Ridge Regression

Nipun Batra

February 4, 2020

IIT Gandhinagar

Introduction

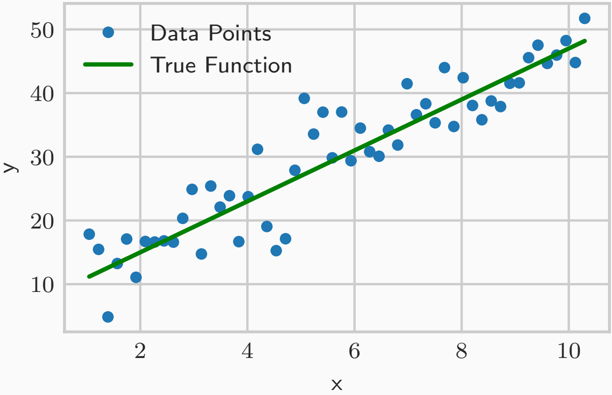
A know measure of over-fitting can be the magnitude of the coefficient.

Introduction

A know measure of over-fitting can be the magnitude of the coefficient.

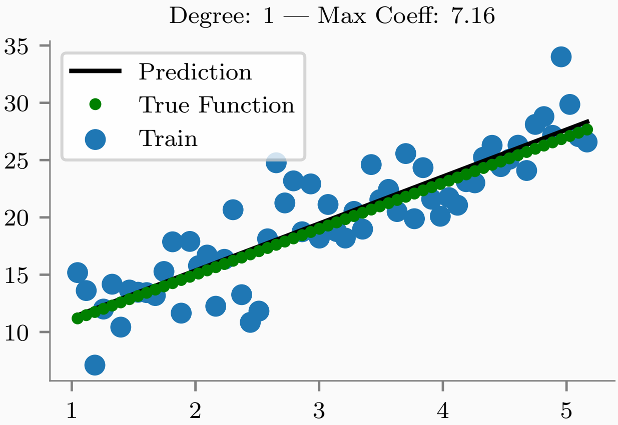
In $f(x) = c_0 + c_1x + c_2x^2 + \dots$ it is $\max|c_i|$

