

$$\bar{x}_1 = \frac{1+3+3+6+8+9}{6} = 5$$

$$\bar{x}_2 = 2+4+4+8+12 = 30$$

$$\bar{x}_3 = 4+7+1+6 = 18$$

$$\log_b b^x = x$$

$$\log_a x = \frac{\log_b x}{\log_b a}$$



$$\begin{aligned} \frac{a}{b} &= \frac{a}{bc} & X^2 - 4X + 5 &\leq 5 \\ \frac{a}{b} &= \frac{ac}{b} & X^2 - 4X &\leq 0 \\ \frac{a}{b} + \frac{c}{d} &= \frac{ad+bc}{bd} & n(B \cap C) &= 22 \\ & & n(B) &= 68 \\ & & n(C) &= 84 \end{aligned}$$

$$n(B \cup C) = n(B) + n(C) - n(B \cap C)$$

$$He = 4.002602$$

$$Na = 22.989769$$

$$Ar = 39.948$$



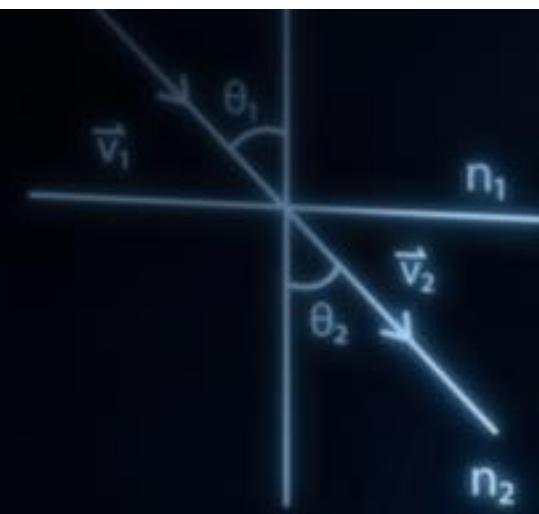
$$M = \frac{0.046765}{30L}$$

$$n(B \cup C) = n(B) + n(C) - n(B \cap C)$$



Recap

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



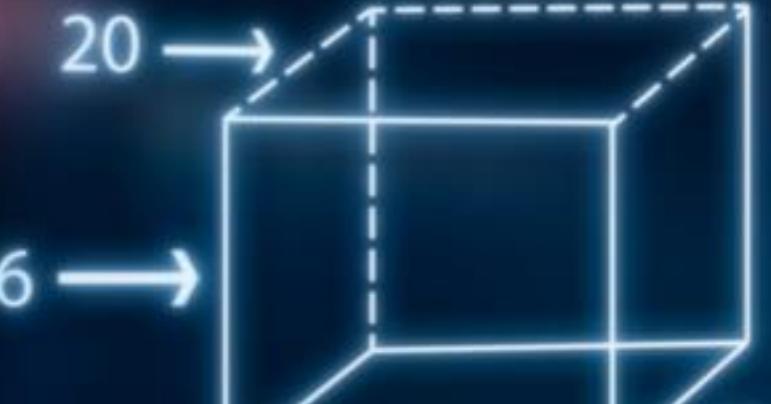
$$\bar{x}_1 = \frac{1+3+3+6+8+9}{6} = 5$$

$$\bar{x}_2 = \frac{2+4+4+8+12}{6} = 30$$

$$\bar{x}_3 = \frac{4+7+1+6}{6} = 18$$

$$\log_b b^x = x$$

$$\log_a x = \frac{\log_b x}{\log_b a}$$



$$\begin{aligned} \frac{a}{b} &= \frac{a}{bc} & X^2 - 4X + 5 &\leq 5 \\ \frac{a}{b} &= \frac{ac}{b} & X^2 - 4X &\leq 0 \\ \frac{a}{b} + \frac{c}{d} &= \frac{ad+bc}{bd} & n(B \cap C) &= 22 \\ & & n(B) &= 68 \\ & & n(C) &= 84 \end{aligned}$$

$$n(B \cup C) = n(B) + n(C) - n(B \cap C)$$

$$He = 4.002602$$

$$Na = 22.989769$$

$$Ar = 39.948$$



$$M = \frac{0.046765}{30L}$$

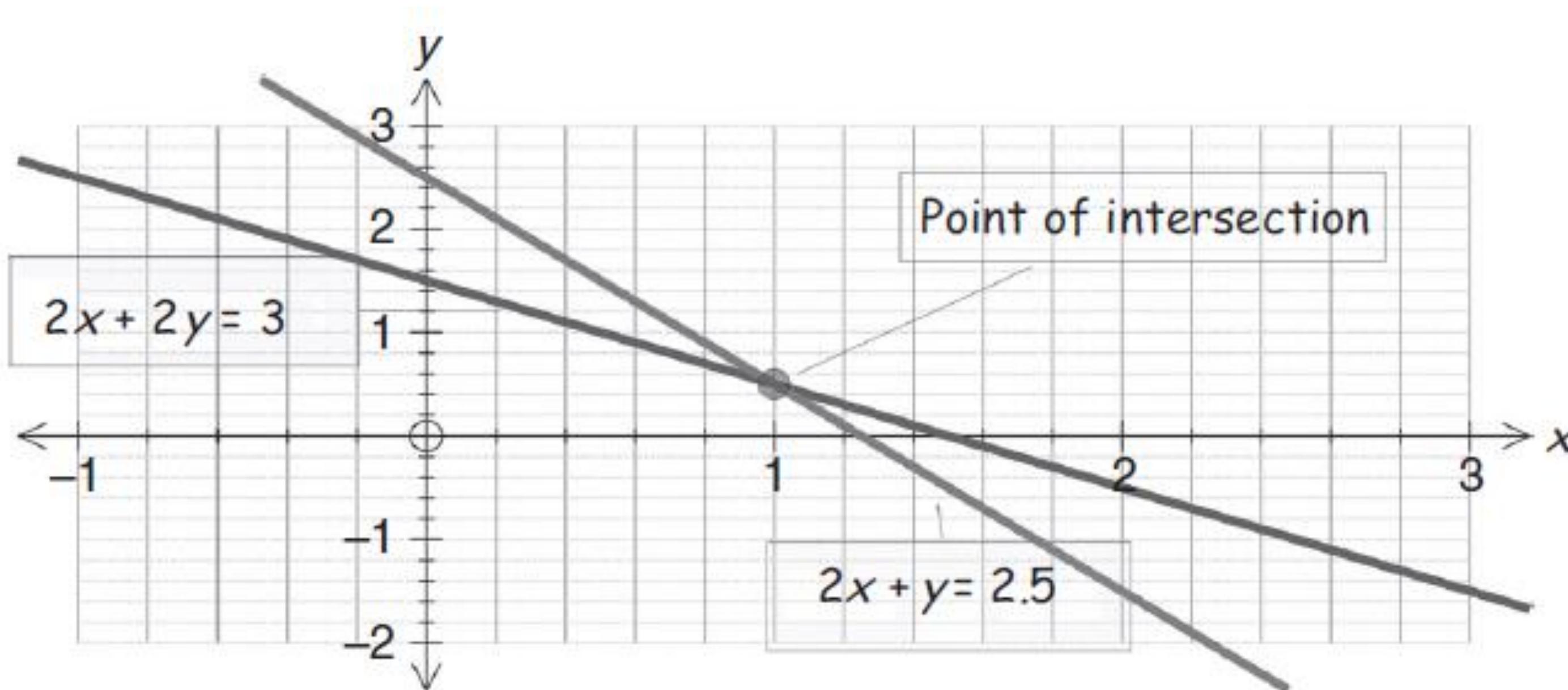


System of Equations

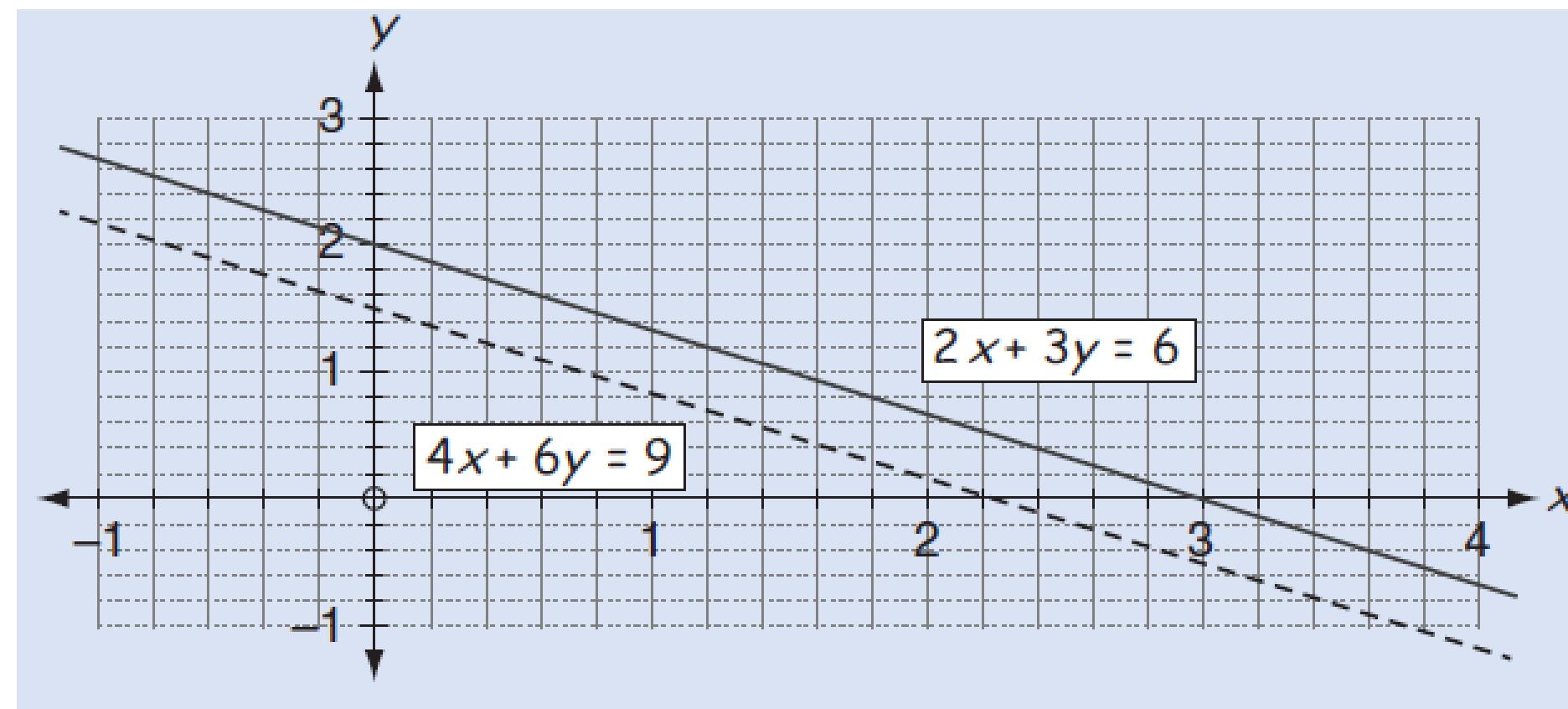
$$2x + 2y = 3$$

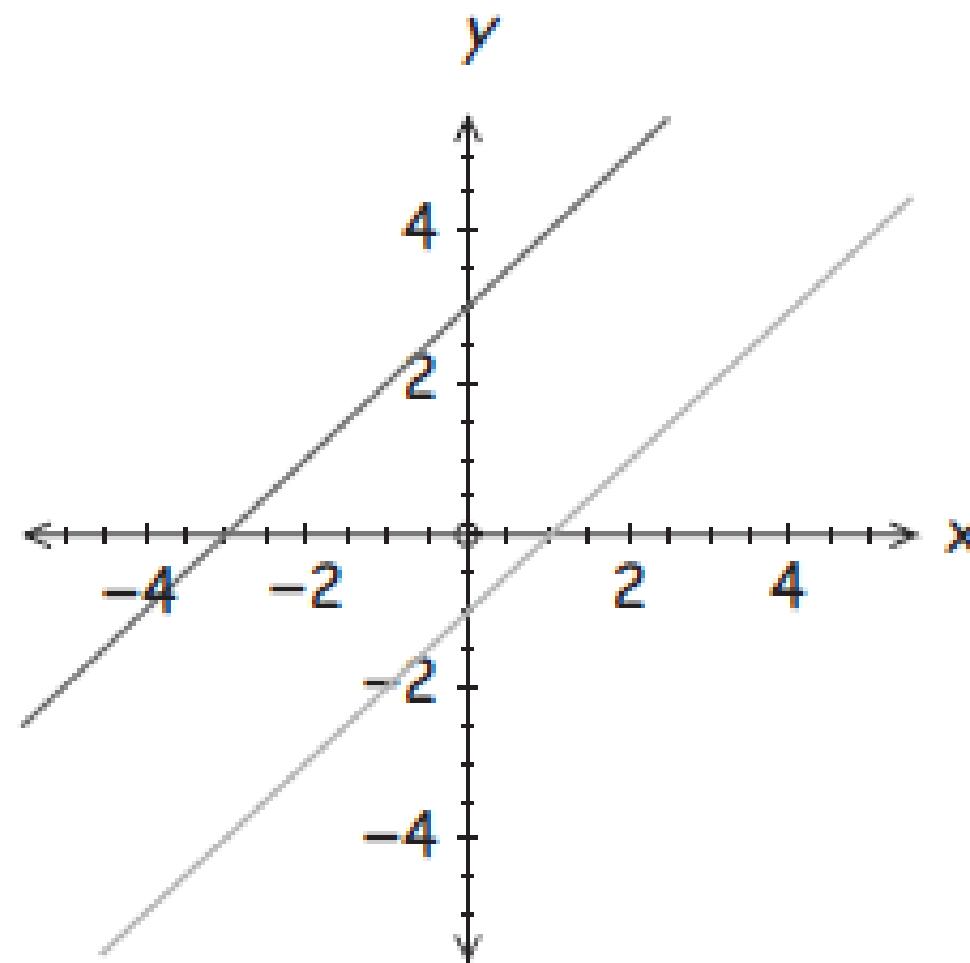
$$2x + y = 2.5$$

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

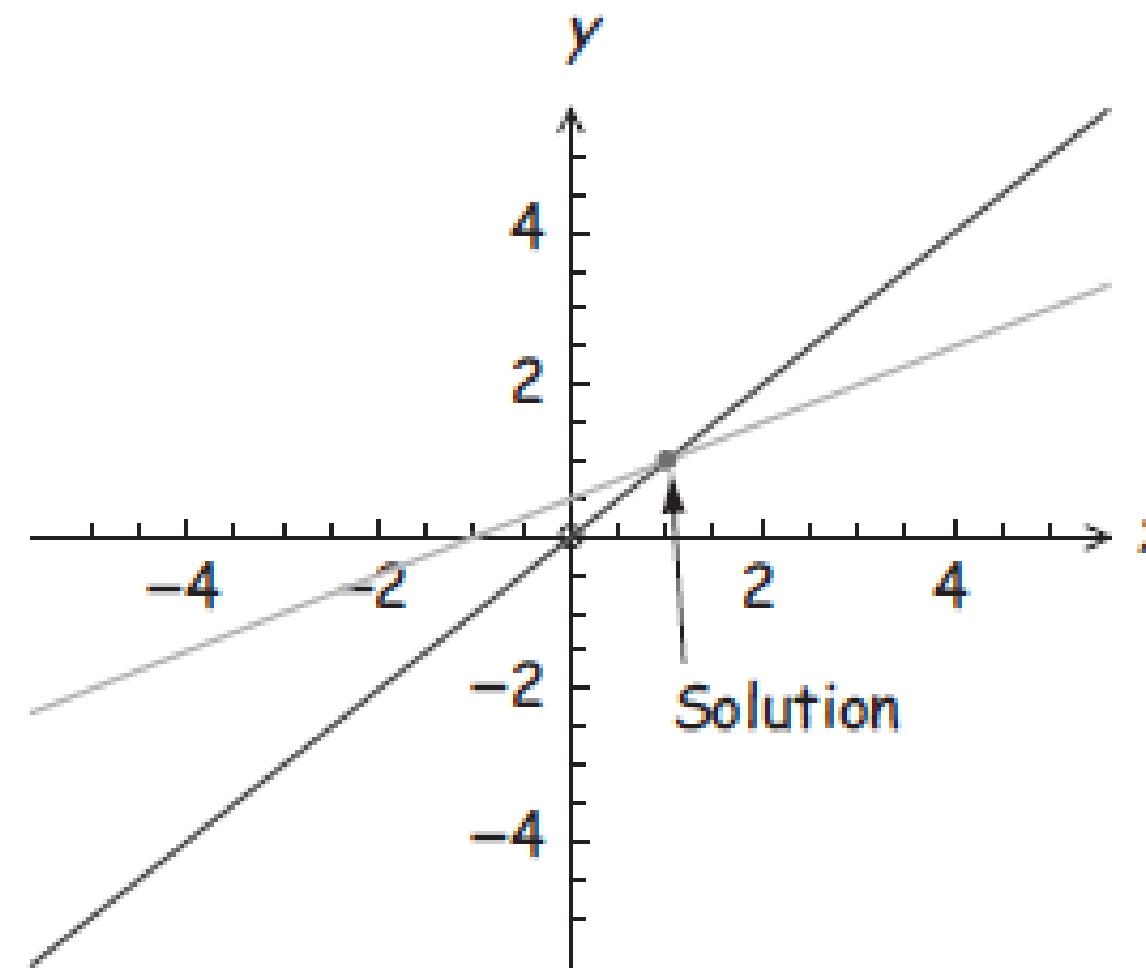


No Solution

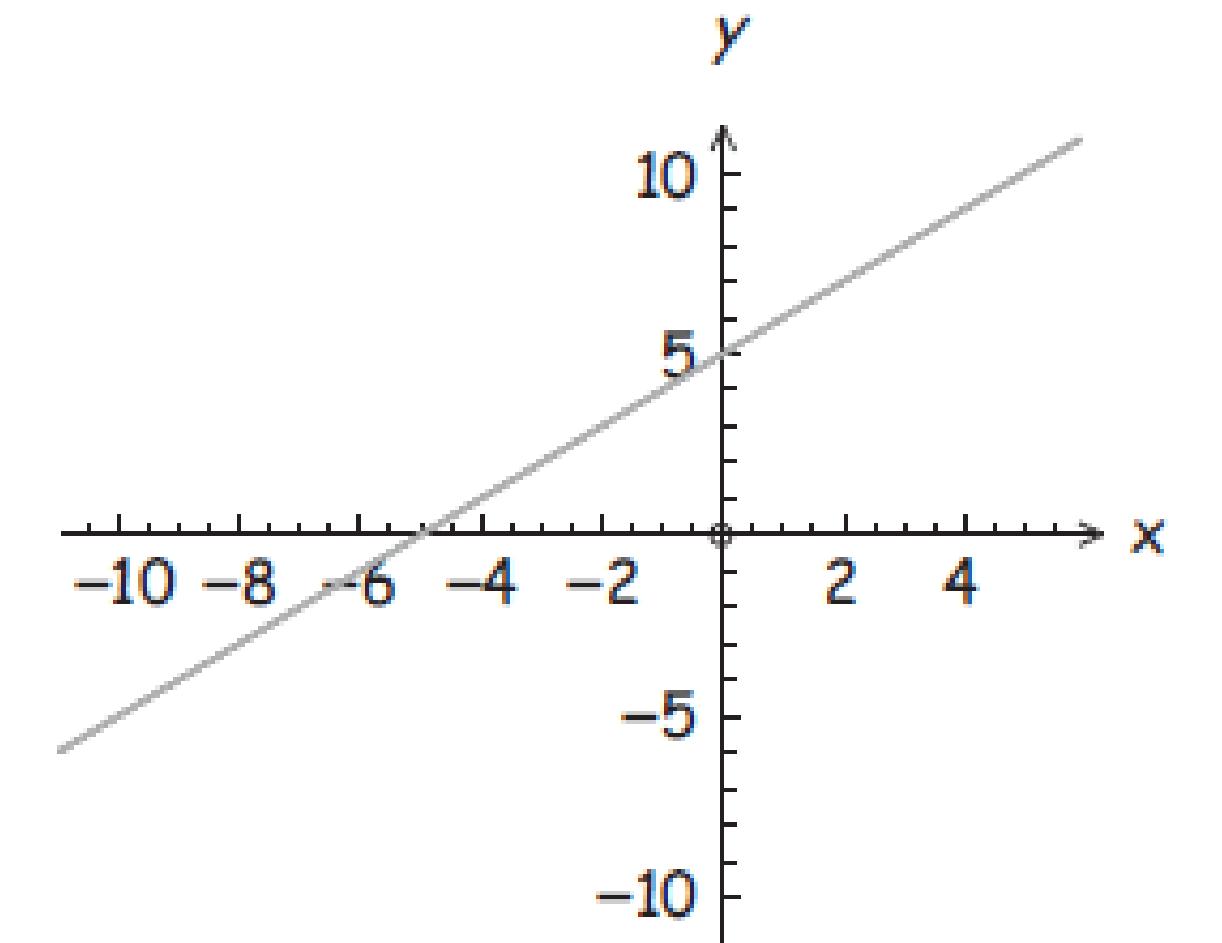




(a) No solution



(b) Unique solution



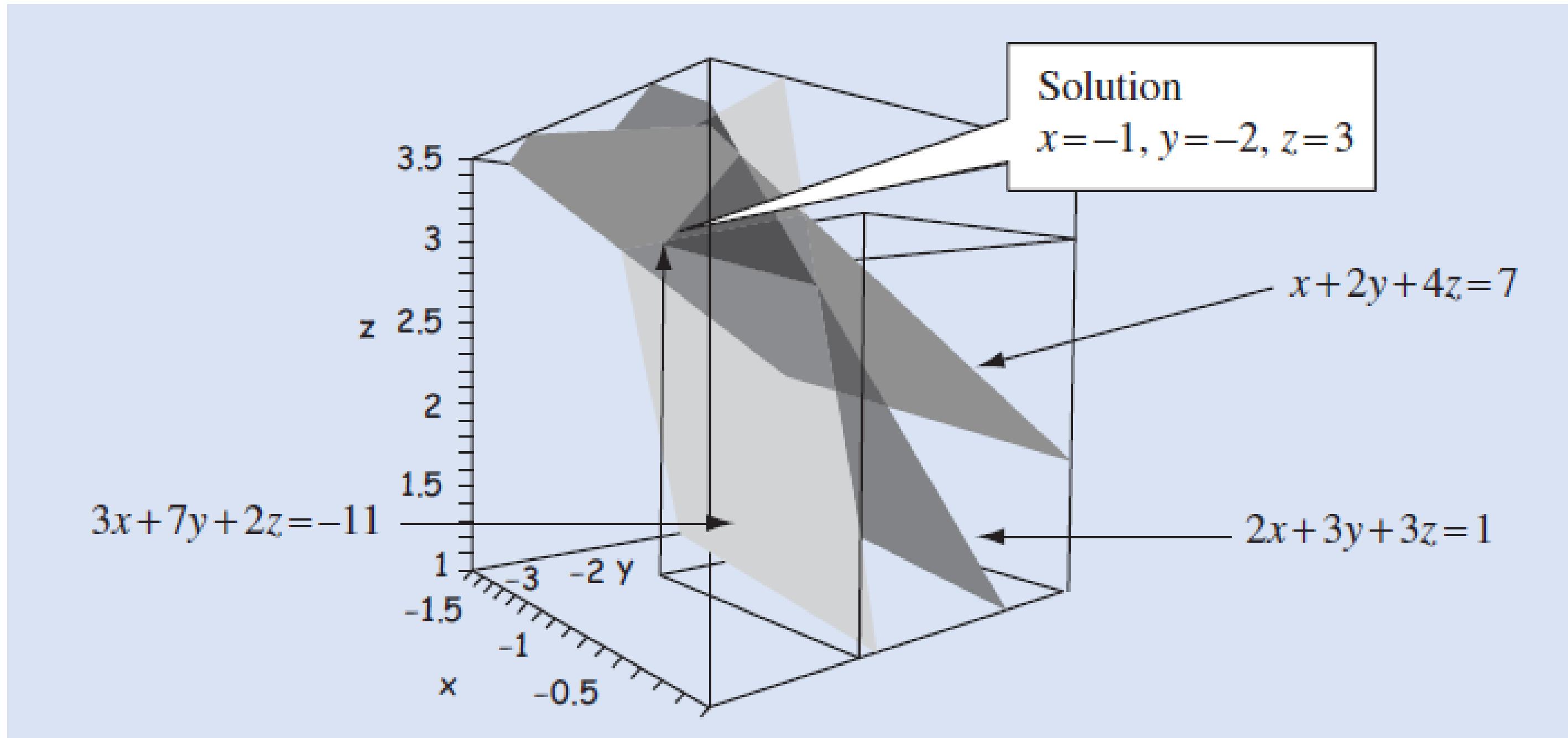
(c) Infinite number of solutions

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

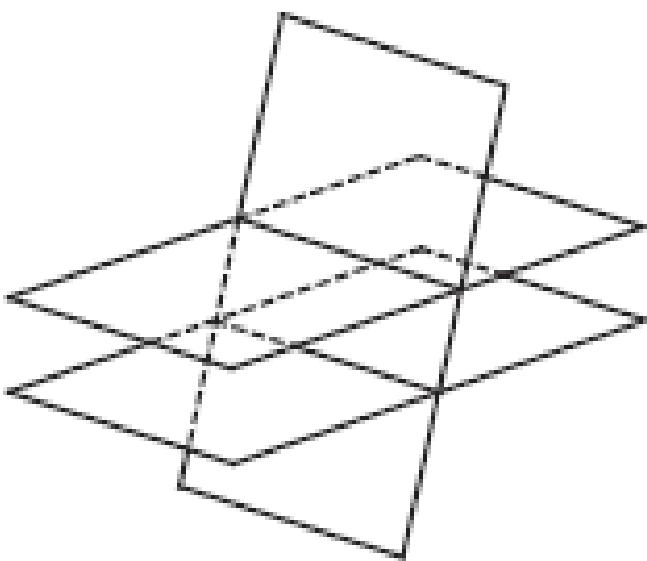
$$3x + 7y + 2z = -11$$

$$2x + 3y + 3z = 1$$

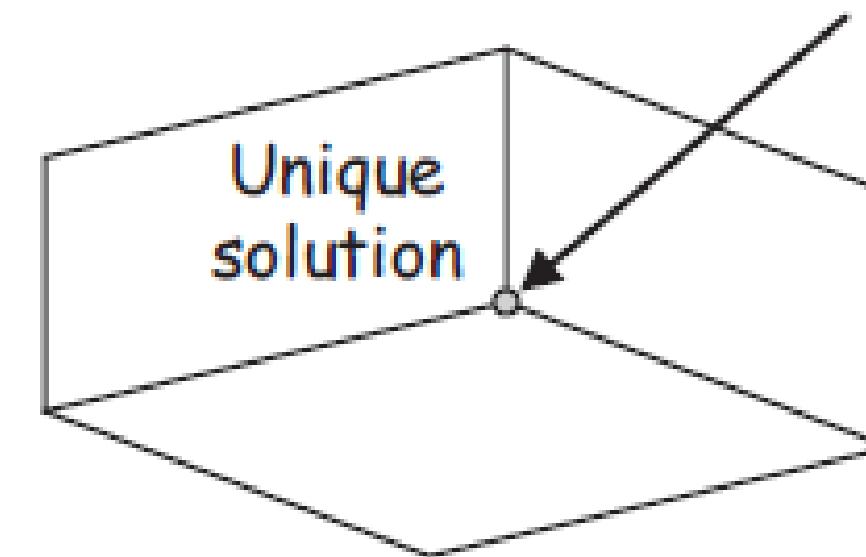
• <https://www.geogebra.org/calculator/xuqtru6e>



