

# Ridge Regression

---

Nipun Batra

IIT Gandhinagar

August 28, 2025

# Introduction

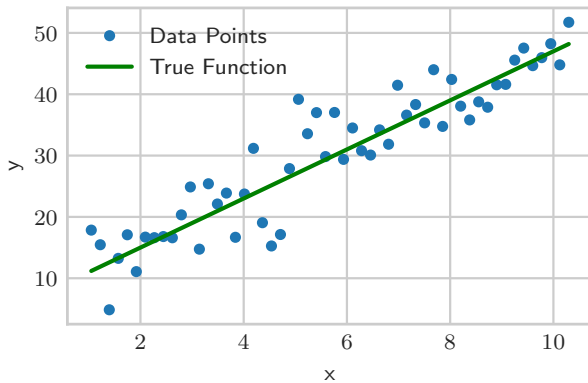
A known measure of overfitting can be the magnitude of the coefficient.

# Introduction

A known measure of overfitting can be the magnitude of the coefficient.

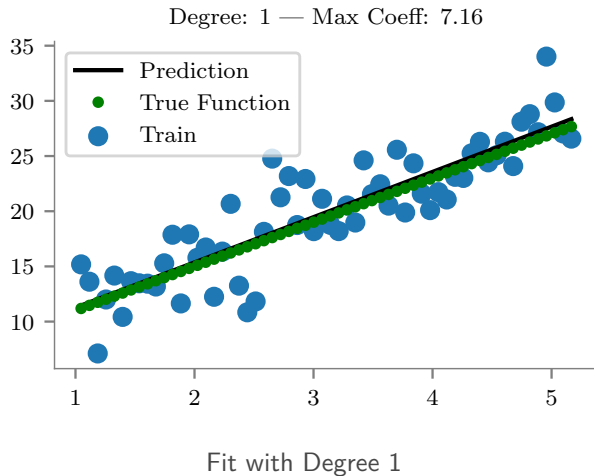
In  $f(x) = c_0 + c_1x + c_2x^2 + \dots$  it is  $\max |c_i|$

