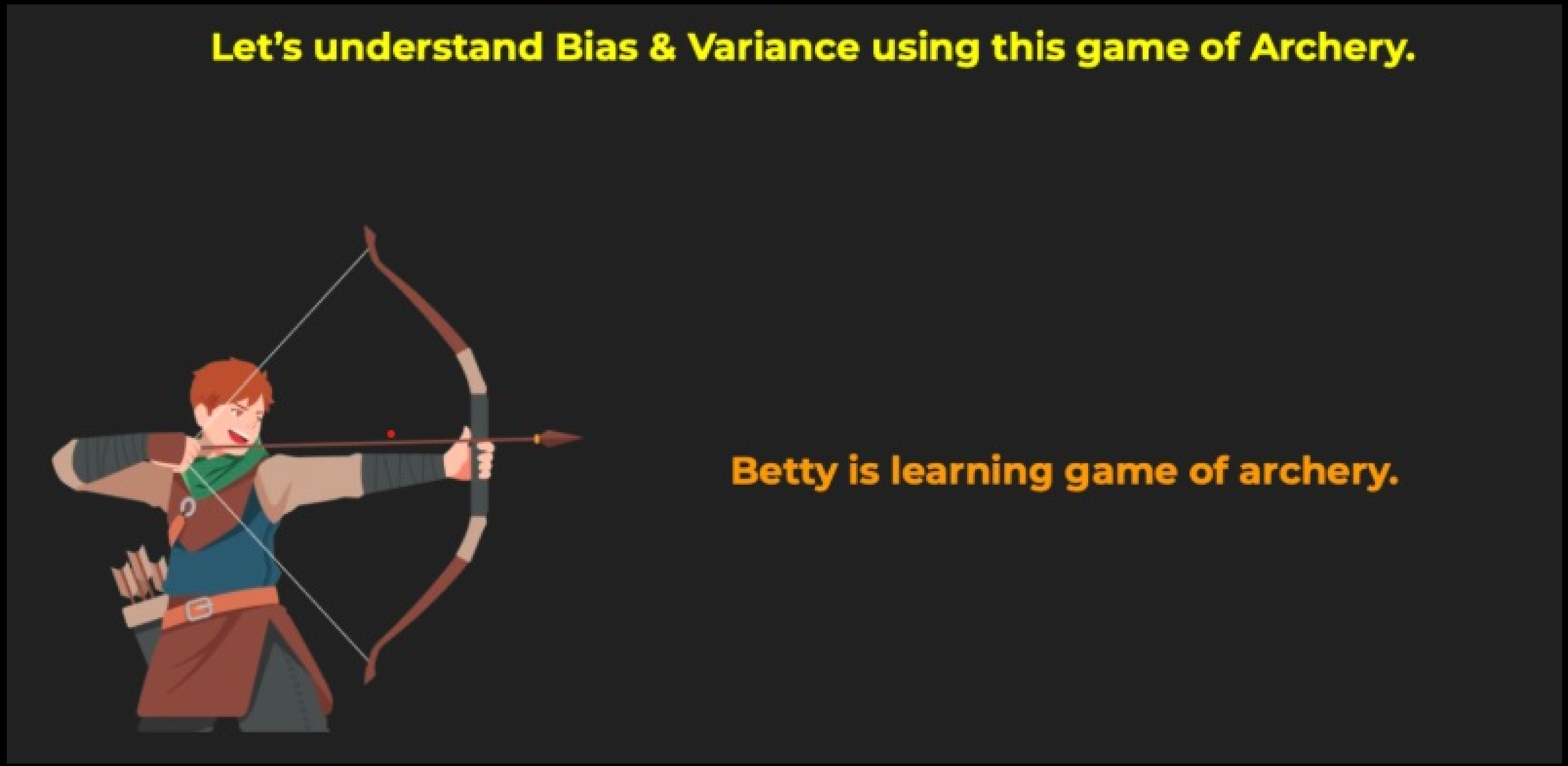
Agenda

1. Bias Variance Tradeoff
2. Regularization
   - L1
   - L2
   - L1+L2
3. Cross Validation

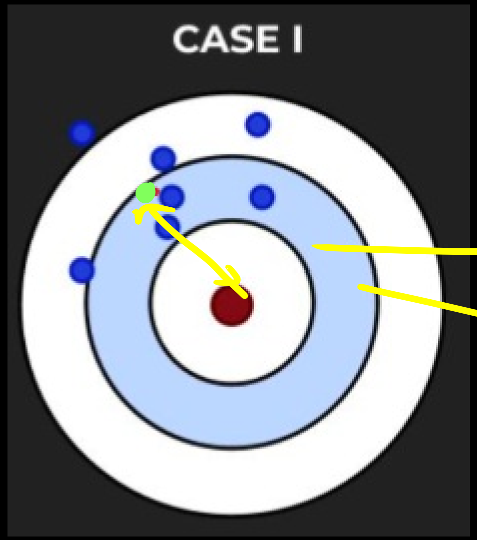


$\hat{f}$ = where you hit

f = Actual target

$$\left\{ \begin{array}{l} \text{Bias} = \left(\text{Mean}(\hat{f}) - f \right) \\ \text{Variance} = \text{Spread of } \hat{f} \end{array} \right.$$

