Bias and Variance Trade-off

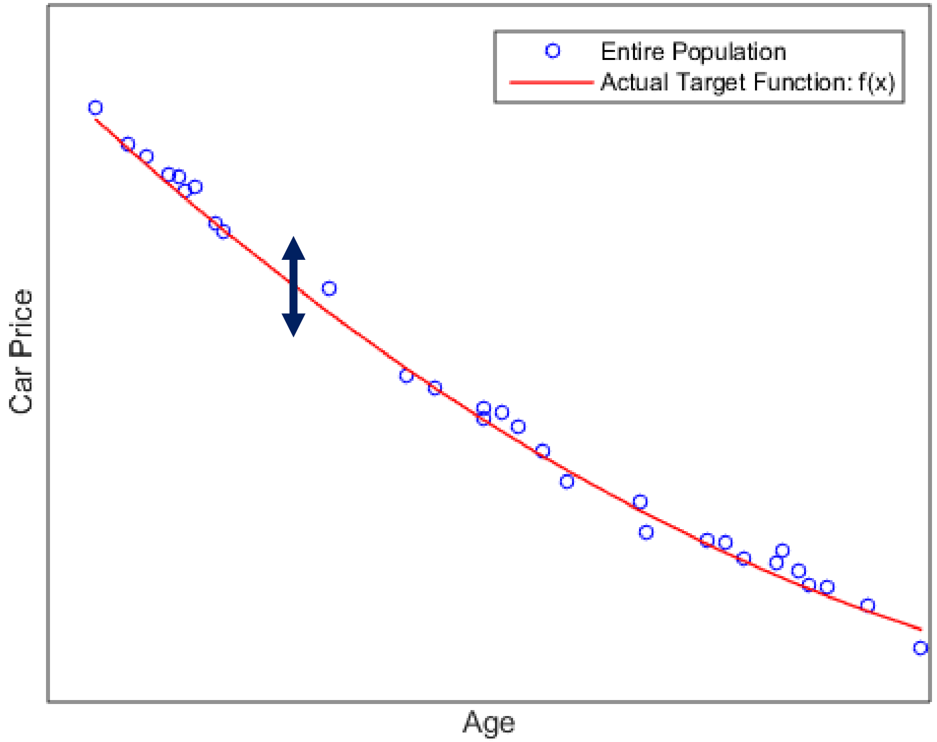
Anahita Zarei, Ph.D.

Overview

- Sources of Error
  - Noise
  - Bias
  - Variance
- Bias and Variance
  - Graphically
  - Conceptually
  - Mathematically
- Reading: 2.3 from Learning from Data

Stochastic Noise

- Stochastic noise is the irreducible error inherent in the data.
$$y = f(x) + \varepsilon$$
- This noise is the property of the data and has nothing to do with the model.
- E.g. the relationship between a car's price and its age is not a perfect relationship.
- No model, can capture the exact relationship.
- The mean of noise is zero, the variance is epsilon.



Why Discussing Bias and Variance?

- Bias-Variance Decomposition is a key component in understanding learning algorithms.
- Understanding how different sources of error lead to bias and variance helps us improve the data fitting process resulting in more accurate models.
- Helps understand and avoid overfitting and underfitting.
- Helps explain why simple models can outperform the more complex ones. For example:
    - A regression model with fewer parameters maybe better than one with more parameters.
    - A neural network model with fewer neurons maybe better than one with more neurons.
    - A simple classifier such as Naïve Bayes maybe better than decision trees.

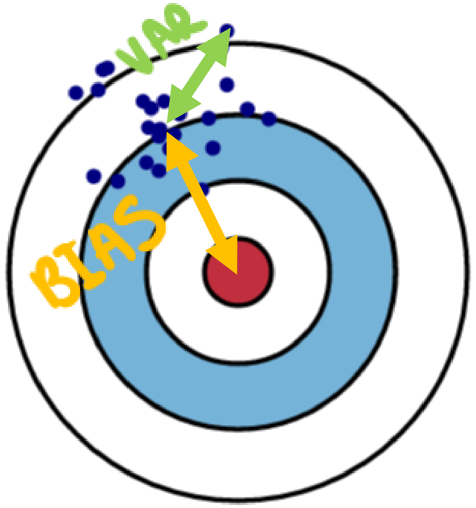
Conceptual Definition of Bias and Variance

Error due to Bias

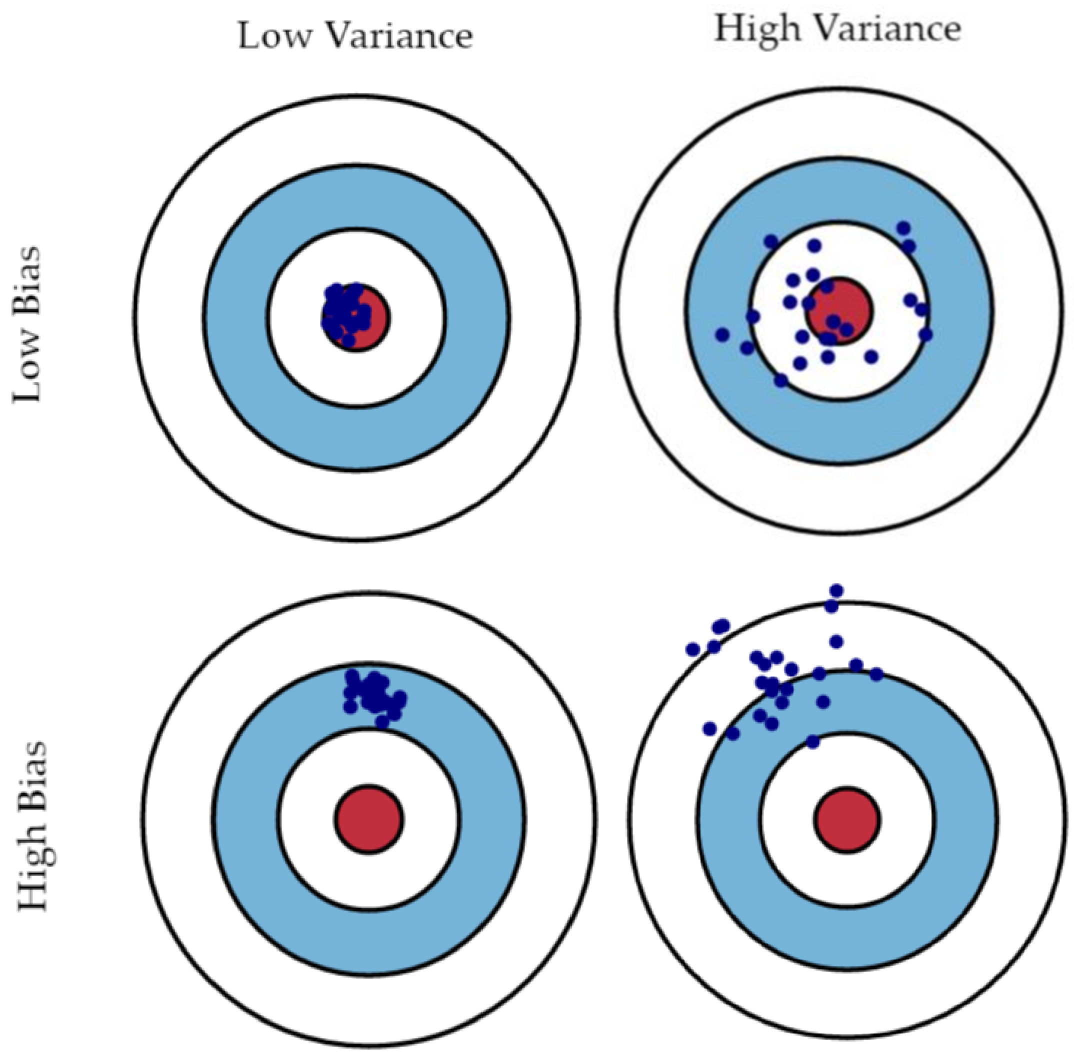
- We assume we could repeat the whole model building process more than once: each time we gather new data and run a new analysis creating a new model. Due to randomness in the underlying data sets, the resulting models will have a range of predictions.
- Bias measures how far off *in general* these models' predictions are from the correct value.
- The error due to bias is taken as the difference between the average prediction of our models and the correct value which we are trying to predict.

