Last class - August 8

1) Quick recap
2) Overview of GD, SGD, MBGD
3) Polynomial Regression
4) Underfit vs Overfit
5) Bias Variance trade-off

Today's class

1) Regularization
2) L1 and L2 Regularization
3) Parameter vs Hyper-parameter
4) Cross - Validation
5) K-fold CV

$m = 2^{10}$

$BS = 2^{5}$

$$N.B = \frac{2^{10}}{2^{5}} = 2^{5}$$

$$\text{num\_iter} = \text{epochs} \times N.B$$

$$= 2 \times 2^{5}$$

Weight update is happening [grad descent loop is running

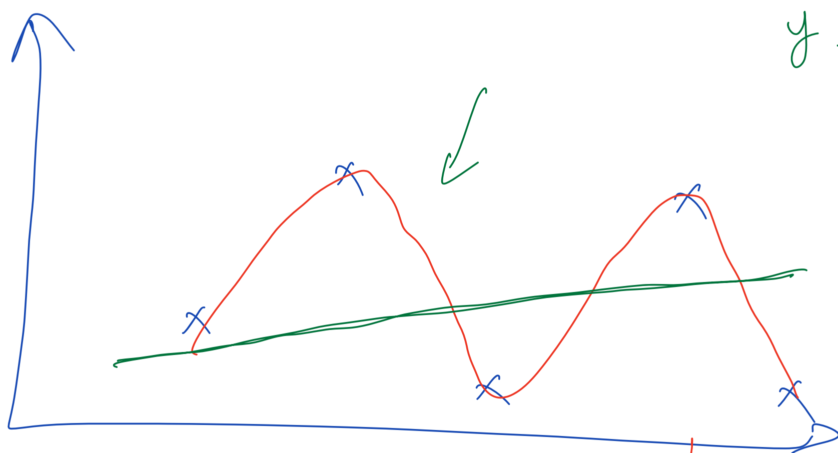
$$x_{ij} \rightarrow \begin{array}{ccc} x_1 & x_1^2 & x_1^3 \\ \downarrow & \downarrow & \downarrow \\ w_1 & w_2 & w_3 \end{array}$$

$$|w_3| > |w_2| > |w_1|$$

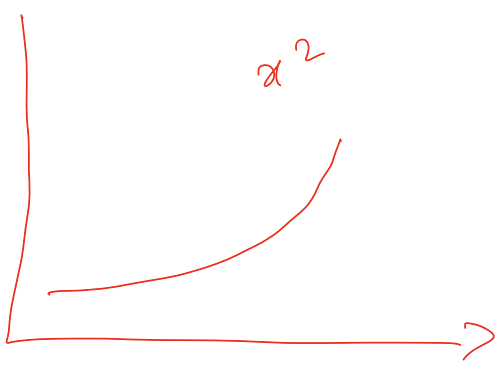
$$SS(x) = x' = \frac{x - \mu}{\sigma}$$

Ridge Regularization

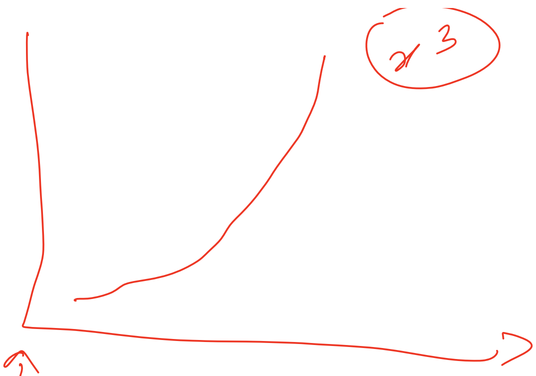
$$y = 1 + x - 0.1x^2 + 0.01x^3$$



$$y = 1 + 5(x) - 5x^2 + 5x^3$$



$$1 + 10x - 10x^2 + 10x^3$$



1 to 2 to 4, 8, 16, 32, . . .