Introduction



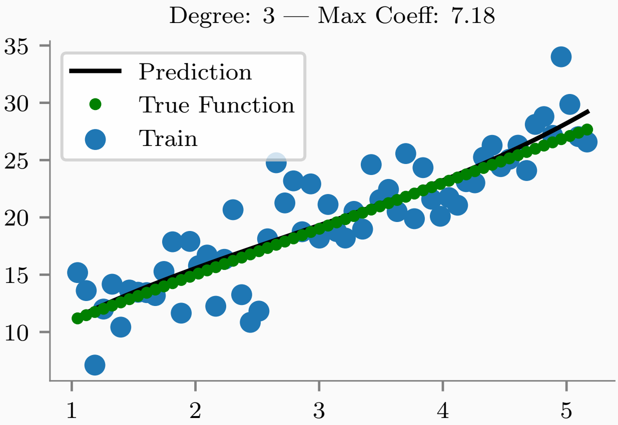
Base Data Set

Introduction

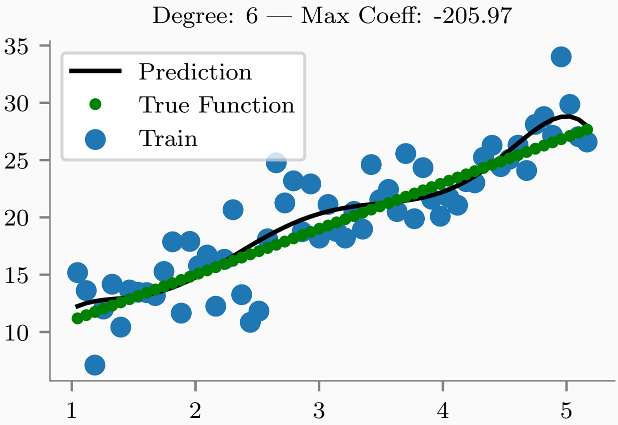


Fit with Degree 1

Introduction

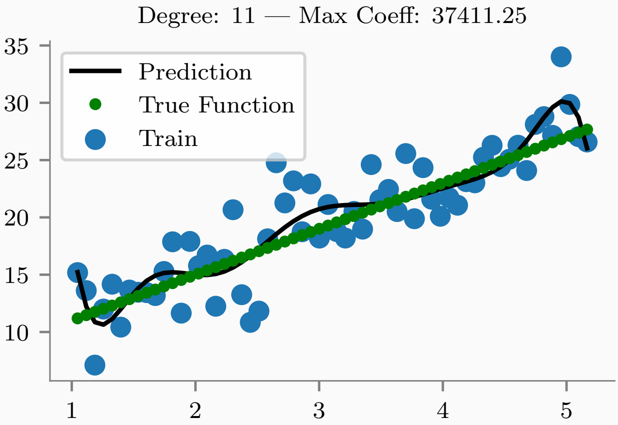


Fit with Degree 3



Fit with Degree 6

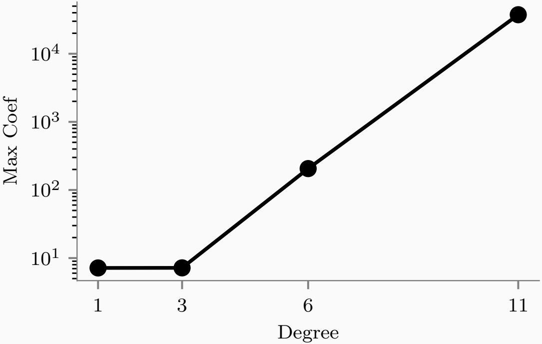
Introduction



Fit with Degree 11

Introduction

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



Trend of the coefficients

Introduction

To prevent over fitting we place penalties on large θ_i

Introduction

To prevent over fitting we place penalties on large θ_i

Objective

$$\text{Minimize } (y - X\theta)^T (y - X\theta)$$
$$\text{s.t. } \theta^T \theta \leq S$$

Introduction

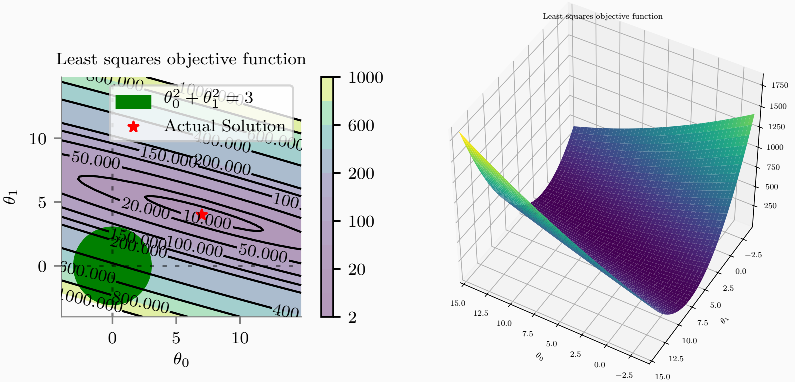
To prevent over fitting we place penalties on large θ_i

Objective

$$\text{Minimize } (y - X\theta)^T (y - X\theta)$$
$$\text{s.t. } \theta^T\theta \leq S$$

This is equivalent to

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



(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

$$\text{Minimize } (y - X\theta)^T (y - X\theta)$$
$$\text{s.t. } \theta^T\theta \leq S$$
$$L(\theta, \mu) = (y - X\theta)^T (y - X\theta) + \mu\left(\theta^T\theta - S\right)$$

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

KKT Conditions

To implement this we use KKT Conditions

$$\text{Minimize } (y - X\theta)^T (y - X\theta)$$
$$\text{s.t. } \theta^T\theta \leq S$$
$$L(\theta, \mu) = (y - X\theta)^T (y - X\theta) + \mu\left(\theta^T\theta - S\right)$$

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

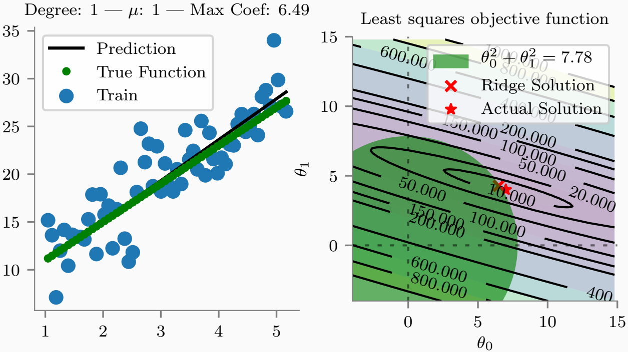
$$\text{Minimize } (y - X\theta)^T (y - X\theta)$$
$$\text{s.t. } \theta^T\theta \leq S$$
$$L(\theta, \mu) = (y - X\theta)^T (y - X\theta) + \mu\left(\theta^T\theta - S\right)$$

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

If $\mu = 0$
There is no regularization
No effect on constraint

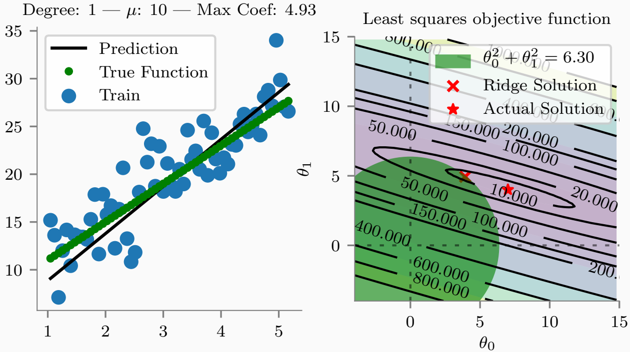
If $\mu \neq 0$
$\implies \theta^T\theta - S = 0$