Error due to Variance

- Again, assume we can repeat the entire model building process multiple times.
- The variance is how much the predictions for a given point vary between different realizations of the model.
- The error due to variance is taken as the difference between the model prediction and average predictions of a given data point.
- Note that variance has nothing to do with where the actual target is.

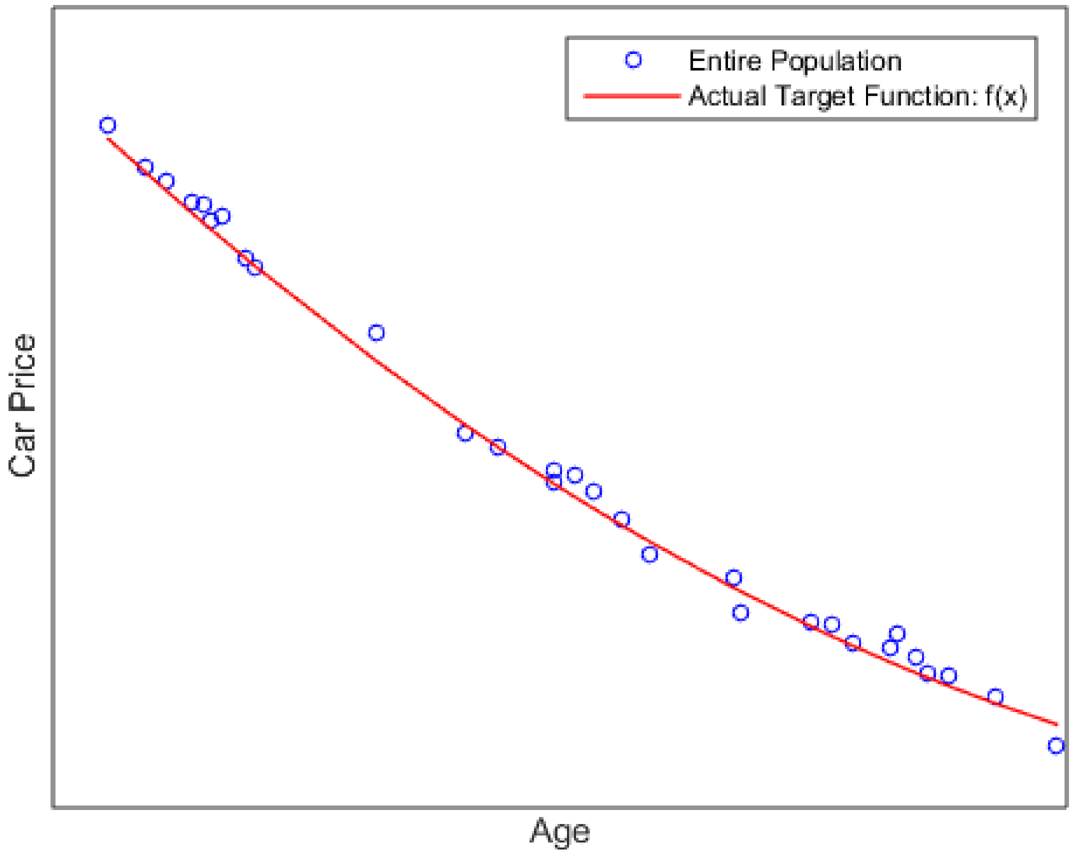


Graphical Illustration of Bias and Variance

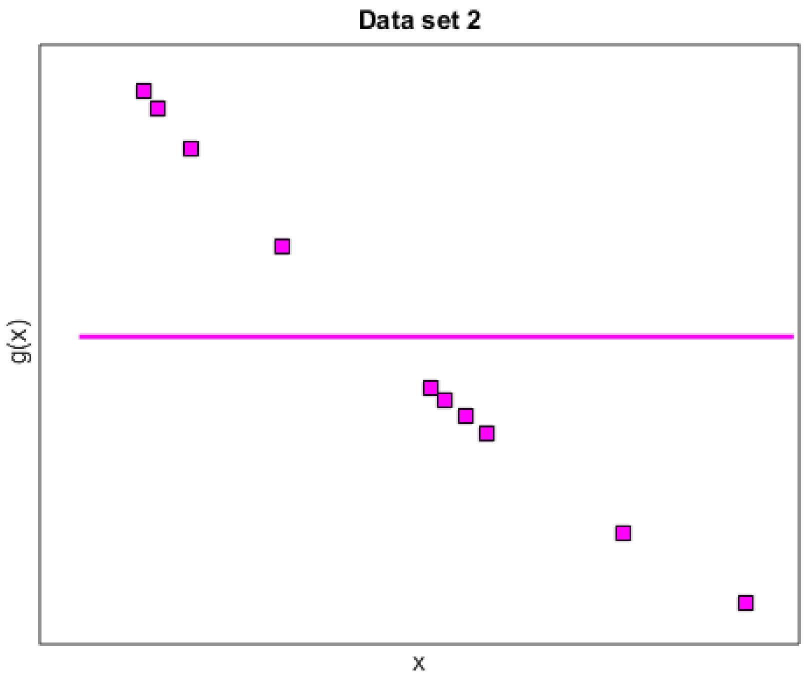


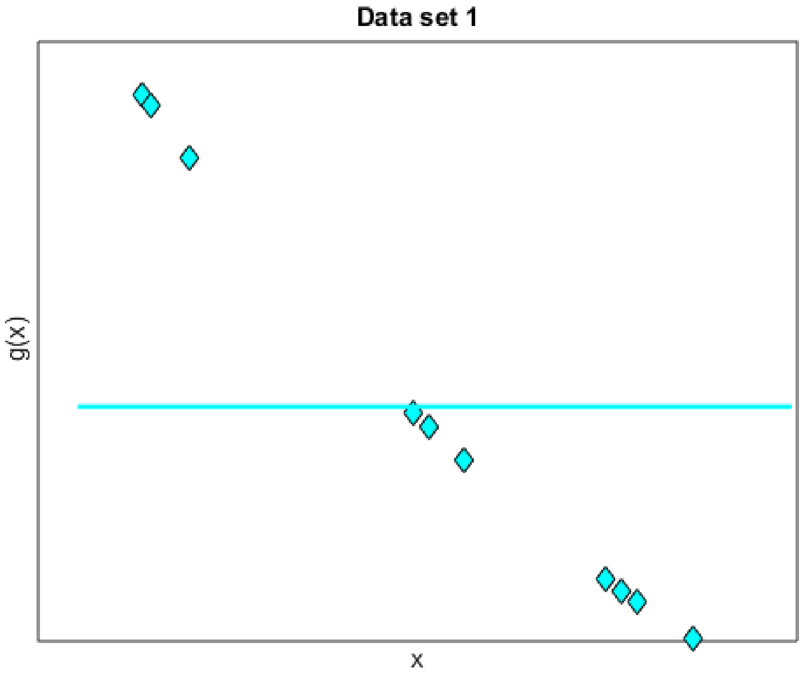
Example

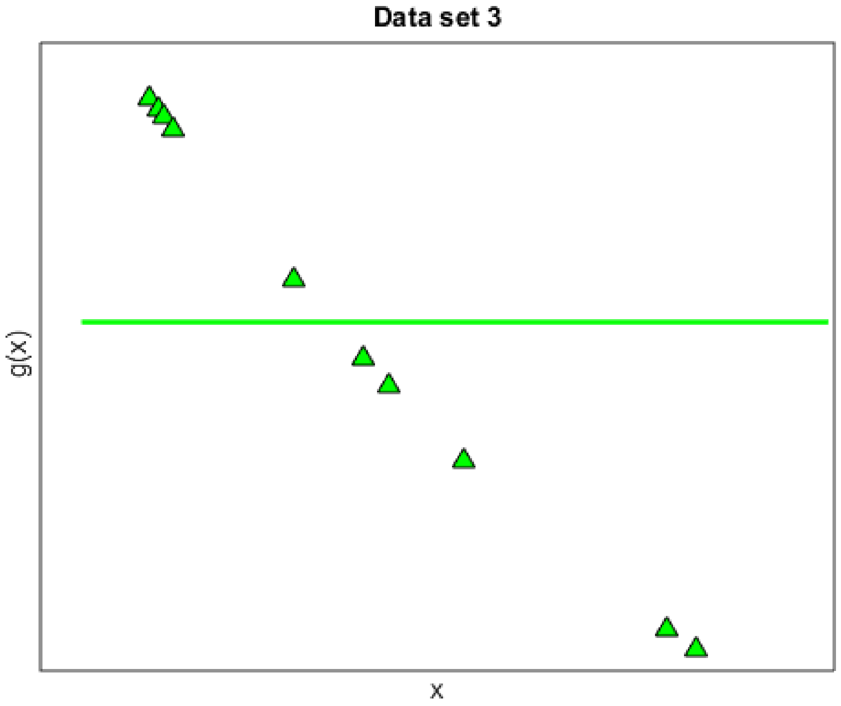
- The objective is to create a model that predicts the price of a car based on its Age.
- The red curve denotes the underlying relationship between the Age and the price of cars in the entire population.



- Assume different groups collect different samples and create a simple constant model based on their data set.