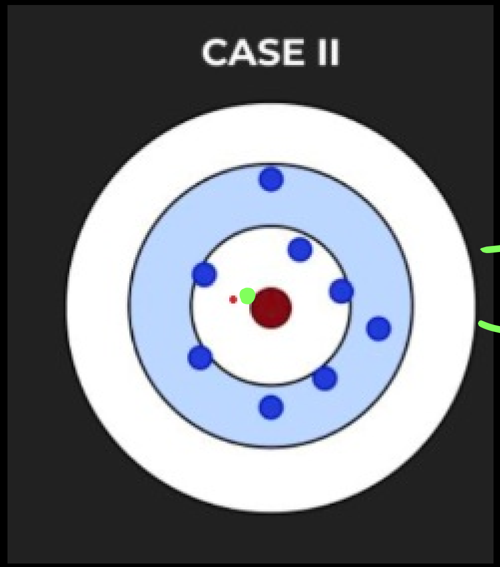
1) Bias
2) Variance



High bias
High Variance

$\hat{f}$ = where you hit
f = Actual target

Overfitting

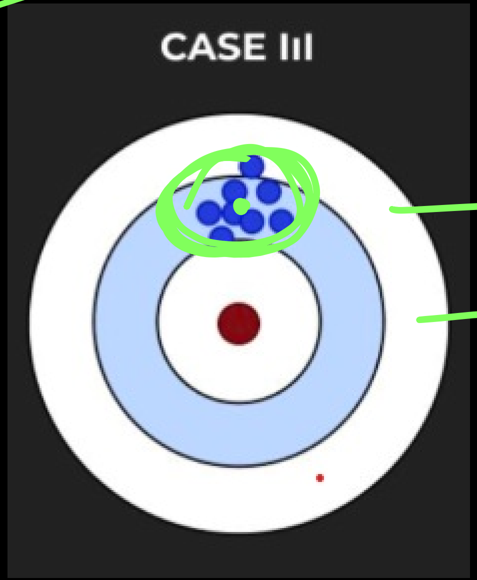


Low bias
High variance

$$\text{Bias} = \left(\underbrace{\text{Mean}(\hat{f}) - f}_{} \right)$$

Variance = Spread of $\hat{f}$

Underfitting



High bias

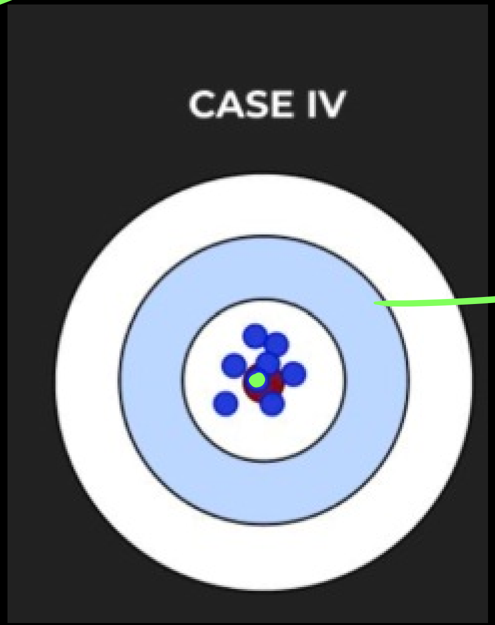
low variance

$\hat{f}$ = where you hit

f = Actual target

$$\text{Bias} = \underbrace{\left(\text{Mean}(\hat{f}) - f\right)}$$

Best fit



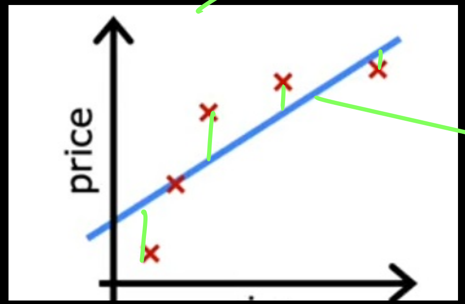
low bias

low Variance

Variance = Spread of $\hat{f}$

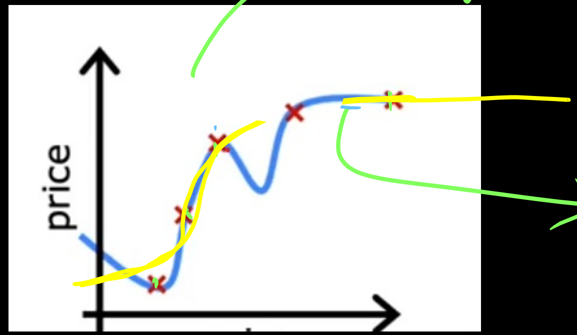
$$\hookrightarrow \frac{1}{m} \sum_{i=1}^{m} \left(\hat{f} - \bar{\hat{f}}\right)^2$$

High bias & low variance (Underfit)



$\hat{y}$

Low bias High variance (Overfit)