Effect of μ



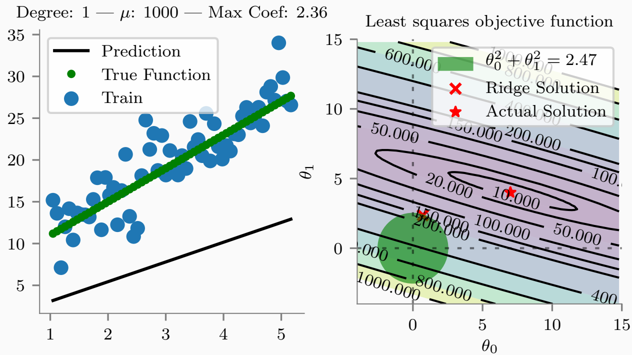
$\mu = 1$

Effect of μ



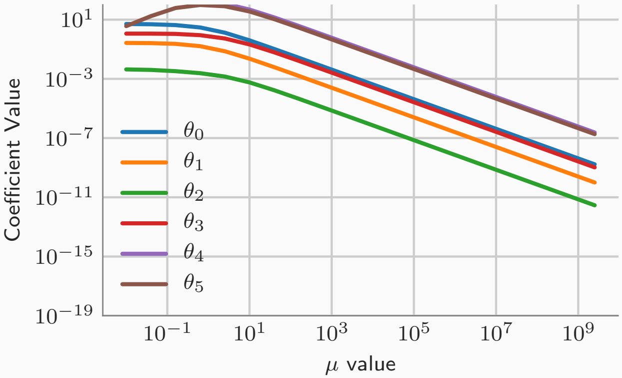
$\mu = 10$

Effect of μ



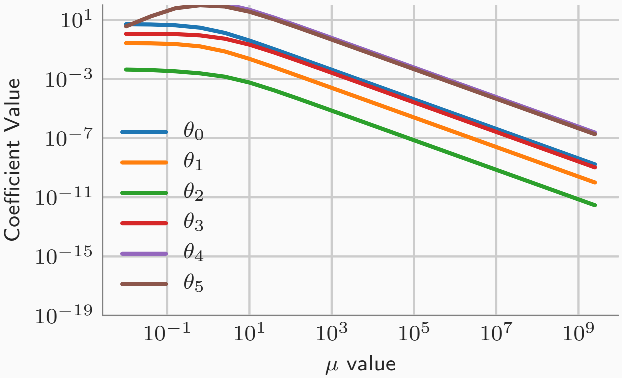
$\mu = 1000$

Effect of μ - Regularization of Parameters



Comparing the magnitudes of the coefficients with varying μ
(on the *Real Estate Data Set*)

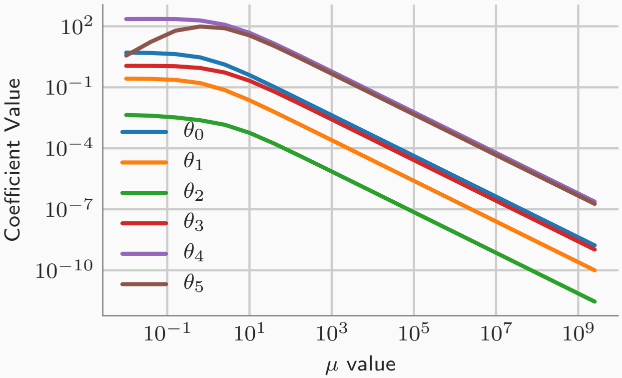
Effect of μ - Regularization of Parameters



Comparing the magnitudes of the coefficients with varying μ (on the *Real Estate Data Set*)

Are θ_i all zero for high μ?

Effect of μ - Regularization of Parameters



Comparing the magnitudes of the coefficients with varying μ (on the *Real Estate Data Set*)

Analytical Method

Ridge Objective

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

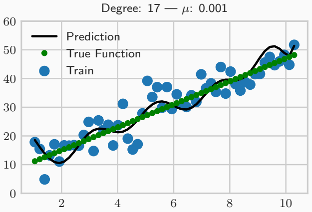
$$\frac{\partial L(\theta, \mu)}{\partial \theta} = 0$$