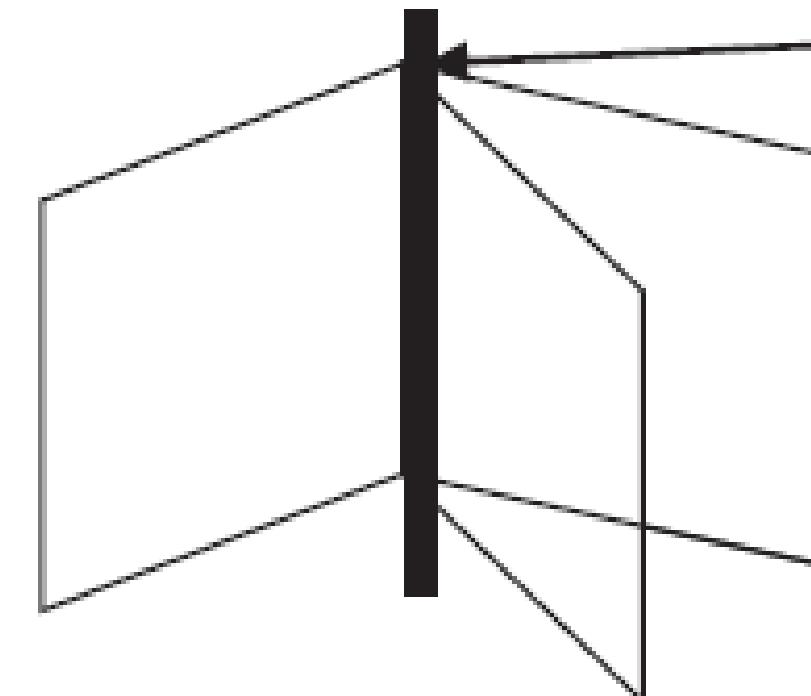
(a) No solution



(b) Unique solution



(c) Infinite number of solutions



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

$$2x + 2y = 3$$

- Equation MAY have exact solution

$$2x + 2y = 9$$

- Need not necessarily be the case

$$2x + 2y = 3$$

$$4x + 4y = 6$$

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

$$230x_1 + 37.8x_2 = 22.1$$

$$44.5x_1 + 39.3x_2 = 10.4$$

$$17.2x_1 + 45.9x_2 = 9.3$$

$$151.5x_1 + 41.3x_2 = 18.5$$

$$180.8x_1 + 10.8x_2 = 12.9$$

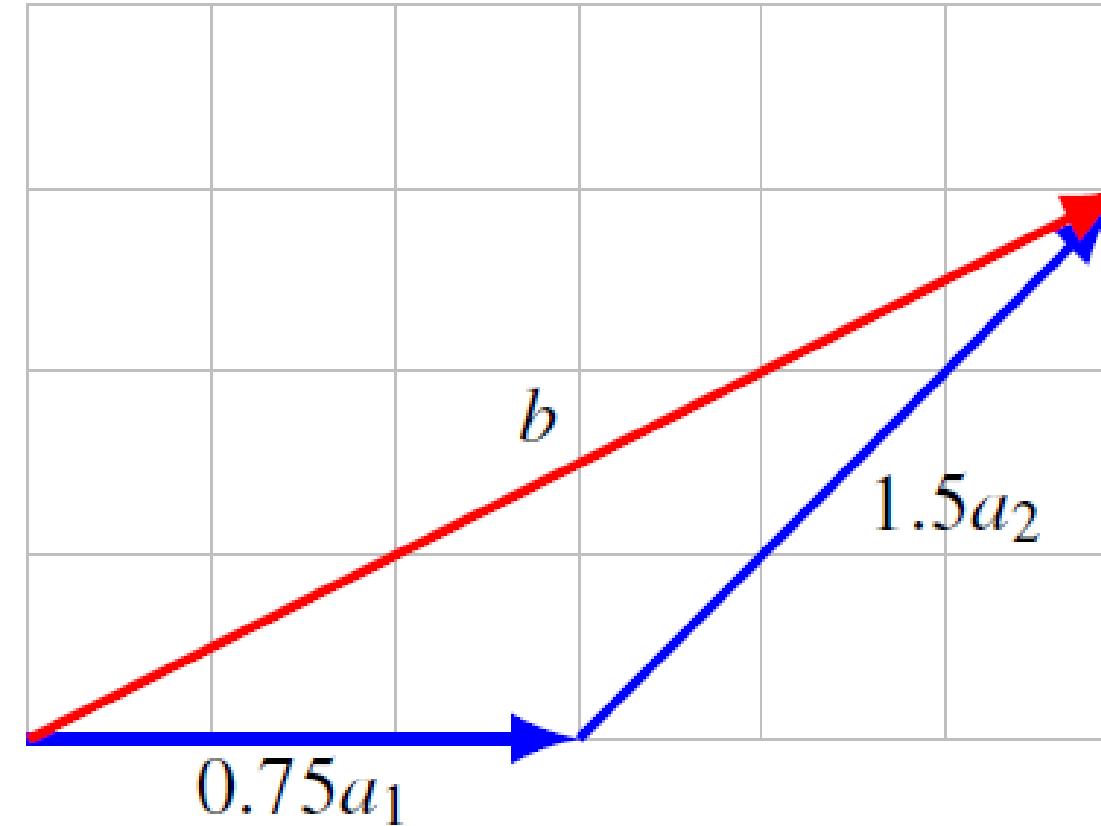
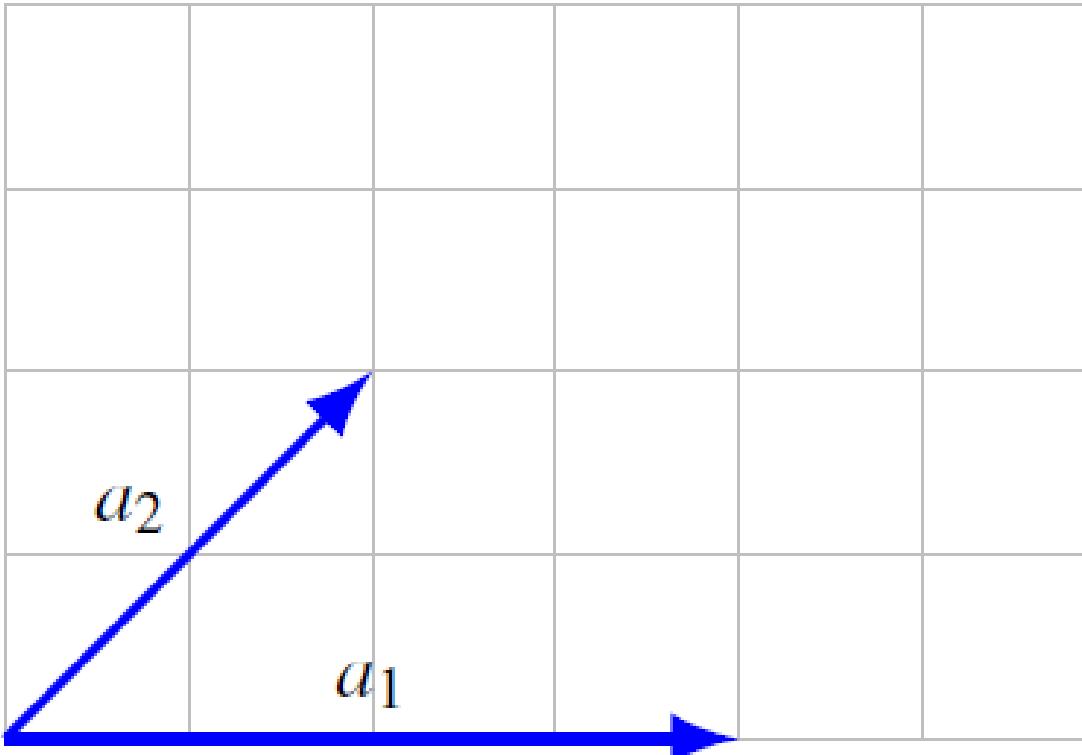
TV	Radio	Sales
230.1	37.8	22.1
44.5	39.3	10.4
17.2	45.9	9.3
151.5	41.3	18.5
180.8	10.8	12.9

Dataset with
m records &
n features

$m \gg n$

Linear Combinations

- Definition: $\beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \dots + \beta_n \mathbf{x}_n$
 - $\beta_1, \beta_2, \dots, \beta_n$ are scalars
 - $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ are vectors
- Simply put: Scale and add vectors



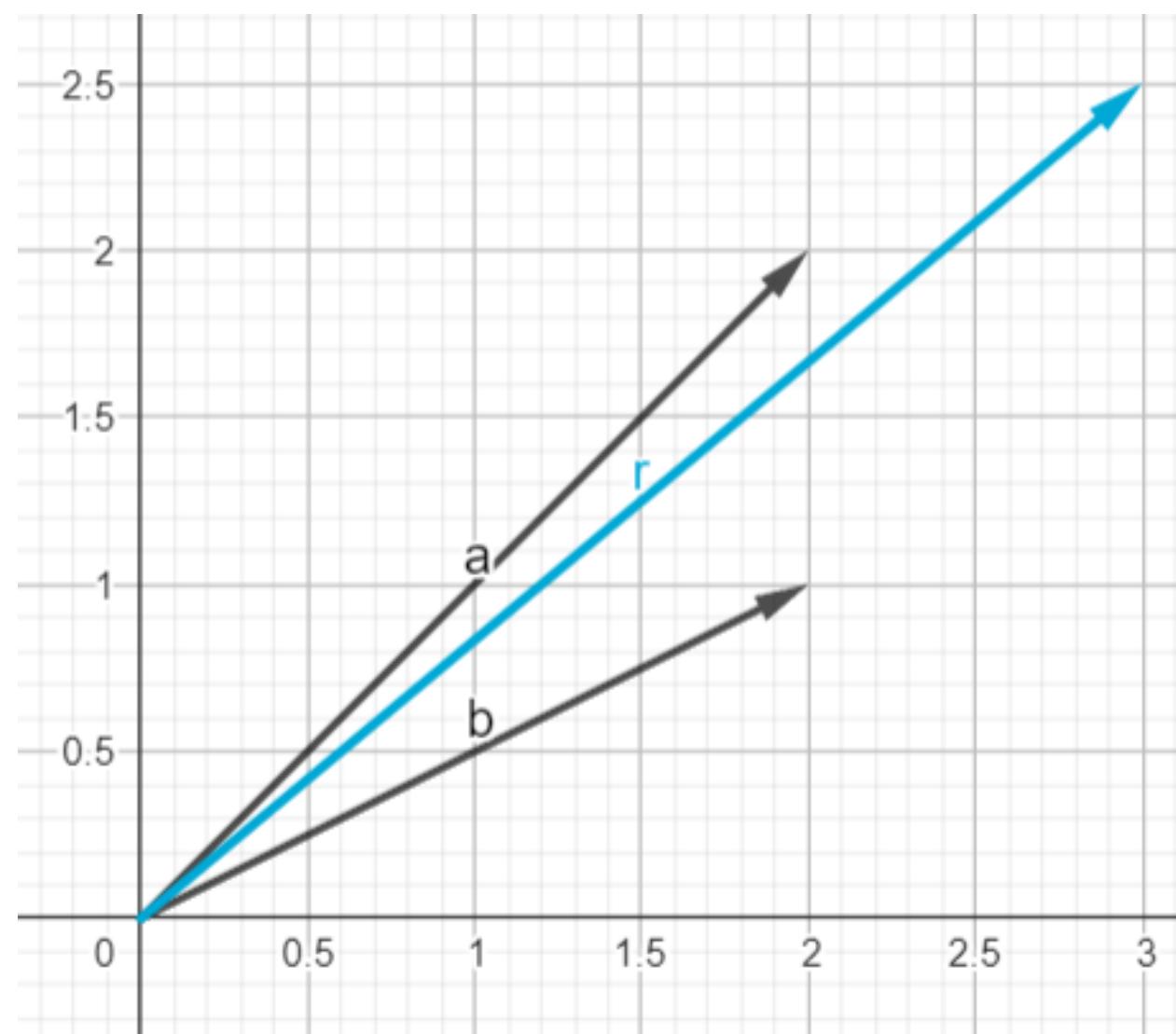
- <https://www.geogebra.org/calculator/eedpbvb9>

System of Equations as linear combination

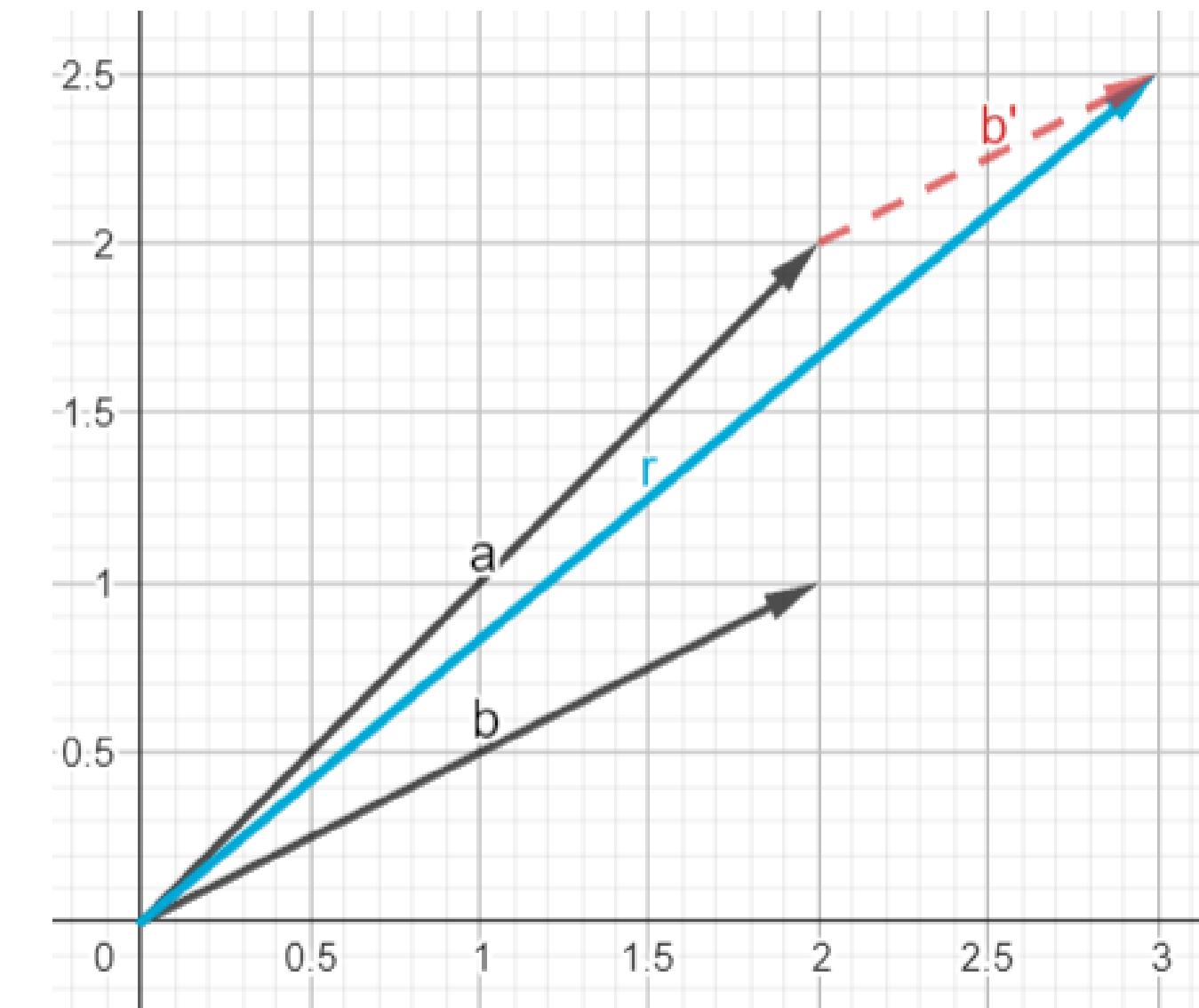
$$2x + 2y = 3$$

$$2x + y = 2.5$$

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



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



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
- SOE can be written as linear combination

$$2x + 2y = 3$$

$$2x + y = 2.5$$

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

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

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

$$\mathbf{Ax} = \mathbf{b}$$

$$\mathbf{Xw} = \mathbf{y}$$

- 3Blue1Brown Essence of Linear Algebra on youtube
 - <https://www.youtube.com/playlist?list=PLZHQB0WTQDPD3MizzM2xVFitgF8hE>

Matrix equation solution & notations

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

Linear Algebra
Convention

$$Ax = b$$

Inverse
of Matrix

$$x = A^{-1}y$$

Statistics
Convention

$$X\beta = y$$

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

Andrew Ng
Convention

$$X\theta = y$$

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

Our uniform
Convention

$$Xw = y$$

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

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

$$230x_1 + 37.8x_2 = 22.1$$

$$44.5x_1 + 39.3x_2 = 10.4$$

$$17.2x_1 + 45.9x_2 = 9.3$$

$$151.5x_1 + 41.3x_2 = 18.5$$

$$180.8x_1 + 10.8x_2 = 12.9$$

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

$$5 \times 2$$

$$2 \times 1$$

$$5 \times 1$$

What linear combinations possible if no exact solution?

How about approximate solution?

Unreachable combinations

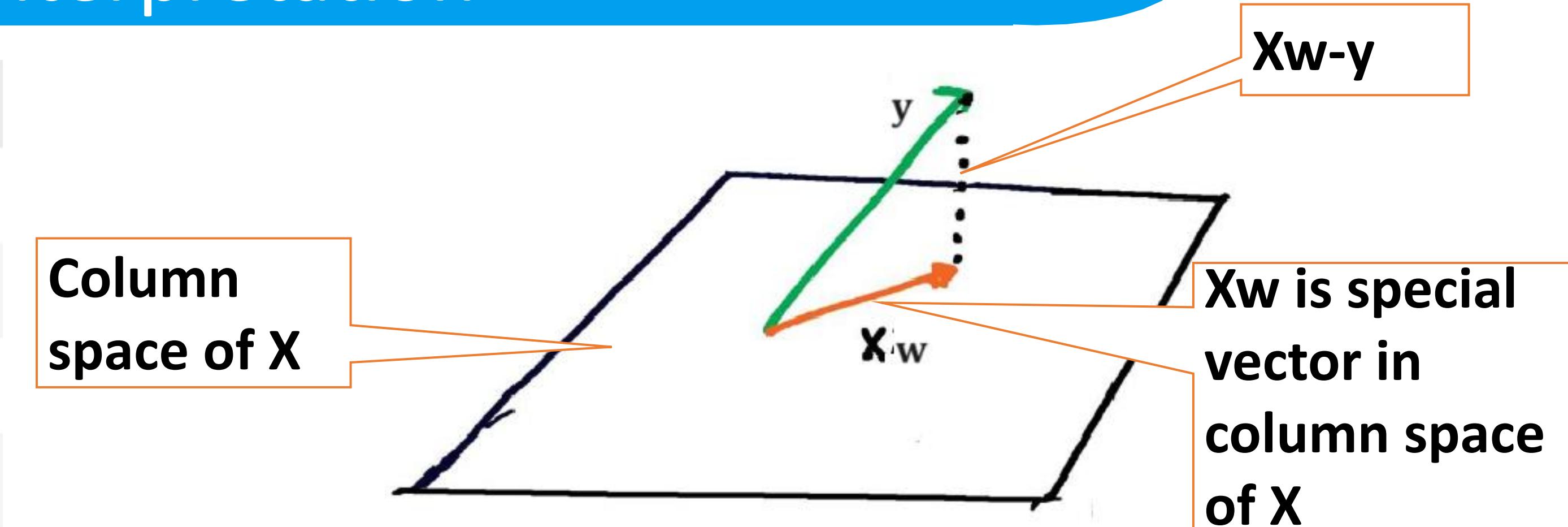
- Recall: Exact solutions to system of equations are vectors reachable by linear combination
- How do unreachable vectors & approximate solutions look like?
 - <https://www.geogebra.org/calculator/xansjfxj>
- Ambient vector space
- Actual vector sub space (column space)

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

$$\begin{bmatrix} 230 & 37.8 \\ 44.5 & 39.3 \\ 17.2 & 45.9 \end{bmatrix}$$

$$\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} 22.1 \\ 10.4 \\ 9.3 \end{bmatrix}$$



$$Xw = \hat{y}$$

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

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

Vectorized closed form solution

$$X = \begin{bmatrix} 1 & x_1^{(1)} \\ 1 & x_1^{(2)} \\ 1 & .. \\ 1 & x_1^{(m)} \end{bmatrix} \quad \mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$$

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

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

Normal
equation

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

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

Compare with non vectorized
solution

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

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

Inverse of
Matrix

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

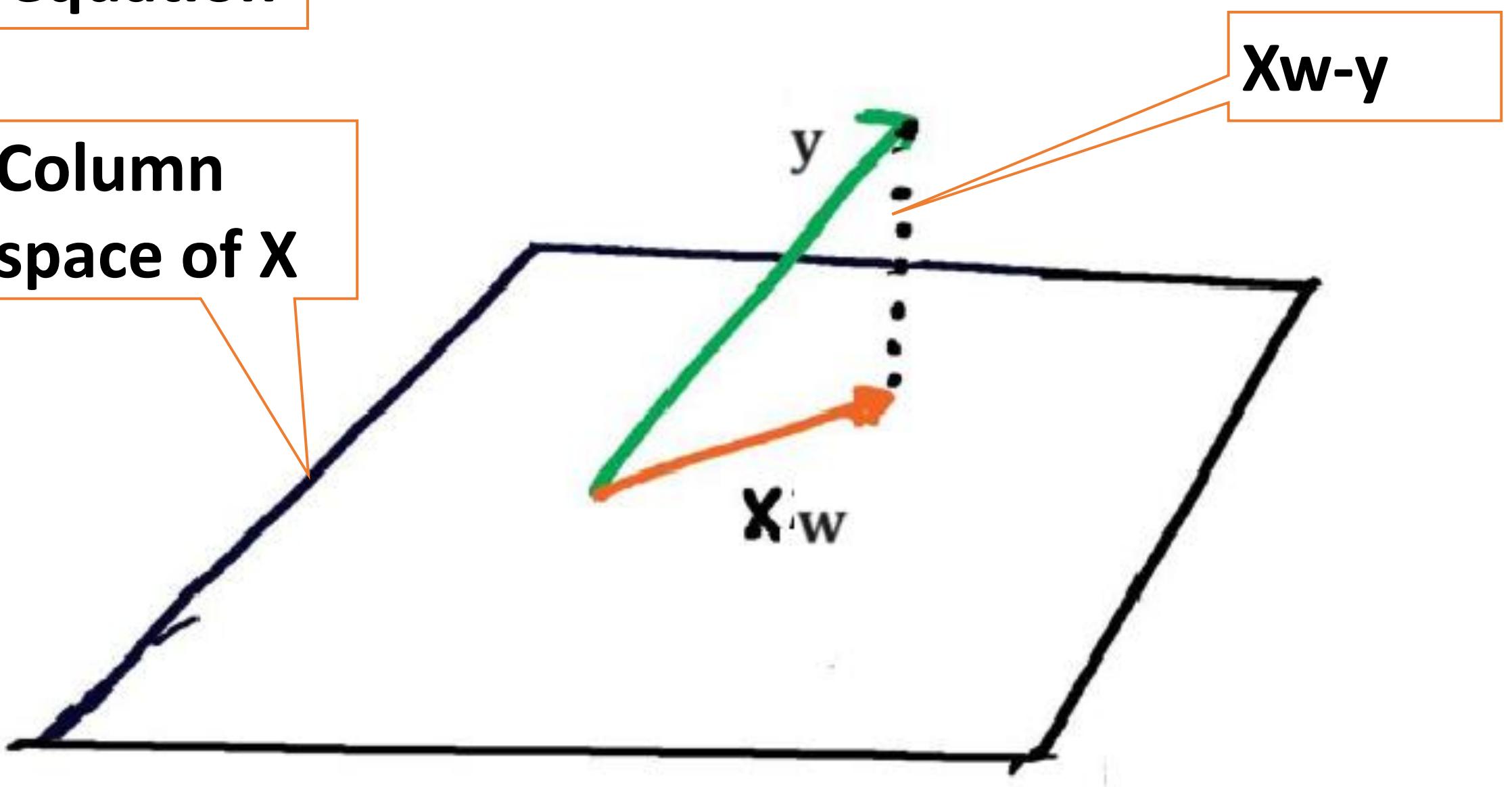
Left Inverse (Moore Penrose Pseudo Inverse)

