Last week: How to make more sophisticated models by engineering features

This week: How to make sure your sophisticated model predicts the variable you want well
( how to ensure your model **generalizes** well)

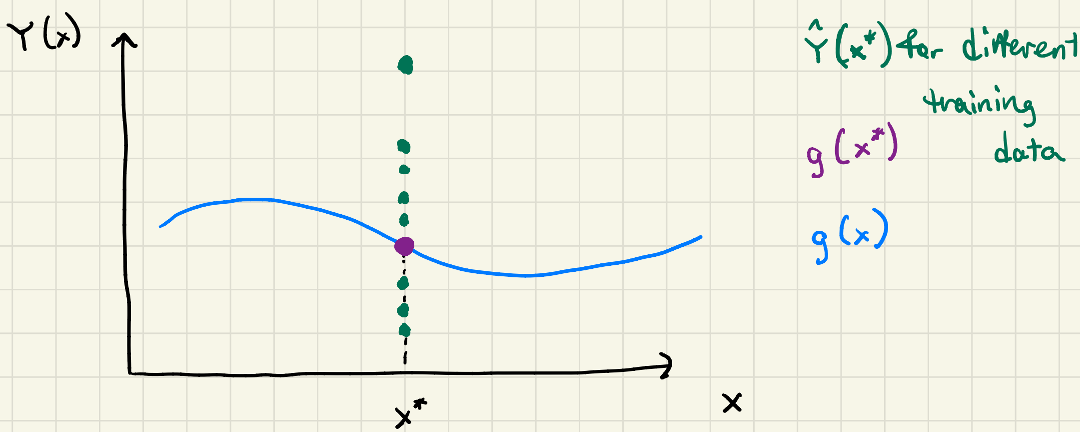
## Bias - Variance Decomposition

we don't know g

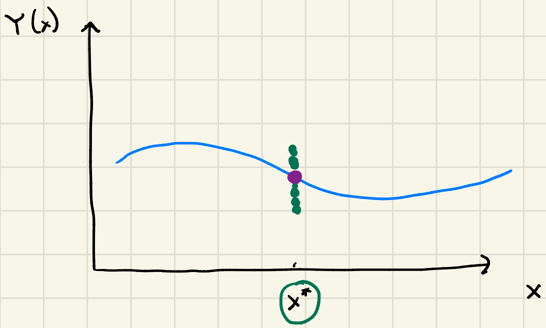
$$E\left[(Y-\hat{Y}(x))^2\right] = \underbrace{E\left[\hat{Y}(x)-g(x)\right]^2}_{\text{model bias}^2} + \underbrace{E\left[\left(\hat{Y}(x)-E[\hat{Y}(x)]\right)^2\right]}_{\text{model variance}} + \underbrace{\sigma^2}_{\text{irreducible error}}$$