# Introduction

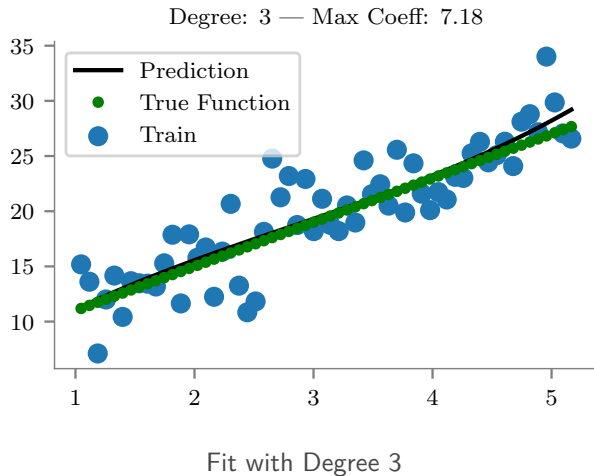


Base Data Set

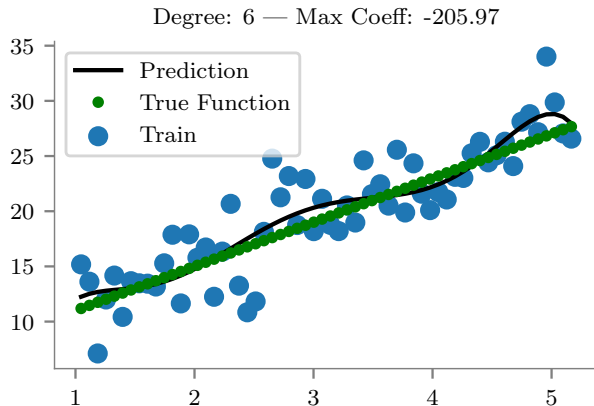
# Introduction



# Introduction

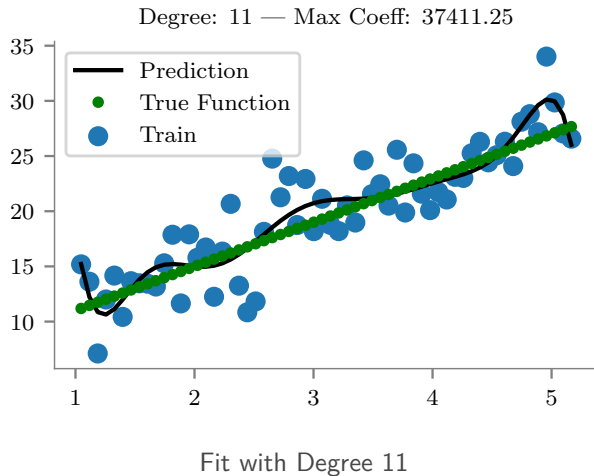


# Introduction



Fit with Degree 6

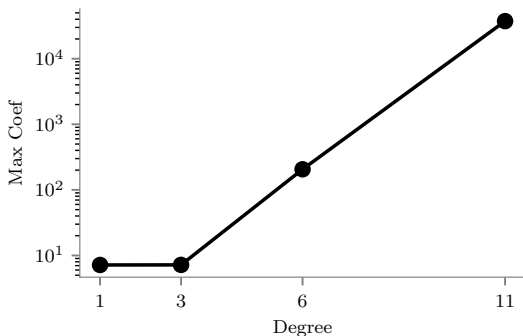
# Introduction





# Introduction

In the examples we notice that as the degree increases (as the prediction starts to overfit the base data), the maximum coefficient also increases.



Trend of the coefficients

# Introduction

To prevent overfitting we place penalties on large  $\theta_i$

# Introduction

To prevent overfitting we place penalties on large  $\theta_i$

**Objective:**

$$\begin{aligned} &\text{Minimise } (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) \\ &\text{s.t. } \boldsymbol{\theta}^T \boldsymbol{\theta} \leq S \end{aligned}$$

# Introduction

To prevent overfitting we place penalties on large  $\theta_i$

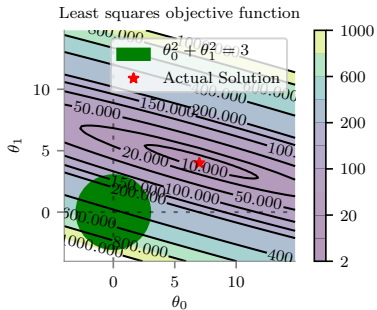
**Objective:**

$$\begin{aligned} &\text{Minimise } (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) \\ &\text{s.t. } \boldsymbol{\theta}^T \boldsymbol{\theta} \leq S \end{aligned}$$

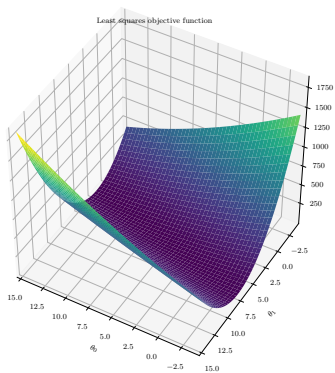
This is equivalent to

$$\text{Minimise } (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) + \delta^2 \boldsymbol{\theta}^T \boldsymbol{\theta}$$

# Introduction



(a) Contour Plot



(b) Surface Plot

Visualization of the Example

# KKT Conditions

To implement this we use KKT Conditions

# KKT Conditions

To implement this we use KKT Conditions

$$\text{Minimise } (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})$$

$$\text{s.t. } \boldsymbol{\theta}^T \boldsymbol{\theta} \leq S$$

$$L(\boldsymbol{\theta}, \mu) = (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) + \mu (\boldsymbol{\theta}^T \boldsymbol{\theta} - S)$$

where,  $\mu \geq 0$  (and  $\mu = \delta^2$ )

# KKT Conditions

To implement this we use KKT Conditions

$$\text{Minimise } (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})$$

$$\text{s.t. } \boldsymbol{\theta}^T \boldsymbol{\theta} \leq S$$

$$L(\boldsymbol{\theta}, \mu) = (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) + \mu (\boldsymbol{\theta}^T \boldsymbol{\theta} - S)$$

where,  $\mu \geq 0$  (and  $\mu = \delta^2$ )