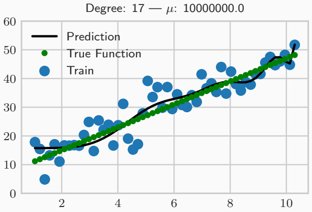
$$\frac{\partial}{\partial \theta} \left\{ y^T y - 2y^T X\theta + \theta^T X^T X\theta \right\} + \frac{\partial}{\partial \theta} \mu\theta^T\theta = 0$$

$$\implies -X^T y + \left(X^T X + \mu I \right) \theta = 0$$

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

Bias/Variance





Fit High Order Polynomial

$\implies$ high variance

$\implies$ $\mu \to 0$

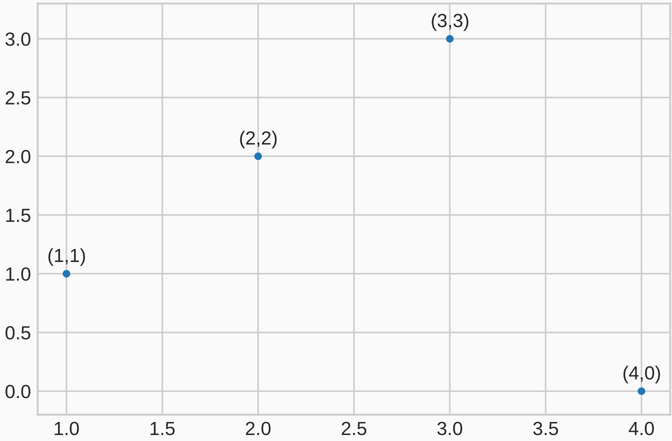
Fit High Order Polynomial

$\implies$ low variance

$\implies$ $\mu \to \infty$

Example

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



Example: Unregularized

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

Example: Unregularized

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

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

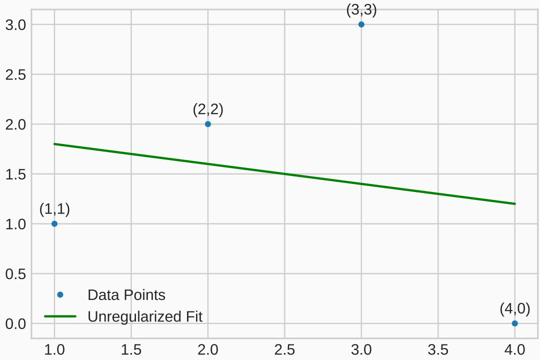
$$(X^T X)^{-1} = \frac{1}{20}\begin{bmatrix} 30 & -10 \\ -10 & 4 \end{bmatrix}$$

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

Example: Unregularized

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

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



Example: Regularized

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

Example: Regularized

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

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

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

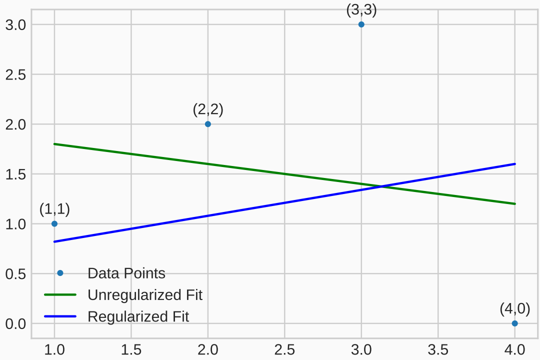
$$(X^T X + \mu I)^{-1} = \frac{1}{92} \begin{bmatrix} 32 & -10 \\ -10 & 6 \end{bmatrix}$$

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

Example: Regularized

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

$$\begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix} = \begin{bmatrix} 0.56 \\ 0.26 \end{bmatrix}$$



Multi-collinearity

$(X^T X)^{-1}$ is not computable when $|X^T X| = 0$.

This was a drawback of using linear regression

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

The matrix X is not full rank.

Multi-collinearity

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

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

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

Multi-collinearity

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

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

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

Another interpretation of “regularisation”