- Every data set results in a slightly different line g(x).
- The predicted hypotheses g(x) for a data set whose cars are worth below the true relationship, is different from a data set where most cars are worth more than the typical values in the population.

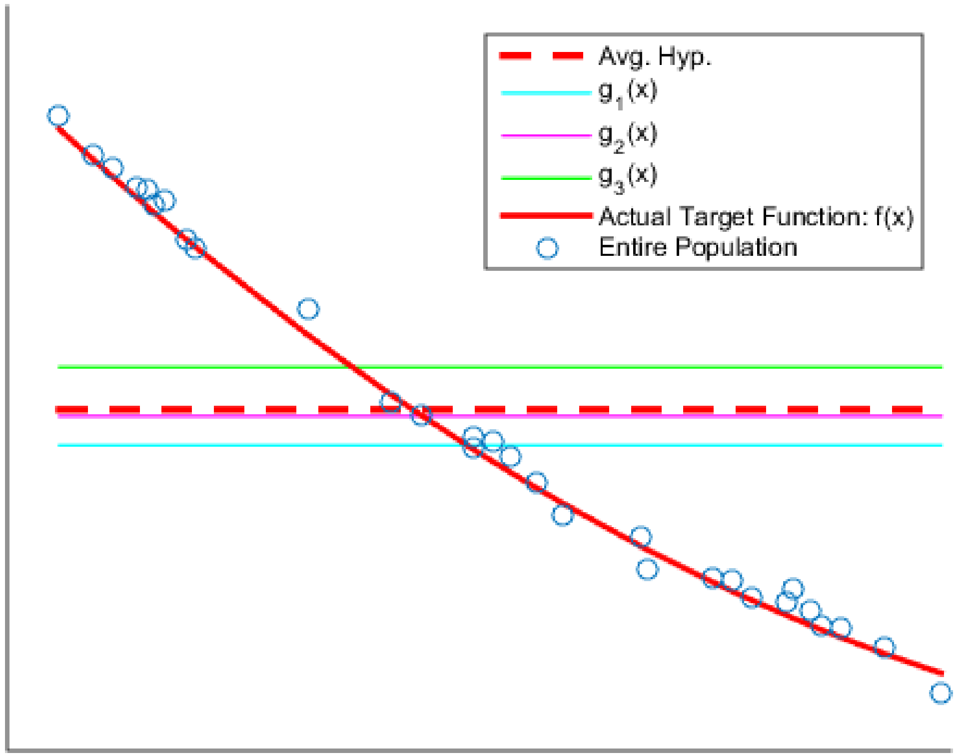






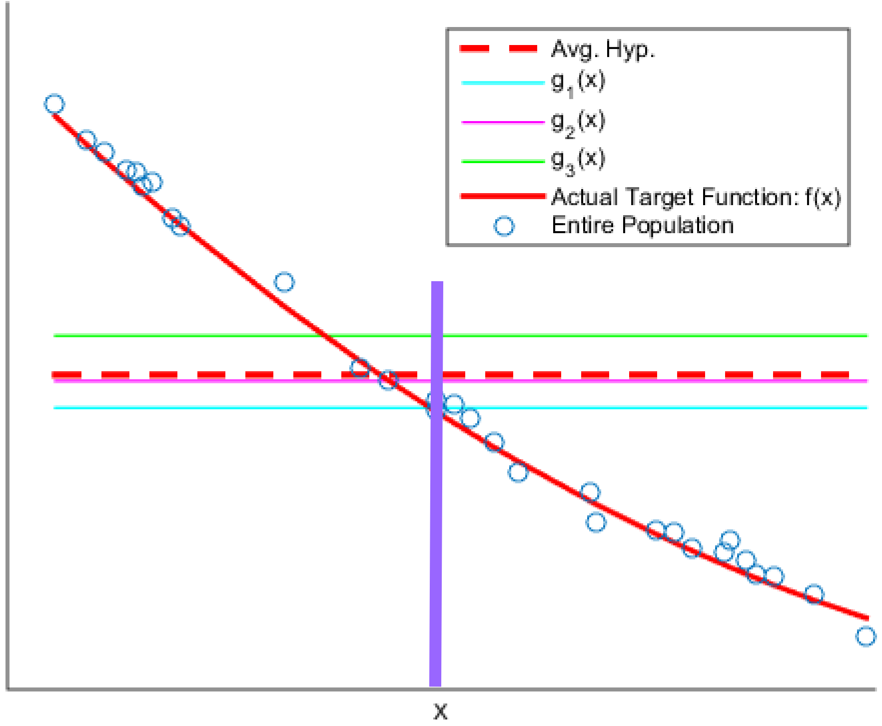
Average Hypothesis - Conceptually

- Every sample data set of car sales results in a different model.
- The question is what the **expected fit will look like, over all possible sets?**
- For a *very large* number of data sets there will be many fits, some of which may underestimate and some that overestimate the true value.
- Finding the **average** over all those data sets gives us the advantage of a much larger data set.
- The dashed red line represents the average fit which is the average over all fits (weighted by how likely they had occurred.)