Normal equation from Linear Algebra

Normal
equation

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

Column
space of X

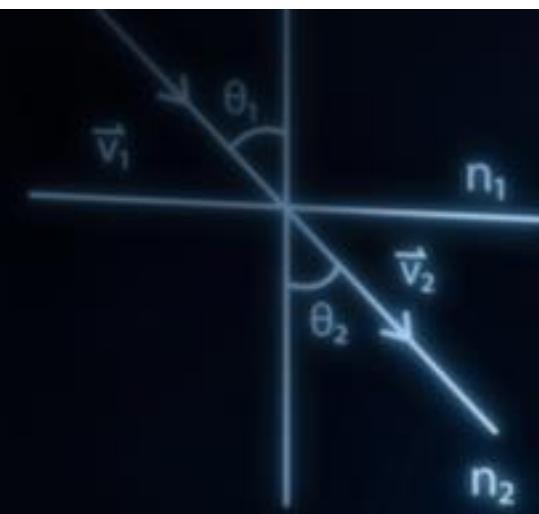


$$X^T(Xw - y) = 0$$

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

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

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



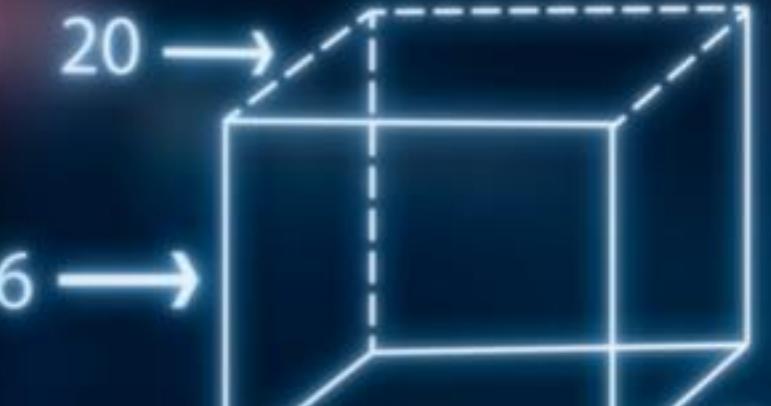
$$\bar{x}_1 = \frac{1+3+3+6+8+9}{6} = 5$$

$$\bar{x}_2 = \frac{2+4+4+8+12}{6} = 30$$

$$\bar{x}_3 = \frac{4+7+1+6}{6} = 18$$

$$\log_b b^x = x$$

$$\log_a x = \frac{\log_b x}{\log_b a}$$



$$\frac{a}{bc} = \frac{a}{bc}$$

$$\frac{a}{b} = \frac{ac}{b}$$

$$\frac{a}{b} + \frac{c}{d} = \frac{ad+bc}{bd}$$

$$X^2 - 4X + 5 \leq 5$$

$$X^2 - 4X \leq 0$$

$$n(B \cap C) = 22$$

$$n(B) = 68$$

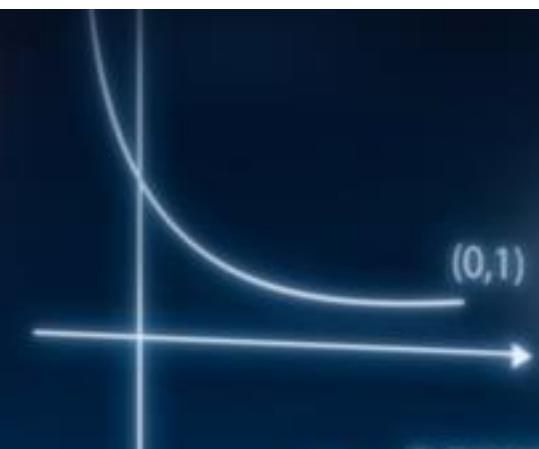
$$n(C) = 84$$

$$n(B \cup C) = n(B) + n(C) - n(B \cap C)$$

$$He = 4.002602$$

$$Na = 22.989769$$

$$Ar = 39.948$$



$$M = 0.046765$$

30L



Interpreting Linear Regression Coefficients

Interpreting Linear Regression Coefficients

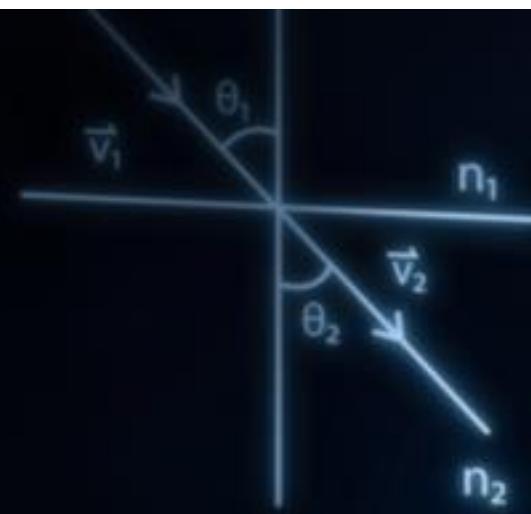
$$\hat{y} = w_0 + w_1 x_1 + w_2 x_2$$

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

Example with Boston house prices

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

True for numeric variables



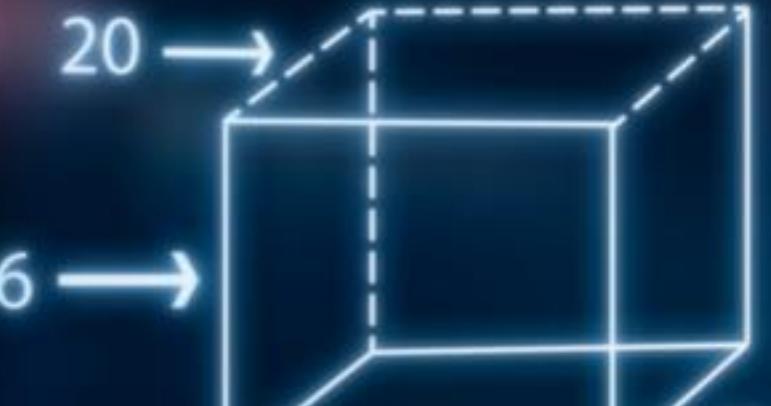
$$\bar{x}_1 = \frac{1+3+3+6+8+9}{6} = 5$$

$$\bar{x}_2 = \frac{2+4+4+8+12}{6} = 30$$

$$\bar{x}_3 = \frac{4+7+1+6}{6} = 18$$

$$\log_b b^x = x$$

$$\log_a x = \frac{\log_b x}{\log_b a}$$



$$\frac{a}{b} = \frac{a}{bc}$$

$$\frac{a}{b} = \frac{ac}{b}$$

$$\frac{a}{b} + \frac{c}{d} = \frac{ad+bc}{bd}$$

$$X^2 - 4X + 5 \leq 5$$

$$X^2 - 4X \leq 0$$

$$n(B \cap C) = 22$$

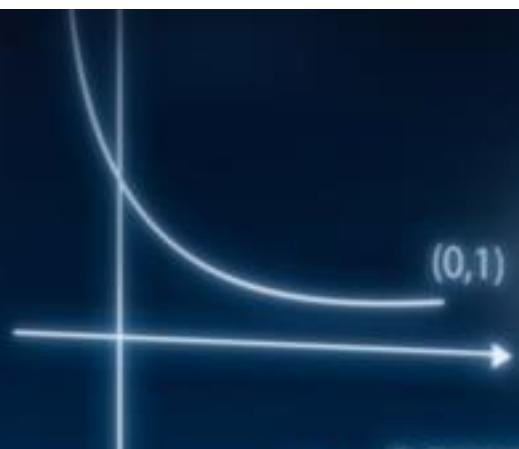
$$n(B) = 68$$

$$n(B \cup C) = n(B) + n(C) - n(B \cap C)$$

$$He = 4.002602$$

$$Na = 22.989769$$

$$Ar = 39.948$$



$$M = \frac{0.046765}{30L}$$

30L

Word Anatomy of Multicollinearity

Multi-col-linear-ity

Referring to the multiple independent variables within multiple regression.

A modification of the prefix co, meaning together or joint. Referencing the linear movement in tandem i.e., correlation.

Occurring within a linear equation.

Suffix meaning the quality or state of.

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

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

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

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

How to find multicollinearity

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

No Inverse

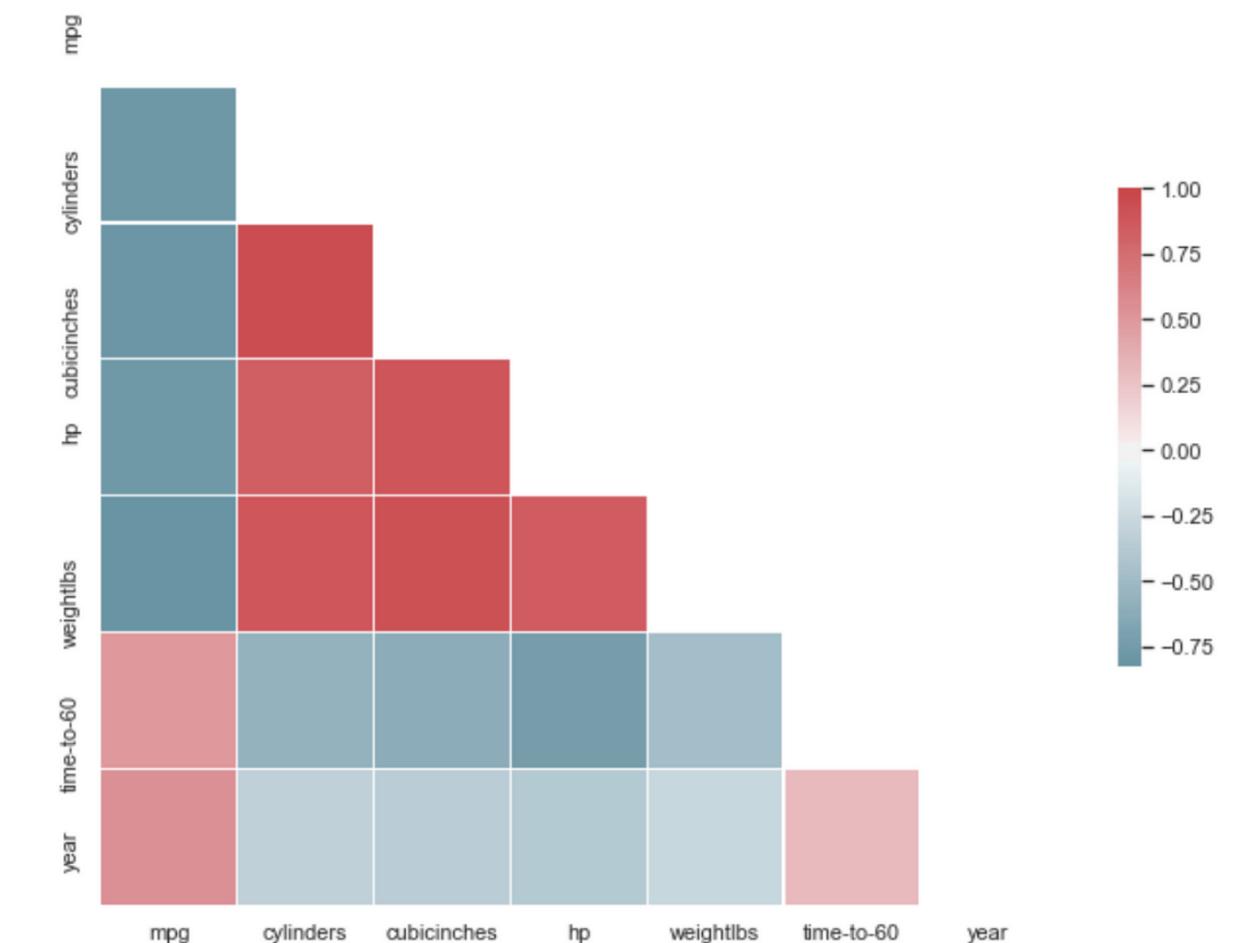
$$\begin{bmatrix} 2 & 33 & 4 \\ 3 & 39 & 6 \\ 1 & 41 & 2 \\ 4 & 46 & 8 \\ 6 & 44 & 12 \end{bmatrix}$$

$X^T X$
Correlation=1

Inverse exists with
determinant close to 0

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

$$\lambda_1 = 3, \lambda_2 = 0.9, \lambda_3 = 0 \quad \lambda_1 = 3, \lambda_2 = 0.9, \lambda_3 = 10^{-4}$$

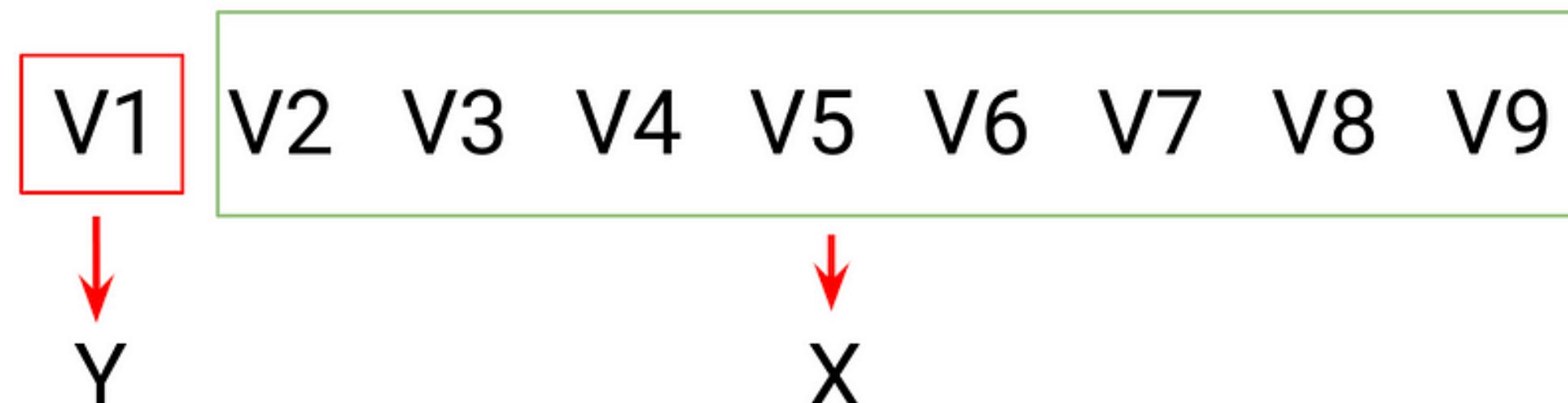


Variance Inflation Factor (VIF)

- Given features (variables)

V1 V2 V3 V4 V5 V6 V7 V8 V9

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



- Calculate R^2 . Good R^2 = good fit. Hence bad predictor

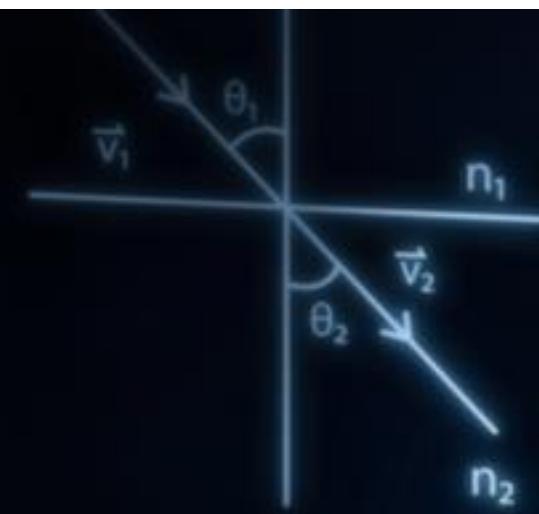
$$R^2 = 1 - \frac{\text{MSE(model)}}{\text{MSE(baseline)}}$$

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

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.



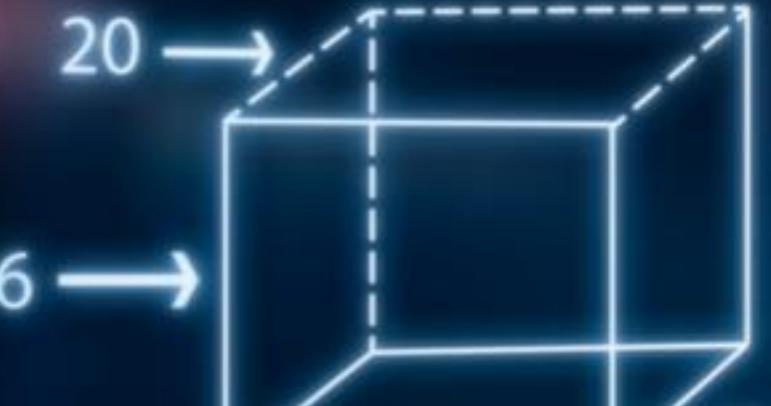
$$\bar{x}_1 = \frac{1+3+3+6+8+9}{6} = 5$$

$$\bar{x}_2 = 2+4+4+8+12 = 30$$

$$\bar{x}_3 = 4+7+1+6 = 18$$

$$\log_b b^x = x$$

$$\log_a x = \frac{\log_b x}{\log_b a}$$



$$\begin{aligned} \frac{a}{b} &= \frac{a}{bc} & X^2 - 4X + 5 \leq 5 \\ \frac{a}{b} &= \frac{ac}{b} & X^2 - 4X \leq 0 \\ \frac{a}{b} + \frac{c}{d} &= \frac{ad+bc}{bd} & n(B \cap C) = 22 \\ & & n(B) = 68 \\ & & n(C) = 84 \end{aligned}$$