Extension of the analytical model

For ridge with no penalty on θ_0

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

where,

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

TRUE FUNCTION: $y = x$

x	y
1	1
2	2

TRUE FUNCTION: $y = 100 + x$

x	y
1	101
2	102

TRUE FUNCTION: $y = x$

x_0	x	y
1	1	1
1	2	2

ADD COLUMN OF 1s

TRUE FUNCTION: $y = 100 + x$

x_0	x	y
1	1	101
1	2	102

TRUE FUNCTION: $y = x$

x_0	x	y
1	1	1
1	2	2

CASE I: $I = I_{2\times2}$

$\mu = 100$

TRUE FUNCTION: $y = 100 + x$

x_0	x	y
1	1	101
1	2	102

TRUE FUNCTION: $y = x$

x_0	x	y
1	1	1
1	2	2

CASE I: $I = I_{2\times 2}$

$\mu = 100$

TRUE FUNCTION: $y = 100 + x$

x_0	x	y
1	1	101
1	2	102

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

$$\hat{\theta} = [0.02 \quad 0.046]^T$$

TRUE FUNCTION: $y = x$

x_0	x	y
1	1	1
1	2	2

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

$$\hat{\theta} = [0.02 \quad 0.046]^T$$

CASE I: $I = I_{2\times 2}$

$\mu = 100$

TRUE FUNCTION: $y = 100 + x$

x_0	x	y
1	1	101
1	2	102

$$\hat{\theta} = [1.9 \quad 2.8]^T$$

TRUE FUNCTION: $y = x$

x_0	x	y
1	1	1
1	2	2

CASE I: $I = I_{2\times2}$

$\mu = 100$

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

$$\hat{\theta} = [0.02 \quad 0.046]^T$$

$$\hat{y}(0) = 0.02$$

TRUE FUNCTION: $y = 100 + x$

x_0	x	y
1	1	101
1	2	102

$$\hat{\theta} = [1.9 \quad 2.8]^T$$

$$\hat{y}(0) = 1.9$$

TRUE FUNCTION: $y = x$

x_0	x	y
1	1	1
1	2	2

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

$$\hat{\theta} = [1.49, 0.0049]^T$$

$$\hat{y}(0) = 1.49$$

CASE 2: USE I^*

$\mu = 100$

TRUE FUNCTION: $y = 100 + x$

x_0	x	y
1	1	101
1	2	102

$$\hat{\theta} = [101, \sim 0]^T$$

$$\hat{y}(0) = 101$$

TRUE FUNCTION: $y = x$

x_0	x	y
1	1	1
1	2	2

CASE 2 : USE I^*

$\mu = 100$

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

$$\hat{\theta} = [1.49 \ , 0.0049]^T$$

$$\hat{y}(0) = 1.49$$

$\Longrightarrow$ TENDS TOWARDS $\bar{y}$

TRUE FUNCTION: $y = 100 + x$

x_0	x	y
1	1	101
1	2	102

$$\hat{\theta} = [101 \ , \sim 0]^T$$

$$\hat{y}(0) = 101$$

TENDS TOWARDS $\bar{y}$

TRUE FUNCTION: $y = x$

x_0	x	y
1	1	1
1	2	2

CASE 2 : USE I^*

$\mu = 100$

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

$$\hat{\theta} = [1.49, 0.0049]^T$$

$$\hat{y}(0) = 1.49$$

$\Rightarrow$ TENDS TOWARDS $\bar{y}$

TRUE FUNCTION: $y = 100 + x$

x_0	x	y
1	1	101
1	2	102

$$\hat{\theta} = [101, \sim 0]^T$$

$$\hat{y}(0) = 101$$

TENDS TOWARDS $\bar{y}$

ALTERNATIVE APPROACH

(1) TRANSFORM $y \rightarrow y'$ s.t. $\overline{y'} = 0$

$$y' = y - \overline{y}$$

ALTERNATIVE APPROACH

① TRANSFORM $y \rightarrow y'$ s.t. $\overline{y'} = 0$

$$y' = y - \overline{y}$$

② TRAIN ON $\{(x_i, y'_i) \; \forall i\}$

ALTERNATIVE APPROACH

① TRANSFORM $y \rightarrow y'$ s.t. $\overline{y'} = 0$

$$y' = y - \overline{y}$$

② TRAIN ON $\{(x_i, y'_i) \ \forall i\}$

③ PREDICT y' ON TEST $x_{TEST}(i)$ AND ADD $\overline{y}$ to get $\hat{y}$

ALTERNATIVE APPROACH

① TRANSFORM $y \rightarrow y'$ s.t. $\overline{y'} = 0$

$$y' = y - \overline{y}$$

② TRAIN ON $\{(x_i, y'_i)\ \forall i\}$

③ PREDICT y' ON TEST $x_{TEST}(i)$ AND ADD $\overline{y}$ to get $\hat{y}$

NO NEED TO USE I^* HERE

TRUE FUNCTION: $y = 100 + x$

x_0	x	y
1	1	101
1	2	102

$\bar{y} = 101.5$

TRUE FUNCTION: $y = 100 + x$

x_0	x	y	y'
1	1	101	-0.5
1	2	102	0.5

TRUE FUNCTION: $y = 100 + x$

x_0	x	y	y'
1	1	101	-0.5
1	2	102	0.5

$$\hat{\theta} = (X^T X + \mu I)^{-1} X^T y'$$

$$= [-0.0001, \; 0.0047]^T$$

TRUE FUNCTION: $y = 100 + x$

x_0	x	y	y'
1	1	101	-0.5
1	2	102	0.5

$$\hat{\theta} = (X^T X + \mu I)^{-1} X^T y'$$

$$= [-0.0001, \quad 0.0047]^T$$

$$\hat{y}'(0) = 0$$

$$\hat{y}(0) = \hat{y}'(0) + \bar{y} = 101.5$$

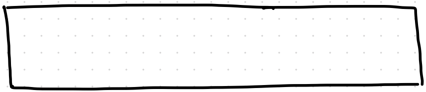
RIDGE REGRESSION

WHAT 'μ' to use?