Average Hypothesis – Mathematically

- Consider a fixed point **x** (e.g. Age = 6)
- Each of these data sets will provide a different prediction for price of the car at this particular **x**.
- i.e. $g(\boldsymbol{x})$ is a random variable and the source of randomness is to due to the different choice of data set.
- The expected value of all *g(**x**)* at a particular **x** will be $\bar{g}(\boldsymbol{x})$.
- $\bar{\boldsymbol{g}}(\boldsymbol{x}) = \boldsymbol{E}_{\boldsymbol{D}}[\boldsymbol{g}^{\boldsymbol{D}}(\boldsymbol{x})]$
- The subscript ***D*** denotes that we take the expected value over all data sets.
- The superscript ***D*** emphasizes that each hypothesis is a function of a particular data set.

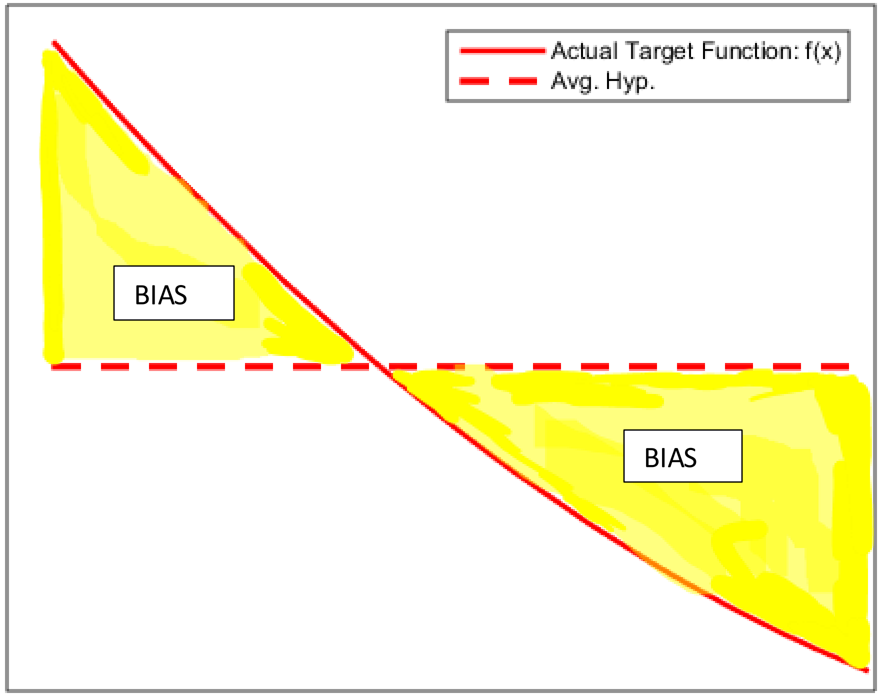


- Assume you have K data sets $D_1, D_2, \ldots D_K$.
- As K goes to infinity (i.e. you have many data sets):

$$\bar{g}(x) \approx \frac{1}{K}\sum_{k=1}^{K} g^{D_k}(x)$$

Bias - Conceptually

- Bias is the difference between the average hypothesis and the true function.
- A small bias indicates that the model on average is flexible enough to capture the true relationship between x and y.
- In our example, we see that a simple constant model is resulting in a large bias.
- This indicates that this low complexity hypothesis set is not flexible enough to capture the relationship between the age and price of a car.
- Therefore, bias leads to errors in future predictions.



Bias – Mathematically

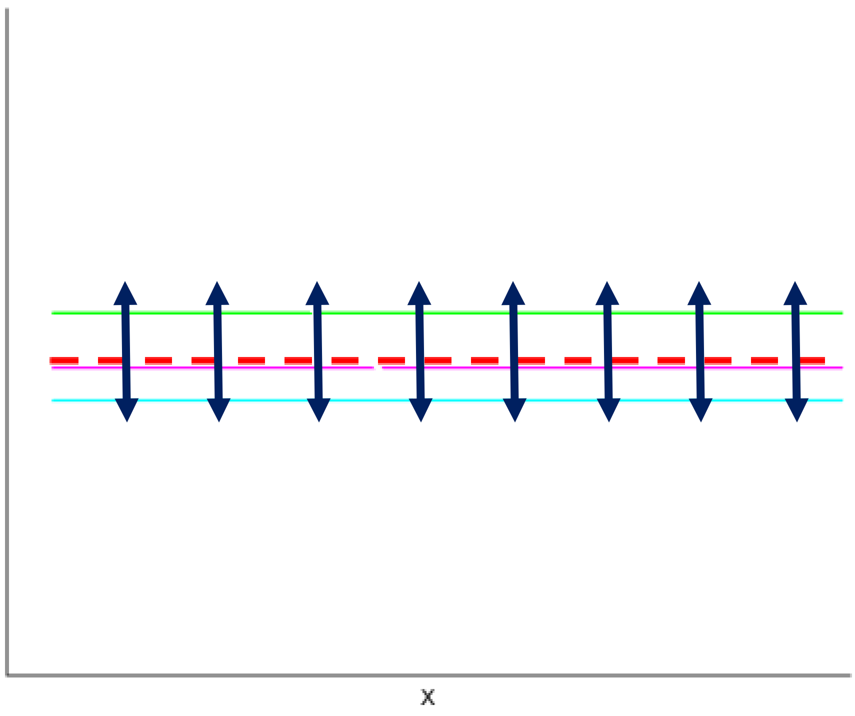
- Bias measures how much the average hypothesis deviates from the true function.
- Based on this definition we will have:

$$\boldsymbol{bias(x) = \bar{g}(x) - f(x)}$$

- Recall that the average hypothesis in a sense is the best fit that hypothesis set could do by taking advantage of unlimited data sets.
- Therefore, if the best fit in that hypothesis set still deviates from the target, that only shows the limitations of that hypothesis set.
- i.e. the learning model is not flexible enough to estimate the target function.

Variance – Conceptually

- Variance shows how different fits to a given data set vary from one another, when considering different possible data sets.
- In our example, we see that although the lines differ from one set to another, but across the space of all possible observations, they're fairly similar.



Variance – Mathematically

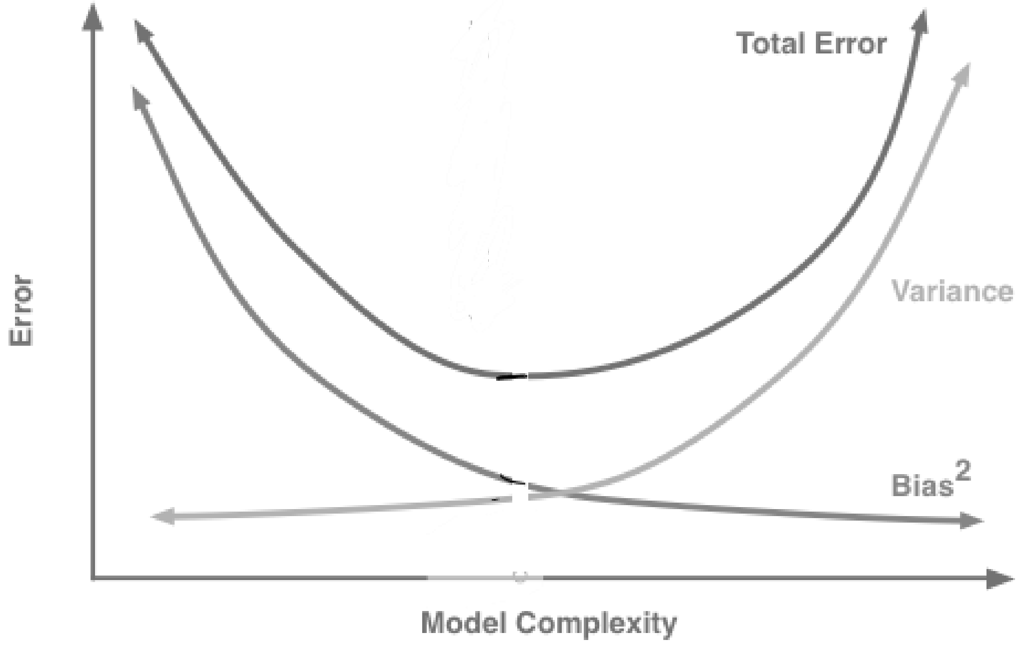
- Variance indicates how different fits vary from the expected fit depending on the data set.
- Based on this definition we will have:

$$var(x) = E_D[\left(g^D(x) - \bar{g}(x)\right)^2]$$

- Note that a large variance indicates that predications at a particular x can vary dramatically from one hypothesis set to another.
- Variance is a measure of instability. It manifests itself in wild reactions to small variations in the data, resulting in vastly different hypotheses.
- This sensitivity to a particular data set makes the predictions unreliable and is a source of error for future predictions.