If  $\mu = 0$

There is no regularization

No effect on constraint



# KKT Conditions

To implement this we use KKT Conditions

$$\text{Minimise } (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})$$

$$\text{s.t. } \boldsymbol{\theta}^T \boldsymbol{\theta} \leq S$$

$$L(\boldsymbol{\theta}, \mu) = (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) + \mu (\boldsymbol{\theta}^T \boldsymbol{\theta} - S)$$

where,  $\mu \geq 0$  (and  $\mu = \delta^2$ )

If  $\mu = 0$

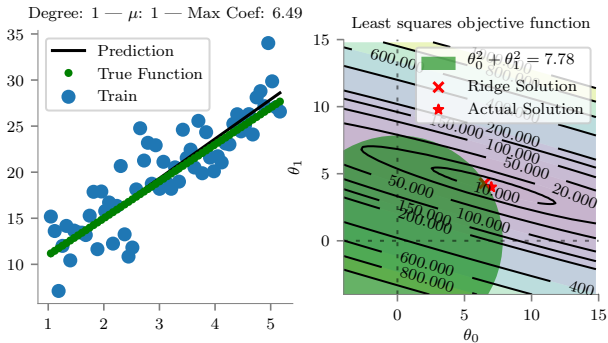
There is no regularization

No effect on constraint

If  $\mu \neq 0$

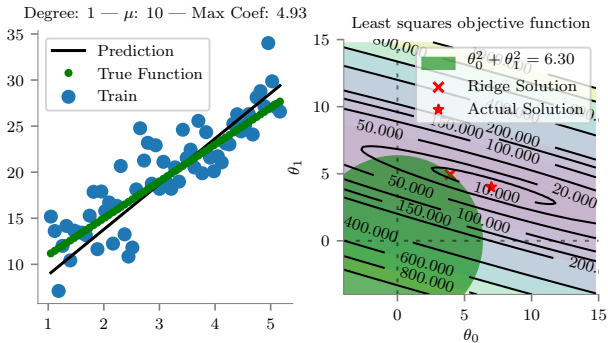
$$\implies \boldsymbol{\theta}^T \boldsymbol{\theta} - S = 0$$

# Effect of $\mu$



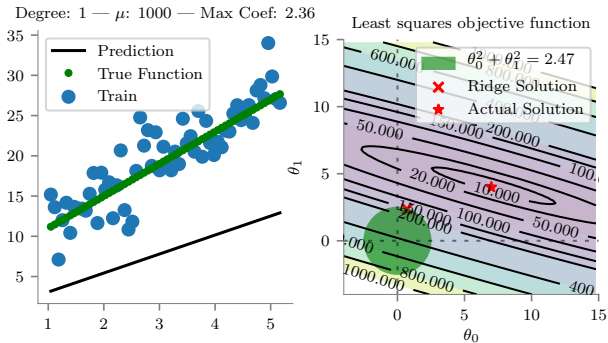
$$\mu = 1$$

# Effect of $\mu$



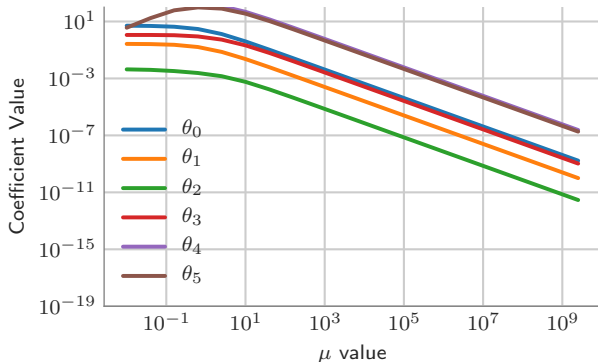
$$\mu = 10$$

# Effect of $\mu$



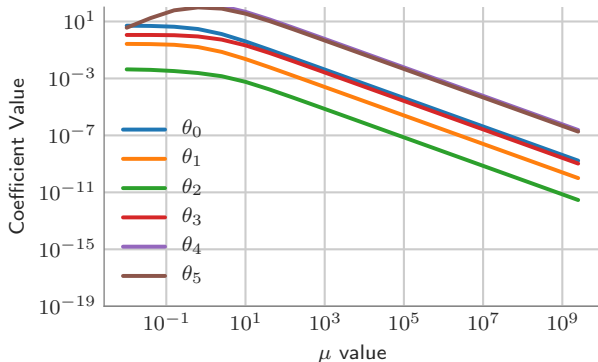
$$\mu = 1000$$

## Effect of $\mu$ - Regularization of Parameters



Comparing the magnitudes of the coefficients with varying  $\mu$   
(on the *Real Estate Data Set*)