$$n(B \cup C) = n(B) + n(C) - n(B \cap C)$$

$$He = 4.002602$$

$$Na = 22.989769$$

$$Ar = 39.948$$

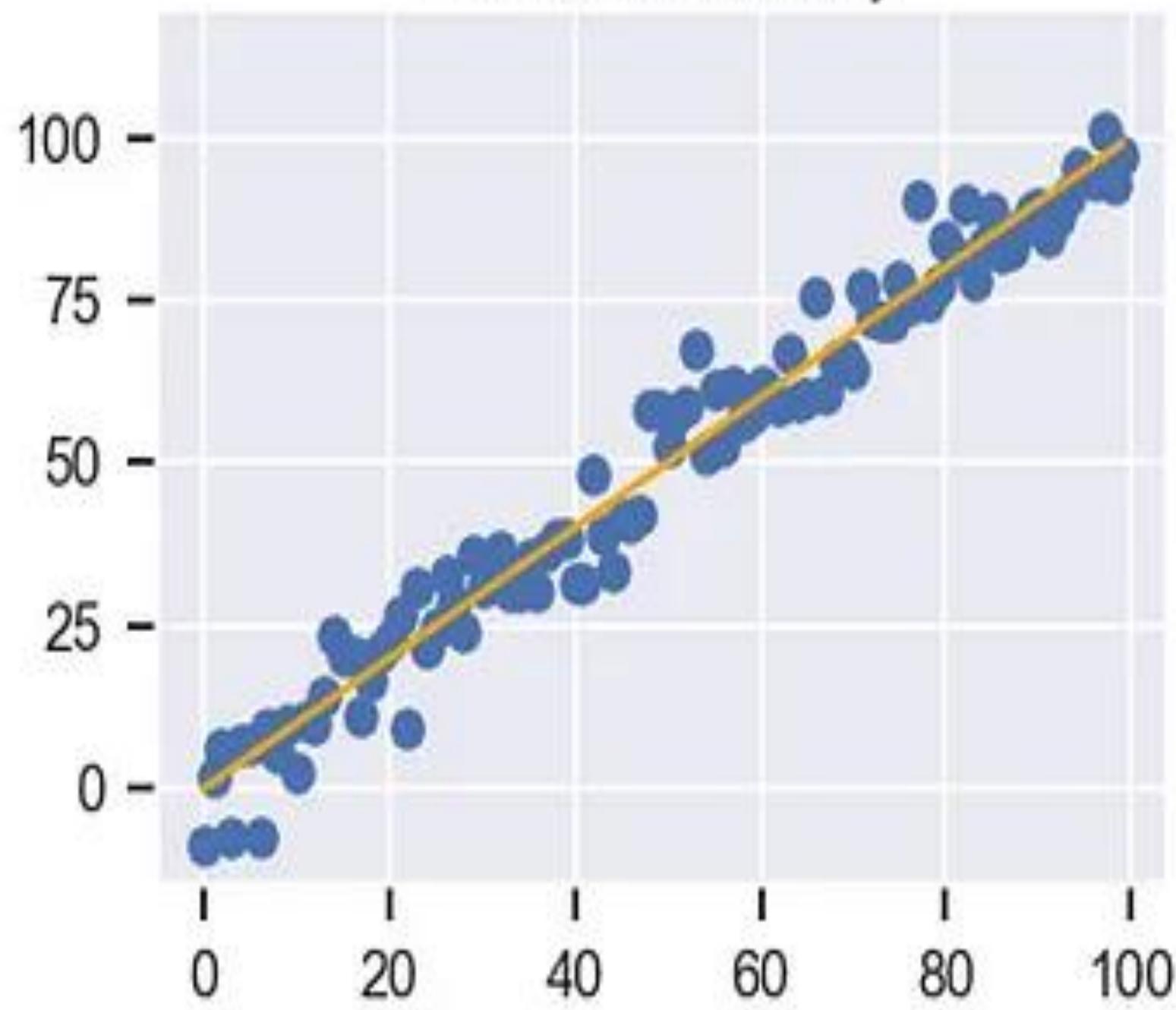


$$M = \frac{0.046765}{30L}$$

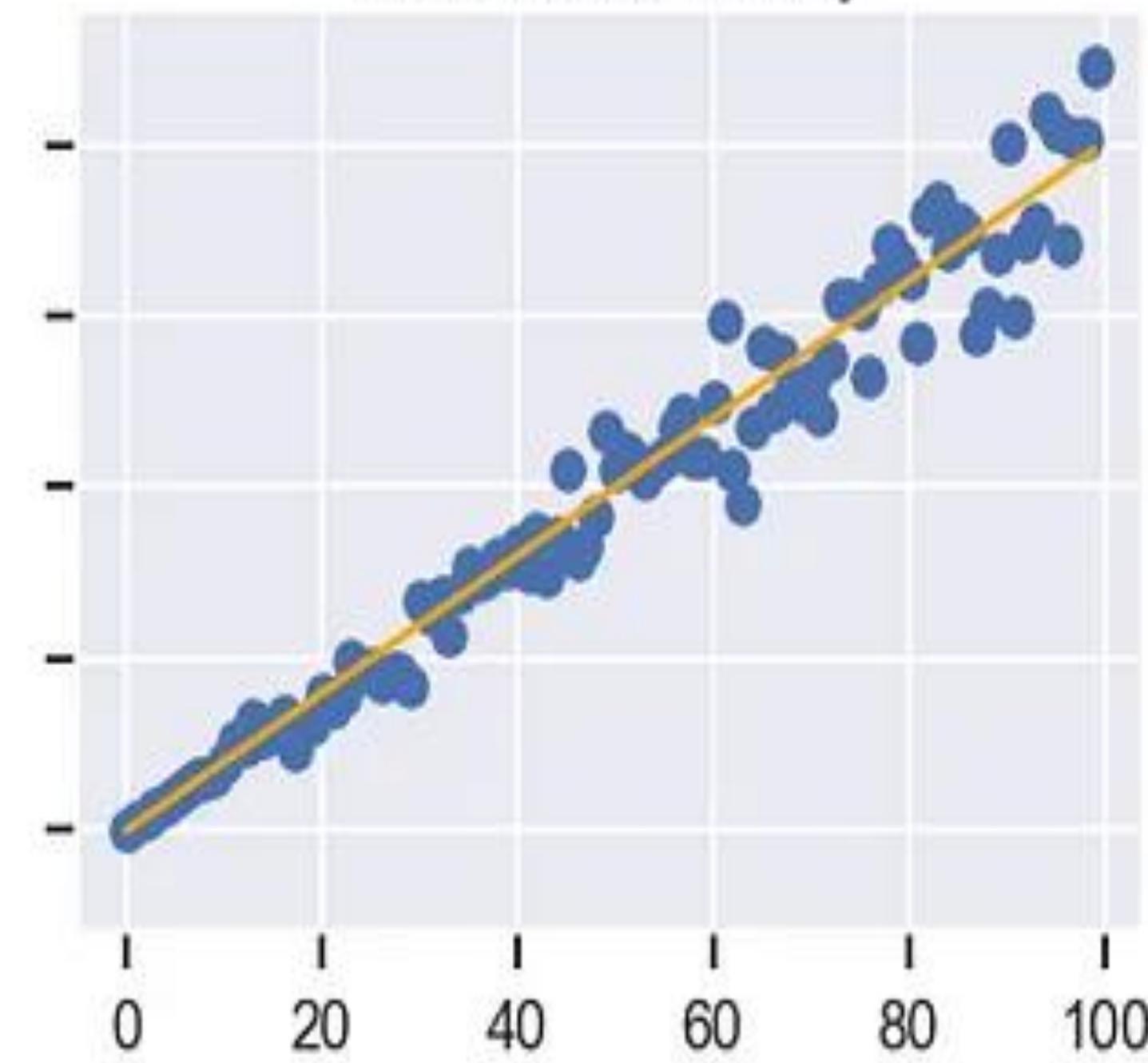


Heteroskedasticity

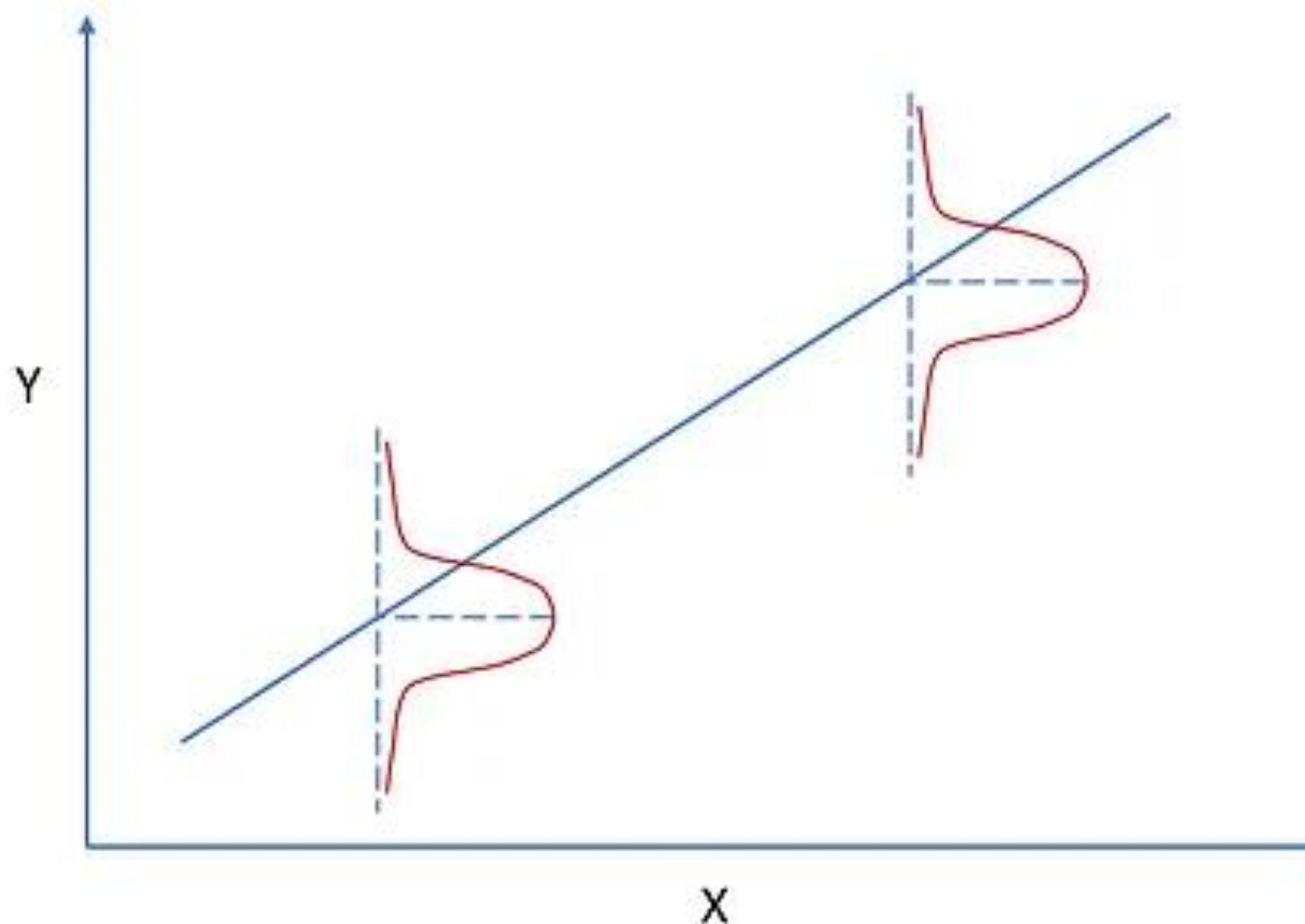
Homoskedasticity



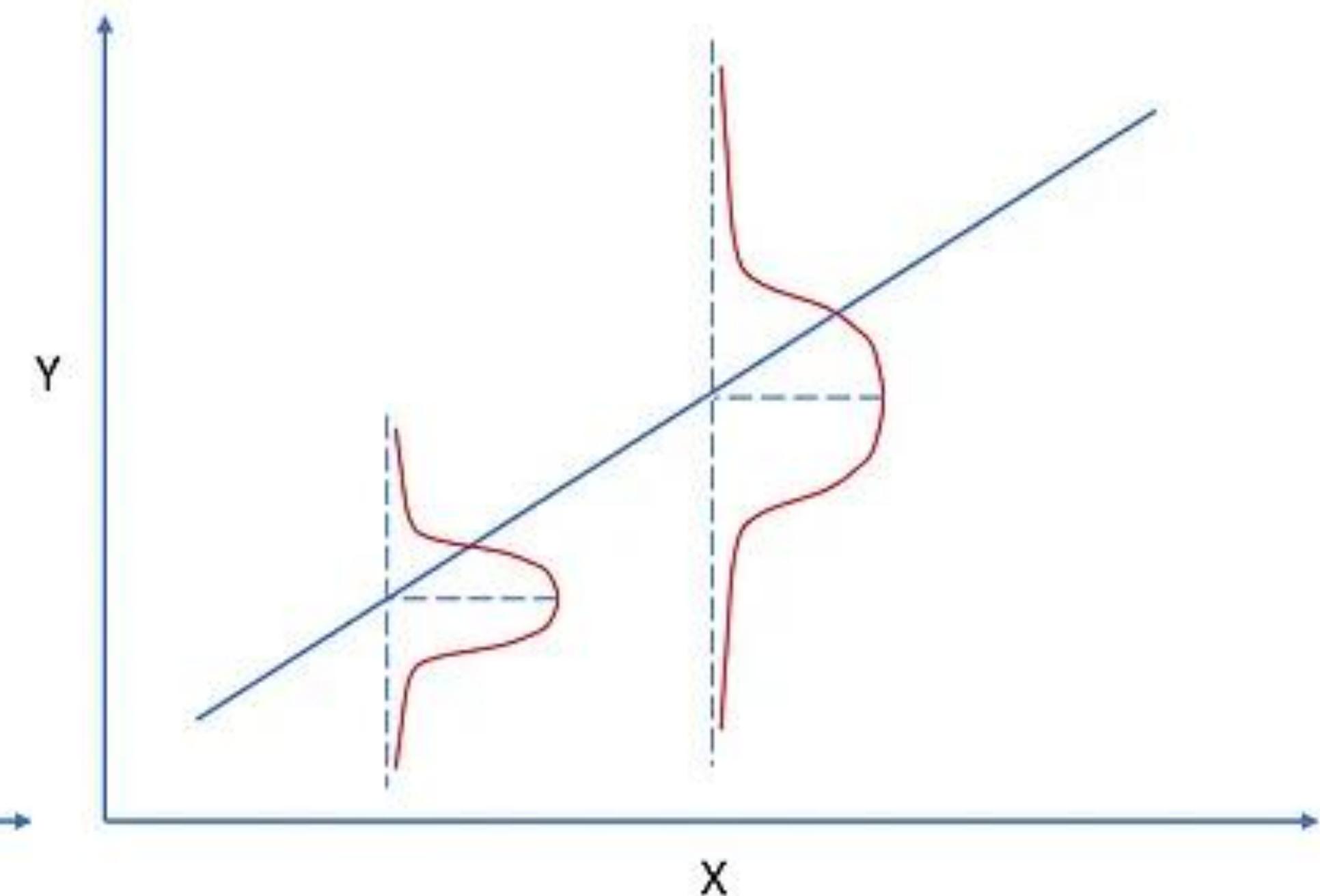
Heteroskedasticity



Homoskedasticity

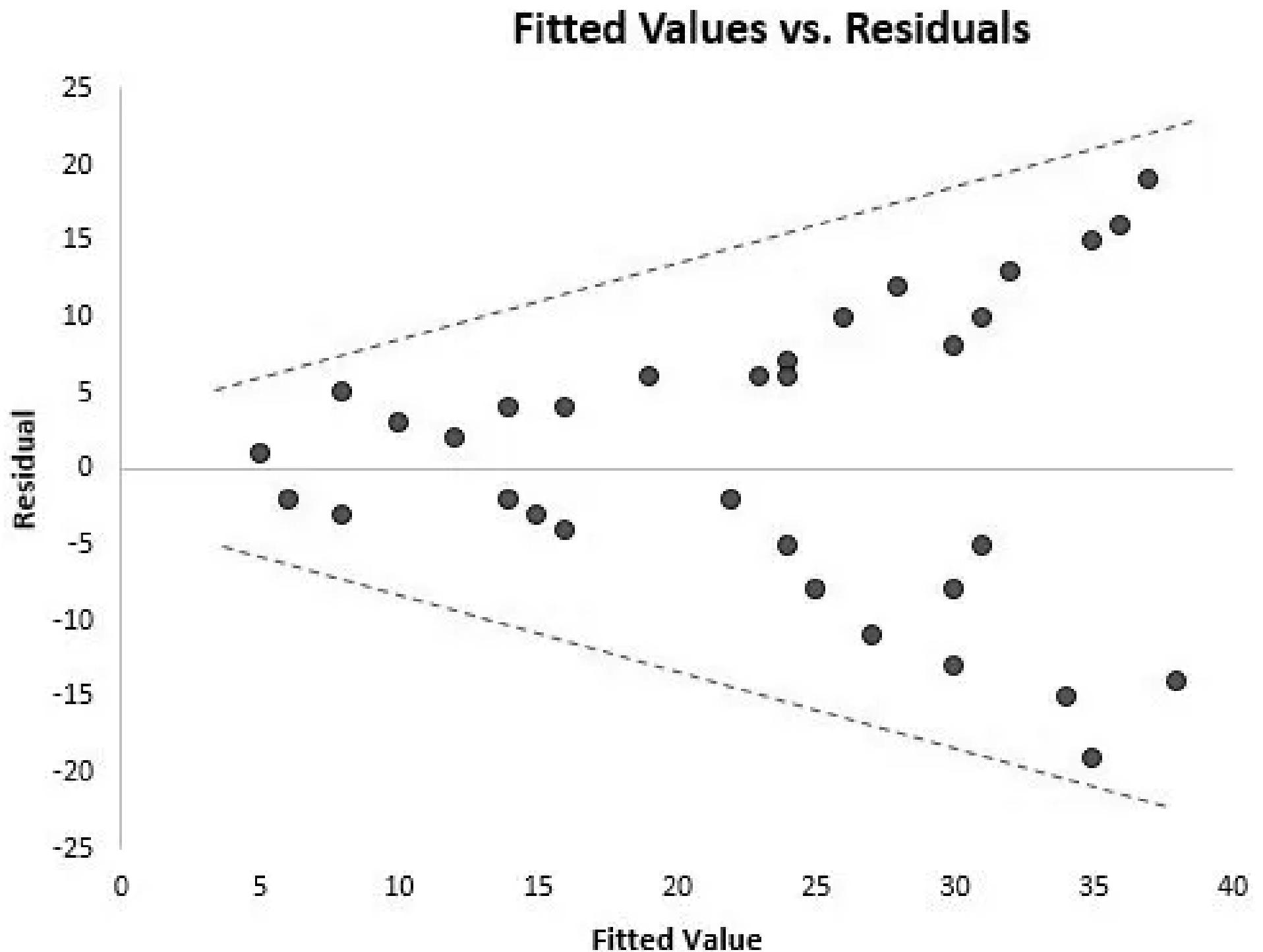


Heteroskedasticity



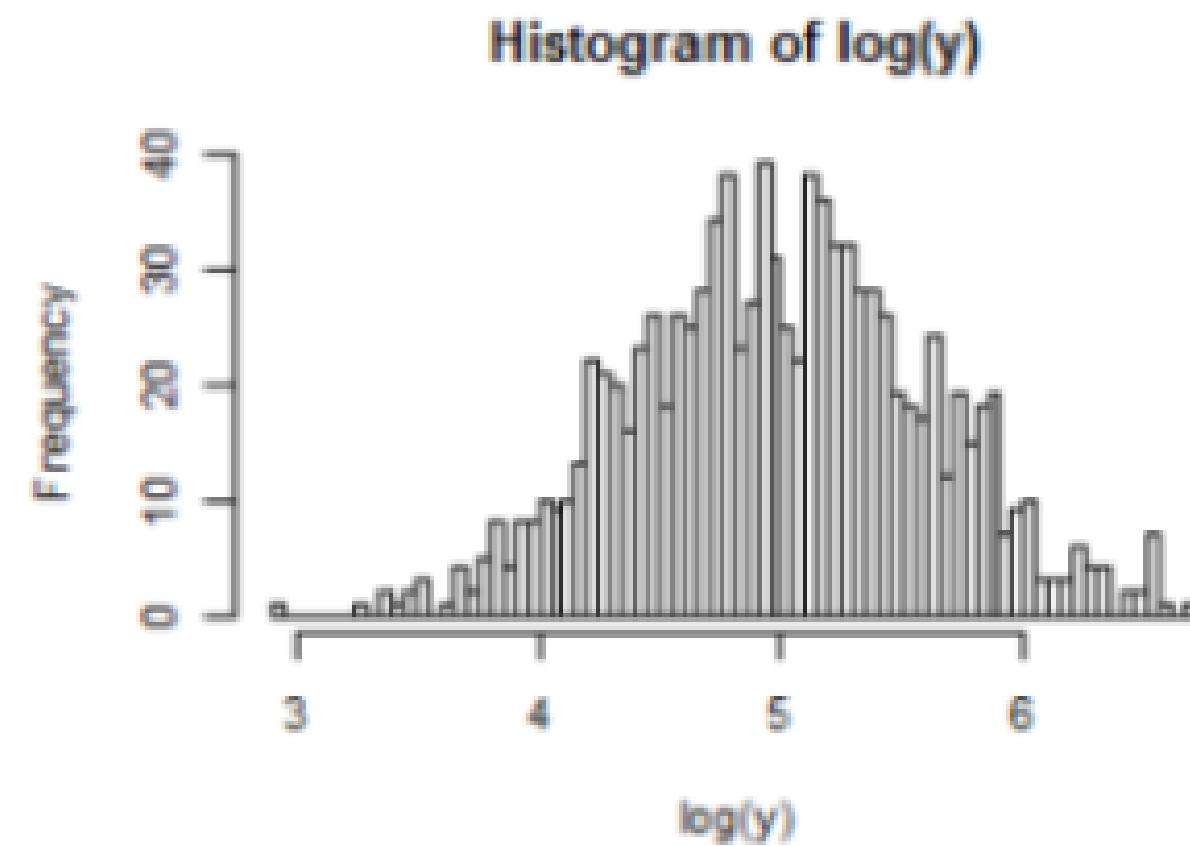
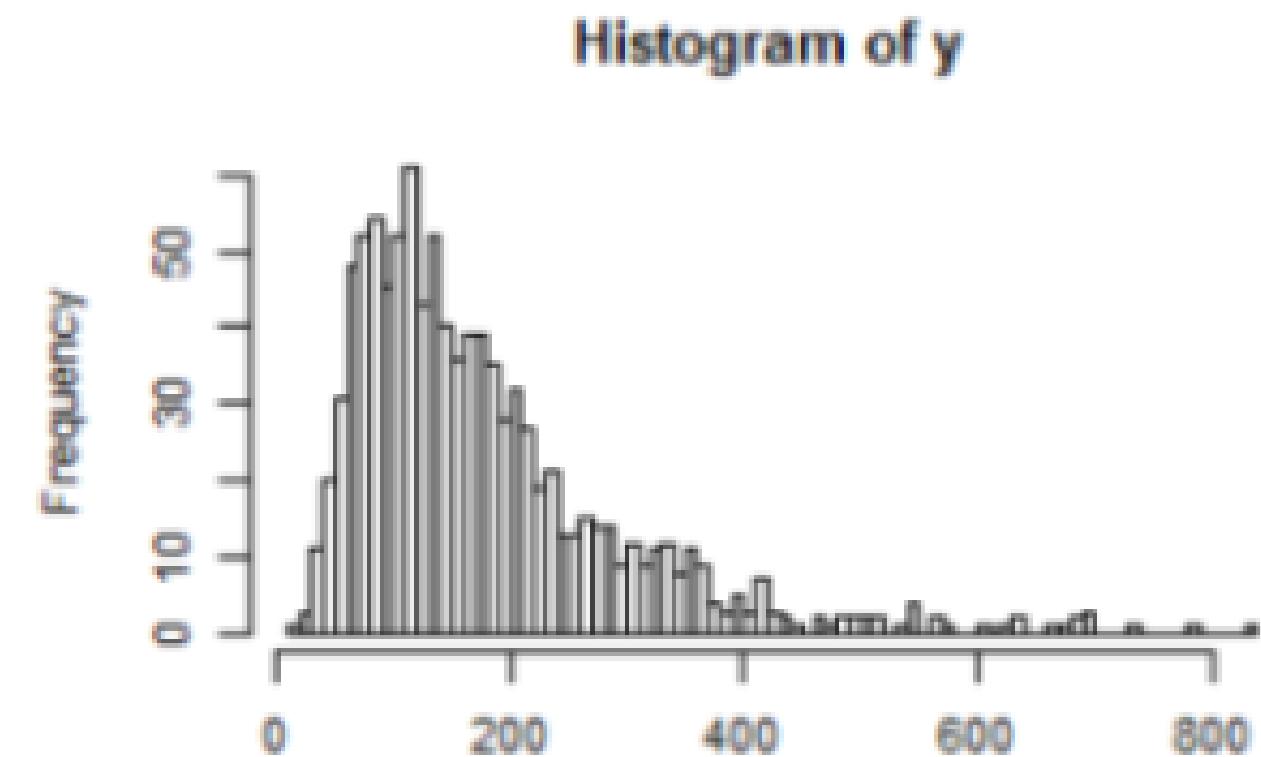
How to find

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



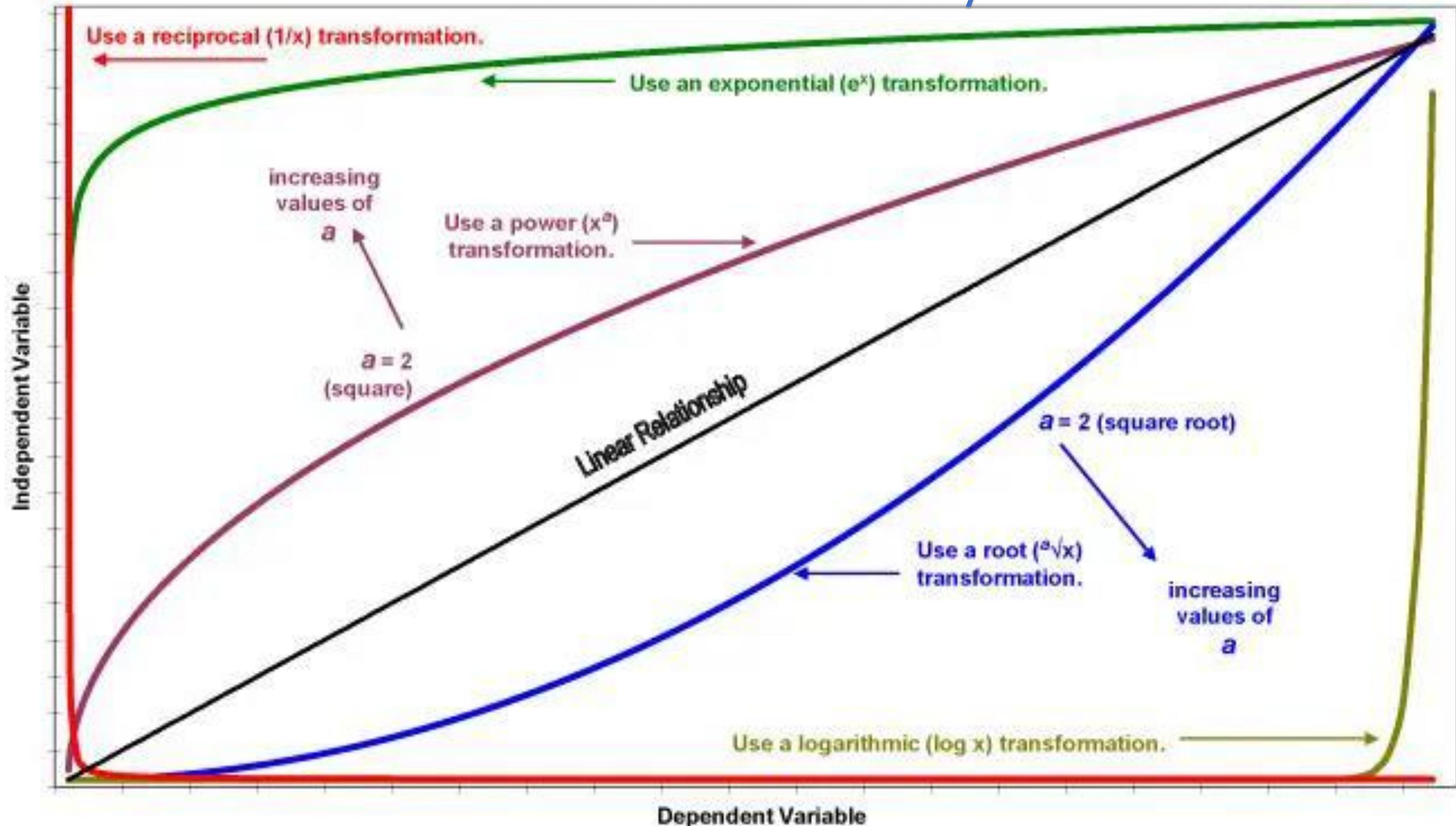
How to fix

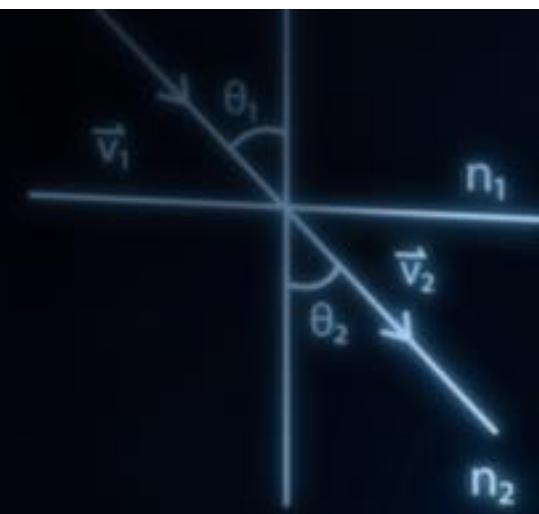
- Fit to $\log y$ instead of y



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

Transformations on y





$$\bar{x}_1 = \frac{1+3+3+6+8+9}{6} = 5$$

$$\bar{x}_2 = \frac{2+4+4+8+12}{6} = 30$$

$$\bar{x}_3 = \frac{4+7+1+6}{6} = 18$$

$$\log_b b^x = x$$

$$\log_a x = \frac{\log_b x}{\log_b a}$$



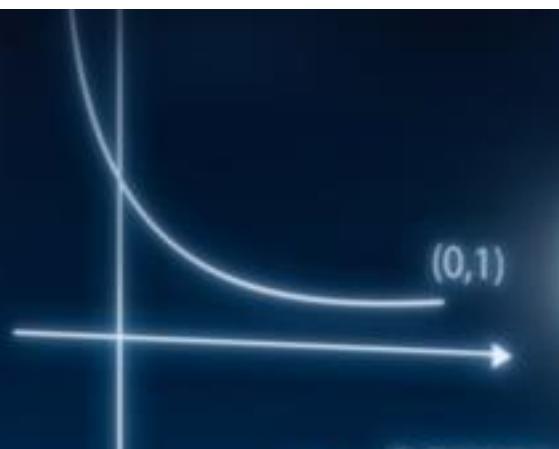
$$\begin{aligned} \frac{a}{b} &= \frac{a}{bc} & X^2 - 4X + 5 &\leq 5 \\ \frac{a}{b} &= \frac{ac}{b} & X^2 - 4X &\leq 0 \\ \frac{a}{b} + \frac{c}{d} &= \frac{ad+bc}{bd} & n(B \cap C) &= 22 \\ & & n(B) &= 68 \\ & & n(C) &= 84 \end{aligned}$$