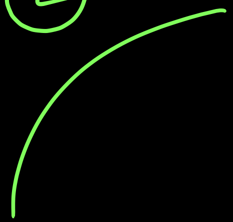


$\hat{y}$

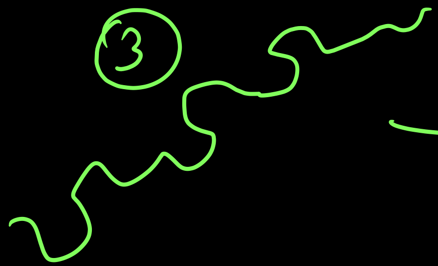
Bias = Error of $(y - \hat{y})$

Variance =

2



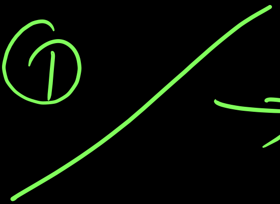
3



→ High variance

1 < 2 < 3

1



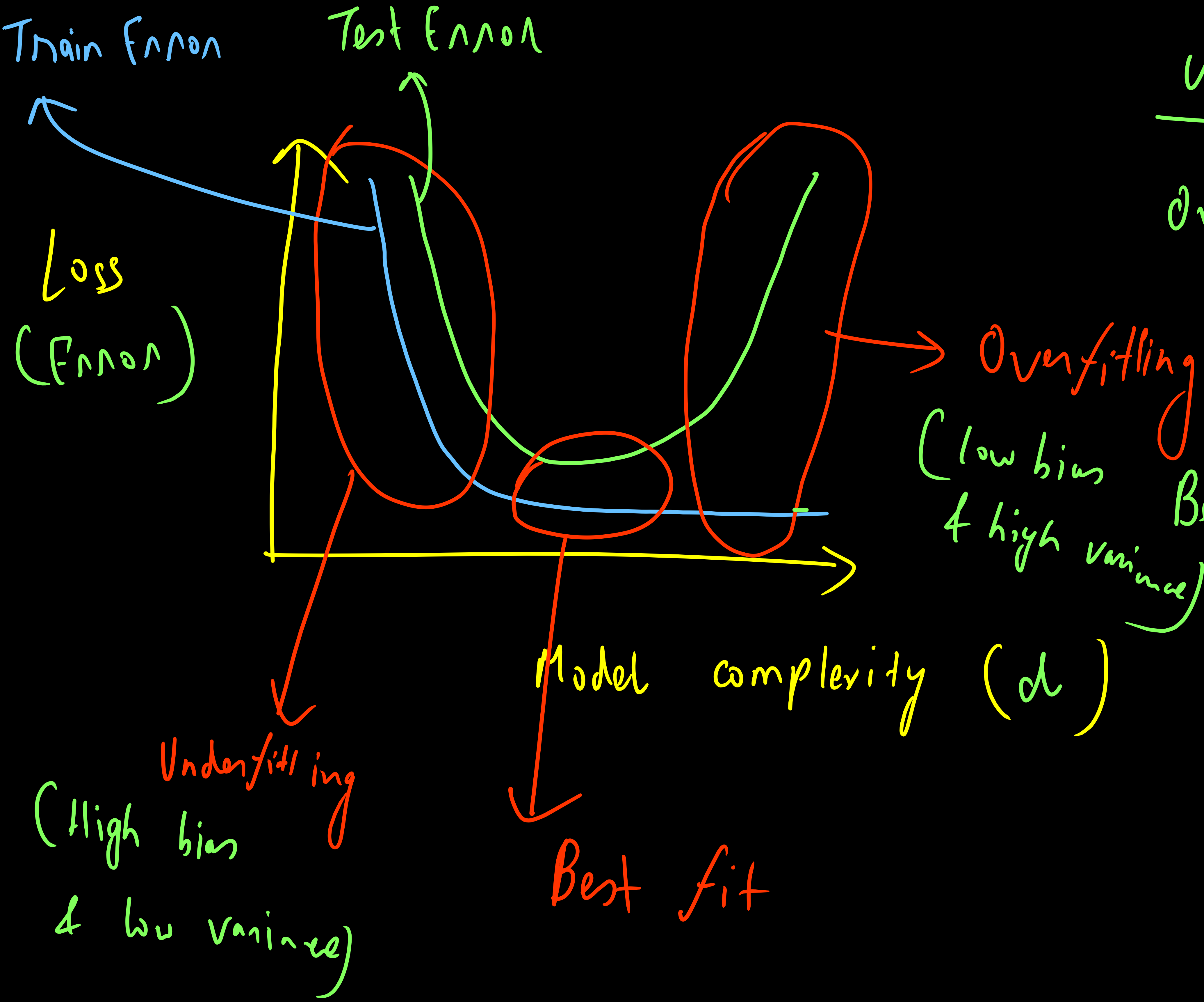
→ low variance

$\hat{y} = w_1 x + w_0$ → Underfit

$\hat{y} = w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^4 + w_0$ → Overfit

$\boxed{\hat{y} = w_1 x + w_2 x^2 + w_0}$ → Best fit

→ $w_1 x_1 + w_2 x_2 + w_{11} x_1^2 + w_{22} x_2^2 + w_0$



Underfitting : Train High
Test High

Overfitting : Train low
Test High

Best fit : Train low
& Test low

→ How to reduce overfitting ?

