GENERALIZATION ERROR

LECTURE 3
SECTION 2
JUNE 5th

INSTITUTE FOR APPLIED
COMPUTATIONAL SCIENCE
AT HARVARD UNIVERSITY

UNIVERSITY of
RWANDA

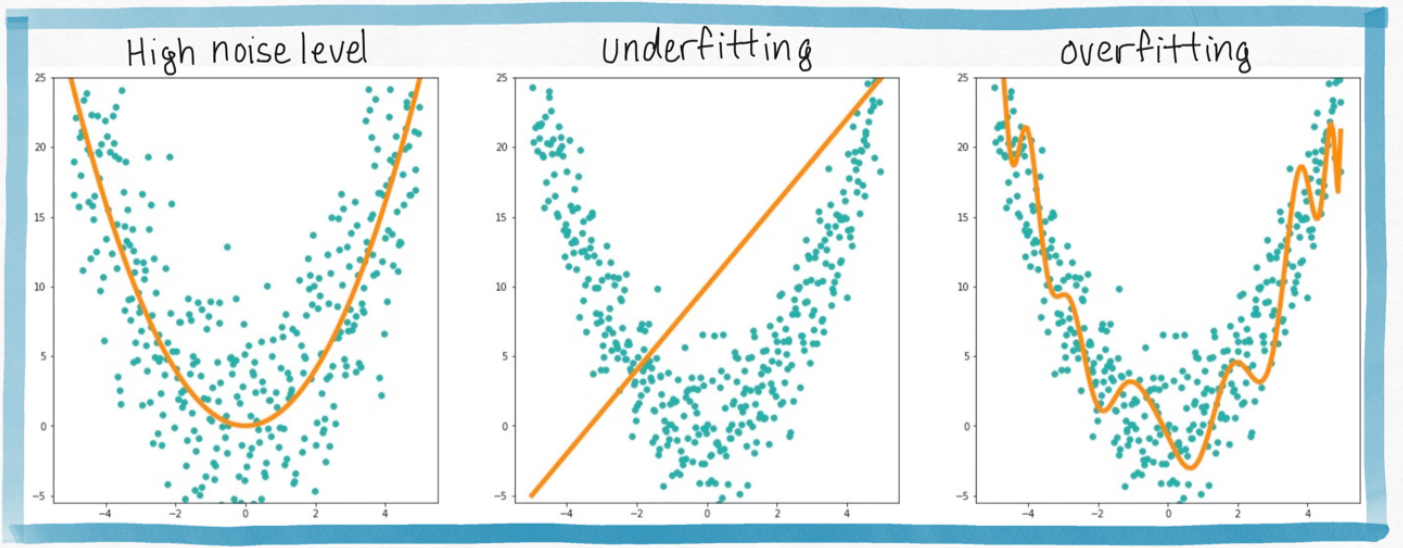
WHAT CONTRIBUTES TO TEST ERROR?

TEST ERROR & GENERALIZATION:

We know to evaluate models on both train and test data, because models do well on training data may do poorly on new data.

The ability of models to do well on new data is called generalization.

We know of at least three ways models can have a high test error, i.e. poor generalization:



But how much does each contribute to generalization error?

IRREDUCIBLE AND REDUCIBLE ERRORS:

It's important to distinguish the contributions of noise, underfitting, and overfitting to the generalization error because:

