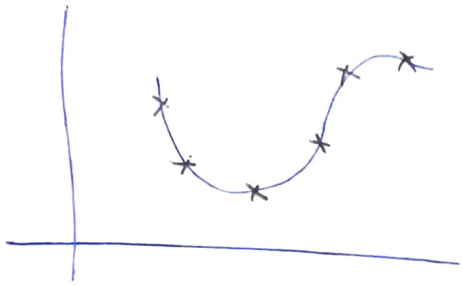
Regularization

As no. of features ↑, $y = w_1 x_1 + w_2 x_2^2 + w_3 x_1 x_2 + w_4 x_4^2$

⇒ Polynomial

Overfitting can happen on data.



→ Fits every pt.
⇒ overfitting

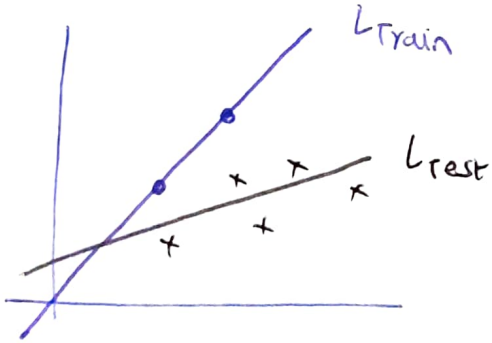
Regularization - Modification we make to the learning algo to prevent overfitting.

Low train error

High test error.

① Ridge (L_2)

Premise of overfitting - Too powerful model fit on too little data.



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

with ridge (L_2) regularization,

✓ ① Reduce power of model → Here.

② ↑ no. of samples.

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

wt. decay

Hyperparam (1 to ∞)

slope

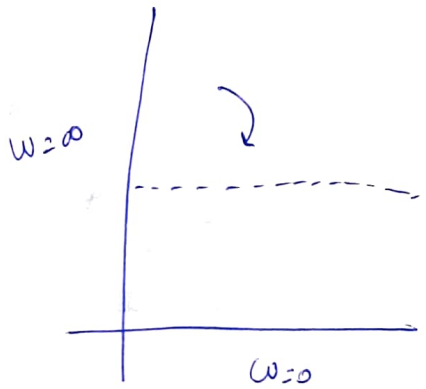
preferring soln where slope is 0.

λ – how much imp. should be given to both parts of addi

1 → equal weightage

≥ 2 → more imp. to wt. decay.

Slope gets closer to 0.



$y = 2x + 0.3$

$y = 1.2x + 0.7$ } slope ↓

slope ↓, bias ↑

Ridge tries to make slope 0 & min. $\|\hat{y} - y\|_2^2$

Outcome – (1) Reduce reliance on traing data

(2) slope tending → 0

Features (multiple)

$y = w_1 x_1 + w_2 x_2 + c$

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

② Lasso (L_1) Regularization

$$L = ||\hat{y} - y||_2^2 + \gamma(w)$$

Multi-Linear

$$y = w_1x_1 + w_2x_2 + w_3x_3$$

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

⇒ while ridge will tend towards 0, lasso will make those terms 0.

Ridge

$w = \frac{-}{- + \gamma}$

as $\gamma \uparrow$, $w \downarrow$

For $w = 0$, $\gamma = \infty$ ↓ not posible.

⇒ value of w can't be 0 ever.

Lasso

$w = \frac{- \pm \gamma}{-}$

as $\gamma \uparrow$, unimportant features will be eliminated.

Lasso
- Helps in feature selection.
- Prevents overfitting.