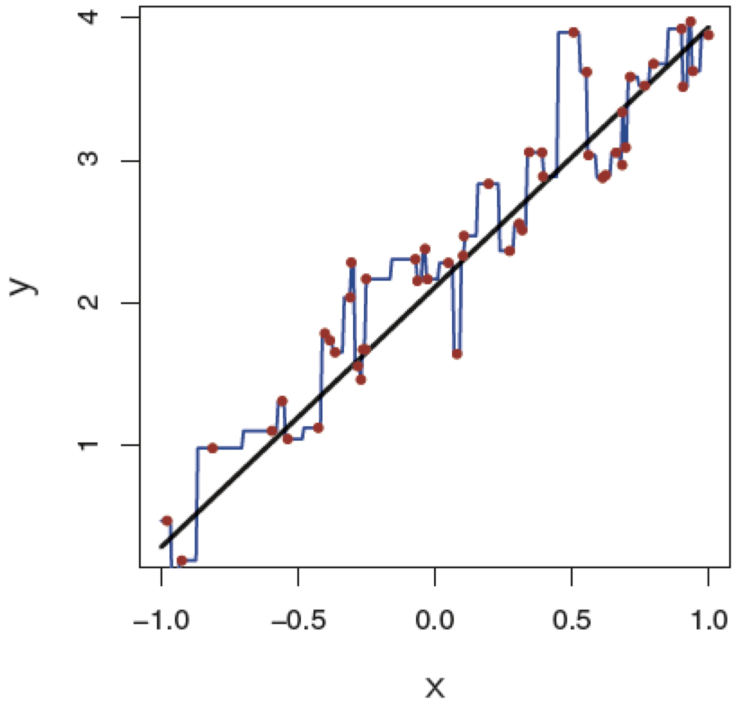
Bias – Variance Trade Off Plot

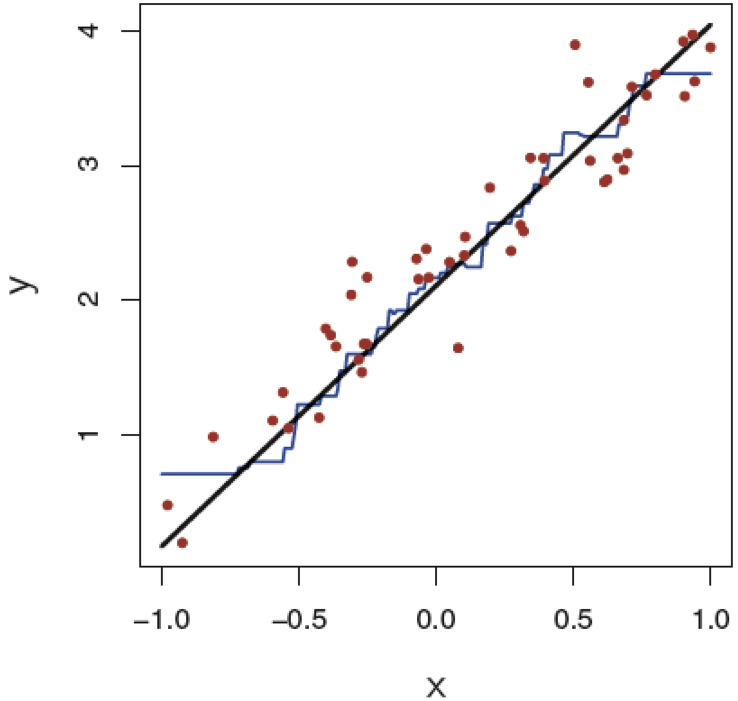
- As model complexity increases, bias decreases, because a better approximation of the true relationship between x and y can be obtained.
- As model complexity increases, variance increases, because the hypothesis is more flexible.
- As we will see, the total of $bias^2$ and variance is the MSE.
- The goal is to minimize the bias without significantly increasing the variance and minimizing the variance without increasing the bias too much.
- The point that minimizes MSE is the optimum point where bias and variance contribute the least to the predication error.



Example

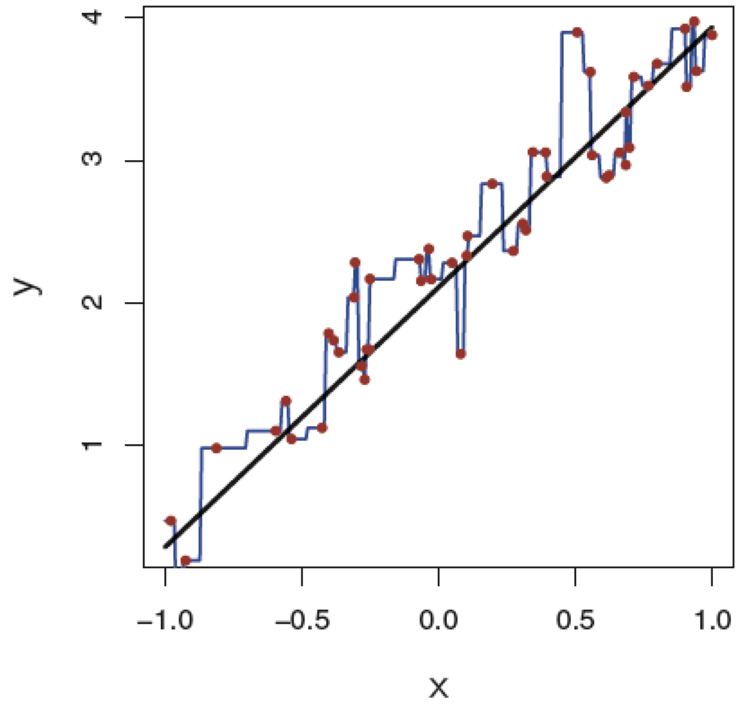
- We used KNN regression on a one-dimensional data set with 100 observations where the the true relationship is linear (black line).
- Top figure: The blue curve corresponds to K = 1 and passes directly through the training data. Bottom figure: The blue curve corresponds to K = 9, and represents a smoother fit.