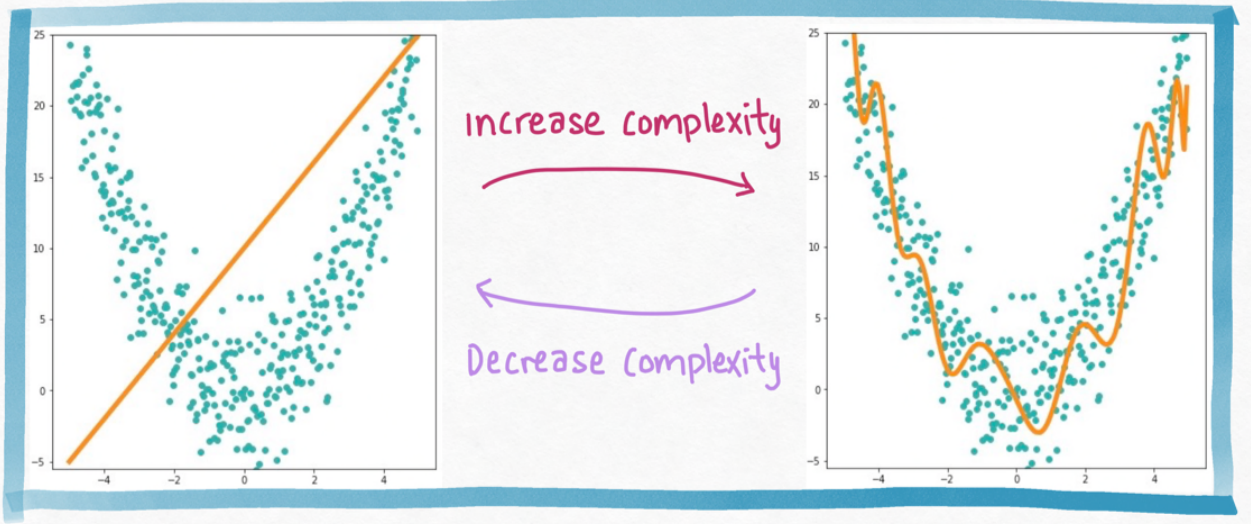
- (irreducible error) we can't do anything to decrease error due to noise
- (reducible error) we can decrease error due to under and overfitting by choosing a different model.

We will focus on decreasing reducible error.

THE BIAS-VARIANCE TRADE-OFF:

Reducible error comes from either underfitting or overfitting.
There is a trade-off between the two sources of errors.

As we increase model complexity, we decrease error due to underfitting but increase risk of overfitting.



As we decrease model complexity we decrease error due to overfitting but increase risk of underfitting.

FORMALIZATION OF GENERALIZATION ERROR

ALL TOGETHER, WITH MORE MATH:

Assume the data comes from: $y = f_w(x) + \varepsilon$, $\varepsilon \sim N(0, \sigma^2)$
Let our estimated model be $f_{\hat{w}}(x)$.

Then the generalization error is the SE averaged over all possible test data $P(x,y)$ and all possiblible samples of training set $D \sim P(D)$:

$$\text{Gen. Error} = \mathop{\mathbb{E}}_{(x,y)\sim P(x,y)} \left[\mathop{\mathbb{E}}_{D\sim P(D)} \left[(y - f_{\hat{w}}(x))^2 \right] \right]$$

We rewrite the above using properties of random variables and algebra:

$$\begin{aligned}
\mathop{\mathbb{E}}_{D\sim P(D)} \left[(y - f_{\hat{w}}(x))^2 \right] &= \mathbb{E}\left[(f_w(x) + \varepsilon - f_{\hat{w}}(x))^2 \right] = \mathbb{E}\left[(f_w(x) + \varepsilon - f_{\hat{w}}(x) - \mathbb{E}[f_{\hat{w}}(x)] + \mathbb{E}[f_{\hat{w}}(x)])^2 \right] \\
&= \mathbb{E}\left[(f_w(x) - \mathbb{E}[f_{\hat{w}}(x)])^2 \right] + \mathbb{E}[\varepsilon^2] + \mathbb{E}\left[(\mathbb{E}[f_{\hat{w}}(x)] - f_{\hat{w}}(x))^2 \right] + 2\mathbb{E}\left[(f_w(x) \right. \\
&\quad \left. - \mathbb{E}[f_{\hat{w}}(x)])\varepsilon \right] + \mathbb{E}\left[\varepsilon (\mathbb{E}[f_{\hat{w}}(x) - f_{\hat{w}}(x)]) \right] \\
&\quad + 2\mathbb{E}\left[(\mathbb{E}[f_{\hat{w}}(x)] - f_{\hat{w}}(x)) \right] (f_w(x) - \mathbb{E}[f_{\hat{w}}(x)]) \\
&\quad \vdots \\
&= \underbrace{(f_w(x) - \mathbb{E}[f_{\hat{w}}(x)])^2}_{\text{Bias}} + \underbrace{\text{Var}[\varepsilon]}_{\sigma^2} + \underbrace{\text{Var}[f_{\hat{w}}(x)]}_{\text{variance}}
\end{aligned}$$

