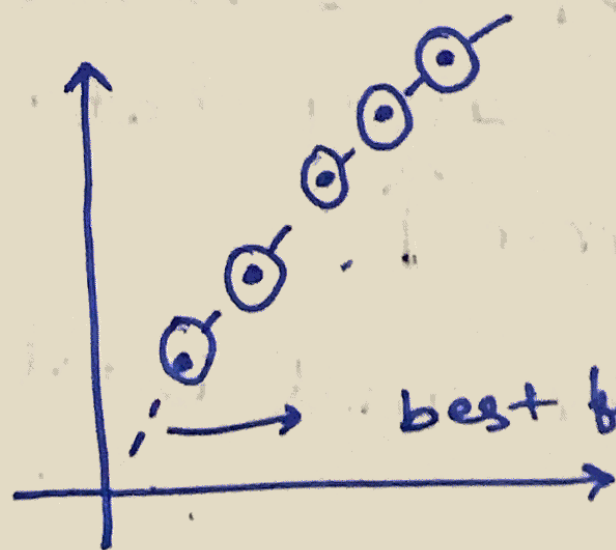


- A.) models which have a high complexity have a high variance. if we have less data for training the model is most likely to overfit the data and perform poorly on the test data. this is because of the limited data the model gets suited to replicate the patterns in a small sample which may not exist in much larger samples. hence if we increase the size of our training data it is guaranteed that the test error or the validation error will reduce.

suppose we have a dataset with $N=5$

Points initially \rightarrow



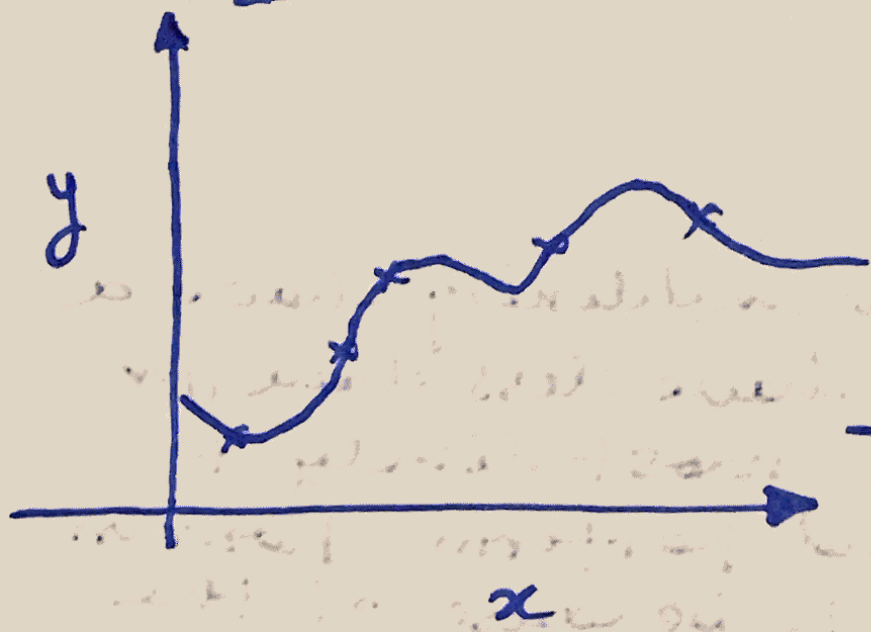
best fit model with low complexity

if we increase the training data it becomes very difficult for the model to fit all the data points hence the training error will increase if sample data \uparrow

high variance

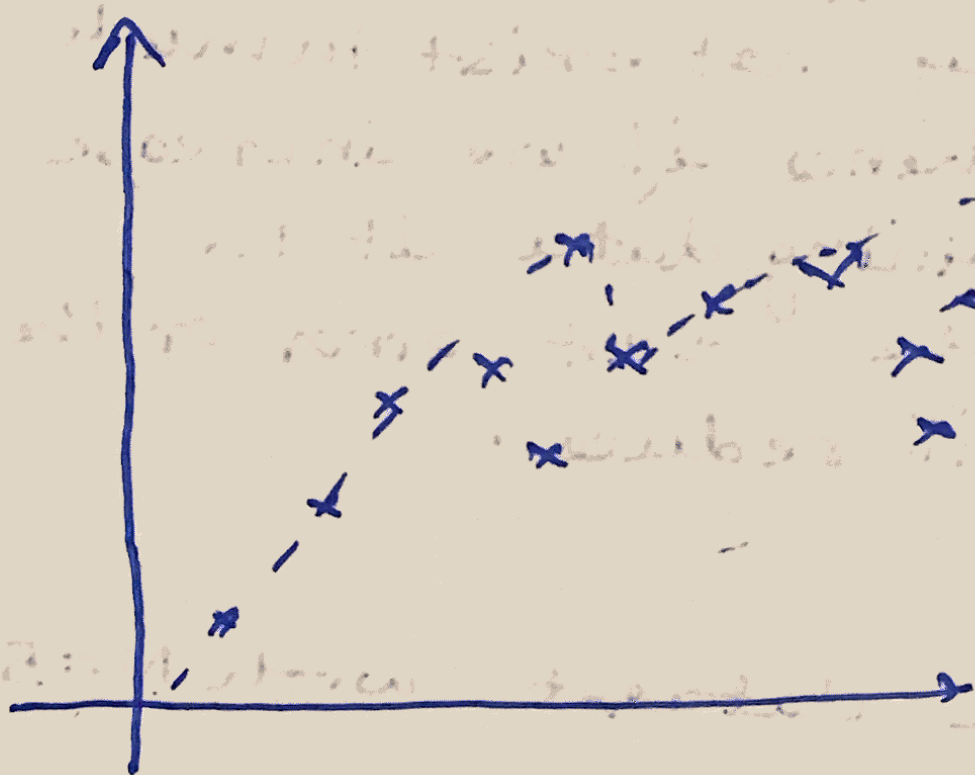
$$f_0(x) = \theta_0 + \theta_1 x + \dots + \theta_n x^n$$

case @ small
data set



→ overfitting if
sample size ↑

as we increase the data set size →



→ it becomes harder for the model
to fit the additional noisy data
hence the training error ↑
→ the cross validation error will still
be going down.

→ the training error plateaus due to
the fact that even if we increase the
training data size the model can't give
a much better fit.