RIDGE REGRESSION

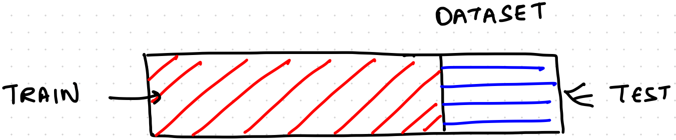
WHAT 'μ' to use?

DATASET



RIDGE REGRESSION

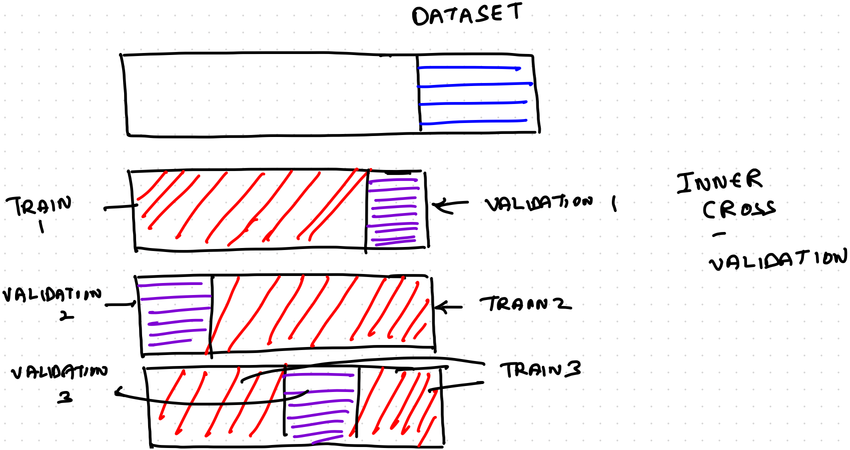
WHAT 'μ' to use?



CROSS-VALIDATN (OUTER LOOP)

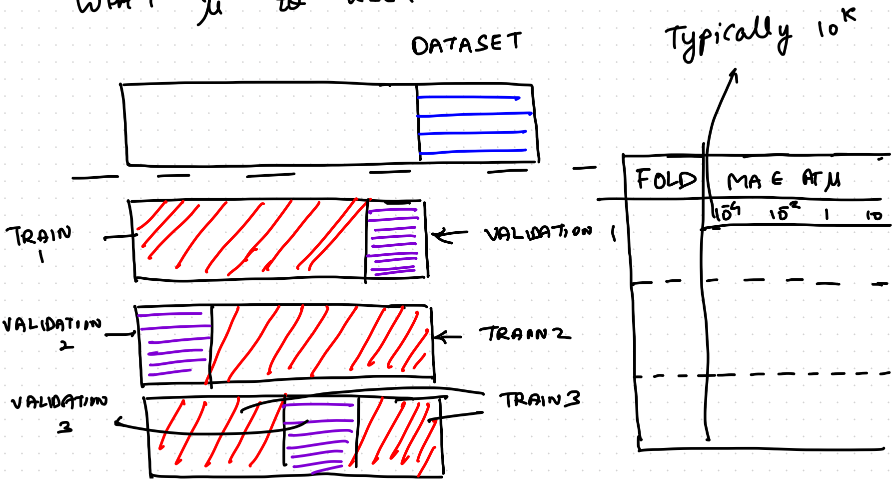
RIDGE REGRESSION

WHAT 'μ' to use?



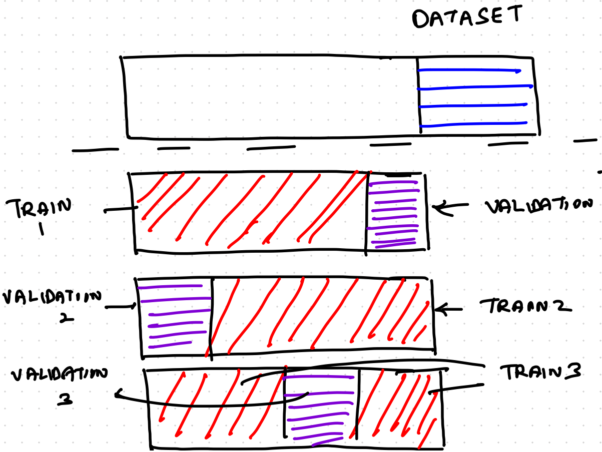
RIDGE REGRESSION

WHAT 'μ' to use?