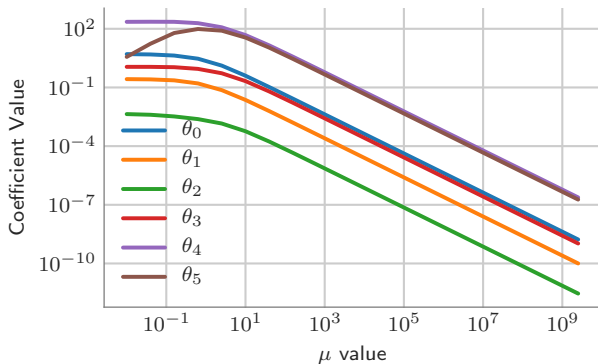
## Effect of $\mu$ - Regularization of Parameters



Comparing the magnitudes of the coefficients with varying  $\mu$   
(on the *Real Estate Data Set*)

Are  $\theta_i$  all zero for high  $\mu$ ?

## Effect of $\mu$ - Regularization of Parameters



Comparing the magnitudes of the coefficients with varying  $\mu$   
(on the *Real Estate Data Set*)

# Analytical Method

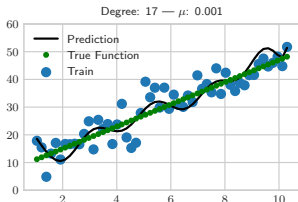
**Ridge Objective:**

$$\min_{\theta} (\mathbf{y} - \mathbf{X}\theta)^T (\mathbf{y} - \mathbf{X}\theta) + \mu \theta^T \theta$$

$$\begin{aligned}\frac{\partial L(\theta, \mu)}{\partial \theta} &= 0 \\ \frac{\partial}{\partial \theta} \left\{ \mathbf{y}^T \mathbf{y} - 2\mathbf{y}^T \mathbf{X}\theta + \theta^T \mathbf{X}^T \mathbf{X}\theta \right\} + \frac{\partial}{\partial \theta} \mu \theta^T \theta &= 0 \\ \implies -\mathbf{X}^T \mathbf{y} + (\mathbf{X}^T \mathbf{X} + \mu \mathbf{I}) \theta &= 0 \\ \implies \theta^* &= (\mathbf{X}^T \mathbf{X} + \mu \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}\end{aligned}$$



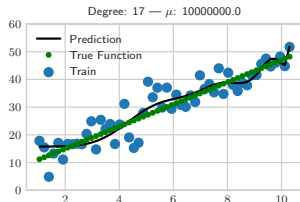
# Bias/Variance



Fit High Order  
Polynomial

$\Rightarrow$  high variance

$\Rightarrow \mu \rightarrow 0$



Fit High Order  
Polynomial

$\Rightarrow$  low variance

$\Rightarrow \mu \rightarrow \infty$

## Example

**Q.)** Solve Regularized ( $\mu = 2$ ) and Unregularized.

## Example: Unregularised

$$\theta = (\mathbf{X}^T \mathbf{X})^{-1} (\mathbf{X}^T \mathbf{y})$$

## Example: Unregularised

$$\theta = (\mathbf{X}^T \mathbf{X})^{-1} (\mathbf{X}^T \mathbf{y})$$

$$\mathbf{X}^T \mathbf{X} = \begin{bmatrix} 4 & 10 \\ 10 & 30 \end{bmatrix}$$

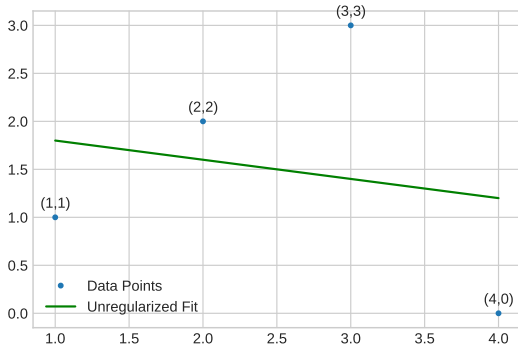
$$(\mathbf{X}^T \mathbf{X})^{-1} = \frac{1}{20} \begin{bmatrix} 30 & -10 \\ -10 & 4 \end{bmatrix}$$