**Model Bias**: How well does my model predict data it's trained on?

**Model Variance**: If I were to change my training data slightly, how much would the model's prediction change?



## Low Bias

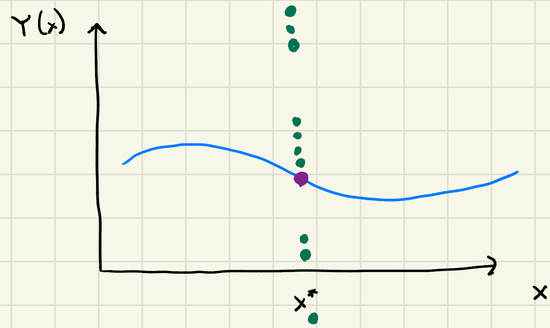



$\hat{Y}(x)$ for different training data

$g(x^*)$

$g(x)$

## High Bias

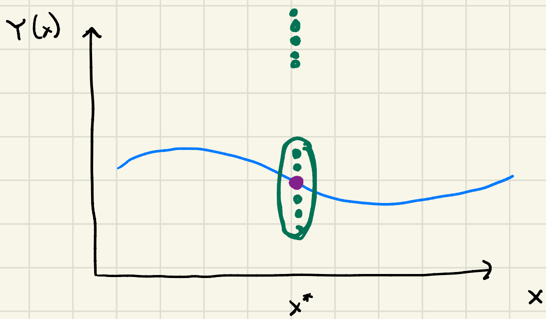



$\hat{Y}(x)$ for different training data

$g(x^*)$

$g(x)$

## Low Variance

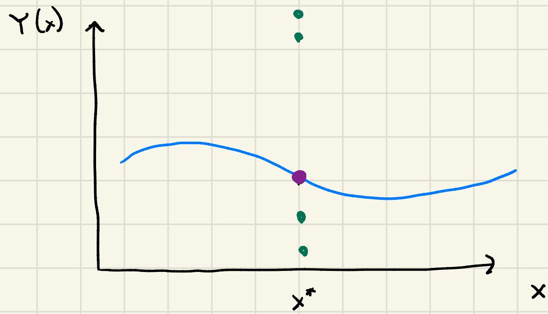



$\hat{Y}(x)$ for different training data

$g(x^*)$

$g(x)$

## High Variance




$\hat{Y}(x)$ for different training data

$g(x^*)$

$g(x)$

We want models with low model bias & low model variance

↑ low training error

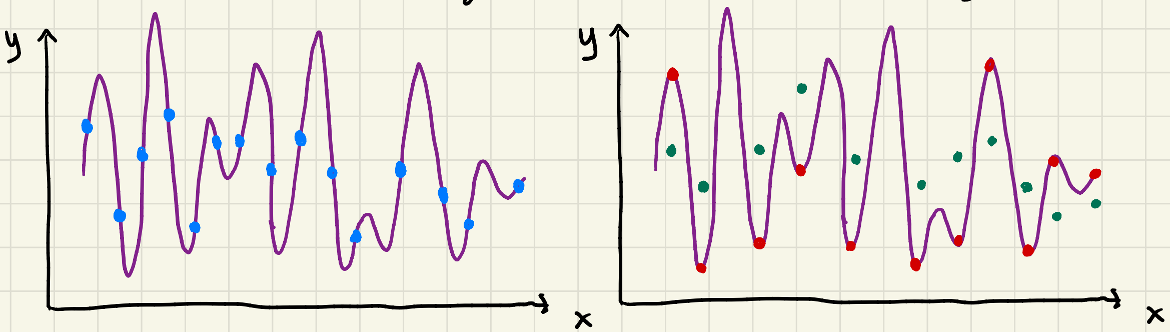
↑ low validation error

## Regularization

high model complexity

- **Motivation:** If we have a lot of features, we might get a model that performs well on the training data, but does not generalize well (low bias, high variance)

model




Training data

Test data

Test predictions

This situation happens because we are giving our model too much freedom ⇒ it is using that freedom to find patterns that exist only in the training data

Need to restrict our model to discourage using all features

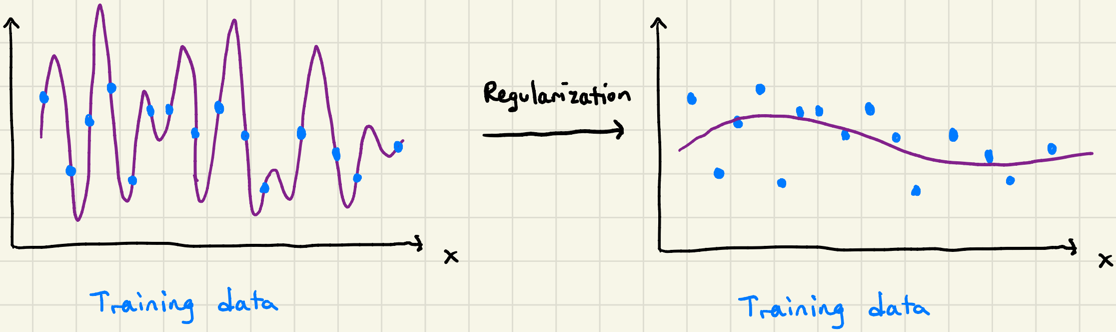
**Before:** $\underset{\theta}{\text{minimize}} \quad MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - f_\theta(x_i))^2$

**Regularization:** $\underset{\theta}{\text{minimize}} \quad \frac{1}{n} \sum_{i=1}^{n} (y_i - f_\theta(x_i))^2 + \lambda \sum_{j=1}^{d} \theta_j^2$

$\|\theta\|_2^2$

$\lambda$ is a scalar that we choose (through cross-validation)

tells the model that choices of $\theta$ with high magnitude are bad!




With regularization, we are sacrificing some training error with the hope that we will get at least as much back on test error (which really matters)