RIDGE REGRESSION

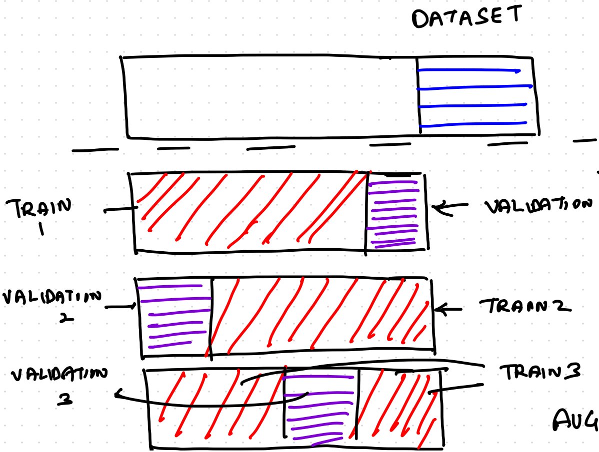
WHAT 'μ' to use?



FOLD	MAE AT μ			
	10^{-4}	10^{-2}	1	10
1	20	15	20	30
2	18	19	20	30
3	12	12	14	30

RIDGE REGRESSION

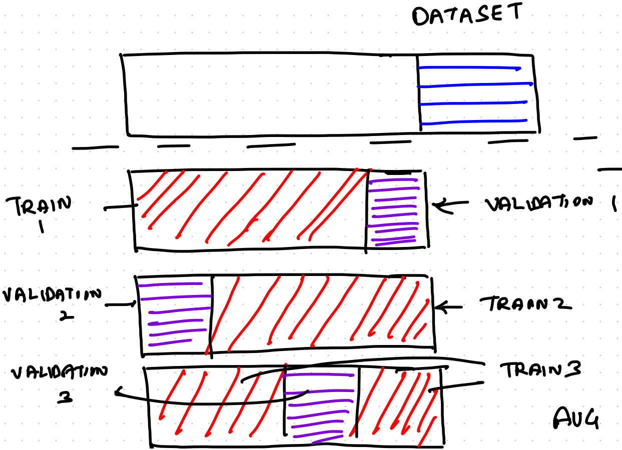
WHAT 'μ' to use?



FOLD	MAE AT μ			
	10^{-4}	10^{-2}	1	10
1	20	15	20	30
2	18	19	20	30
3	12	12	14	30
AVG	17	15	18	30

RIDGE REGRESSION

WHAT μ to use?

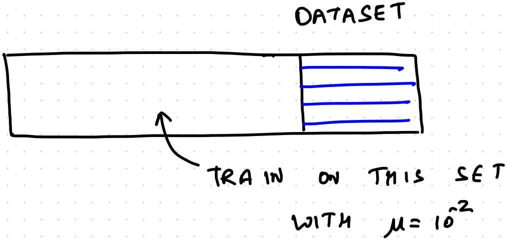


$\mu = 10^{-2}$ GIVES LOWEST VALIDATION ERROR

FOLD	MAE AT μ			
	10^{-4}	10^{-2}	1	10
1	20	15	20	30
2	18	19	20	30
3	12	12	14	30
AVG	17	15	18	30

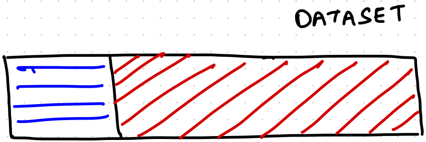
RIDGE REGRESSION

WHAT μ to use?



RIDGE REGRESSION

WHAT 'μ' to use?



REPEAT PROCEDURE

WITH OTHER 'OUTER LOOP' FOLDS

Ridge Solution using Gradient Descent

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

Ridge Solution using Gradient Descent

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

Ridge Solution using Gradient Descent

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

Ridge Solution using Gradient Descent

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

Ridge Solution using Gradient Descent

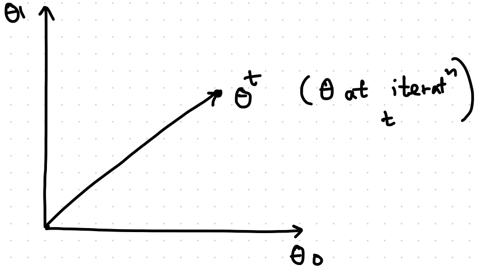
- $\theta = \theta - \alpha \frac{\partial}{\partial\theta}((y - X\theta)^\top (y - X\theta) + \mu\theta^\top\theta)$
- $\theta = \theta - \alpha(-2X^\top y + 2X^\top X\theta + 2\mu I\theta)$
- $\theta = (1 - 2\alpha\mu I)\theta - \alpha(-2X^\top y + 2X^\top X\theta)$
- $\theta = \underbrace{(1 - 2\alpha\mu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