$$n(B \cup C) = n(B) + n(C) - n(B \cap C)$$

$$He = 4.002602$$

$$Na = 22.989769$$

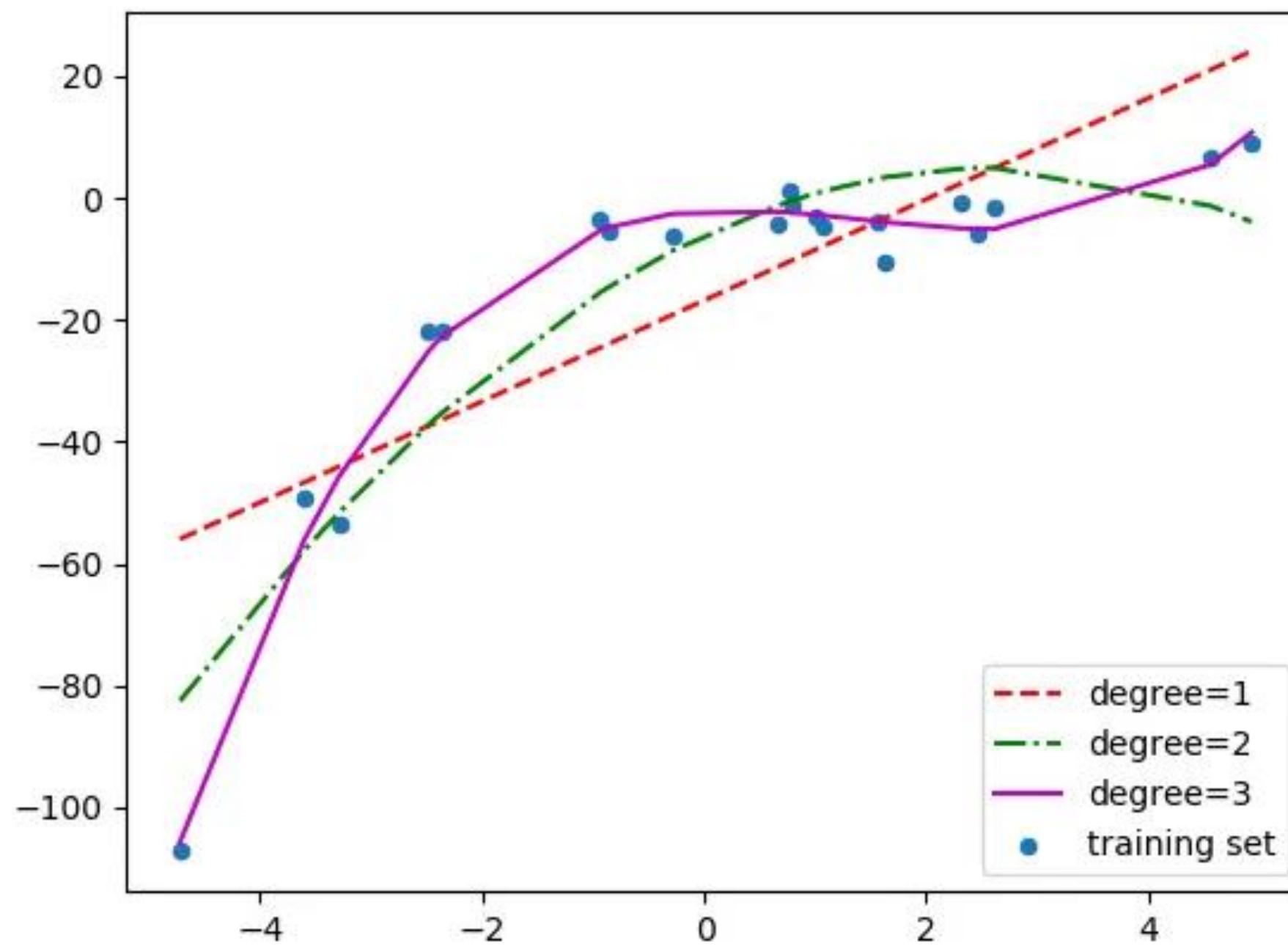
$$Ar = 39.948$$



$$M = \frac{0.046765}{30L}$$

30L

Polynomial Regression



$$y = \mathbf{w}^T \mathbf{x} + b$$
$$y = w_1 x + w_2 x^2 + w_3 x^3 + b$$
$$y = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix}^T \begin{bmatrix} x \\ x^2 \\ x^3 \end{bmatrix} + b$$

Features

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

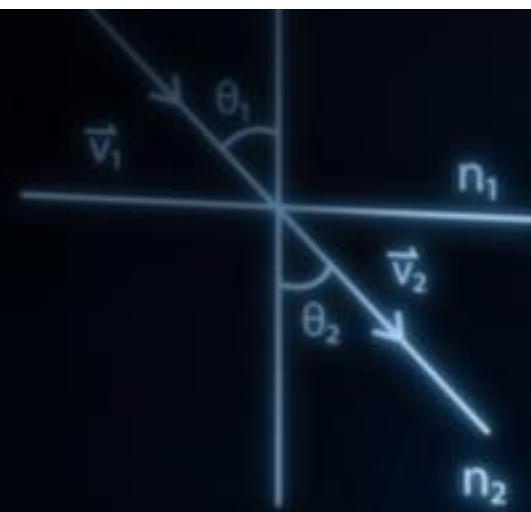
- Feature Cross grows exponentially
- `sklearn.preprocessing.PolynomialFeatures`

```
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.]])
```



$$\bar{x}_1 = \frac{1+3+3+6+8+9}{6} = 5$$

$$\bar{x}_2 = 2+4+4+8+12 = 30$$

$$\bar{x}_3 = 4+7+1+6 = 18$$

$$\log_b b^x = x$$

$$\log_a x = \frac{\log_b x}{\log_b a}$$



$$\frac{(a)}{(b)} = \frac{a}{bc}$$

$$\frac{a}{(b)} = \frac{ac}{b}$$

$$\frac{a}{b} + \frac{c}{d} = \frac{ad+bc}{bd}$$

$$X^2 - 4X + 5 \leq 5$$

$$X^2 - 4X \leq 0$$

$$n(B \cap C) = 22$$

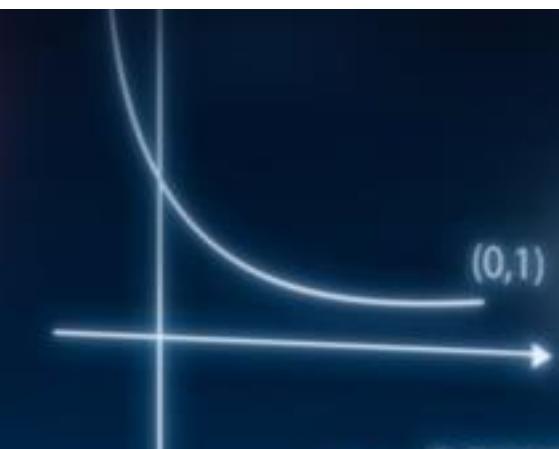
$$n(B) = 68$$

$$n(B \cup C) = n(B) + n(C) - n(B \cap C)$$

$$He = 4.002602$$

$$Na = 22.989769$$

$$Ar = 39.948$$



$$M = \frac{0.046765}{30L}$$



$$f = \frac{R}{2}$$

$$4 \frac{10}{15} \cdot 4 \frac{2}{5} + 5 \frac{1}{3} = \frac{(15 \times 4) + 10}{15}$$



B

C

$$= i \cdot a, a \geq 0$$

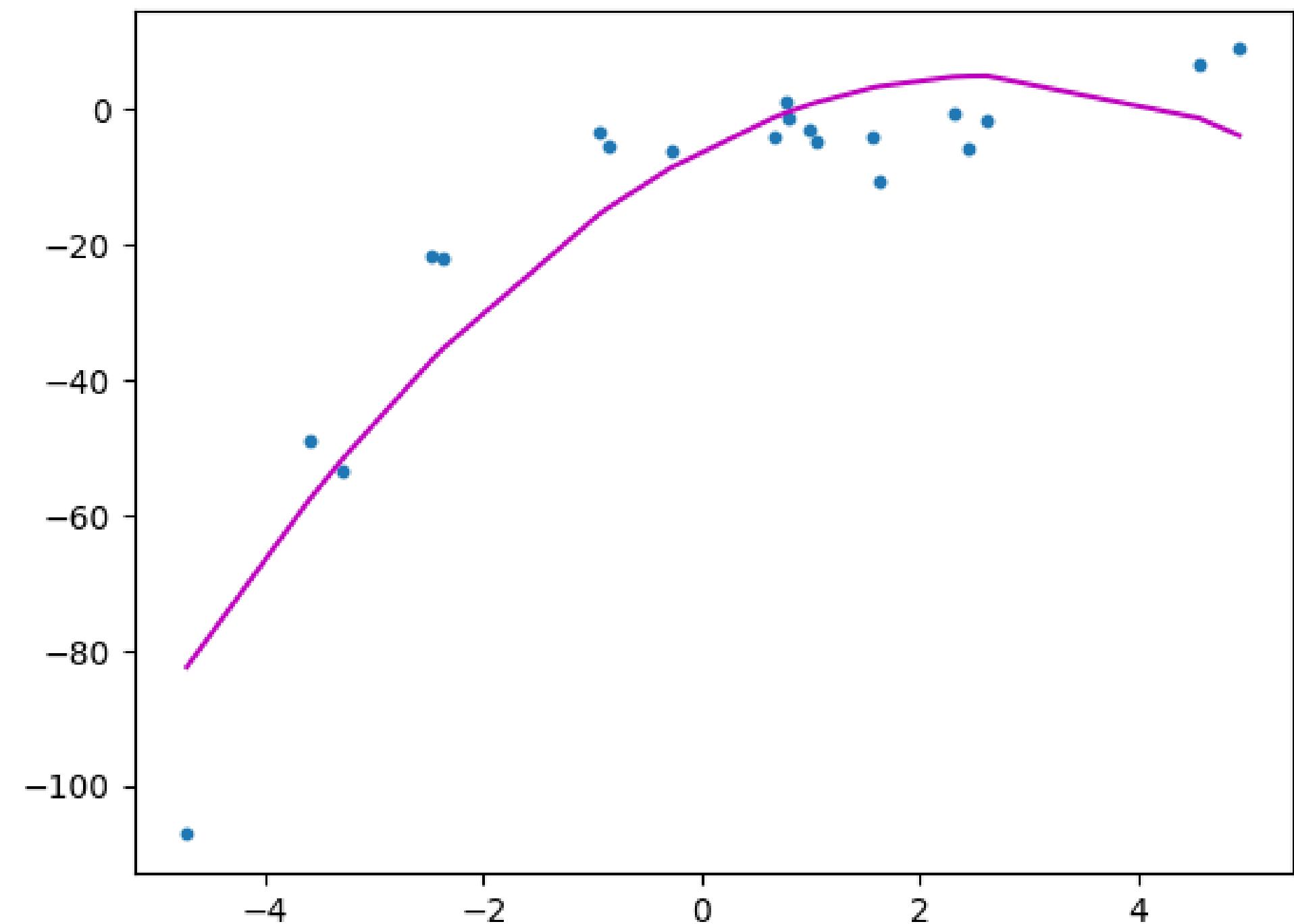
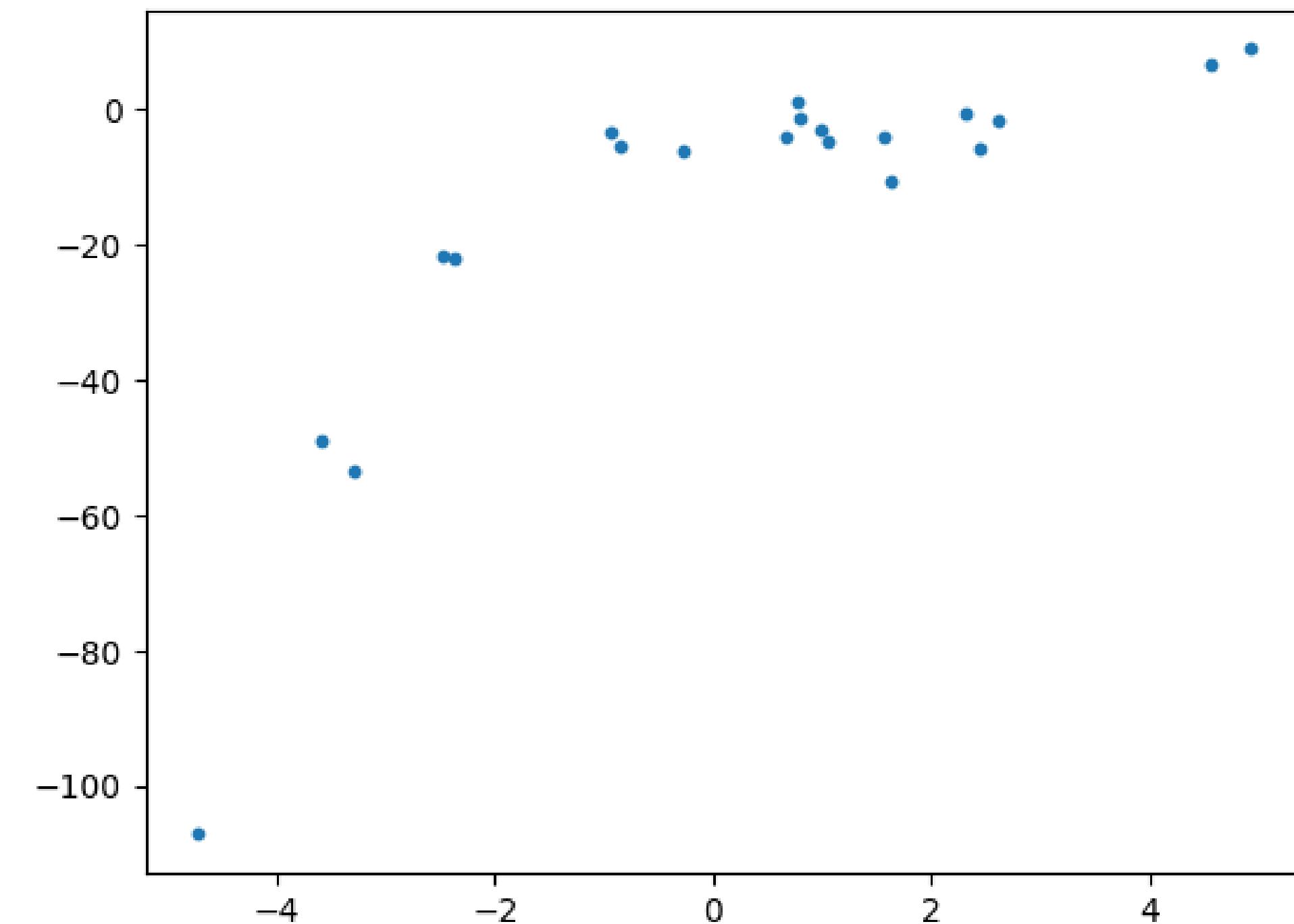
$$(bi)+(c+di) = a+c+(b+d)i$$

$$+bi)-(c+di) = a-c+(b-d)i$$

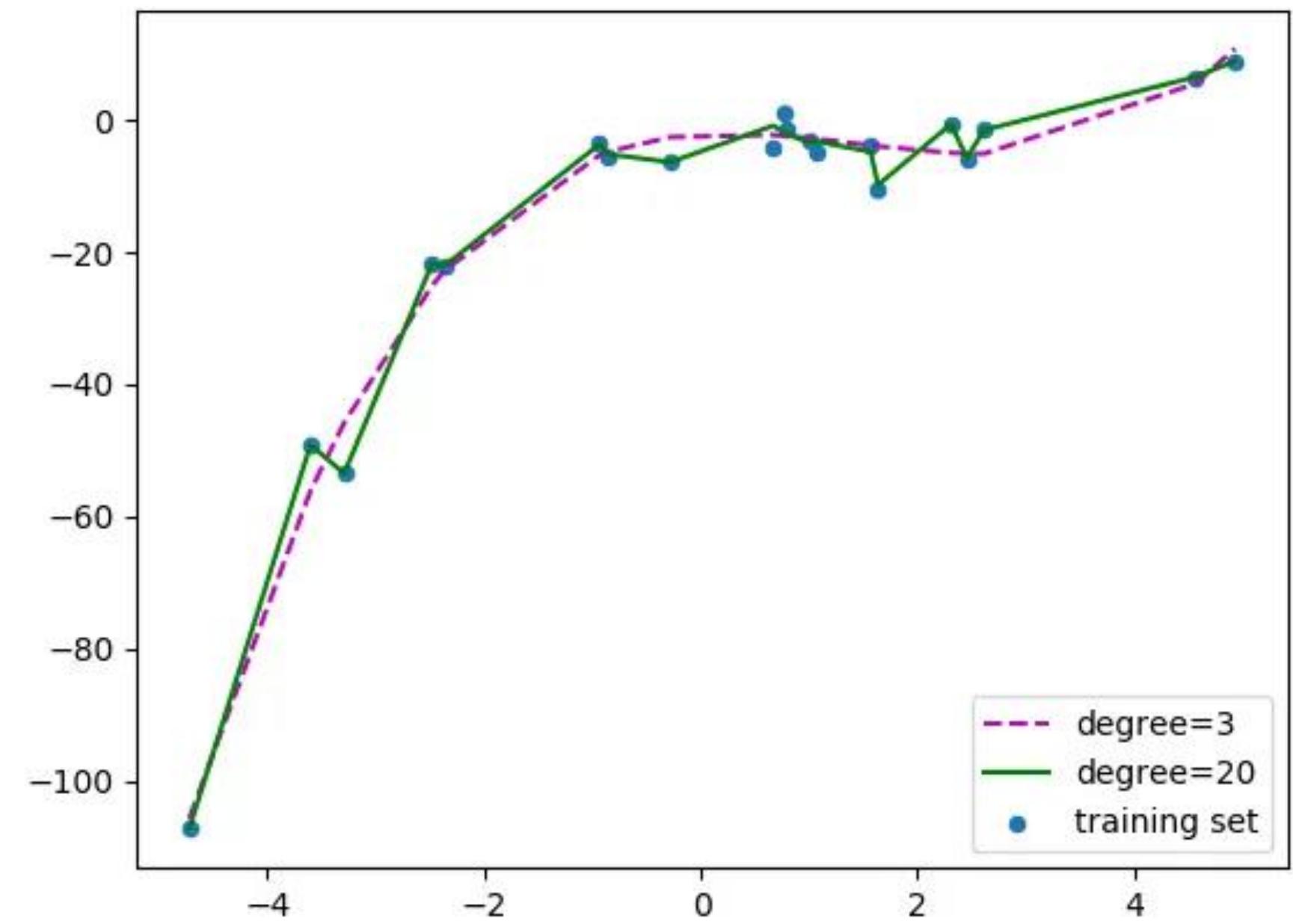
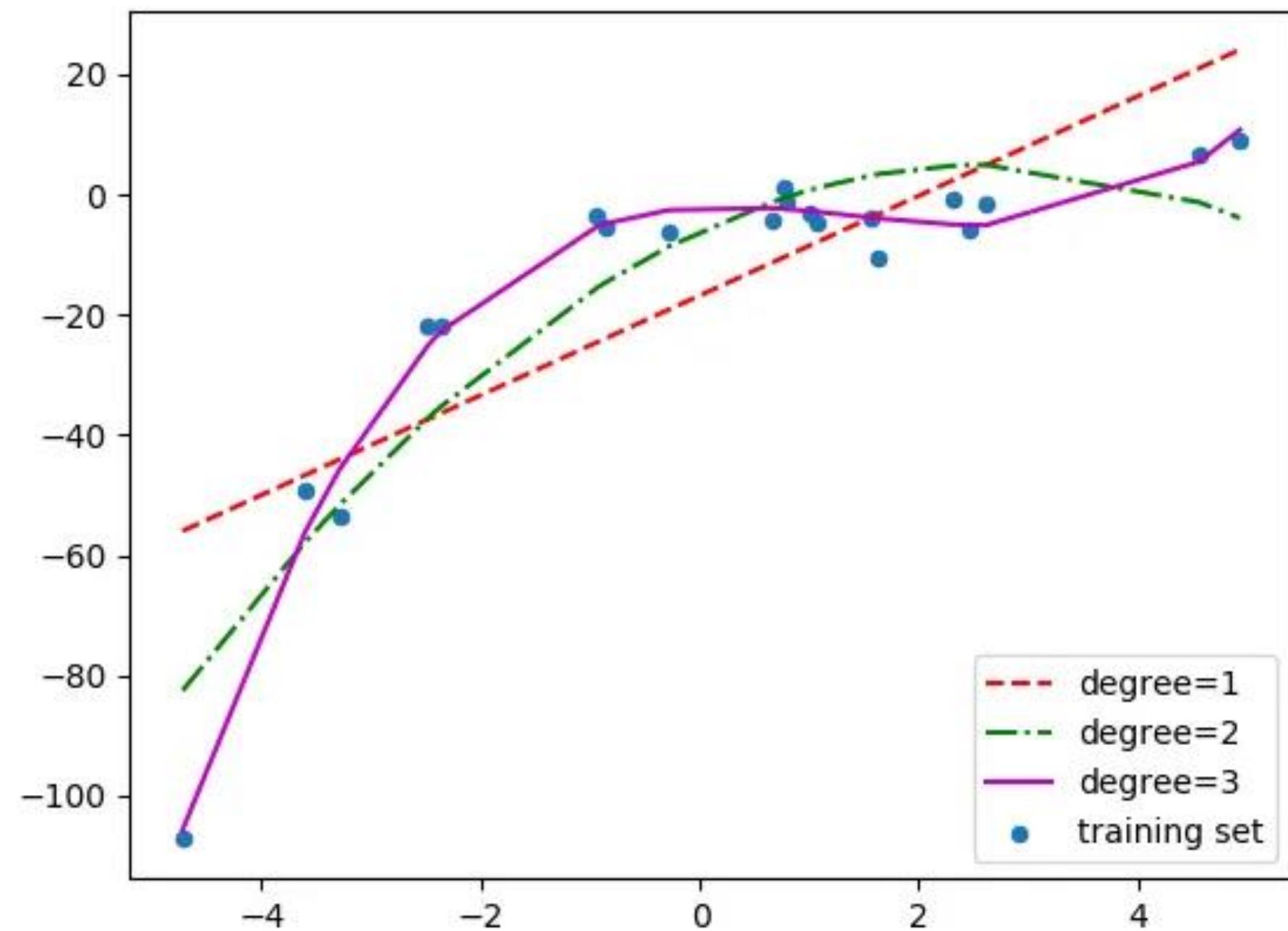
$$(a+bi)(c+di) = ac+bd+(ad+bc)i$$