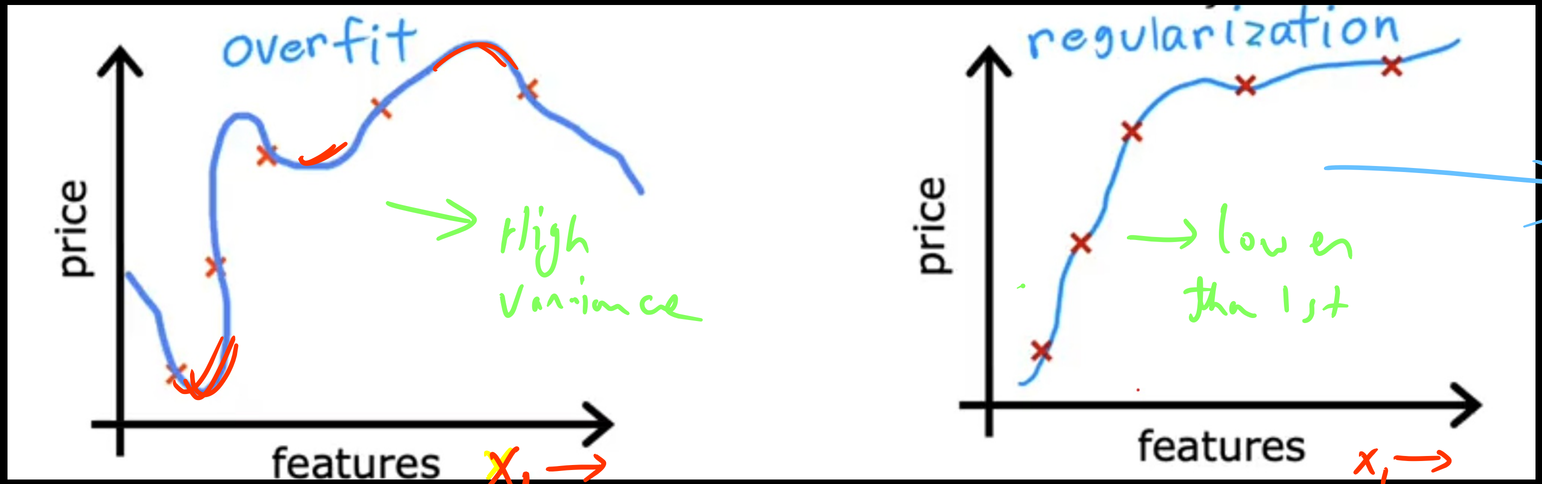
↳ Regularization

→ What is Regularization?

$w_1x + w_0$

$w_1x + w_2x^2 + w_0$

$w_1x + w_2x^2 + w_3x^3 + w_4x^4$



Quadratic

$$\hat{y} = 28x_1 - 385x_1^2 + 39x_1^3 - 174x_1^4 + 10$$

$$\hat{y} = 13x_1 - 0.23x_1^2 + 0.000014x_1^3 - 0.0001x_1^4 + 10$$

$$\hat{y} = w_1x_1 + w_2x_2 \;--\; w_dx_d$$

x_5 $\quad$ x^5 $\quad$ x_1^5

$28x_1 + 28x_2$

Model is Overfitting

↓

High order features are contributing a lot

↓

Reduce the influence of high order feature

↓

w_j associated high order

→ How to achieve Regularization ?

