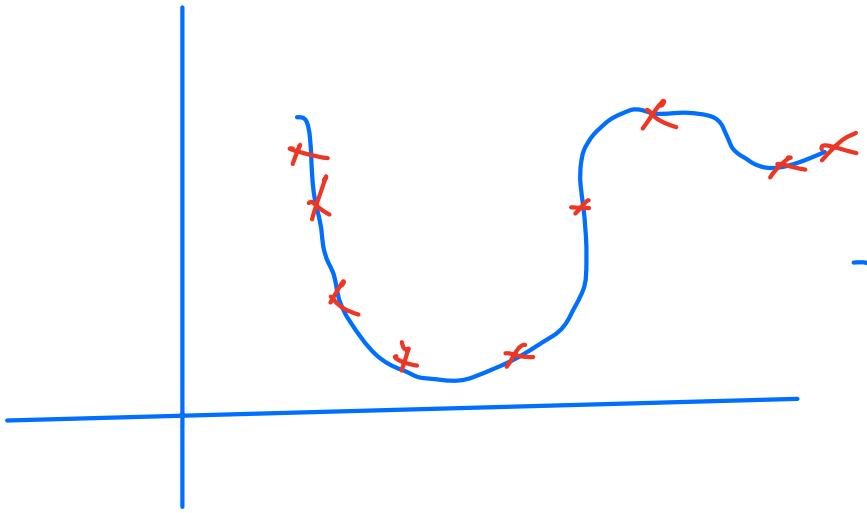
REGULARIZATION

$$y = w_1 x_1 + w_2 x_2^2 + w_3 x_1 x_2 + w_4 x_4^2$$



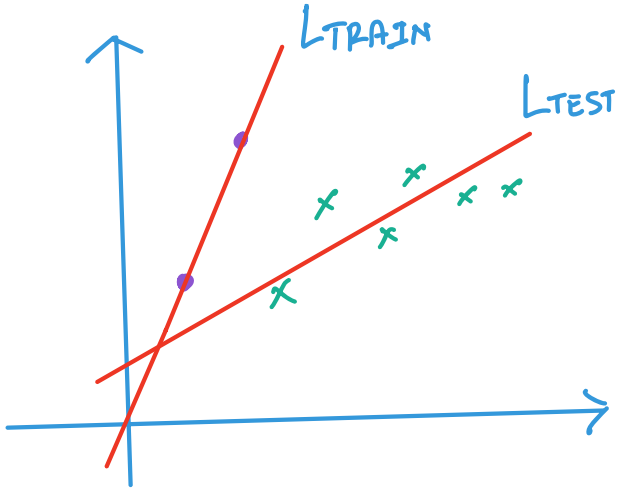
→ Can start overfitting !

- Regularization : Modification we make to the learning algorithm to prevent overfitting

  - Low train error
  - High test error

① Ridge (L2)

Premise of overfitting : Too powerful model fit on too little data



● : Train data

× : Test Data

$$L = \sum_{i=1}^{\hat{} } (y_i - \hat{y}_i)^2$$

With ridge (L2) regularization,

$$L' = \sum_{i=1}^{\hat{} } (y_i - \hat{y}_i)^2 + \lambda w^2$$

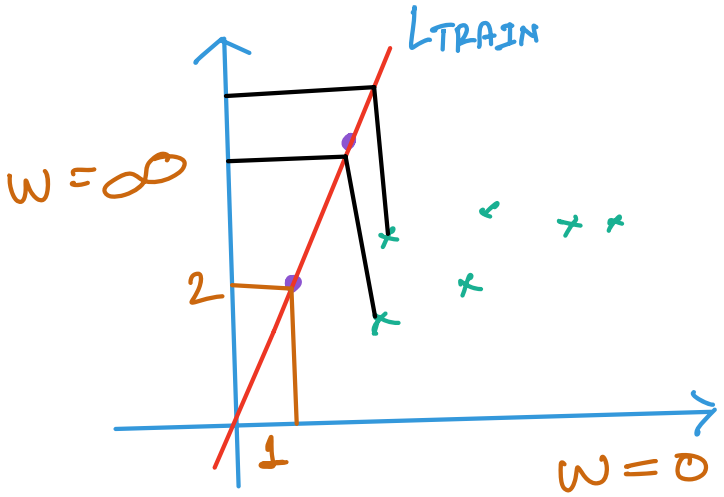
Weight Decay

λ : Hyper param (1 to ∞)