Underfitting, Overfitting, Bias Variance trade off

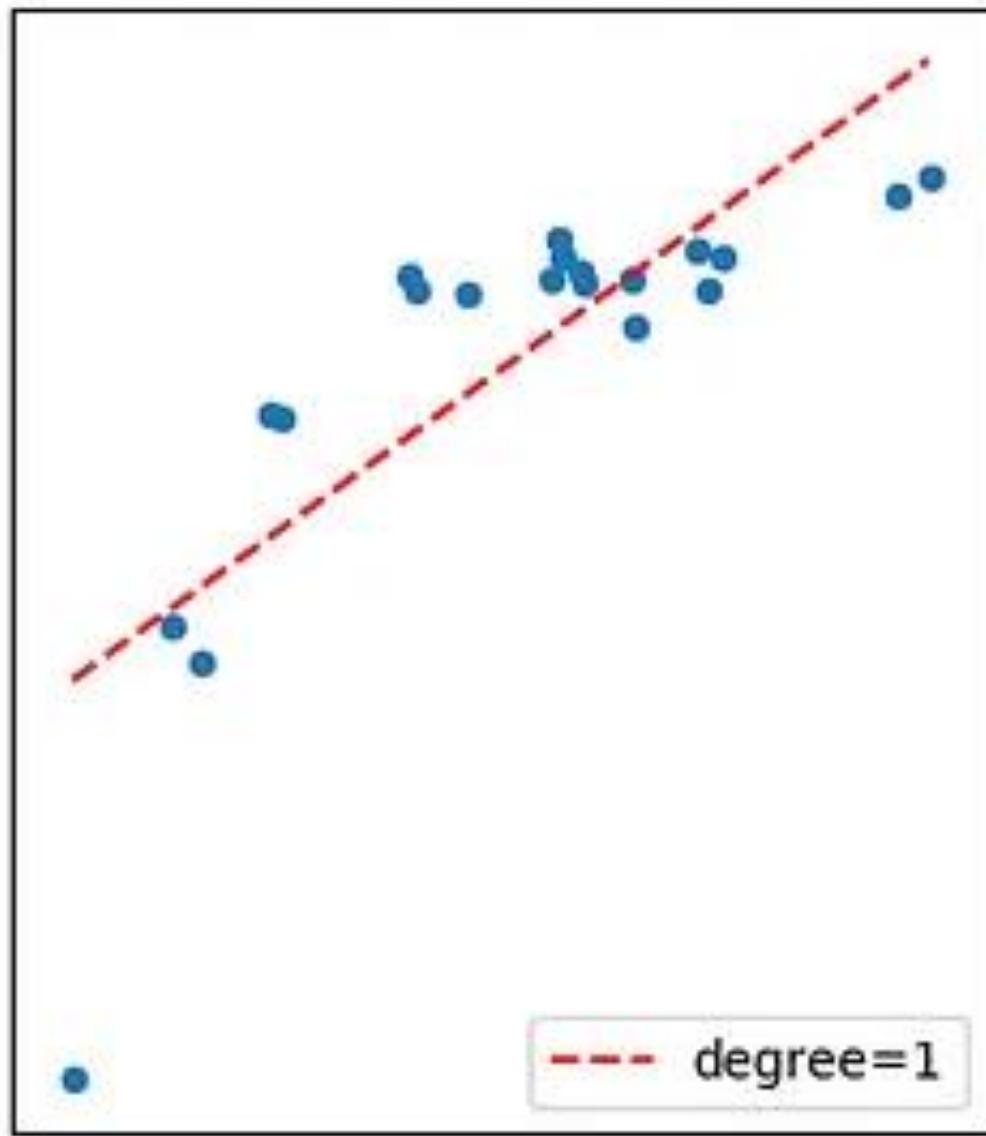
Fitting a curve



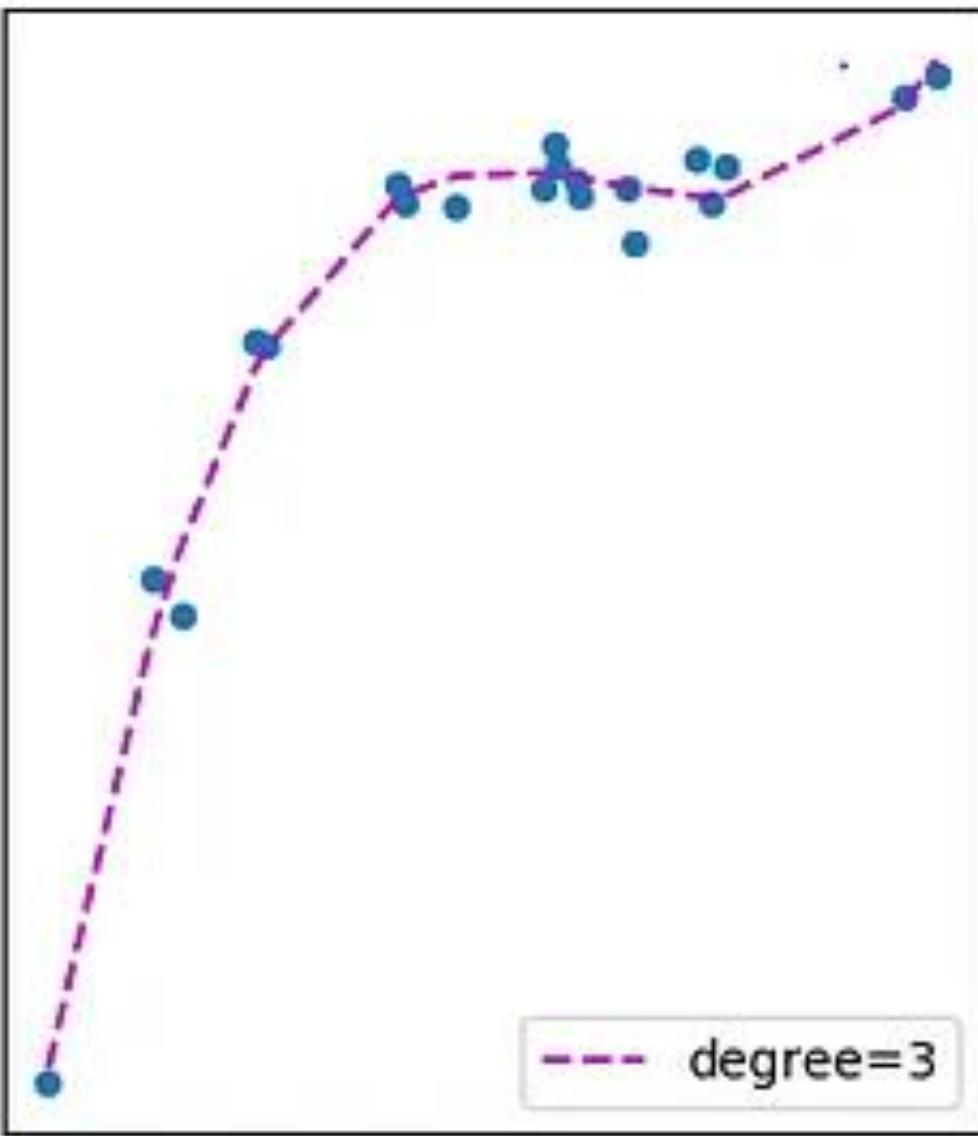
How many degrees is good?



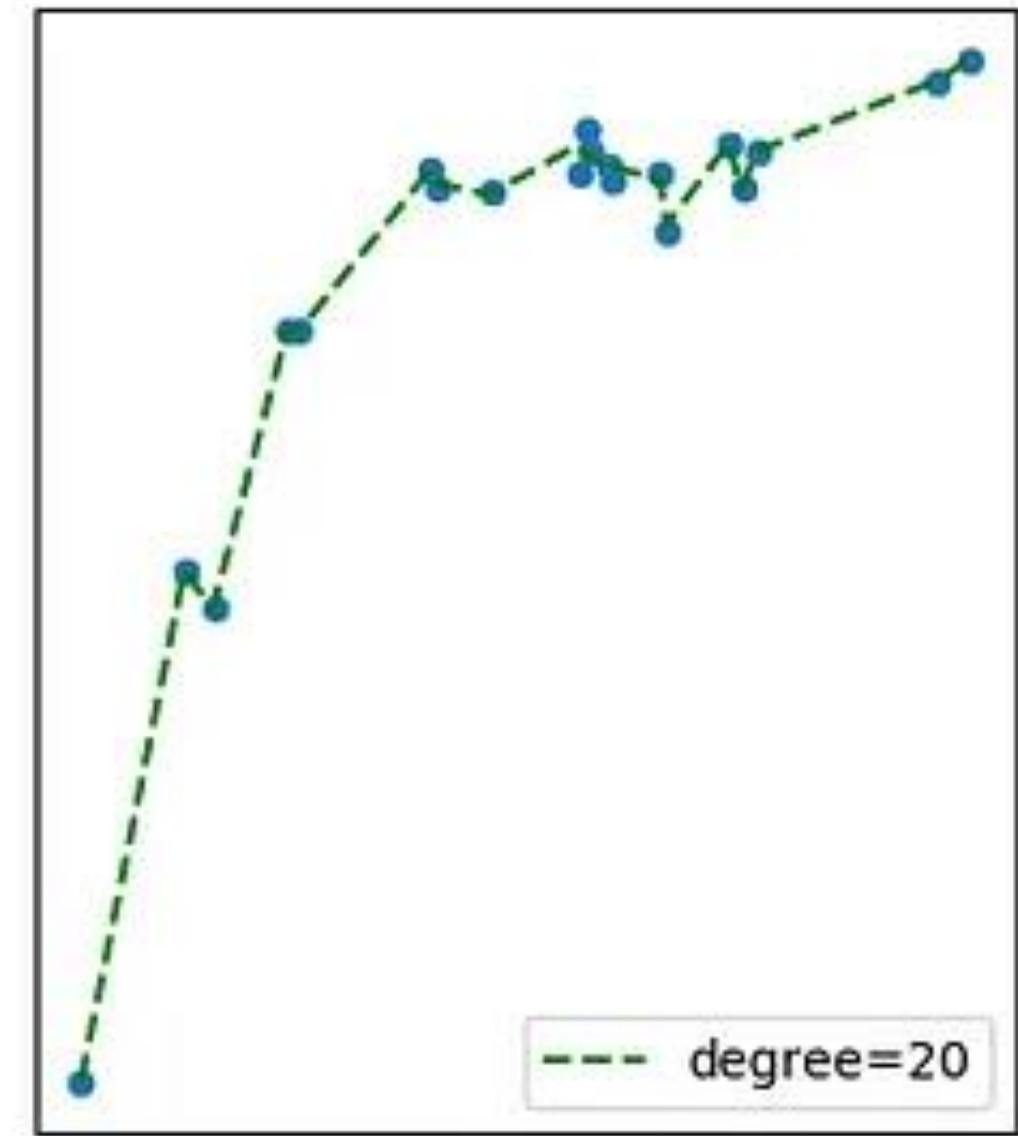
Underfitting & overfitting



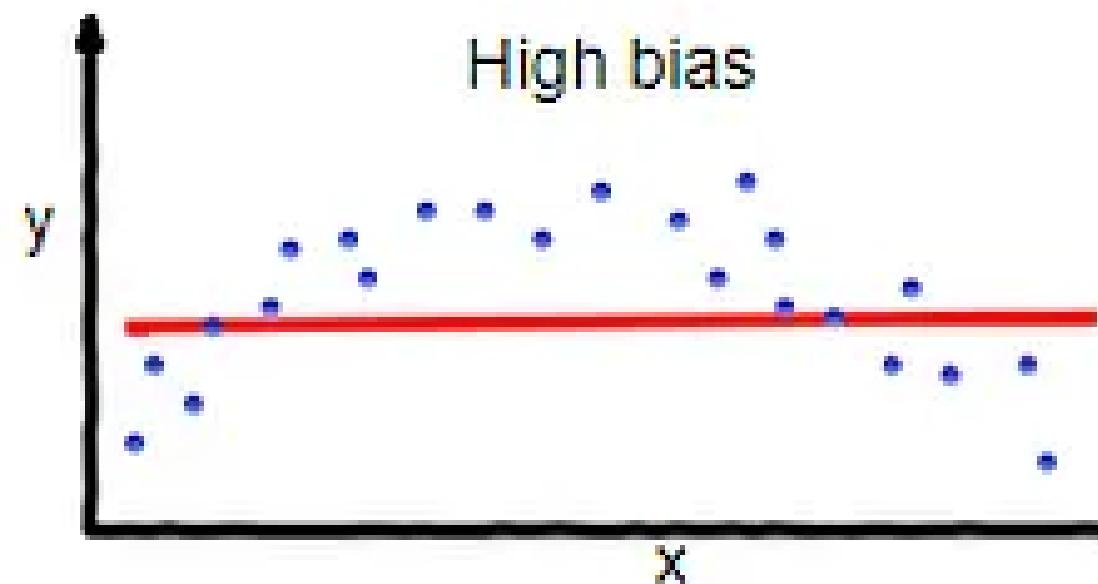
Underfit
High Bias
Low Variance



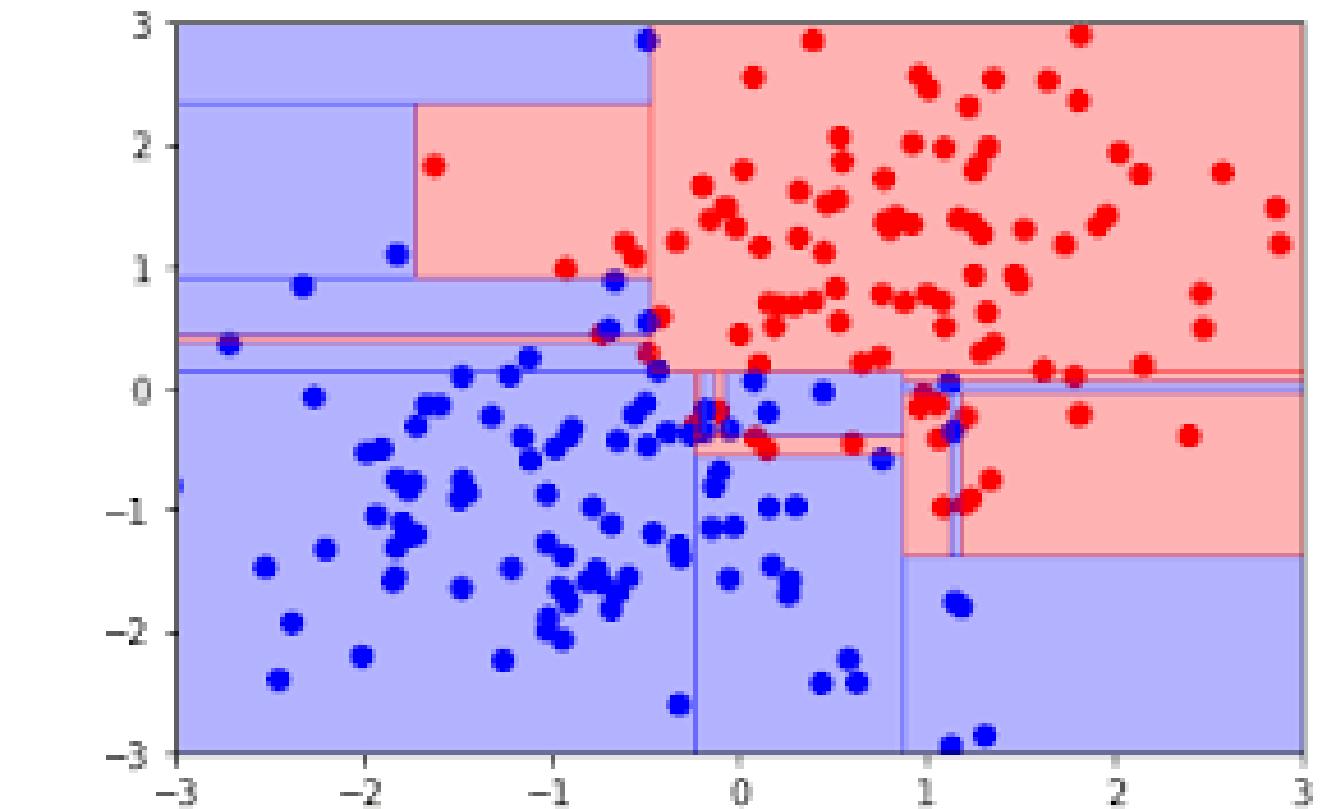
Correct Fit
Low Bias
Low Variance



Overfit
Low Bias
High Variance



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



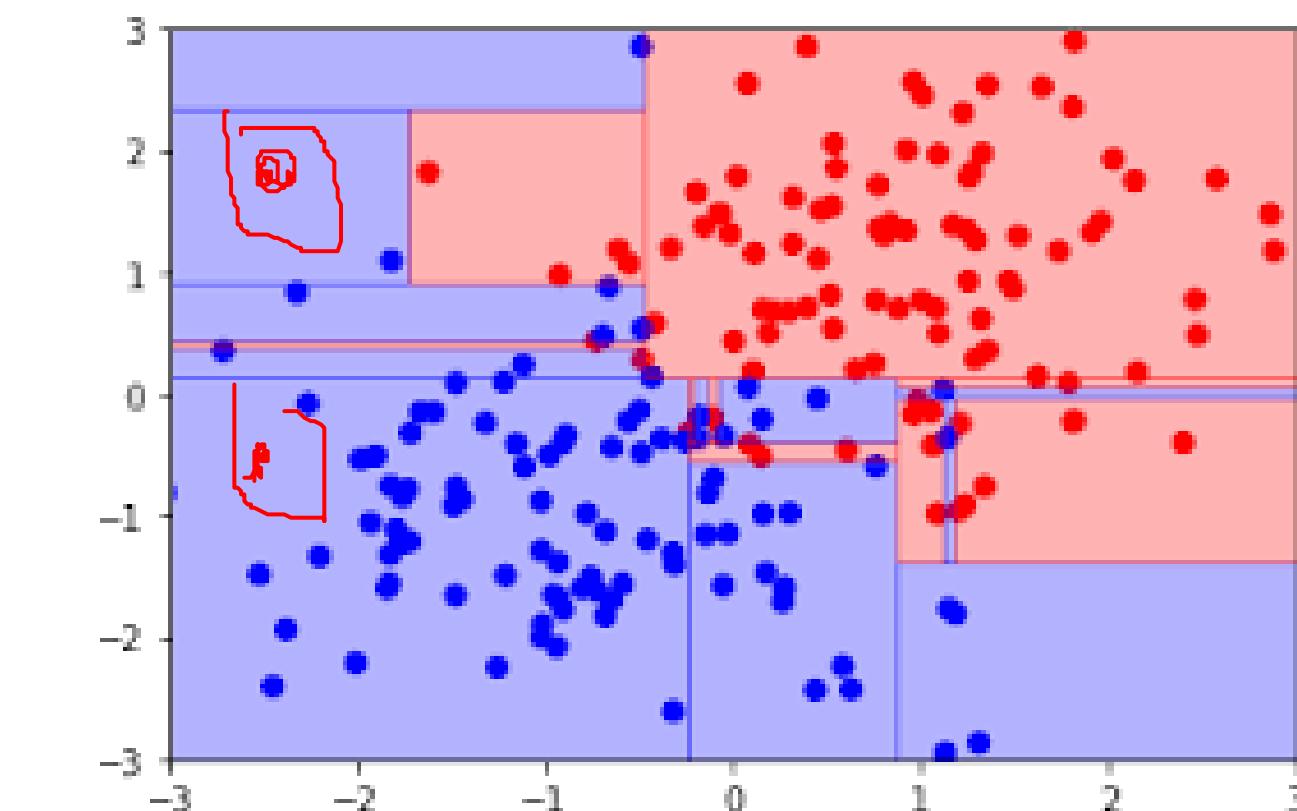
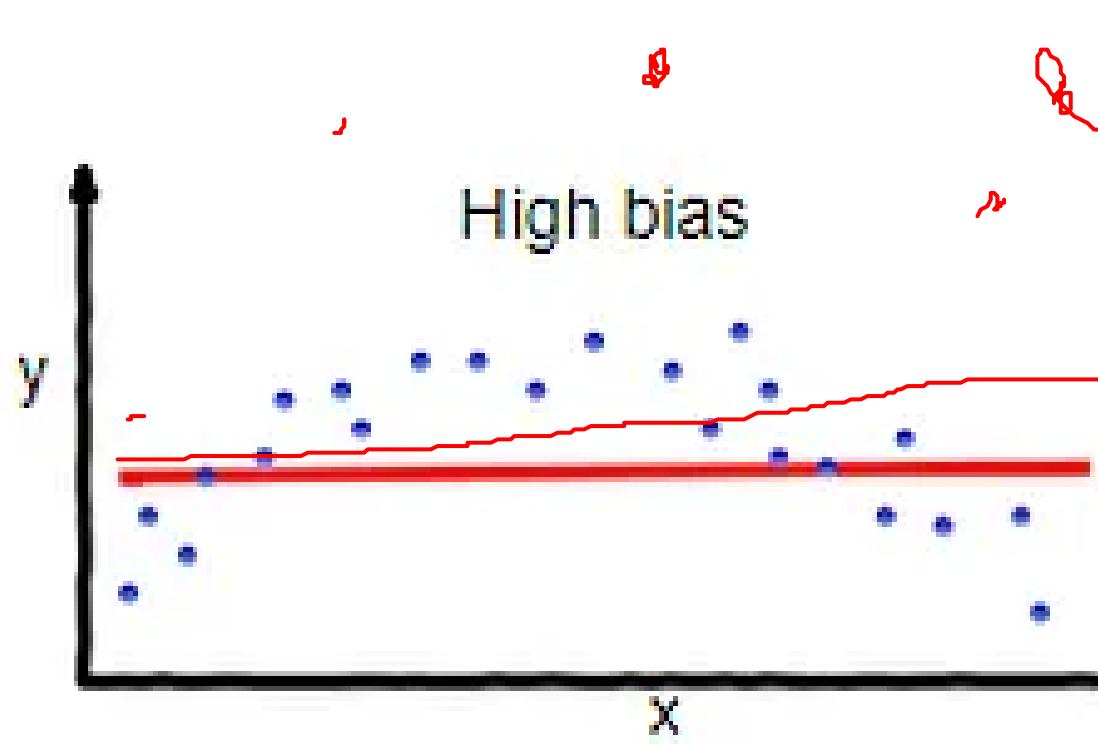
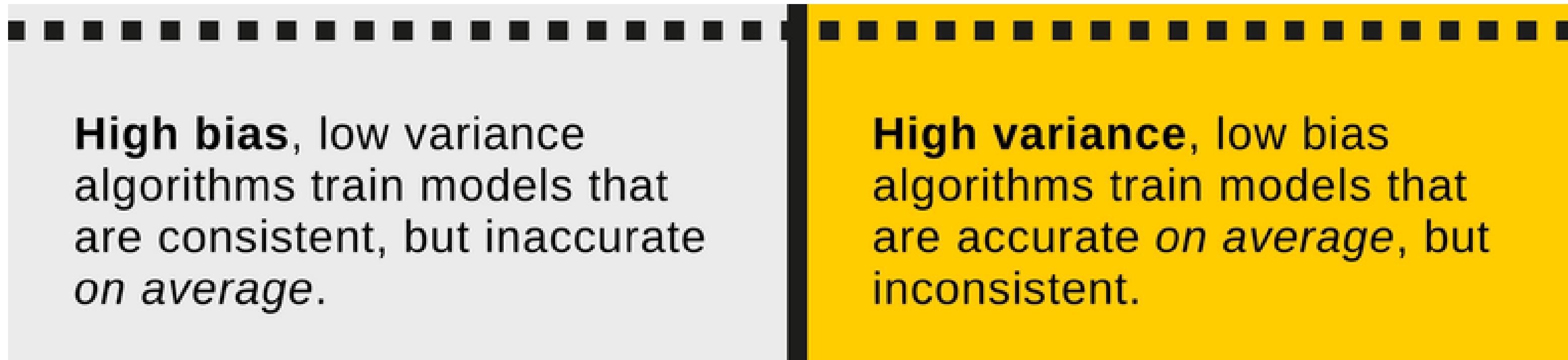
Think Linear Regression

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

Think Decision Tree

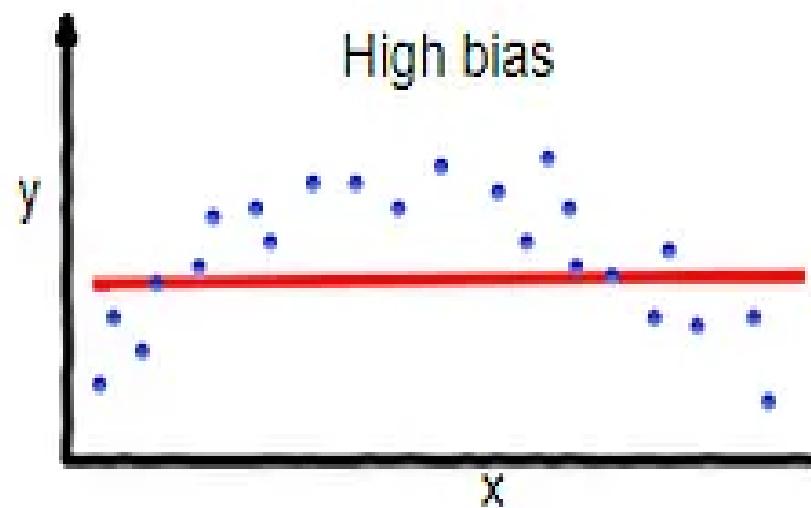
Variance refers to an algo's *sensitivity* to specific sets of training data.

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

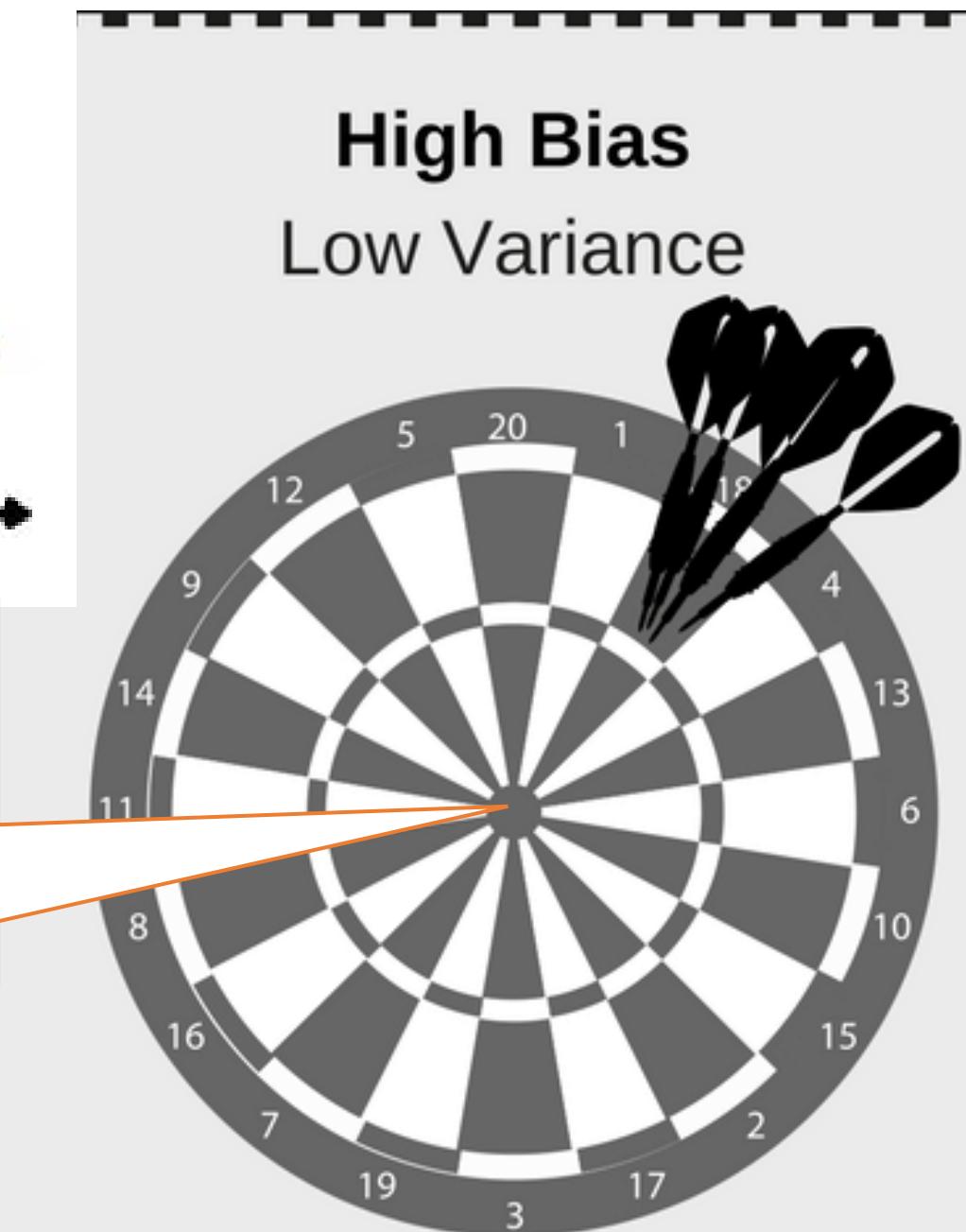


High bias, low variance
algorithms train models that
are consistent, but inaccurate
on average.