$$\mathbf{X}^T \mathbf{y} = \begin{bmatrix} 6 \\ 14 \end{bmatrix}$$

## Example: Unregularised

$$\theta = (\mathbf{X}^T \mathbf{X})^{-1} (\mathbf{X}^T \mathbf{y})$$

$$\begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix} = \begin{bmatrix} 2 \\ -1/5 \end{bmatrix}$$



## Example: Regularised

$$\theta = (\mathbf{X}^T \mathbf{X} + \mu \mathbf{I})^{-1} (\mathbf{X}^T \mathbf{y})$$

## Example: Regularised

$$\theta = (\mathbf{X}^T \mathbf{X} + \mu \mathbf{I})^{-1} (\mathbf{X}^T \mathbf{y})$$

$$\mathbf{X}^T \mathbf{X} = \begin{bmatrix} 4 & 10 \\ 10 & 30 \end{bmatrix}$$

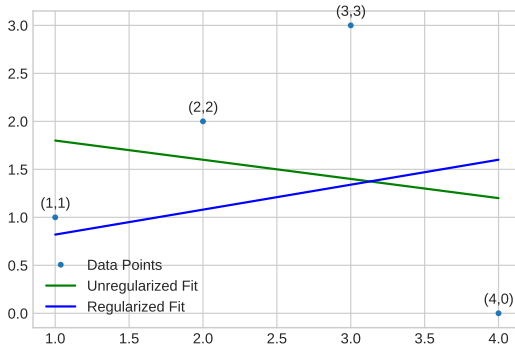
$$\mathbf{X}^T \mathbf{X} + \mu \mathbf{I} = \begin{bmatrix} 6 & 10 \\ 10 & 32 \end{bmatrix}$$

$$(\mathbf{X}^T \mathbf{X} + \mu \mathbf{I})^{-1} = \frac{1}{92} \begin{bmatrix} 32 & -10 \\ -10 & 6 \end{bmatrix}$$

$$\mathbf{X}^T \mathbf{y} = \begin{bmatrix} 6 \\ 14 \end{bmatrix}$$

## Example: Regularised

$$\theta = (\mathbf{X}^T \mathbf{X} + \mu \mathbf{I})^{-1} (\mathbf{X}^T \mathbf{y})$$
$$\begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix} = \begin{bmatrix} 0.56 \\ 0.26 \end{bmatrix}$$





# Multi-collinearity

$(\mathbf{X}^T \mathbf{X})^{-1}$  is not computable when  $|\mathbf{X}^T \mathbf{X}| = 0$ .  
This was a drawback of using linear regression

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 2 & 4 \\ 1 & 3 & 6 \end{bmatrix}$$

The matrix  $\mathbf{X}$  is not full rank.

## Multi-collinearity

But with ridge regression, the matrix to be inverted is  $\mathbf{X}^T\mathbf{X} + \mu\mathbf{I}$  and not  $\mathbf{X}^T\mathbf{X}$ .

$$\mathbf{X}^T\mathbf{X} + \mu\mathbf{I} = \begin{bmatrix} 3 + \mu & 6 & 12 \\ 6 & 14 + \mu & 28 \\ 12 & 28 & 56 + \mu \end{bmatrix}$$

The matrix  $\mathbf{X}^T\mathbf{X}$  would be full rank for  $\mu > 0$ .