w^2 : Slope

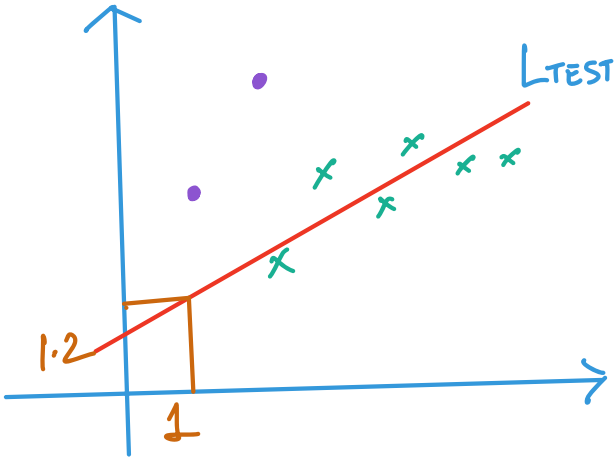
L_{TRAIN} : $y = \underbrace{2x}_{w=2} + 0.3$

$L_{TRAIN} = 0$



L_{TEST} = ↑

L_{TEST} : $y = \underbrace{1.2x}_{\text{Slope has reduced!}} + 0.7$



Old line

$y = 2x + 0.3$

Say $\lambda = 1$,

$L' = m^2 = 4$

New line

$y = 1.2x + 0.7$

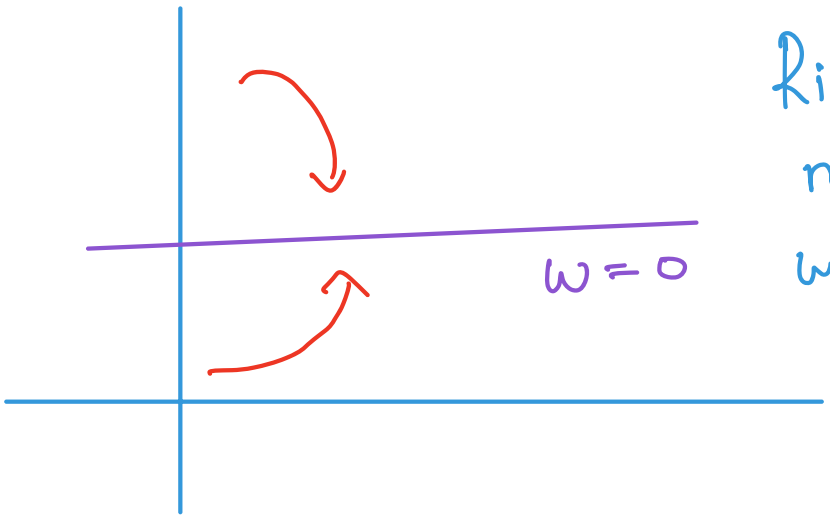
Say $\lambda = 1$,

$L' =$ small error $+ (1.2)^2$

$= 1.5 + 1.44$

$= 2.94$

$\therefore$ New-line preferred !



Ridge tries to make slope 0 while also minimizing $\|\hat{y} - y\|_2^2$

Outcome : ① Reduce reliance on training data

② Slope tending towards 0

$\Rightarrow$ tending towards horizontal line

Multiple features

$$y = w_1 x_1 + w_2 x_2 + c$$

$$L^1 = \| \hat{y} - y \|_2^2 + \lambda (w_1^2 + w_2^2)$$

Lasso (L1) Regularization

$$L = \|\hat{y} - y\|_2^2 + \lambda |w|$$

Multi - Linear

$$y = w_1 x_1 + w_2 x_2 + w_3 x_3$$

$$L = \|\hat{y} - y\|_2^2 + \lambda |w_1 + w_2 + w_3|$$

→ While Ridge will tend towards 0, Lasso will make those terms 0

Ridge

$m = \frac{-}{- + \lambda}$

Lasso

$m = \frac{- \pm \lambda}{-}$

$\Rightarrow$ If $w_2, w_3 = 0$

$\Rightarrow$ We have eliminated unimportant features

$\therefore$ Feature Selection !

Elastic Net

$$L = \|\hat{y} - y\|_2^2 + \lambda_1 \|w\|^2 + \lambda_2 \|w\|$$

Hyperparams		Default
$\lambda = \lambda_1 + \lambda_2$		1
$l1\_ratio = \frac{\lambda_1}{\lambda_1 + \lambda_2}$		0.5

Used when :

1. Unsure which one to use - L1 or L2 ?
2. Large Datasets
3. Multi-collinearity