Ridge Solution using Gradient Descent

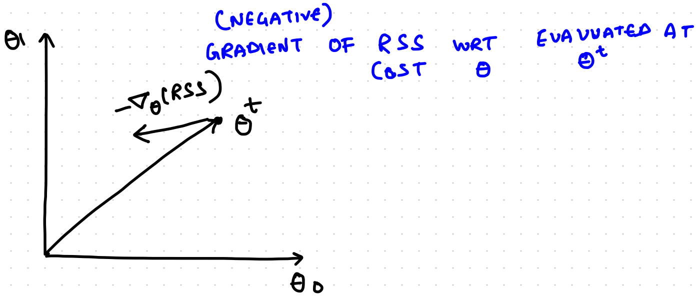
- $\theta = \theta - \alpha \frac{\partial}{\partial \theta}((y - X\theta)^\top (y - X\theta) + \mu\theta^\top\theta)$
- $\theta = \theta - \alpha(-2X^\top y + 2X^\top X\theta + 2\mu I\theta)$
- $\theta = (1 - 2\alpha\mu I)\theta - \alpha(-2X^\top y + 2X^\top X\theta)$
- $\theta = \underbrace{(1 - 2\alpha\mu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

- Contrast with update equation for unregularised regression:
- $\theta = \underbrace{\theta}_{\text{No Shrinking } \theta} - \alpha(-2X^\top y + 2X^\top X\theta)$

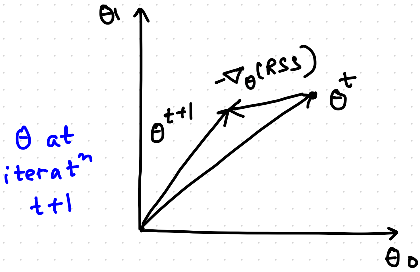
GD UPDATE FOR
UNREG. LINEAR REG.



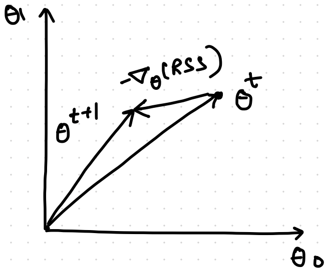
GD UPDATE FOR
UNREG. LINEAR REG.



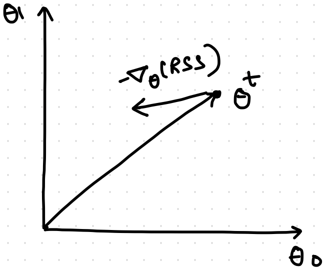
GD UPDATE FOR UNREG. LINEAR REG.



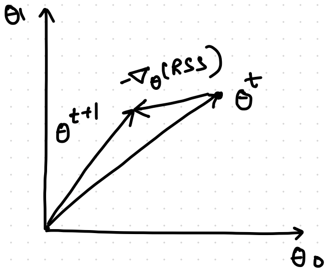
GD UPDATE FOR UNREG. LINEAR REG.



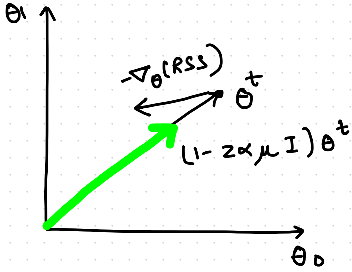
GD UPDATE FOR RIDGE REGRESSION



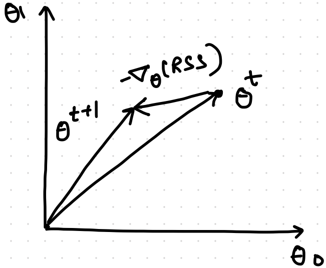
GD UPDATE FOR UNREG. LINEAR REG.



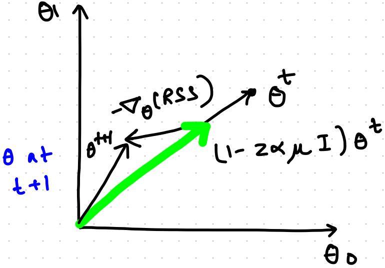
GD UPDATE FOR RIDGE REGRESSION



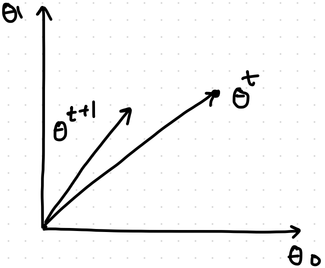
GD UPDATE FOR UNREG. LINEAR REG.



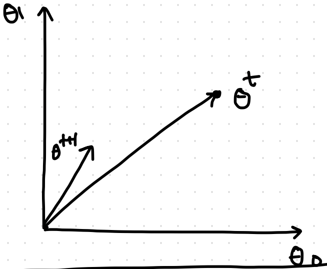
GD UPDATE FOR RIDGE REGRESSION



GD UPDATE FOR UNREG. LINEAR REG.



GD UPDATE FOR RIDGE REGRESSION



Clearly, $\|\theta^{t+1}_{RIDGE}\|_2 \leq \|\theta^{t+1}_{UNREG}\|_2^2$