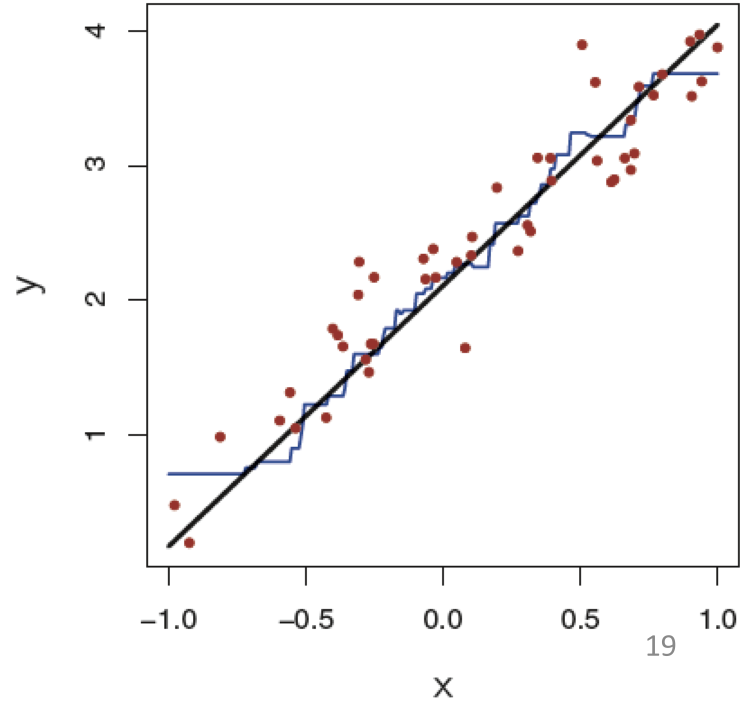




Example

- How do bias and variance compare in 2 figures?
- A small value for K provides the most flexible fit, which will have low bias but high variance. This variance is due to the fact that the prediction in a given region is entirely dependent on just one observation.
- In contrast, larger values of K provide a smoother and less variable fit; the prediction in a region is an average of several points, and so changing one observation has a smaller effect. However, the smoothing may cause bias by masking some of the structure in f (X).
- In general, the optimal value or K will depend on the bias-variance tradeoff.
- Note that since the true relationship is linear, it is hard for a non-parametric approach to compete with linear regression: **a non-parametric approach incurs a cost in variance that is not offset by a reduction in bias.**





Learning Curves

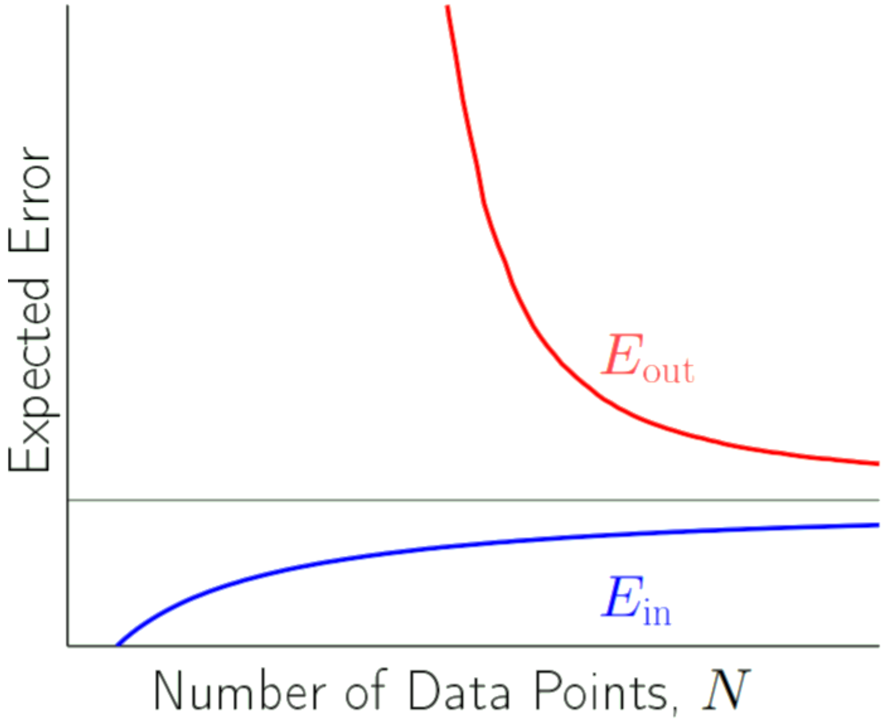
How do E_{in} and E_{out} vary as the size of data set, N changes?

Two things to note here:

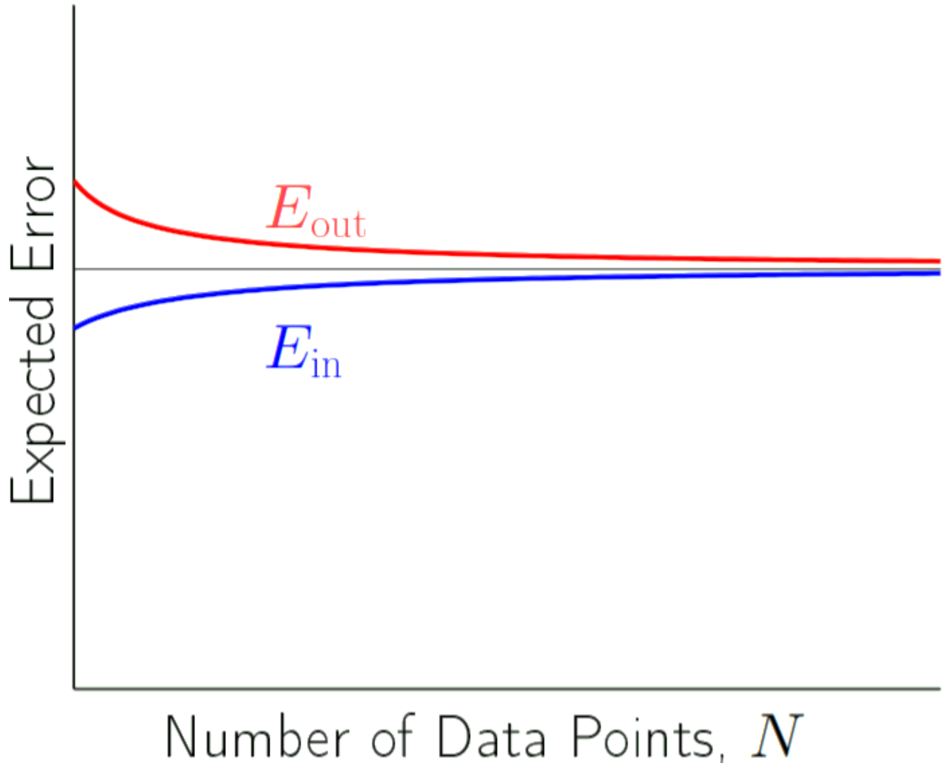
The learning curve of simple models converge more quickly.

The asymptotic value they approach to (final error) is smaller for the complex models.