regularization process

overfitting

$w_1 \quad w_2 \quad w_3 \quad w_4 \qquad \sum w_j \quad \times$

$\boxed{100 \quad -50 \quad -25 \quad -25} \to 0$

$\sum w_j^4 \qquad \sum w_j^2 \qquad \sum w_j^3 \quad \times$

Occam's razor

$$w = 1.5, \quad 2.0, \quad 3.5, \quad 4$$

$$\downarrow \qquad \downarrow \qquad \downarrow \qquad \downarrow$$

$$0.1 \quad 0.2 \quad 0.3 \quad 0.4$$

$L2$

$$w = 1.5, \quad 2.0, \quad 3.5, \quad 4$$

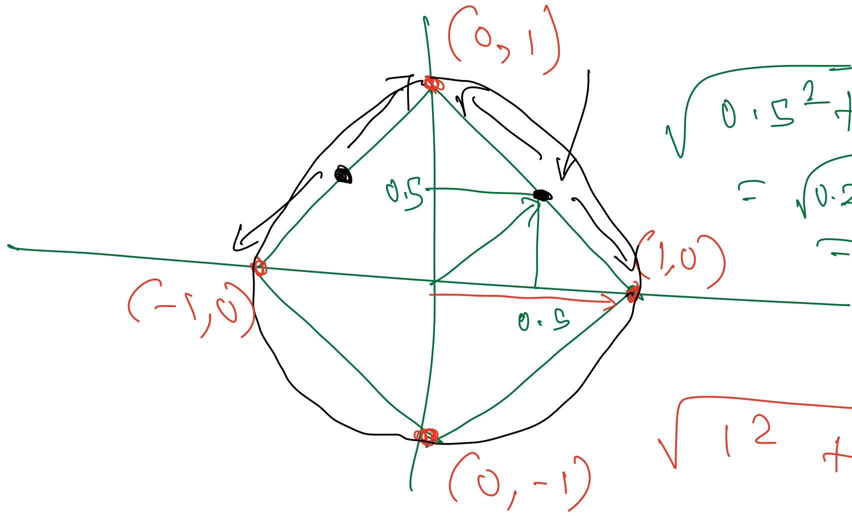
$$\downarrow \qquad \downarrow \qquad \downarrow \qquad \downarrow$$