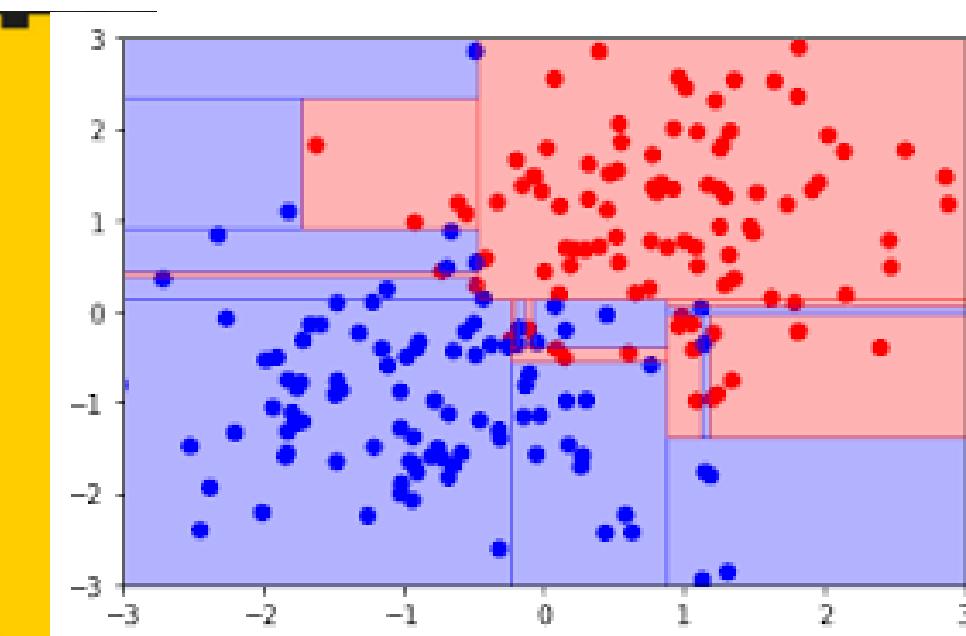
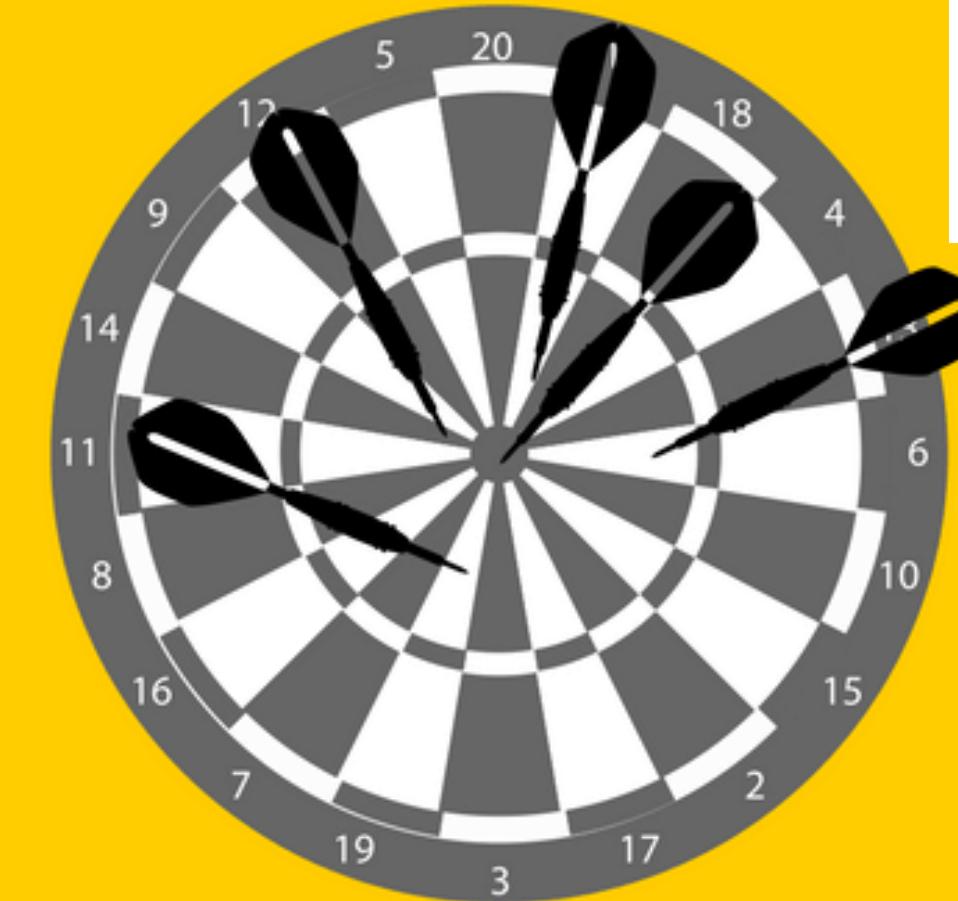
High variance, low bias
algorithms train models that
are accurate *on average*, but
inconsistent.



**Bulls eye means
best model &
consistent model**



**High Variance
Low Bias**



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*

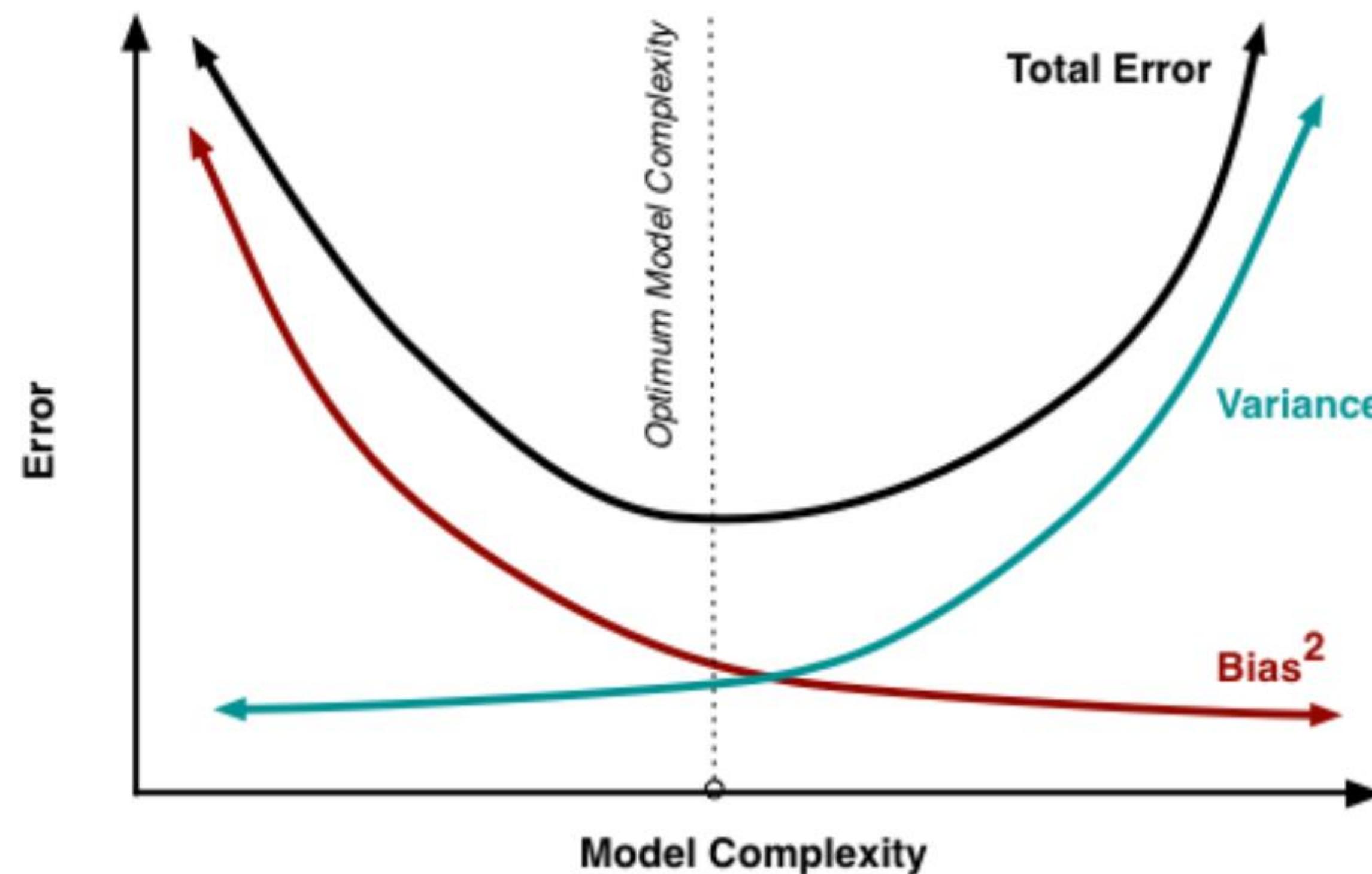
Within each algo family, there's a tradeoff too...

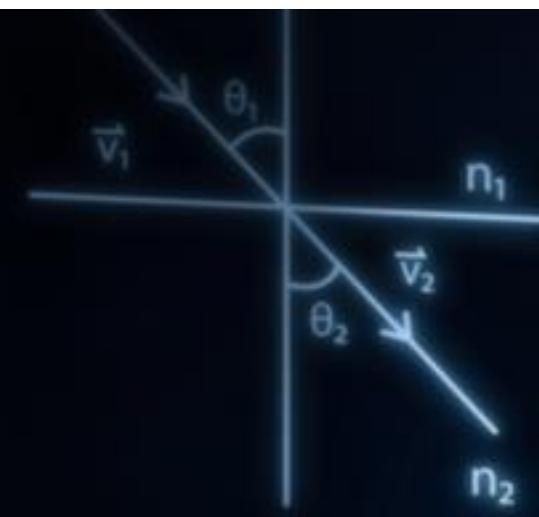
For example, regression can be **regularized** to further reduce complexity.

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

Bias Variance Plot

Total Error = Bias² + Variance + Irreducible Error





$$\bar{x}_1 = \frac{1+3+3+6+8+9}{6} = 5$$

$$\bar{x}_2 = 2+4+4+8+12 = 30$$

$$\bar{x}_3 = 4+7+1+6 = 18$$

$$\log_b b^x = x$$

$$\log_a x = \frac{\log_b x}{\log_b a}$$



$$\frac{a}{b} = \frac{a}{bc}$$

$$\frac{a}{b} = \frac{ac}{b}$$

$$\frac{a}{b} + \frac{c}{d} = \frac{ad+bc}{bd}$$

$$X^2 - 4X + 5 \leq 5$$

$$X^2 - 4X \leq 0$$

$$n(B \cap C) = 22$$

$$n(B) = 68$$

$$n(C) = 84$$

$$n(B \cup C) = n(B) + n(C) - n(B \cap C)$$

$$He = 4.002602$$

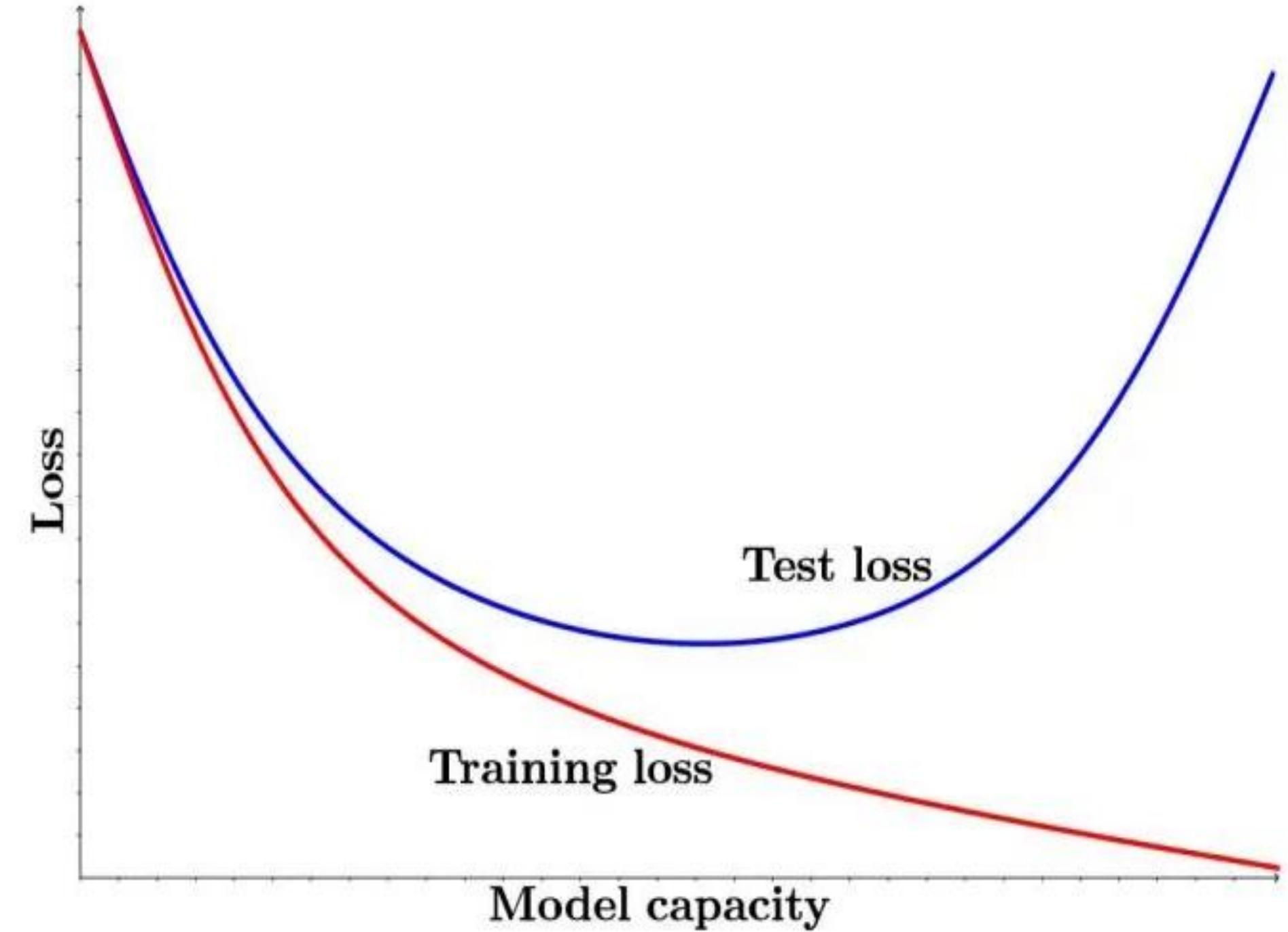
$$Na = 22.989769$$

$$Ar = 39.948$$



$$M = \frac{0.046765}{30L}$$

30L



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

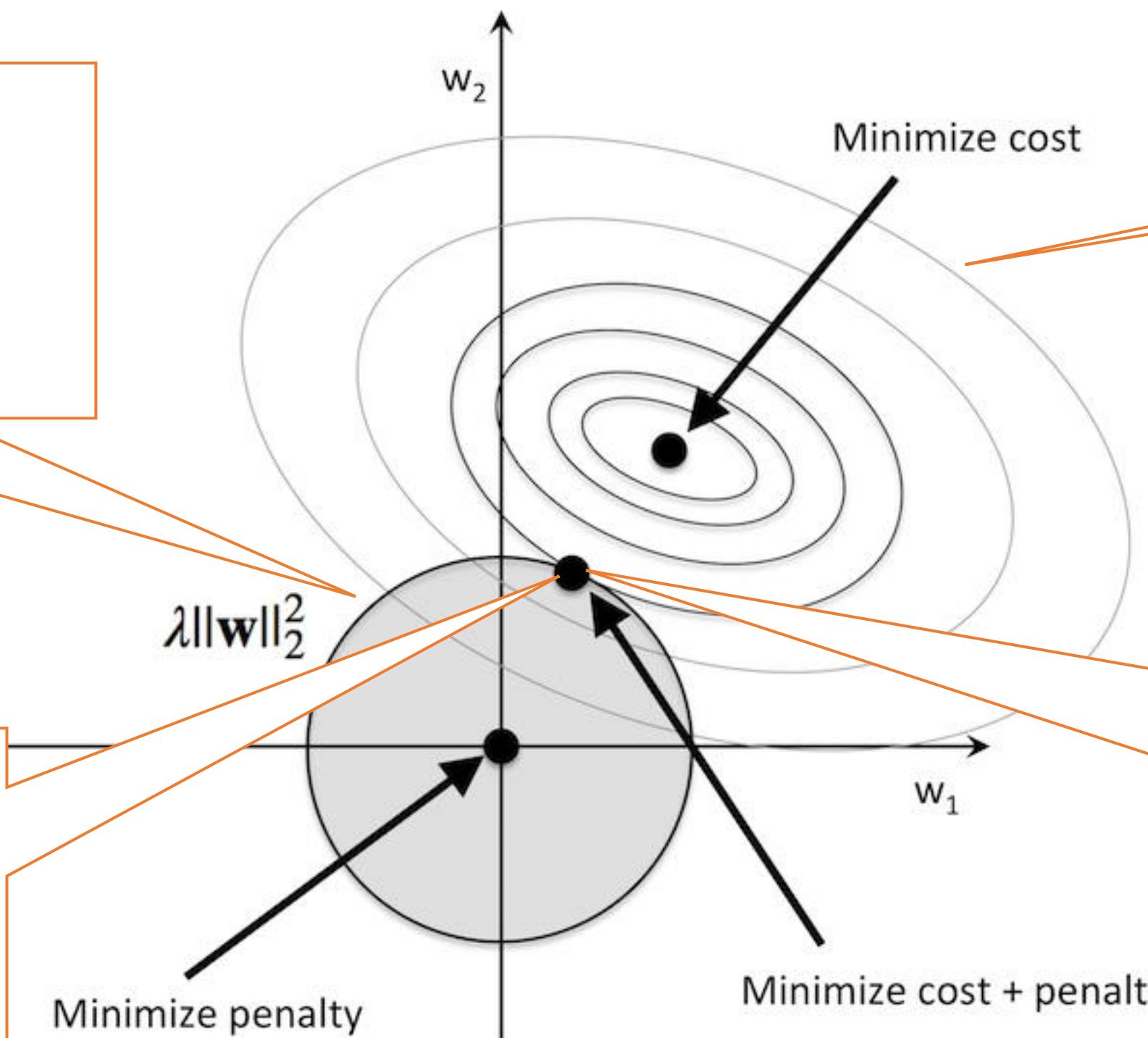
L2 Regularization

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

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

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

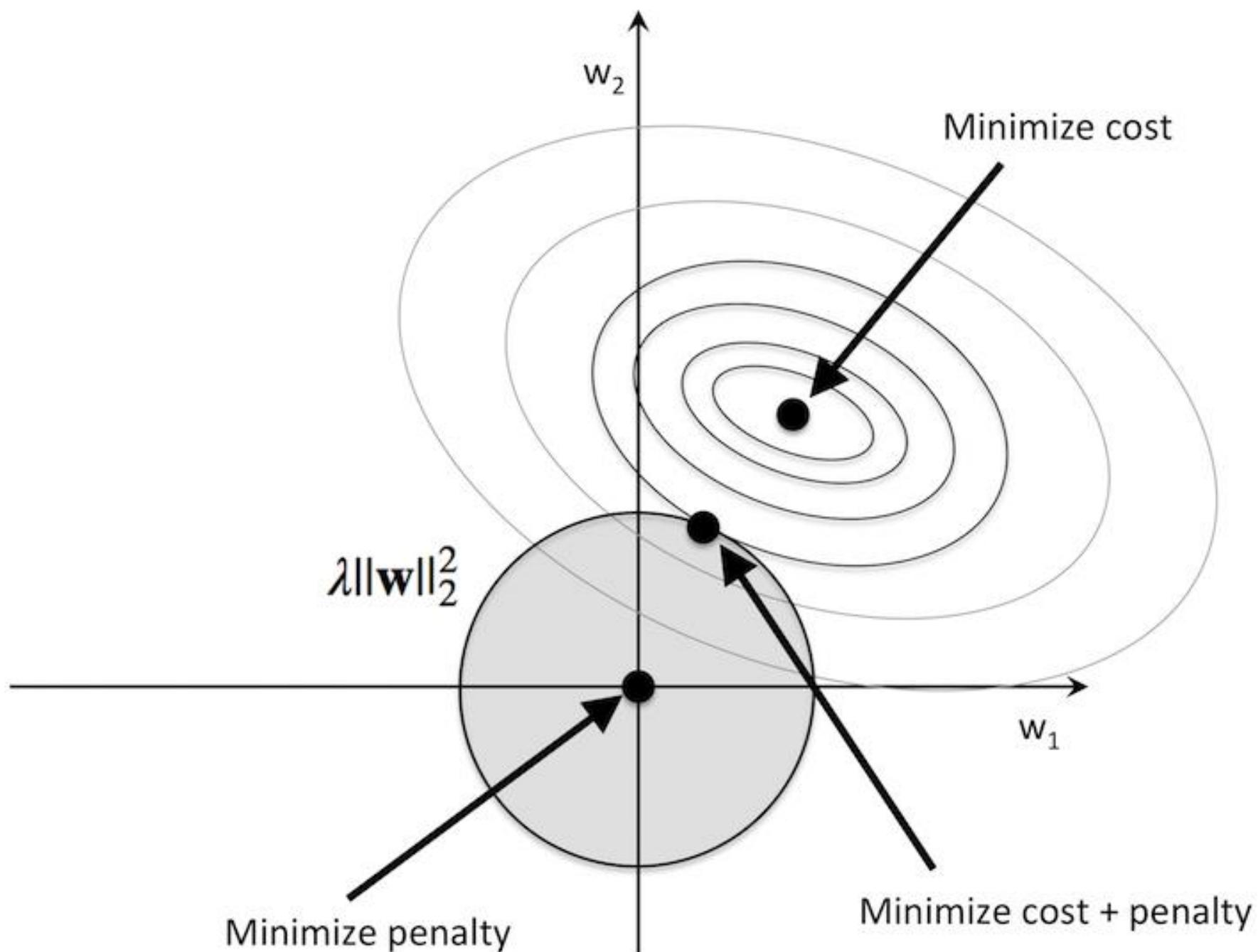
Common tangent implies gradient vector of one is multiple of another