Complex Model



Simple Model



Mathematical Decomposition of Bias and Variance

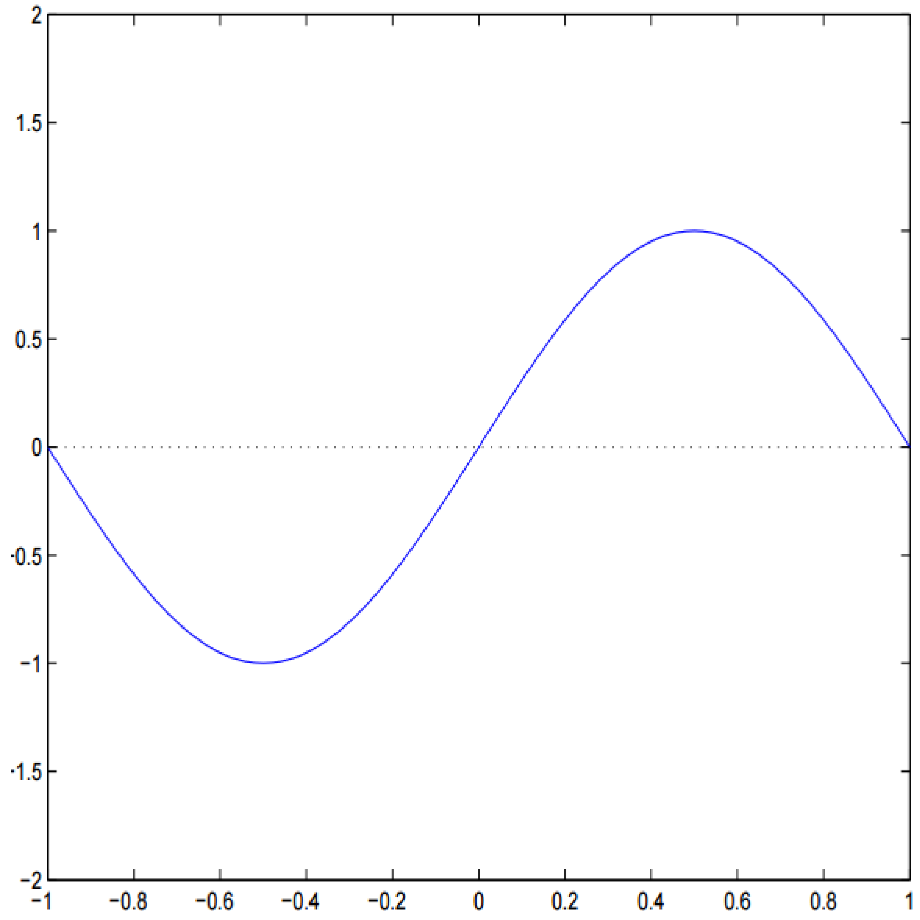
- $E_D\left[\left(g^D(x) - f(x)\right)^2\right] =$
- $E_D\left[\left(g^D(x) - \bar{g}(x) + \bar{g}(x) - f(x)\right)^2\right] =$
- $E_D\left[\left(g^D(x) - \bar{g}(x)\right)^2\right] + 2E_D\left[\left(g^D(x) - \bar{g}(x)\right)\left(\bar{g}(x) - f(x)\right)\right] + E_D\left[\left(\bar{g}(x) - f(x)\right)^2\right]$
- Note that:
- $2E_D\left[\left(g^D(x) - \bar{g}(x)\right)\left(\bar{g}(x) - f(x)\right)\right] = \left(\bar{g}(x) - f(x)\right)2E_D\left[\left(g^D(x) - \bar{g}(x)\right)\right] = \left(\bar{g}(x) - f(x)\right)\left(\bar{g}(x) - \bar{g}(x)\right) = 0$
- Therefore:
- $E_D\left[\left(g^D(x) - \bar{g}(x)\right)^2\right] + 2E_D\left[\left(g^D(x) - \bar{g}(x)\right)\left(\bar{g}(x) - f(x)\right)\right] + E_D\left[\left(\bar{g}(x) - f(x)\right)^2\right] = E_D\left[\left(g^D(x) - \bar{g}(x)\right)^2\right] + E_D\left[\left(\bar{g}(x) - f(x)\right)^2\right]$
- $var(x) + \left(\bar{g}(x) - f(x)\right)^2 = var(x) + bias(x)^2$

Mathematical Decomposition of Bias and Variance

- $E_D\left[\left(g^D(x)-\bar{g}(x)\right)^2\right]$
  $+2E_D\left[\left(g^D(x)-\bar{g}(x)\right)\left(\bar{g}(x)-f(x)\right)\right]+E_D\left[\left(\bar{g}(x)-f(x)\right)^2\right]=$
  $E_D\left[\left(g^D(x)-\bar{g}(x)\right)^2\right]+E_D\left[\left(\bar{g}(x)-f(x)\right)^2\right]$
- $var(x)+\left(\bar{g}(x)-f(x)\right)^2=var(x)+bias(x)^2$
- $E_X\left[var(x)+bias(x)^2\right]=var+bias^2$

Bias – Variance Trade Off Example

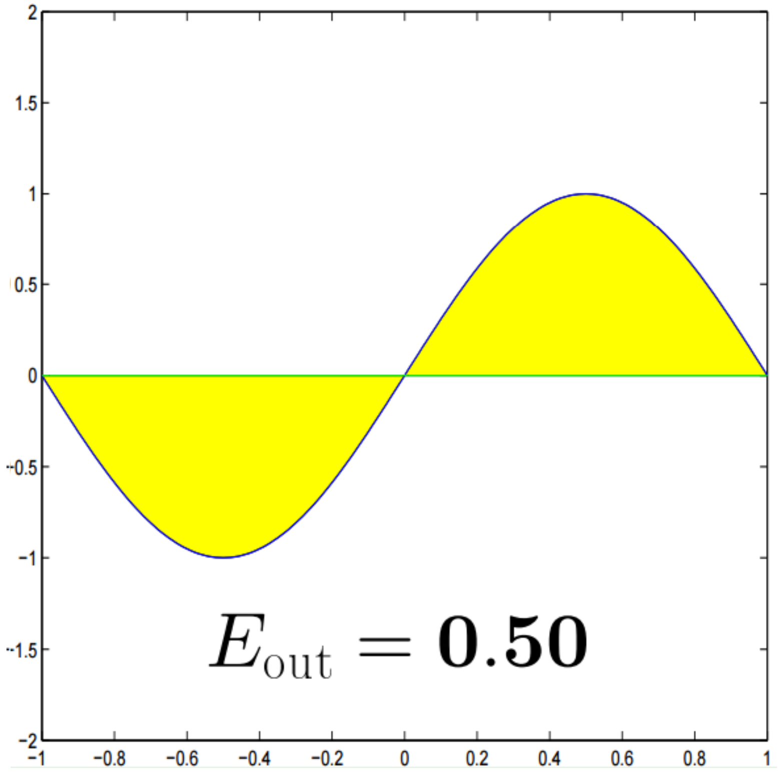
- We will consider two scenarios:
  1. Approximating a sinusoid
  2. Learning a sinusoid
- In each case you are limited to two sets of hypotheses:
  1. Constant model
  2. Linear model
- In each scenario determine which hypothesis is better.



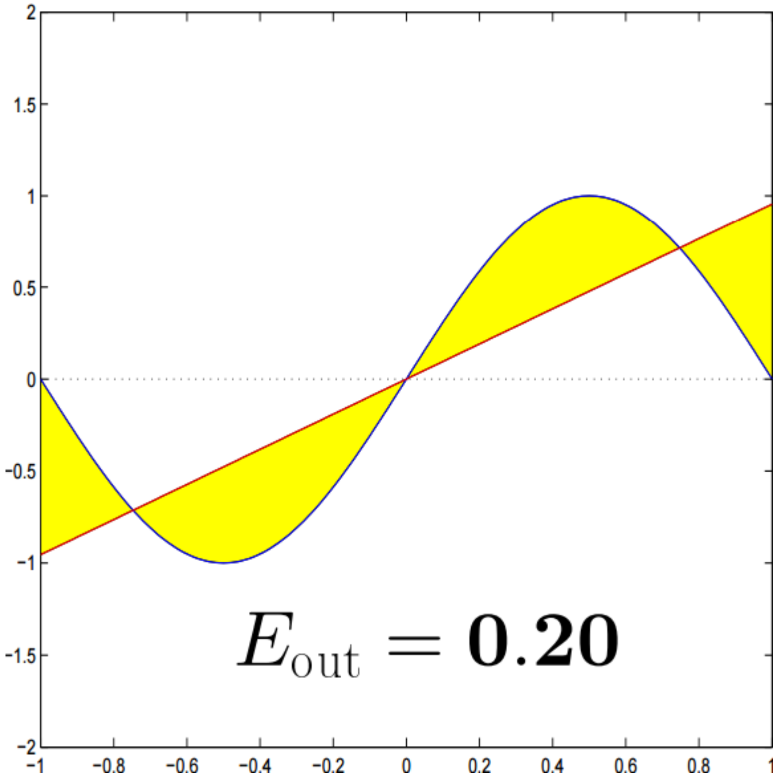
Scenario 1: Approximation

You **know** the target function. Do the best approximation using $h(x) = b$ and $h(x) = mx + b$ models.