$$0 \quad 2.0 \quad 0 \quad 4.0$$

$1.5 + 2 + 3.5 + 4 = 11$

$$\sum_{j=1}^{d} |w_j|$$

$\rightarrow 2 + 4 = 6$



$\sqrt{0.5^2 + 0.5^2}$

$= \sqrt{0.25 + 0.25}$

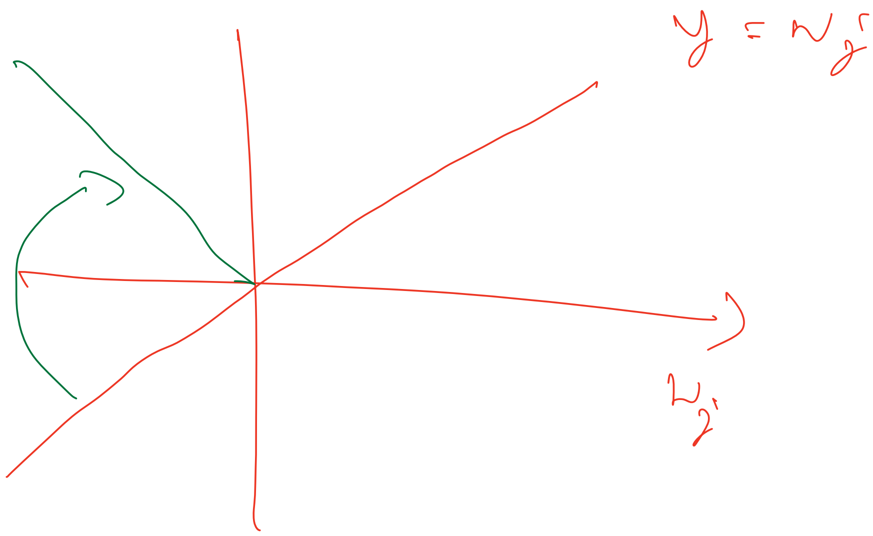
$= \sqrt{0.5}$

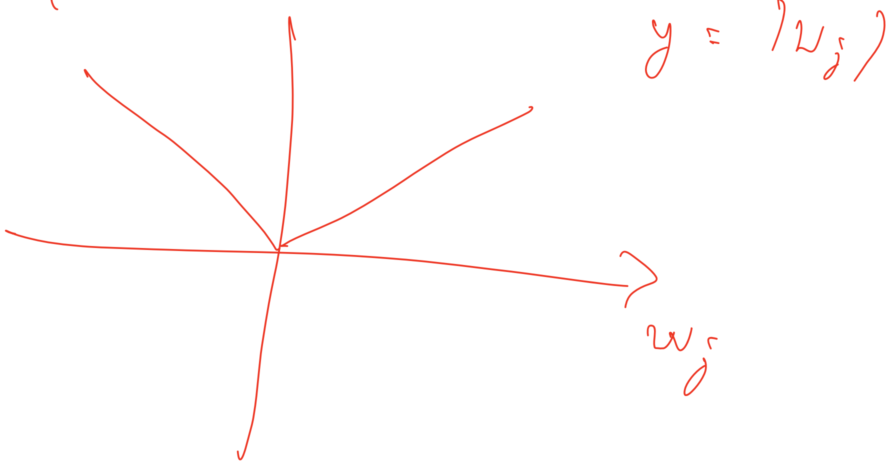
$=$

$\sqrt{1^2 + 0^2}$

$= \sqrt{1} = 1$

$1 > \sqrt{0.5}$ ✓





21

$W_j = 0.5$

$\boxed{\eta = 0.1}$

w_j :

$0.5 - 0.1 \times 1$

$= 0.4$

$$\frac{\partial l(w_j)}{\partial w_j}$$

$0.4 - 0.1 \times 1$

$= 0.3$

$0.3 - 0.1 \times 1 = 0.2$

$0.2 - 0.1 \times 1 = 0.1$

$0.1 - 0.1 \times 1 = 0$

$$\frac{\partial w_j^2}{\partial w_j} = 2 w_j$$

<u>L2:</u>

$w_j = 0.5 - 0.1 \times (2 \times 0.5) \downarrow$

$= 0.5 - 0.1$

$= 0.4$

$0.4 - 0.1 \times (2 \times 0.4) \downarrow$

$0.4 - 0.08 = 0.32$

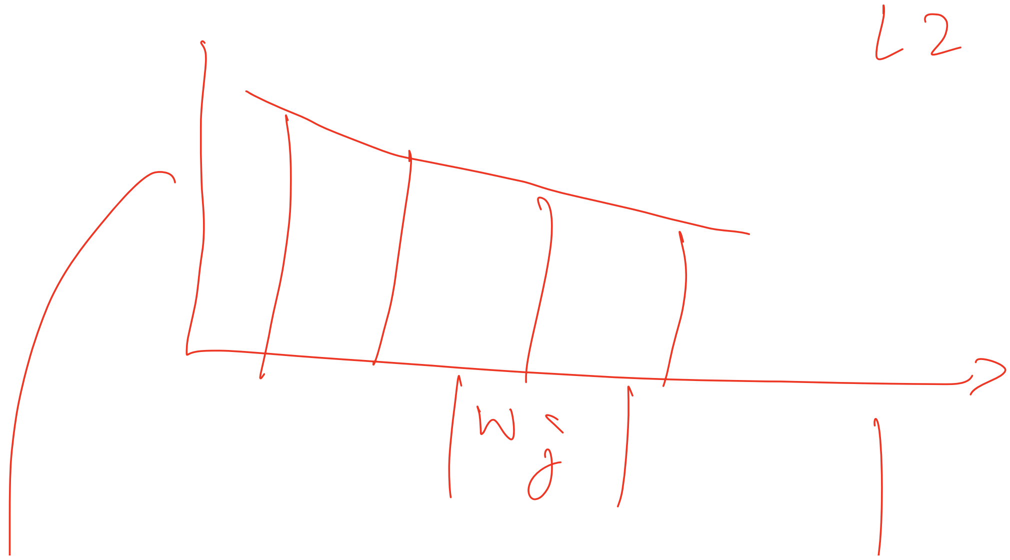
$0.32 - 0.1 \times (2 \times 0.32) \downarrow$