"Bias" measures the difference between the true function & the estimated function

"σ^2" is the variance of the output noise

"Variance" is the variance of our estimate

THE BIAS-VARIANCE TRADE-OFF AGAIN:

We found that generalization error can be decomposed:

$$\mathop{\mathbb{E}}_{(x,y)\sim P(x,y)}\left[\mathop{\mathbb{E}}_{D\sim P(D)}\left[(y-f_{\hat{w}}(x))^2\right]\right] = \mathop{\mathbb{E}}_{(x,y)\sim P(x,y)}\Big[\underbrace{\left(f_w(x)-\mathbb{E}[f_{\hat{w}}(x)]\right)^2}_{\text{Bias}} + \underbrace{\text{Var}[\varepsilon]}_{\sigma^2} + \underbrace{\text{Var}[f_{\hat{w}}(x)]}_{\text{Variance}}\Big]$$

σ^2 represents irreducible error

Bias and Variance represent reducible error, since we can change $f_{\hat{w}}(x)$.

Bias measures the difference between the true function and our estimate $f_{\hat{w}}$ (averaged over all samples of training sets).

- Bias will be high when our function is too simple and underfits.
- Bias will be low when our function is complex and easily fit the data.

Variance measures the variance of our estimate when we chage the training set.

- Variance will be high when our function is too complex and overfits to the noise in each training set D.
- Variance will be low when our function is simple and insensitive to noise.

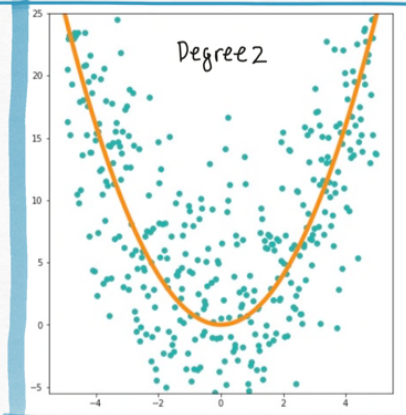
The Bias-Variance Trade-Off: when we decrease bias by making our functions complex, we increase variance. When we decrease variance by making our functions simple we increase bias.

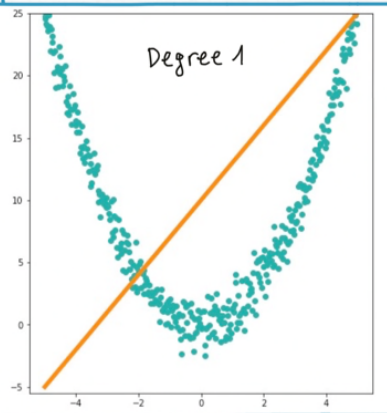
SUMMARIZING IN INTUITIVE TERMS

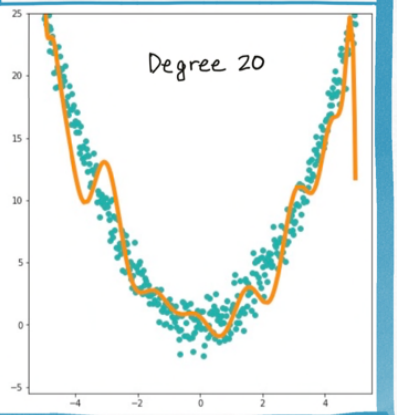
REASONING ABOUT GENERALIZATION ERROR:

All of the following models have the same generalization error, but the sources of error for each model is different.

Irreducible Error	High	Low	Low
Bias	Low	High	Low
Variance	Low	Low	High







Error cannot be reduced

Error may be reduced by increasing the degree.

Error may be reduced by decreasing the degree.