Contour plot

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

Cost function adjusted for L2 Regularization



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

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

Lagrange Multiplier

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

Objective function in Lagrangian notation

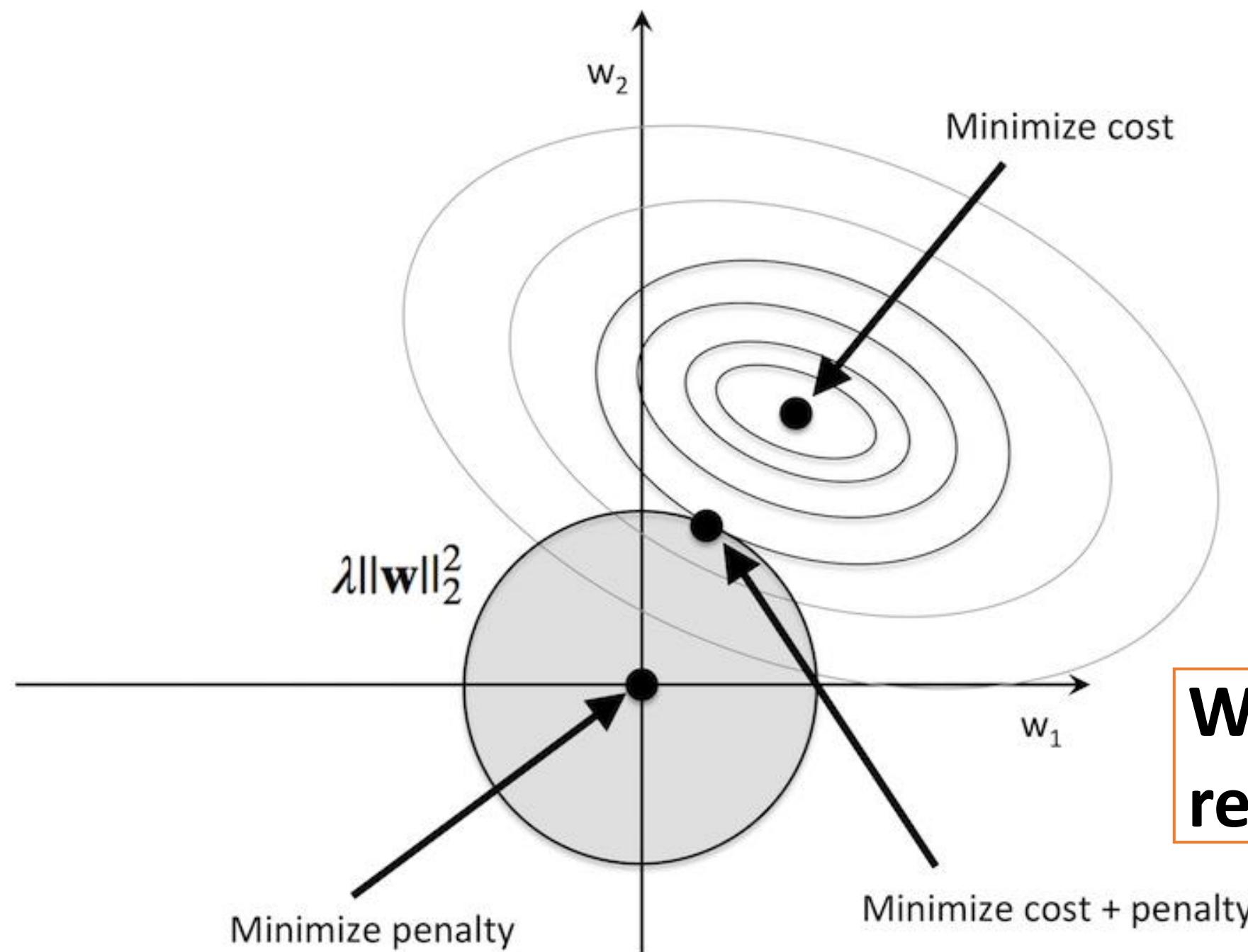
$$\nabla_w J + \lambda \nabla_w \|w\|^2 = 0$$

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

Without w0

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

Cost function adjusted for L2 Regularization



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

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

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

$$w = w - \eta \nabla_w \mathcal{J}$$

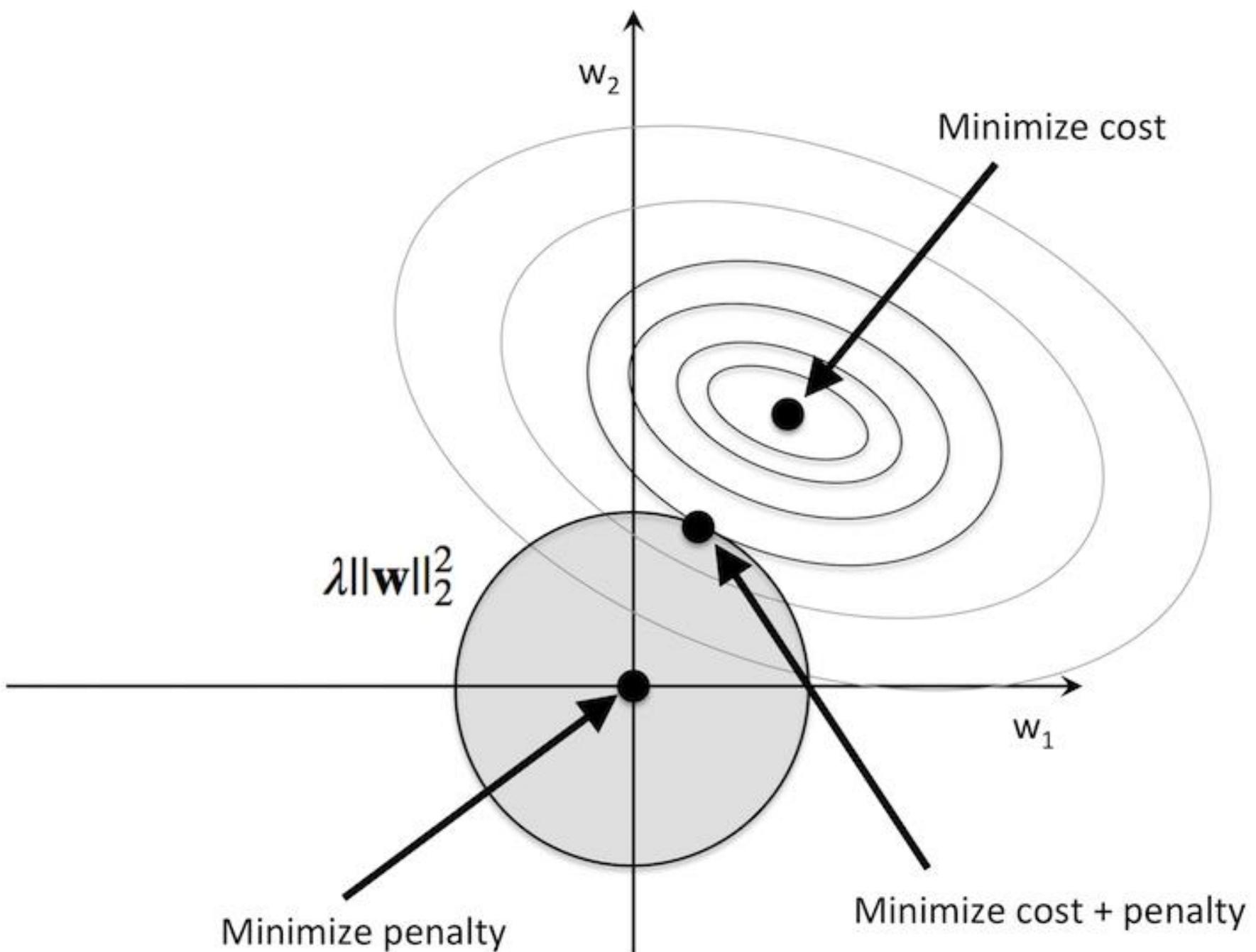
$$w = w - \eta \nabla_w \mathcal{J} - 2\eta \lambda w$$

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

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

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



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

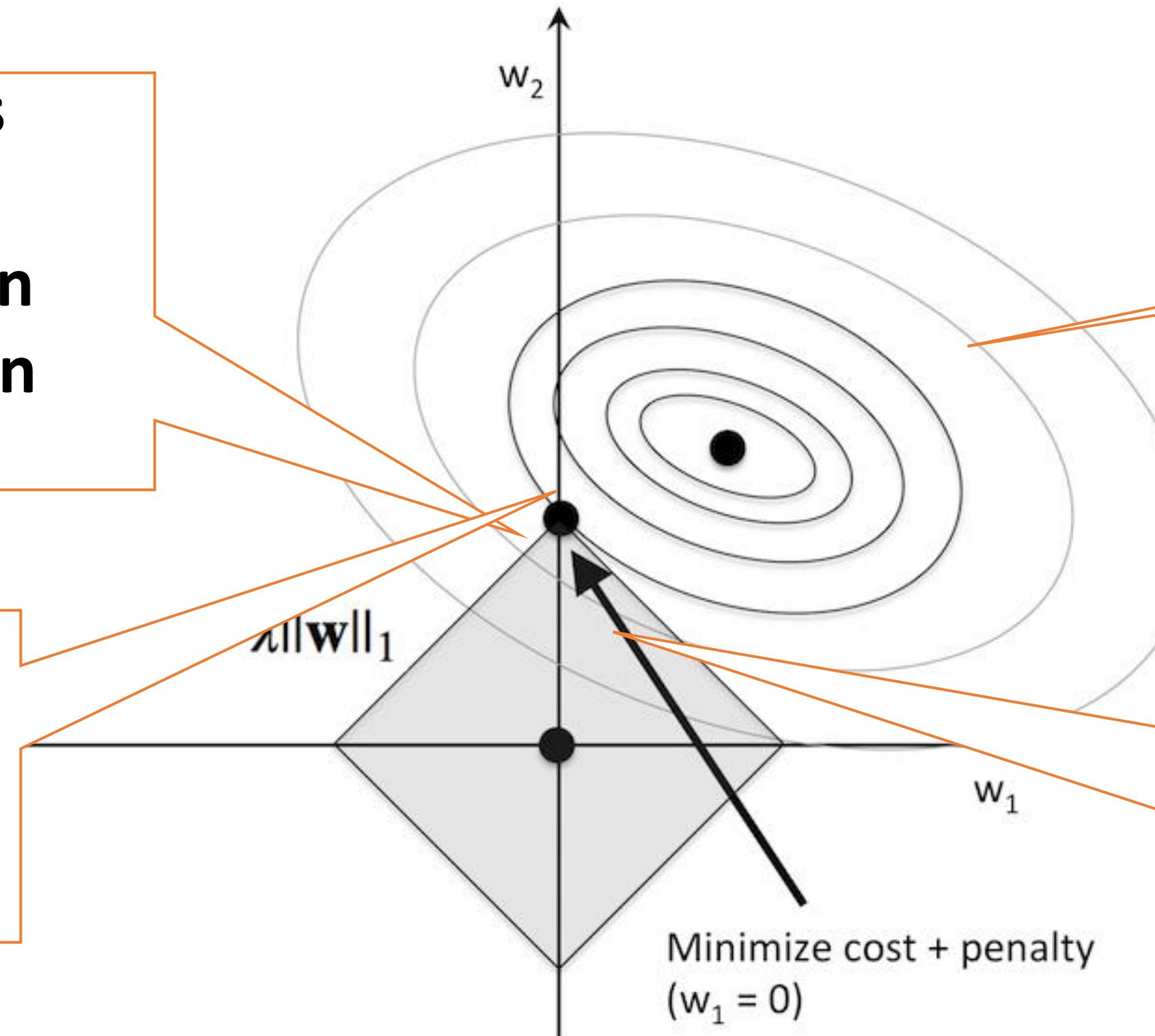
- What happens when
 - Lambda = 0
 - Lambda is very high
- Lambda=0 is equivalent to no regularization
- High lambda overrides objective function

L1 Regularization

- Stop (Adjust) model before it overfits

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

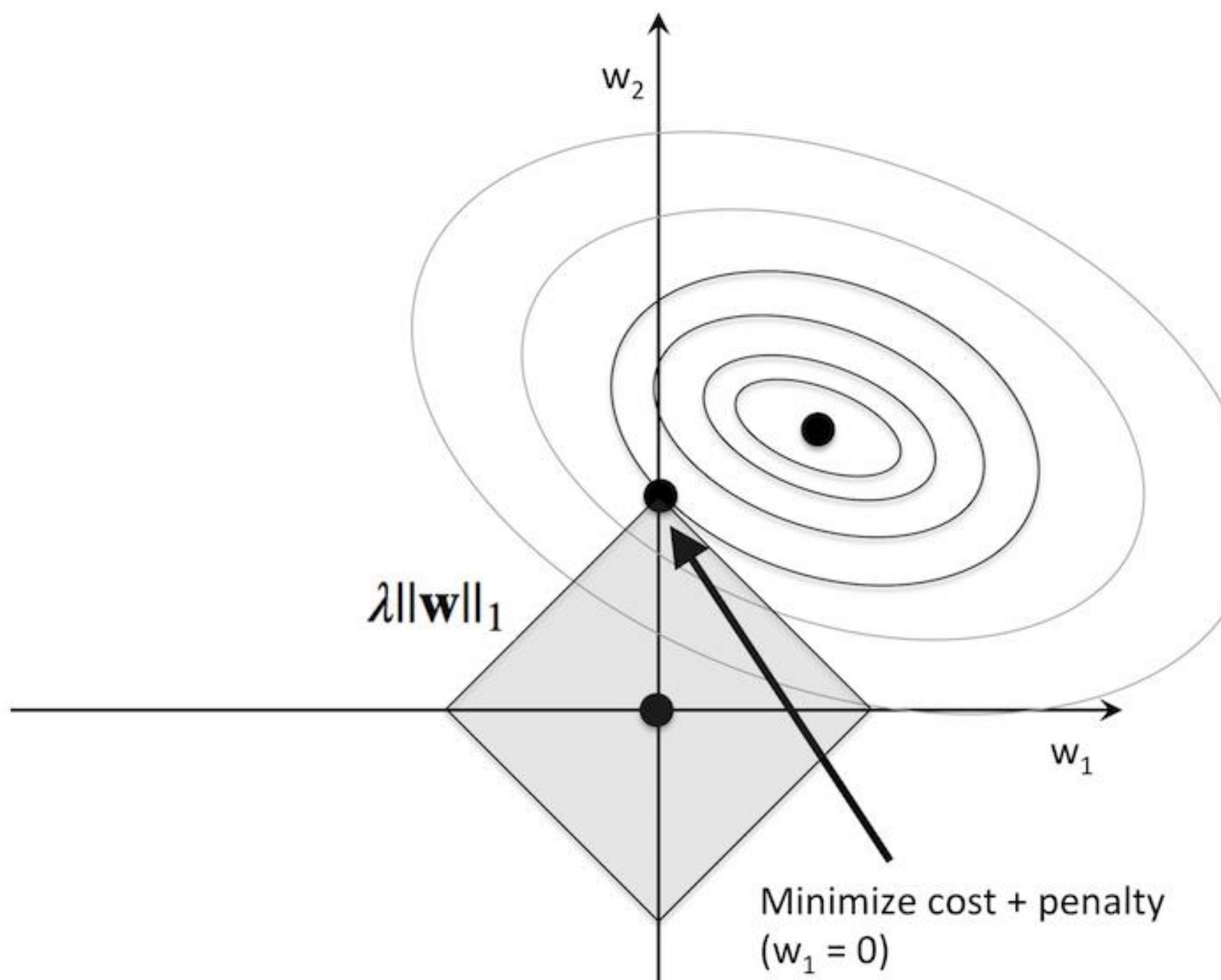
Common tangent implies gradient vector of one is multiple of another



Contour plot

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

Cost function adjusted for L1 Regularization



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

Lagrange Multiplier

$$\nabla_{\mathbf{w}} \mathcal{J} = \lambda \nabla_{\mathbf{w}} \|\mathbf{w}\|_1$$

Objective function in Lagrangian notation

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

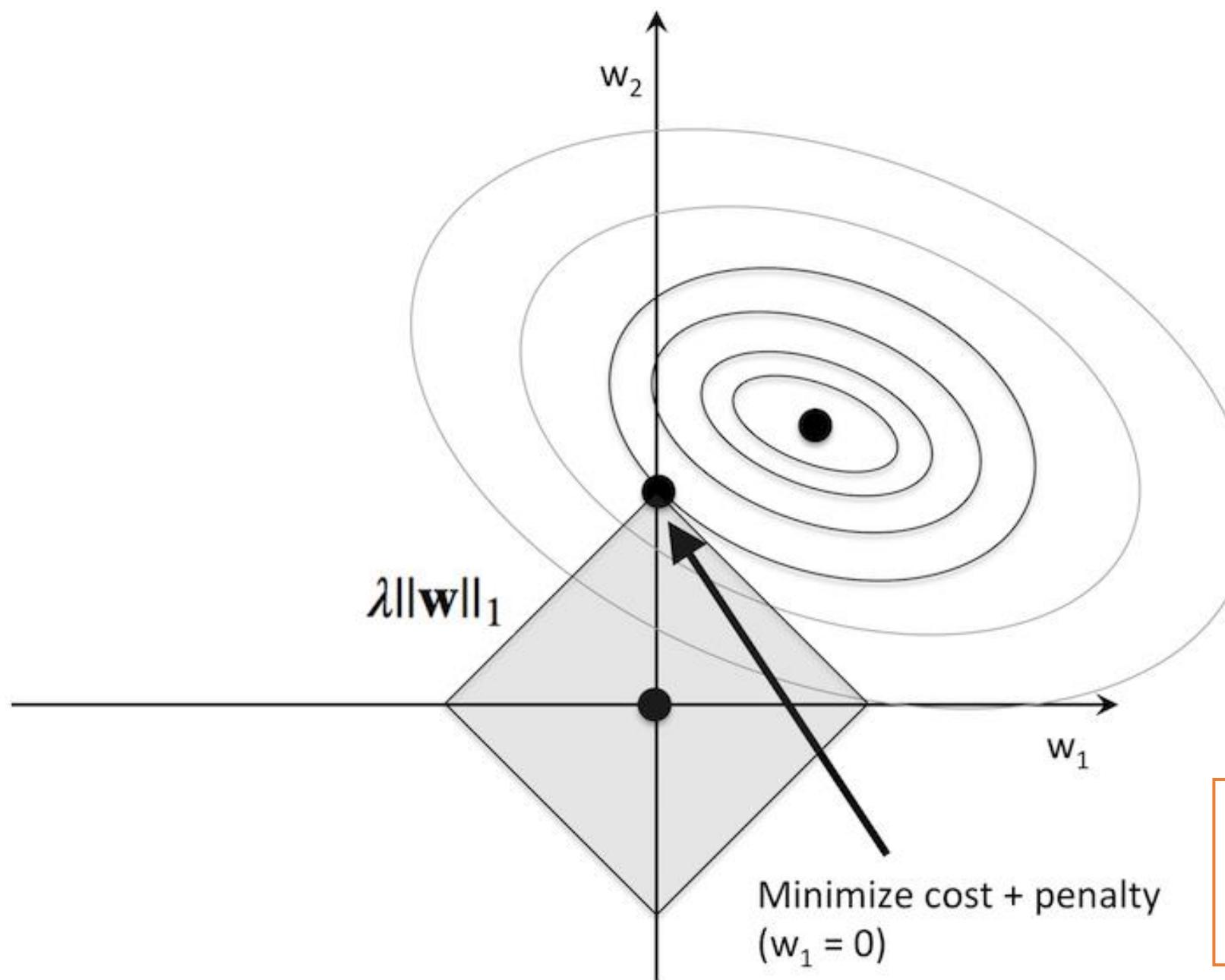
$$\arg \min_{\mathbf{w}} \nabla_{\mathbf{w}} \mathcal{J} + \lambda \nabla_{\mathbf{w}} \|\mathbf{w}\|_1$$

Without w_0

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

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

Cost function adjusted for L1 Regularization



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

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

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

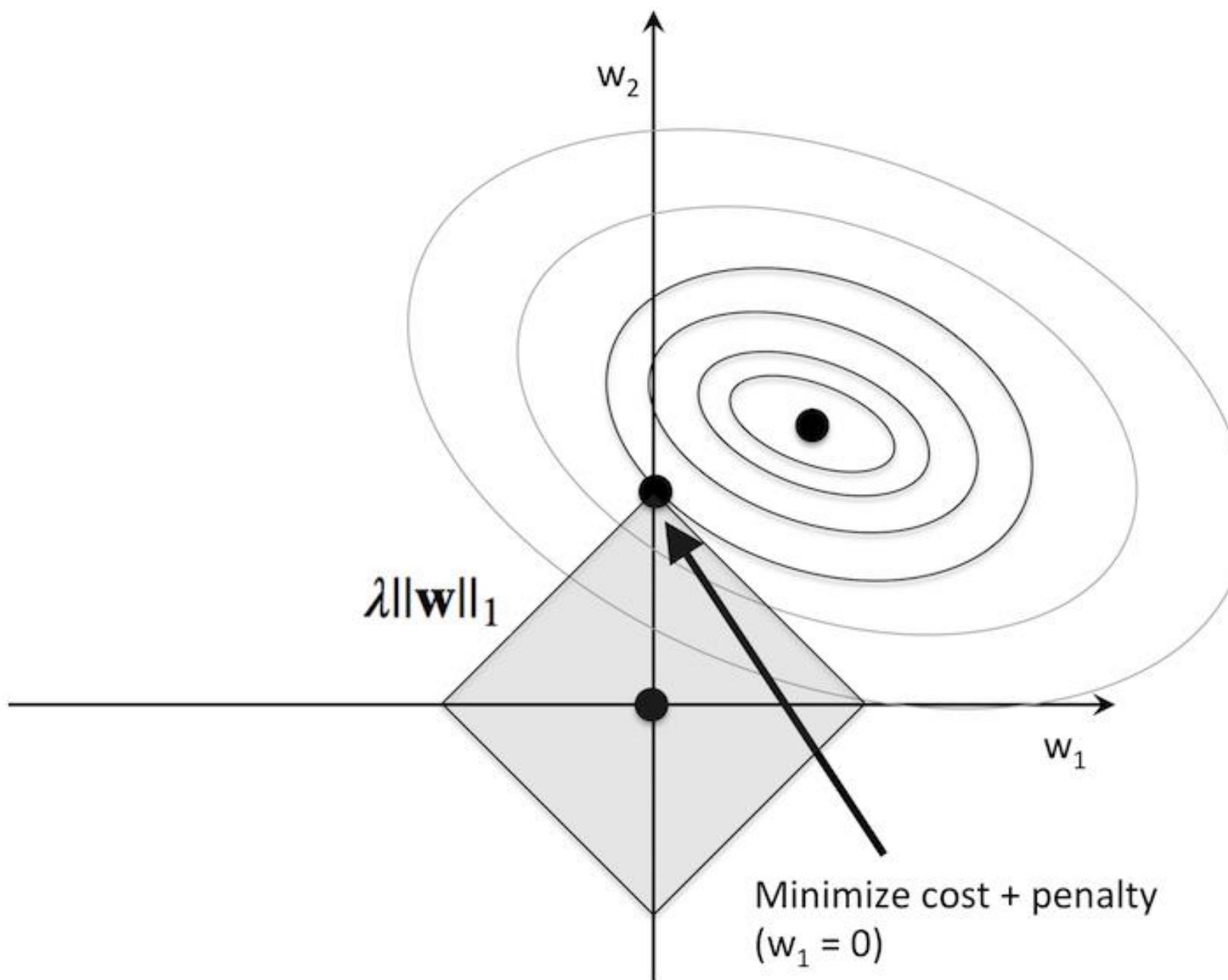
Without regularization

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

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

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

L2 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 $w = 0$

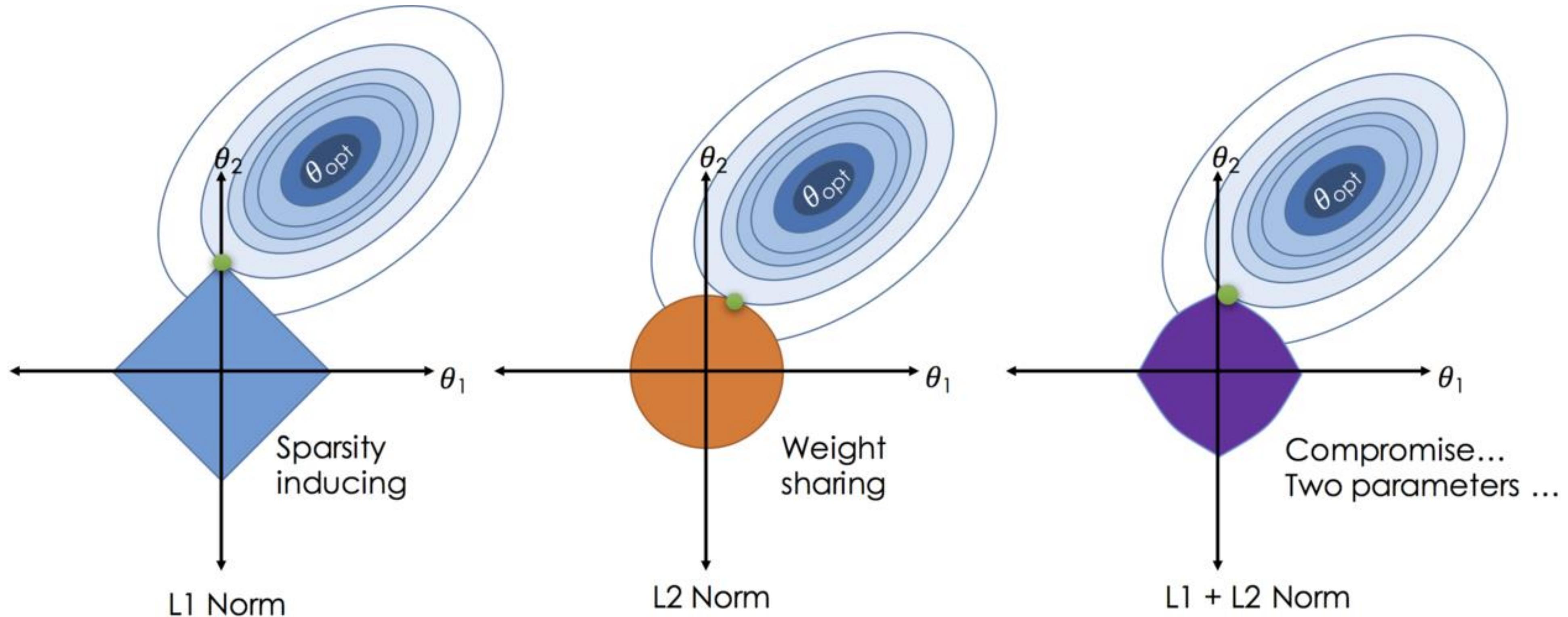
L2 and L1 regularization

- Regularization not applied to bias term (Why not?)
- L2 Regression also called Ridge Regression
- Reduces w_1 thru w_n , but does not make 0
- L1 Regularization also called Lasso Regularization
- Makes some of w_1 thru w_n 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