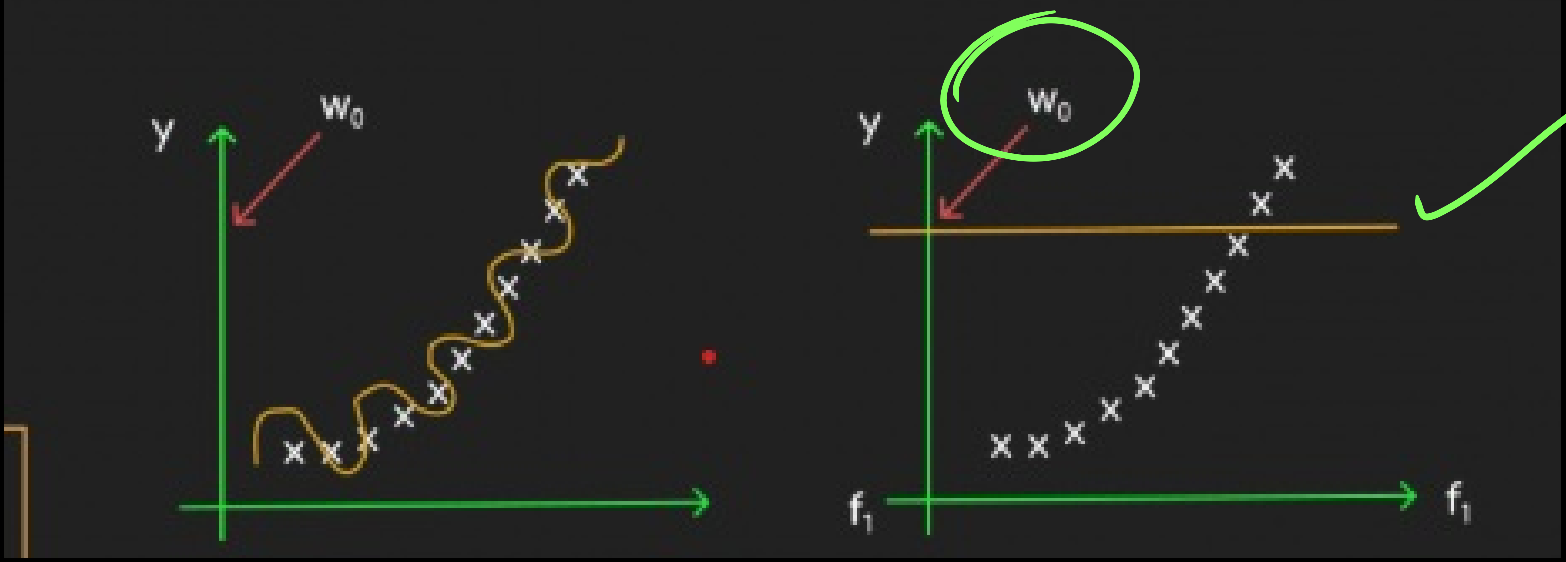
$$\hat{y}^{(i)} = \overset{0}{w_1} x_1 + \overset{0}{w_{12}} x_1^2 + \overset{0}{w_{13}} x_1^3 - - \overset{0}{w_{1d}} x_1^d + w_0$$

$$loss = min_{w_j} \frac{1}{n}\left[\sum_{i=1}^{n}(y^{(i)} - \hat{y}^{(i)})^2 + \sum_{j=1}^{d} w_j^2\right]$$

Updates w_j such that MSE is least

Makes w_j su

value ≈ 0, exc



L2 regularization

$$\|w\|_2^2$$

$$\sum_{j=1}^{d} w_j^2$$

$$\boxed{\min \sum_{j=1}^{d} w_j^2}$$

0

MSE loss

Regularization loss

$$\min \text{ loss} = \min \left[\underbrace{\frac{1}{m} \sum_{i=1}^{m} \left(y^{(i)} - \hat{y}^{(i)}\right)^2}_{\text{MSE loss}} + \lambda \underbrace{\sum_{j=1}^{d} w_j^2}_{\text{Regularization loss}} \right]$$

Regularization Hyperparameter

$\lambda = 0.0001$

→ ↑ MSE than Reg

→ Model can beome very Complex

$\hat{y} = 28x - 385x^2 + 39x^3 - 174x^4 + 10$

→ Overfit

$\lambda = 10^4$

→ Reg ↑ than MSE

→ Model is simple as high weight reduced to small value

$\hat{y} = 13x - 0.023x^2 + 0.014x^3 + 0.001x^4 + 10$

→ Underfit

two

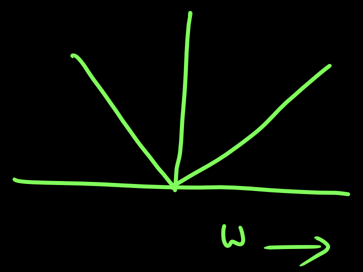
→ Break until 22:30 PM

L1 reg (lasso)