# Multi-collinearity

But with ridge regression, the matrix to be inverted is  $\mathbf{X}^T\mathbf{X} + \mu\mathbf{I}$  and not  $\mathbf{X}^T\mathbf{X}$ .

$$\mathbf{X}^T\mathbf{X} + \mu\mathbf{I} = \begin{bmatrix} 3 + \mu & 6 & 12 \\ 6 & 14 + \mu & 28 \\ 12 & 28 & 56 + \mu \end{bmatrix}$$

The matrix  $\mathbf{X}^T\mathbf{X}$  would be full rank for  $\mu > 0$ .  
Another interpretation of “regularisation”

## Extension of the analytical model

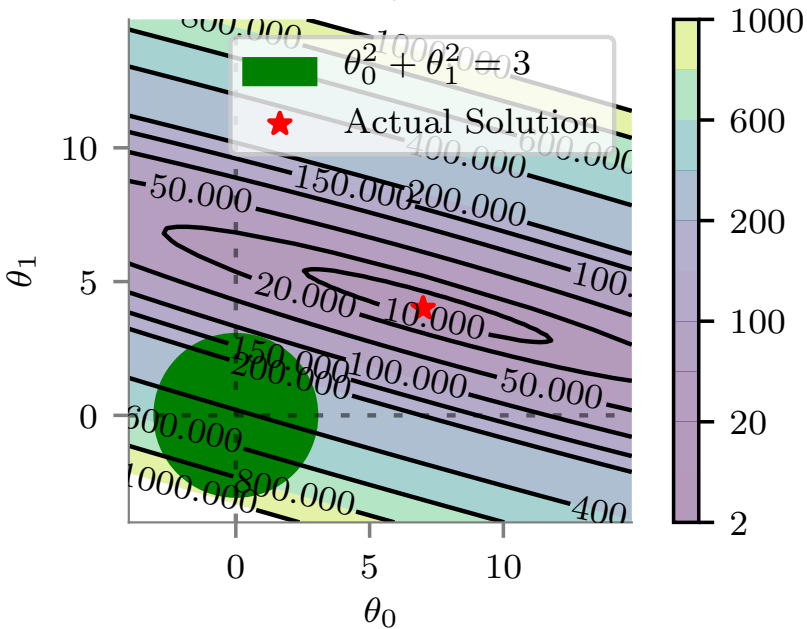
For ridge with no penalty on  $\theta_0$

$$\hat{\boldsymbol{\theta}} = \left( \mathbf{X}^T \mathbf{X} + \mu \mathbf{I}^* \right)^{-1} \mathbf{X}^T \mathbf{y}$$

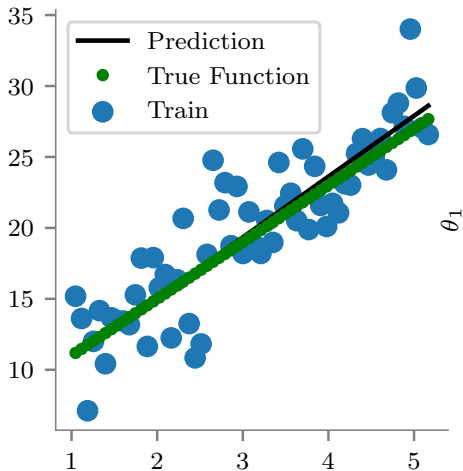
where,

$$\mathbf{I}^* = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}$$

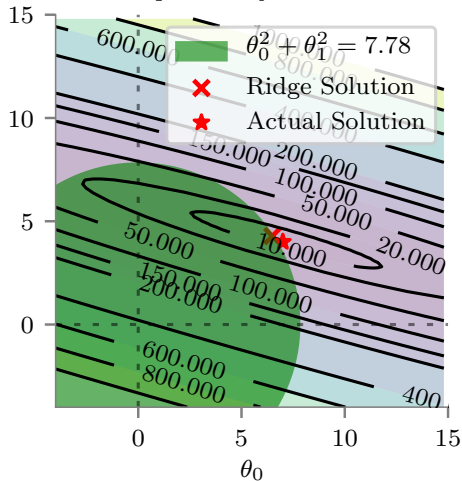
### Least squares objective function



Degree: 1 —  $\mu$ : 1 — Max Coef: 6.49



Least squares objective function



## Ridge Solution using Gradient Descent

- $\theta = \theta - \alpha \frac{\partial}{\partial \theta} ((\mathbf{y} - \mathbf{X}\theta)^\top (\mathbf{y} - \mathbf{X}\theta) + \mu \theta^\top \theta)$

