Regularization :-

* It is a technique to prevent the model from overfitting by adding extra information to it.

* If the model gets overfitted, then it has low bias with training data, but there will be high variance with testing data.

* If there is high variance with testing data then predicting accuracy will be very poor.

* So to overcome this we use regularization in order to reduce the magnitude of the features

* In layman terms, the process of reducing the steepness or slope of best fit line to make best fit line as generalized line.

* There are 2 types of regularization technique

i) Ridge regression

ii) Lasso regression.

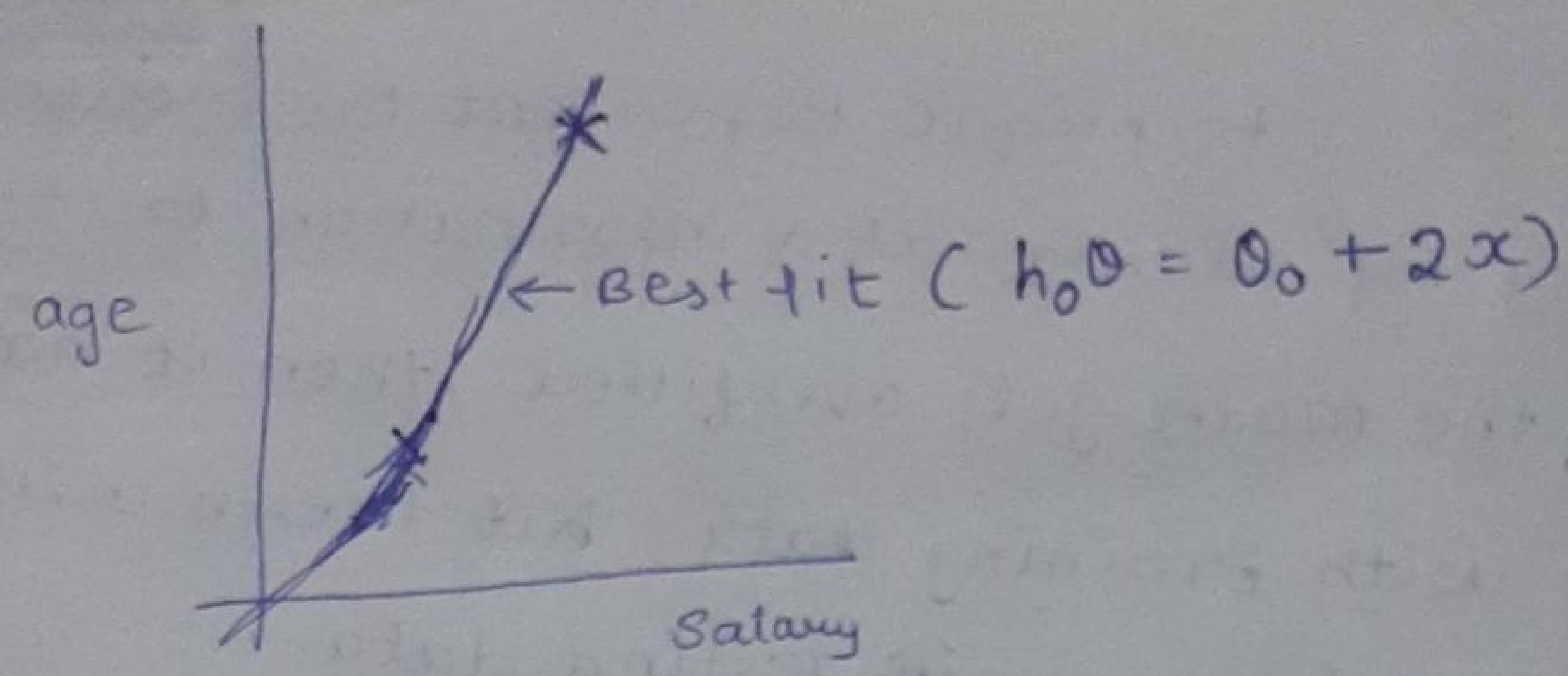
Ridge regression :- (L2 regularization)

* Here we will add small amount of bias to the cost function.

* The cost function is altered by adding the penalty term to it.

* The amount of bias added to the model is called ridge regression penalty

Scanned by TapScanner



* Lets take example of Linear regression, here the best fit line is overfitted.

* For this the cost function

$$J(\theta_0, \theta_1) = 0$$

* So now we want to make this overfitting to generalised one.

* cost function in ridge reg :-

$$\boxed{\underbrace{\sum_{i=1}^{n}\left[y_i - \theta_0 - \sum_{j=1}^{p}\theta_j x_{ij}\right]^2}_{\text{RSS}} + \underbrace{\lambda}_{\text{hyper Parameter}} \underbrace{\sum_{j=1}^{p}\theta_j^2}_{\text{adding Penalty or bias}}}$$

~~the hyperparameter λ can be tuned as per our wish but λ>0 λ≠0~~

* In the above example of age vs salary, the best fit line pass through all the pts.

