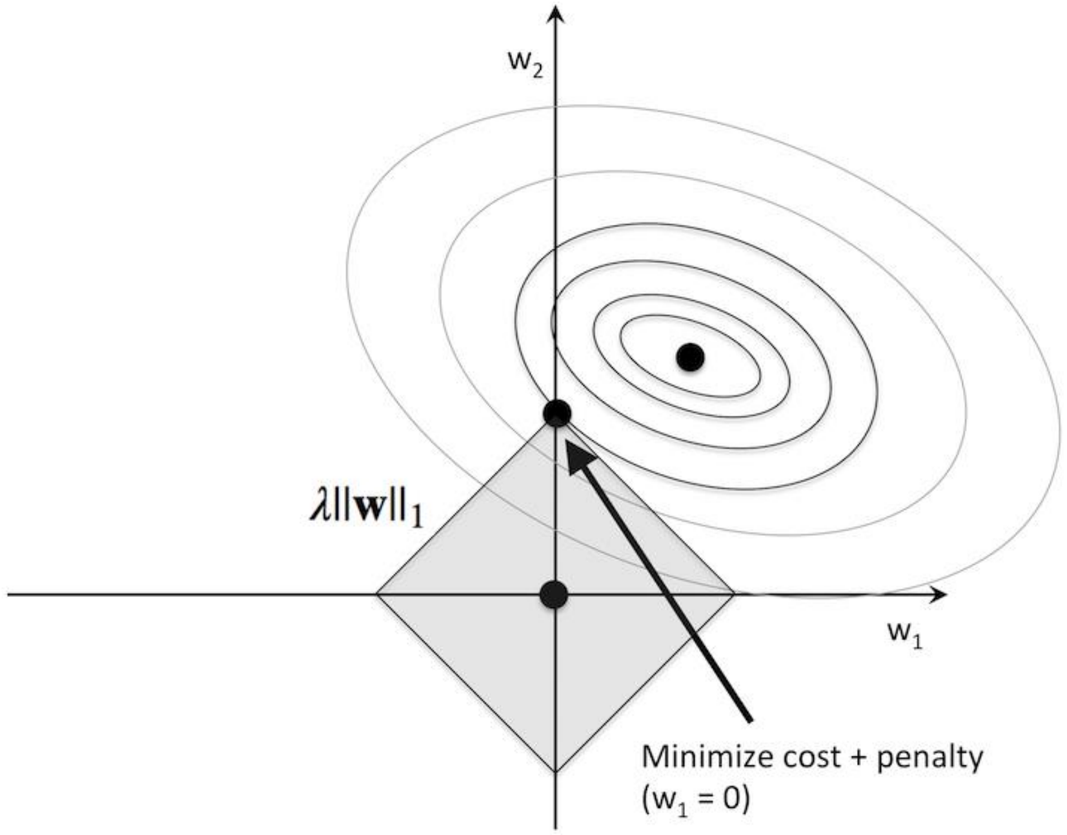
$$\mathbf{w} = \mathbf{w} - \eta \nabla_w \mathcal{J} \qquad \mathbf{w} = \mathbf{w} - \eta \nabla_w \mathcal{J} - \eta\lambda$$

Without regularization

$$\mathbf{w} = (\mathbf{w} - \eta\lambda) - \eta \nabla_w \mathcal{J}$$

A FIXED small number keeps getting subtracting from a small w. Net effect w becomes 0

L1 Regularization - Hyperparam Lambda



$$\arg\min_{w} \nabla_w \mathcal{J} + \lambda \nabla_w \|w\|_1$$

- What happens when
  - Lambda = 0
  - Lambda is very high

- Lambda=0 is equivalent to no regularization
- High lambda overrides objective function and makes all w = 0

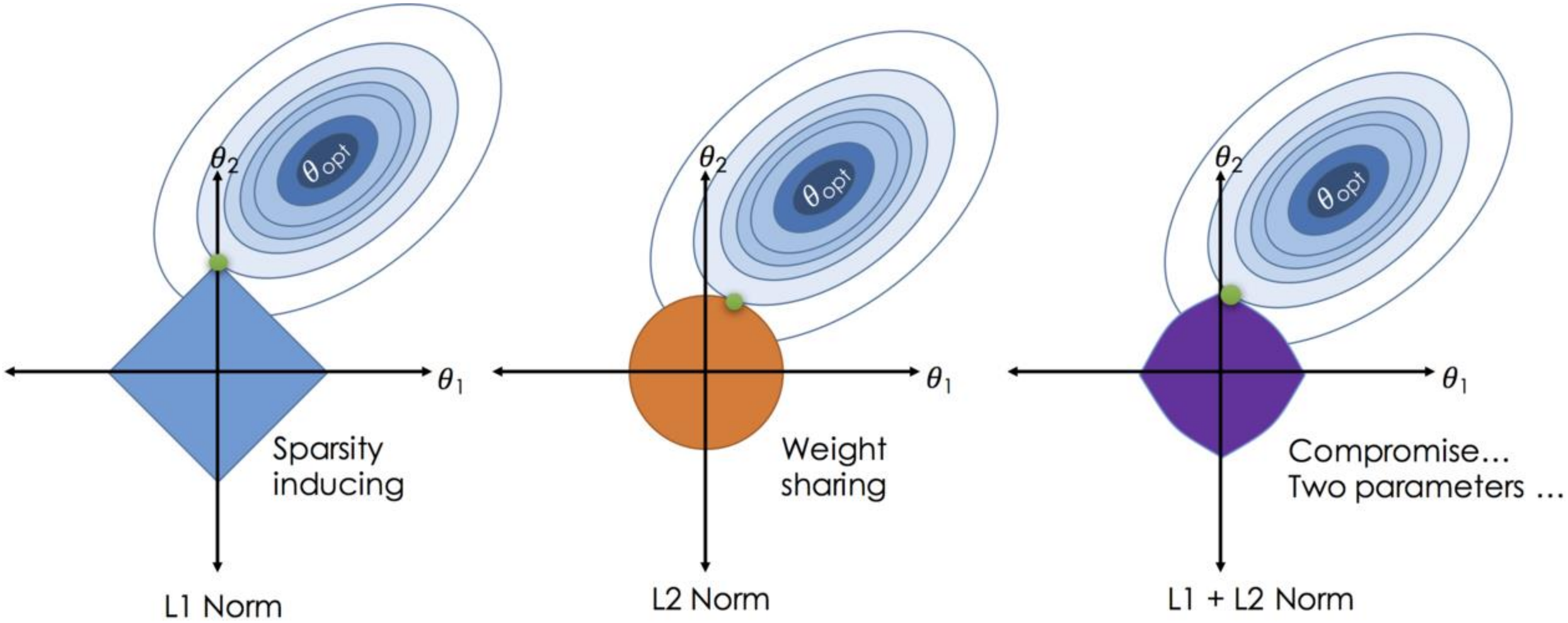
L2 and L1 regularization

- Regularization not applied to bias term (Why not?)
- L2 Regression also called Ridge Regression
- Reduces w1 thru wn, but does not make 0
- L1 Regularization also called Lasso Regularization
- Makes some of w1 thru wn 0 in the process of reducing value
- Used more for Feature Elimination

L1 regularization and Sparsity

- LASSO – Least Absolute Shrinkage and Selection Operator
- LASSO makes the weight coefficients sparse
- Because it puts 0 corresponding to features with low importance

Elastic Net Regularization = L1 + L2



$$\arg\min_{w} \nabla_w \mathcal{J} + \lambda_1 \nabla_w \|w\|_1 + \lambda_2 \nabla_w \|w\|^2$$


QUESTIONS