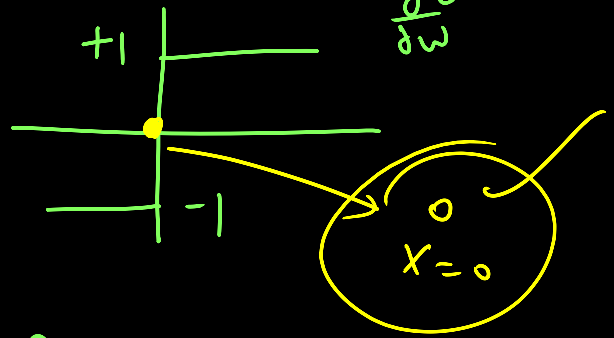
1. $\|w\|_1^2 = \sum_{j=1}^{d} |w_j|$

$|w| = L$

2.



3.



$\frac{\partial L}{\partial w}$

4. Sparsity (Lot of weight reduced to 0)

$w_1 x_1 + w_2 x_1^2 \;--\; w_d x_1^d$

$f_1 - f_2 \quad \cancel{x_{10}} \;--\; \cancel{x_{100}}$

L2 reg (Ridge)

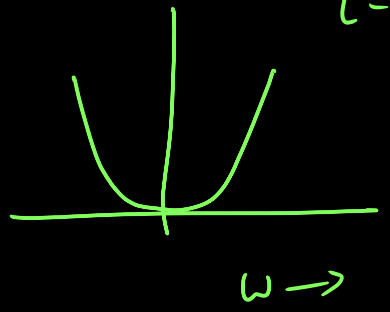
$\|w\|_2^2 = \sum_{j=1}^{d} w_j^2$

$L = w^2$

$\frac{\partial L}{\partial w} = w$

$\boxed{w = w - \eta \frac{\partial L}{\partial w}}$, $\eta = 0.5$

2.

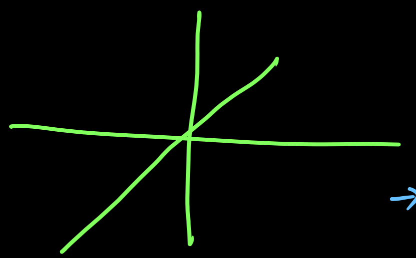


→ $w = 1000 \rightarrow 500$

$w = 1000 - 0.5 \times 1000$

$\hookrightarrow = 500$

3.



→ $w = 1 \rightarrow 0.5$

$w = 1 - 0.5 \times 1$

$= 0.5$

4. Weights close to 0 but not 0.

$x_1 \;---\; x_{10}$

→ Elastic Reg

↳ L1 + L2

$$min_{w_j}\,[MSE + \underbrace{\underbrace{\lambda_1 \sum_{j=1}^{d} |w_j|}_{L1} + \underbrace{\lambda_2 \sum_{j=1}^{d} w_j^2}_{L2}}_{\text{Elastic Reg}}]$$

Parameter	Hyperparameter
→ Learned from data during training	→ Set before training
w, b	→ Learning rate, λ,
$w = w - \eta \frac{\partial L}{\partial w}$	α

Hyperparameter tuning

best R^2 =

1. for λ in [0.01, 0.1, 10, 100, 1000]:

→ fit the model (λ)

→ Check R^2 value

→ best R^2 =

← $\boxed{\lambda_{best}}$ =

R^2	$\lambda = 0.1$	$\lambda = 0.2$	$\lambda = 10$	$\lambda = 1000$
	0.52	0.22	0.88	0.4