So, $J(\theta_0, \theta_1) = 0$

* we can get λ from cross validation
* Lets consider $\theta_1 = 2$ as it passes through origin

So, $J(\theta_0, \theta_1) = 0$

* Lets say $\lambda = 1$

$$= \sum_{i=1}^{n} \left[y_i - \theta_0 - \sum_{j=1}^{p} \theta_j x_{ij}\right]^2 + \lambda \sum_{j=1}^{p} \theta_j^2$$

$$= \underbrace{[y_1 - \theta_0 - \theta_1 x]^2}_{=0} + \lambda \theta_1^2$$

$$= 0 + \lambda \theta_1^2$$

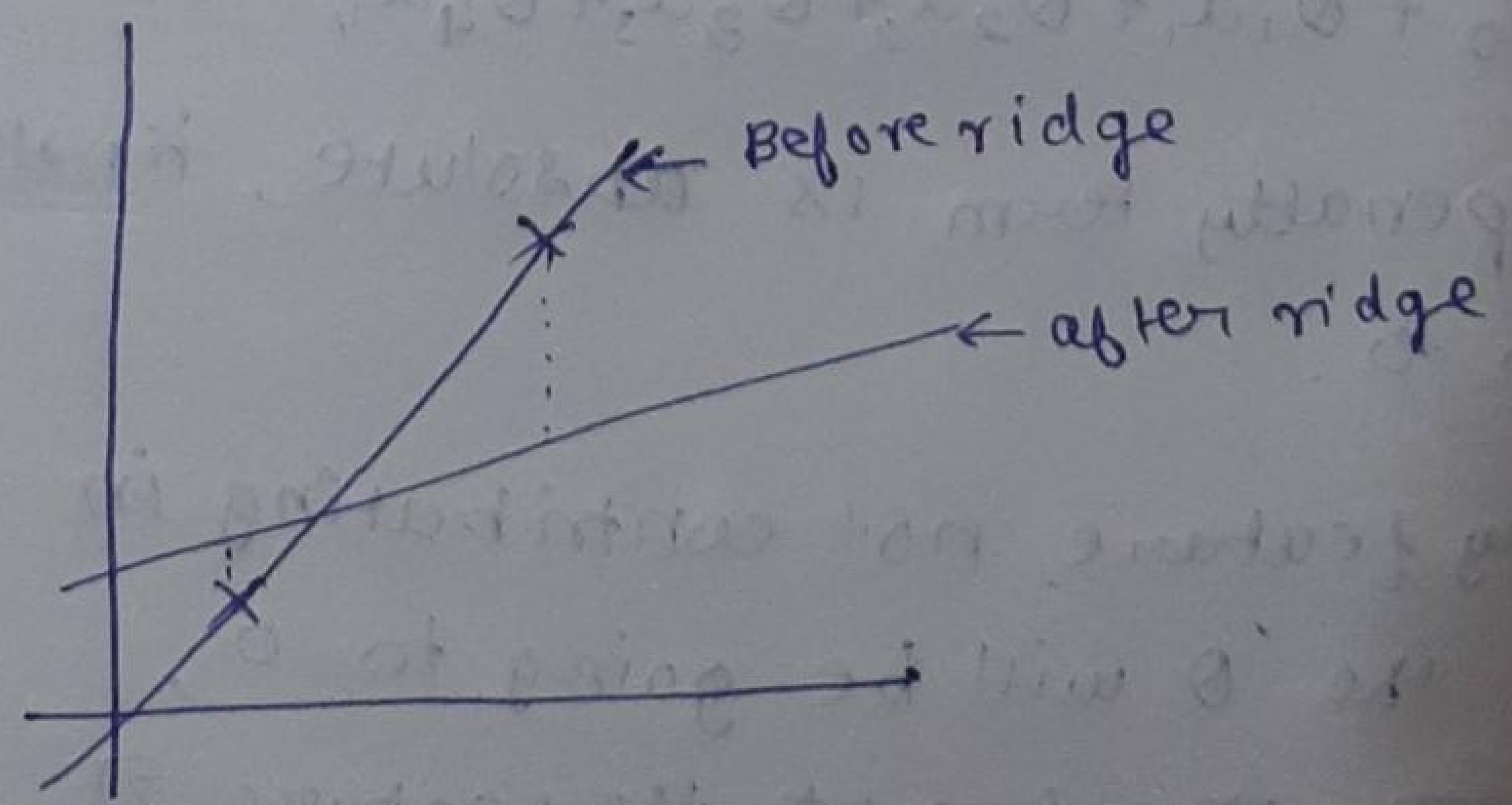
$$= (1)(2)^2$$

$$= 4$$

* Now we want to reduce this 4 nearer to Zero, but not = 0. because there is θ_1^2, so for that cost function

& convergence alg will take several iteration.

* like that it will get reduce upto some extent



* The slope will get reduced
* This solves overfitting
* Solves high collinearity

Lasso regression :- (L1 regularization)

* The aim is same as ridge regression, but it has only a small difference in the penalty term.

* The cost function for Lasso reg is

$$\boxed{\sum_{i=1}^{n} \left[y_i - \theta_0 - \sum_{j=1}^{p} \theta_j x_{ij}\right] + \lambda \sum_{j=1}^{p} |\theta_j|}$$

* It uses magnitude of slopes, so this alg can take slope to 0.

* Like wise it solves overfitting same as ridge.

* But Lasso regression will also perform feature selection automatically in the process.

* Suppose we 4 indep features & 1 depend feature, the best fit equation will be

$$h_\theta O = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4$$

* Since the penalty term is absolute, its slope can reach to 0.

* So if any feature not contributing in prediction the θ will be going to 0, so it will cancel out the feature & reduce the complexity

• For example in reducing process θ_3 & θ_4 became 0 then

$$\boxed{h_\theta \theta = \theta_0 + \theta_1 x_1 + \theta_2 x_2}$$

• ~~Here~~ Hence, slope reduced, as well as feature selection has been done.

Scanned by TapScanner