Linear Regression / (Bias – Variance)

→ $Y/X \sim N(\theta^T X, \sigma^2)$

→ $\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\ RSS(\theta)$; $RSS(\theta) = \|Y - X\theta\|^2$

$= \sum_{i=1}^{n} (Y^i - x^{(i)}\theta^T)^2$

⇒ Equivalent to maxm conditional likelihood Estimate

$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \log(P(Y^{(i)}/X^{(i)}))$

→ If the Residual plot has a pattern, then the model is not right.

→ linear transformation may be used.

i.e Y is linear in X, but X could be non-linear transformation of original data X

→ Linear Transformation (LT) will not impact the linear relationship b/w Variables.

$Corr(Y, X) = Corr(Y, LT(X))$

Non-Linear Transformation may increase/decrease the linear relationship

→ Plot Residual plots to verify Non-linear Transformation help linear relationship.

→ Training Data

$$(x^{(i)}, y^{(i)}) \overset{iid}{\sim} p(x, y)$$

$$= p(x)\, p(y/x)$$

$p(x)$: arbitrary model; $p(y/x)$: linear regression model

→ Gradient Descent in Linear Regression — is used because it is computationally cheaper.

Normal Eqn — for matrix inversion — $O(n^3)$ & very Expensive.

→ $E(\hat{\theta}/X) = (X^T X)^{-1} X^T E(y/X)$

($\hat{\theta} = (X^T X)^{-1} X^T y$)

$E(y/X) = X\theta$

$= (X^T X)^{-1} (X^T X)\theta = \theta$

$$Var(\hat{\theta}/X) = \sigma^2 (X^T X)^{-1}$$

$$\hat{\theta}/X \sim N(\theta, \sigma^2 (X^T X)^{-1})$$

Visit [towardsdatascience.com/mse and bias variance decomposition]

$y = f(x) + \epsilon$

Bias-Variance Decomposition

MSE: $E[(y - \hat{f}_S(x))^2]$

Expected Value Squared prediction Errors

$= Var[f(x) - \hat{f}_S(x)] + Var(\epsilon) + (E[f(x)] - E[\hat{f}_S(x)])^2$

S = training data

Variance → How jumpy is the gap b/w the real model & the predictors is depending on the training data S and test data (x, y)

Bias → Shows whether our model approximates the real model well. i.e If the model passes through all the data points then it has Low Bias

Irreducible noise

Batch Regularization DVA.

features ↑ → samples.

→ If $D << N$, MLE overfits

→ MLE of linear regression Estimator = 0

$E[\hat{\theta}] = \theta$

→ Bias/Intercept different from statistical bias (DL, 110 page)
└ $[E[\hat{\theta}] - \theta]$

→ **Bias** in Linear Regression
- Adds flexibility to the model
- i.e without bias the line has to pass through origin, limiting or Constraining the model.

Ex: If all features = 0, then prediction must be zero irrespective of the data.

→ bias parameter of affine transformation

If $Y \sim N(XW, \sigma^2 I)$

for new (x_0, y_0) $y_0 \sim N(x_0^T W, \sigma^2)$

$$E\left[(y_0 - x_0^T \hat{w})^2 / X, x_0\right]$$

Squared Bias

$$= \underbrace{\sigma^2}_{\text{noise}} + x_0^T (w - E[\hat{w}])(w - E[\hat{w}])^T x_0 + x_0^T \underbrace{Var(\hat{w})}_{\text{Variance}} x_0$$

different from bias/intercept is Statistical Bias

Model Bias: how close to the solution we expect to be on average.

→ how sensitive is our solution to the data.

⇒ LS Solution :-

unbiased, but potentially high variance

Ridge Regression :-

(RR) biased, but lower variance than LS.

Which is preferable?

Ultimately depends on how well our model generalize the data.

→ Using proxy $E\left[\left[y_0 - x_0^T \hat{w}\right]^2 / X, x_0\right]$

LS :- Bias = 0

In RR, we add Bias to lower Variance.

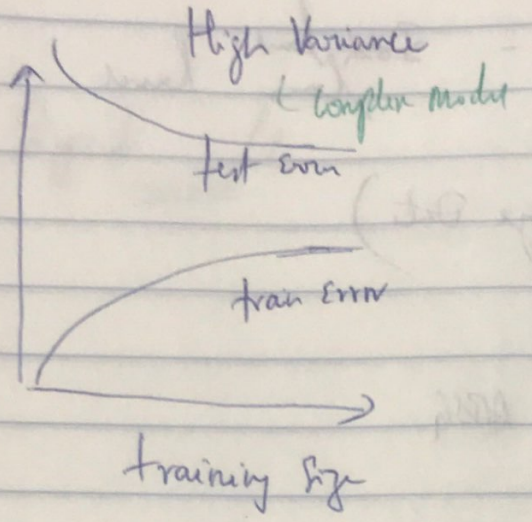
Squared Bias = ~~$x_0^T \left(E[\hat{w}] - E[w]\right)$~~

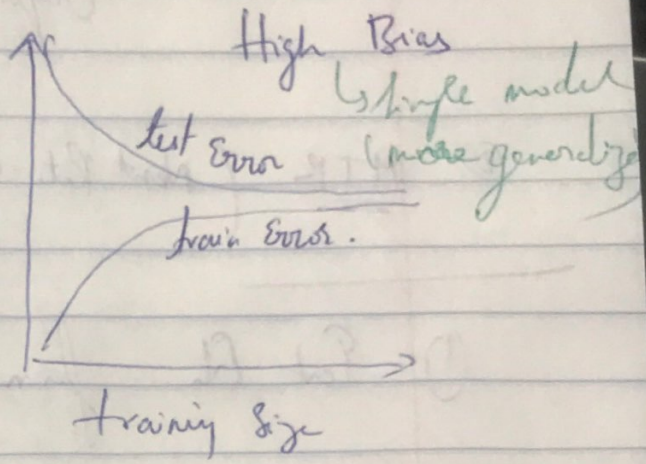
$= x_0^T \left(E[\hat{w}] - E[\hat{w}]\right)\left(w - E[\hat{w}]\right)^T x_0$

In RR we don't w (true weights), can't calculate Bias.

But we could make some statements about what $\hat{w}$ should be?

Ch - Learning Curves (cont.)





→ Difference b/w train Error & test Error