$= 0.32 - 0.064$

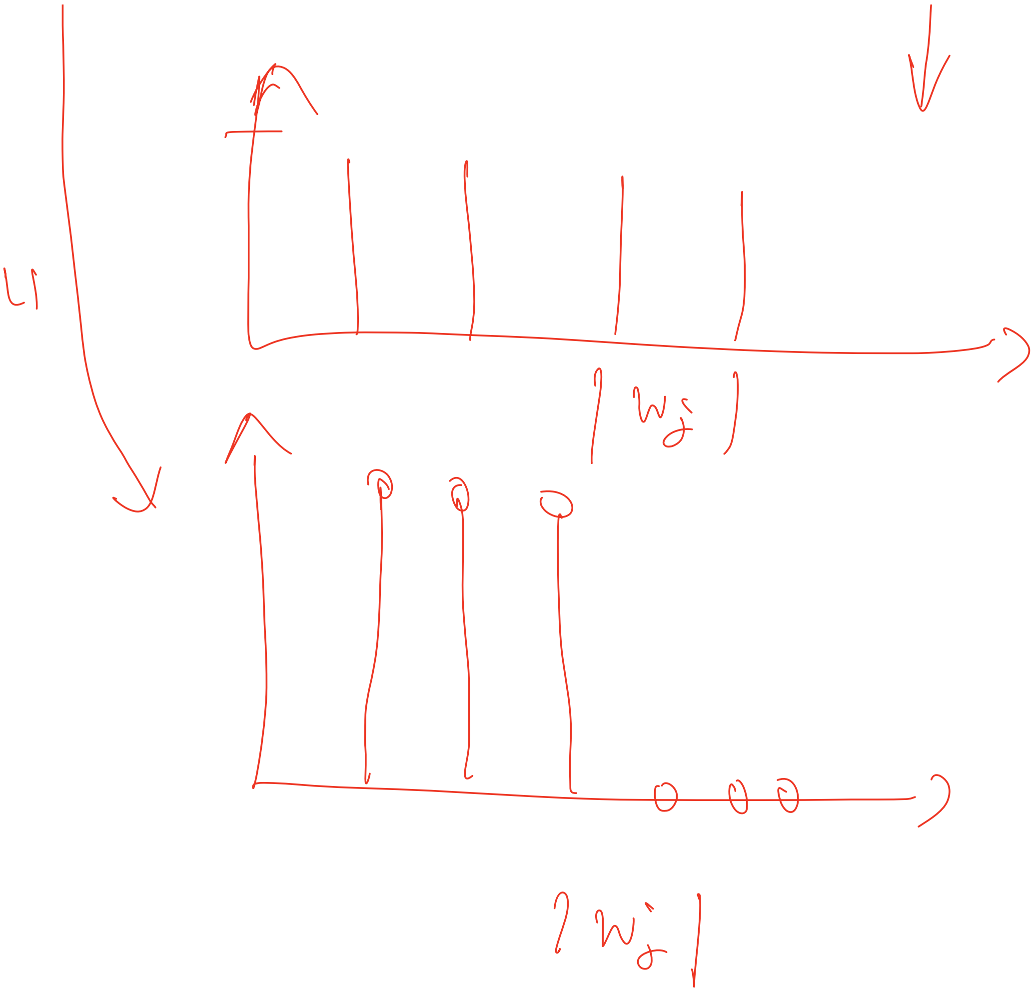
$= \sim 0.2\cancel{8}.$

Elastic net: combination of L1 & L2

L2



L2



Loss = MSE

$$\sum (\hat{y} - y)^2 + \lambda_2 \left(\sum_{j=1}^{d} w_j^2 \right)$$

$$\sum (\hat{y} - y)^2 + \lambda_1 \sum_{j=1}^{d} |w_j|$$

$\lambda = 0.1$

1

2

$$\boxed{\lambda_2 = 10^3}$$

$$\boxed{\lambda_2 = 10^6}$$

λ high $\rightarrow$ underfitting

very less learning

λ low $\rightarrow$ overfitting

too much learning

$\lambda_1 / \lambda_2 = [0.01, 0.05, 0.1, 0.5, 1.0]$

$\lambda_1 \rightarrow$ min RMSE error on test data

parameters $\rightarrow$ $w_0, \underbrace{w_1, w_2, \ldots w_d}_{w}$

w_0: b

hyper-params → η, no. of epochs for gradient descent, λ_1, λ_2, degree d of model