## Ridge Solution using Gradient Descent

- $\theta = \theta - \alpha \frac{\partial}{\partial \theta} ((\mathbf{y} - \mathbf{X}\theta)^\top (\mathbf{y} - \mathbf{X}\theta) + \mu \theta^\top \theta)$
- $\theta = \theta - \alpha (-2\mathbf{X}^\top \mathbf{y} + 2\mathbf{X}^\top \mathbf{X}\theta + 2\mu \mathbf{I}\theta)$



## Ridge Solution using Gradient Descent

- $\theta = \theta - \alpha \frac{\partial}{\partial \theta} ((\mathbf{y} - \mathbf{X}\theta)^\top (\mathbf{y} - \mathbf{X}\theta) + \mu \theta^\top \theta)$
- $\theta = \theta - \alpha (-2\mathbf{X}^\top \mathbf{y} + 2\mathbf{X}^\top \mathbf{X}\theta + 2\mu \mathbf{I}\theta)$
- $\theta = (1 - 2\alpha\mu \mathbf{I})\theta - \alpha (-2\mathbf{X}^\top \mathbf{y} + 2\mathbf{X}^\top \mathbf{X}\theta)$

## Ridge Solution using Gradient Descent

- $\theta = \theta - \alpha \frac{\partial}{\partial \theta} ((\mathbf{y} - \mathbf{X}\theta)^\top (\mathbf{y} - \mathbf{X}\theta) + \mu \theta^\top \theta)$
- $\theta = \theta - \alpha (-2\mathbf{X}^\top \mathbf{y} + 2\mathbf{X}^\top \mathbf{X}\theta + 2\mu \mathbf{I}\theta)$
- $\theta = (1 - 2\alpha\mu \mathbf{I})\theta - \alpha (-2\mathbf{X}^\top \mathbf{y} + 2\mathbf{X}^\top \mathbf{X}\theta)$
- $\theta = \underbrace{(1 - 2\alpha\mu \mathbf{I})\theta}_{\text{Shrinking } \theta} - \alpha (-2\mathbf{X}^\top \mathbf{y} + 2\mathbf{X}^\top \mathbf{X}\theta)$

## Ridge Solution using Gradient Descent

- $\theta = \theta - \alpha \frac{\partial}{\partial \theta} ((\mathbf{y} - \mathbf{X}\theta)^\top (\mathbf{y} - \mathbf{X}\theta) + \mu \theta^\top \theta)$
- $\theta = \theta - \alpha (-2\mathbf{X}^\top \mathbf{y} + 2\mathbf{X}^\top \mathbf{X}\theta + 2\mu \mathbf{I}\theta)$
- $\theta = (1 - 2\alpha\mu \mathbf{I})\theta - \alpha (-2\mathbf{X}^\top \mathbf{y} + 2\mathbf{X}^\top \mathbf{X}\theta)$
- $\theta = \underbrace{(1 - 2\alpha\mu \mathbf{I})\theta}_{\text{Shrinking } \theta} - \alpha (-2\mathbf{X}^\top \mathbf{y} + 2\mathbf{X}^\top \mathbf{X}\theta)$

## Ridge Solution using Gradient Descent

- $\theta = \theta - \alpha \frac{\partial}{\partial \theta} ((\mathbf{y} - \mathbf{X}\theta)^\top (\mathbf{y} - \mathbf{X}\theta) + \mu \theta^\top \theta)$
- $\theta = \theta - \alpha (-2\mathbf{X}^\top \mathbf{y} + 2\mathbf{X}^\top \mathbf{X}\theta + 2\mu \mathbf{I}\theta)$
- $\theta = (1 - 2\alpha\mu \mathbf{I})\theta - \alpha (-2\mathbf{X}^\top \mathbf{y} + 2\mathbf{X}^\top \mathbf{X}\theta)$
- $\theta = \underbrace{(1 - 2\alpha\mu \mathbf{I})\theta}_{\text{Shrinking } \theta} - \alpha (-2\mathbf{X}^\top \mathbf{y} + 2\mathbf{X}^\top \mathbf{X}\theta)$

- Contrast with update equation for unregularised regression:

- $\theta = \underbrace{\theta}_{\text{No Shrinking } \theta} - \alpha (-2\mathbf{X}^\top \mathbf{y} + 2\mathbf{X}^\top \mathbf{X}\theta)$

Least squares objective function