regularization

Big data

→ 1M data-points

60% → train
20% → val
20% → test

randomly

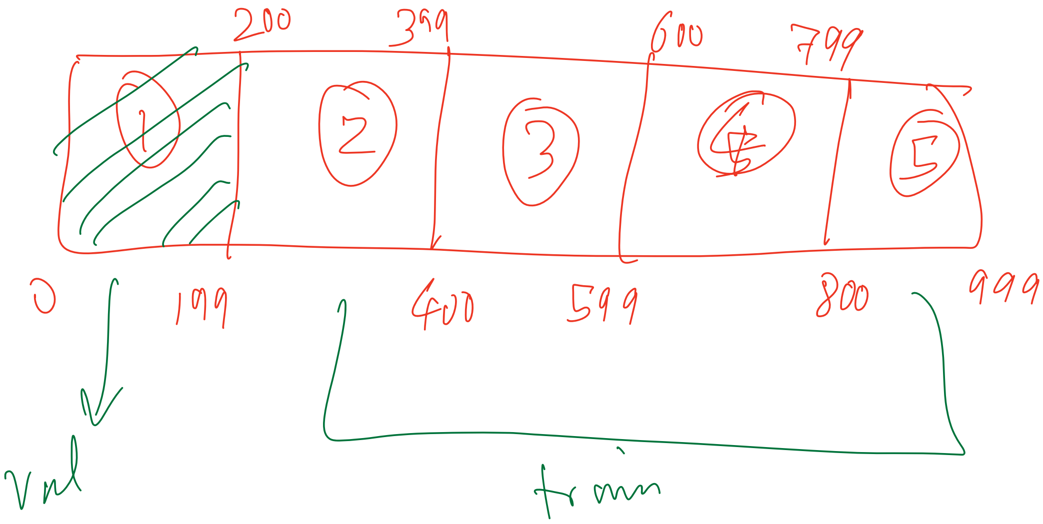
1000 data-points

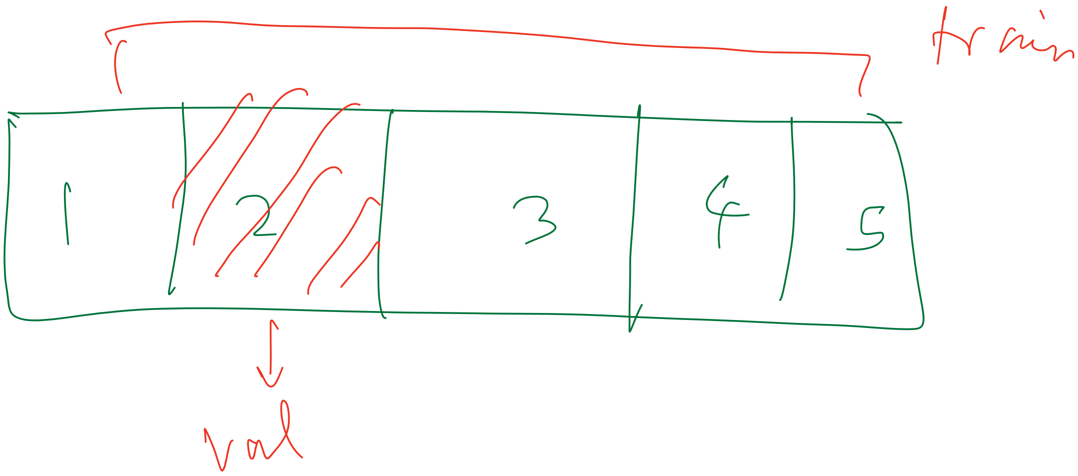
~~1600~~ → train
$\boxed{200}$ → val
$\boxed{200}$ → test

1000 data points 0 to 999

↓ 5 folds

5-fold CV





$$\left[\lambda_1,\ \lambda_2,\ \eta,\ d,\ \text{epoch}\right]$$

↓

hyper-param set

for every

5/K

avg of val fold errors

K

1) Train on entire 1000 train data with optimal hyper-param

2) Train on 4 folds individually [4 out of 5 folds] with optimal hyper-param, & then finally avg out the lin reg coefficients.