Constant Model

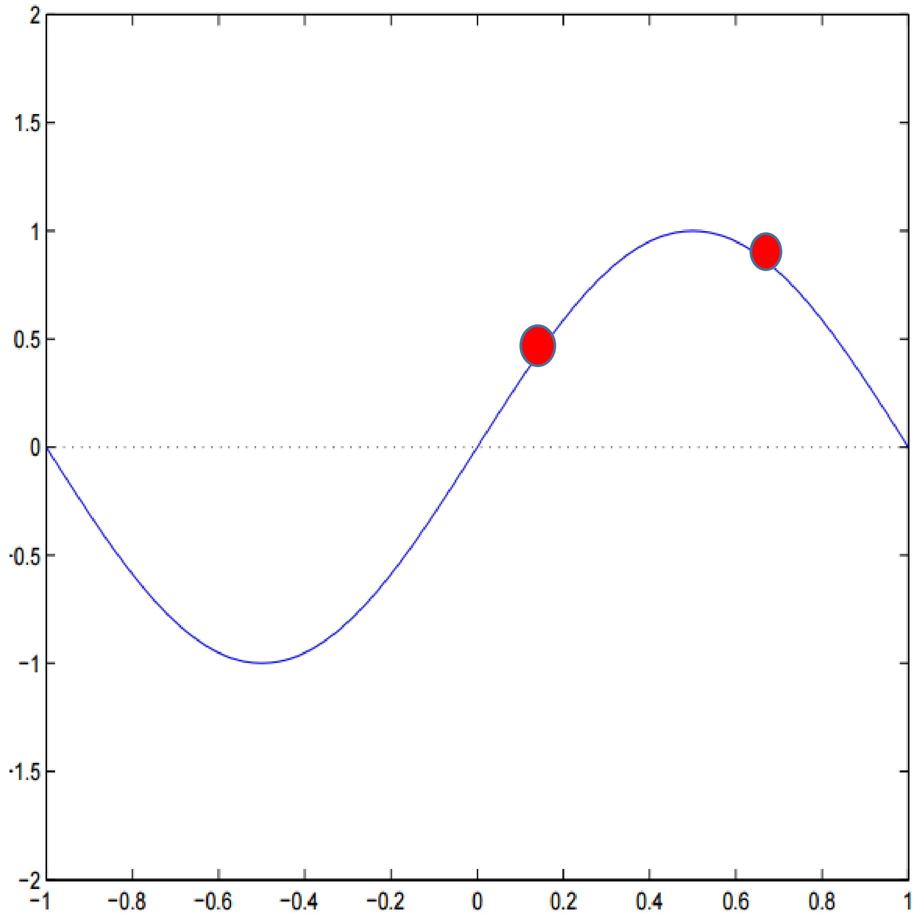


Linear Model



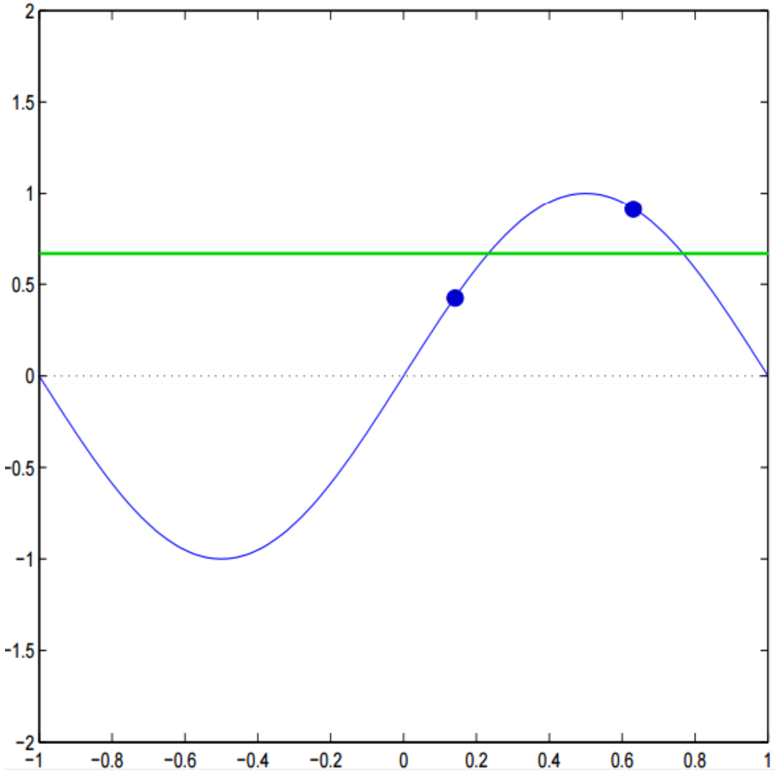
Scenario 2: Learning

- You **don't know** the target function.
- You must use your data set of size N=2 to learn the target function.
- Your hypotheses sets are
  - $h(x) = b$
  - $h(x) = mx + b$

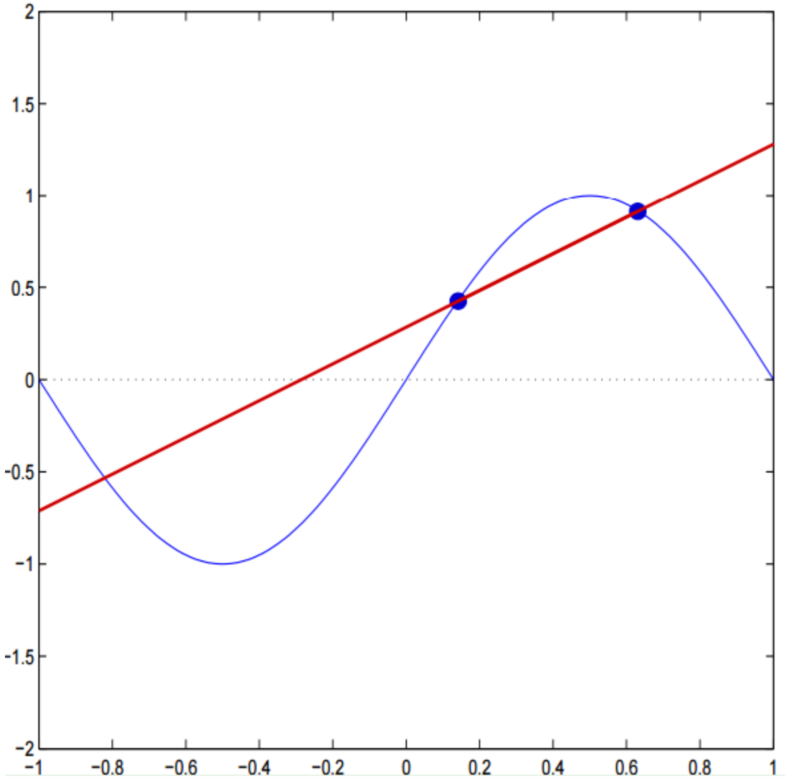


Example – cont.

Constant Model

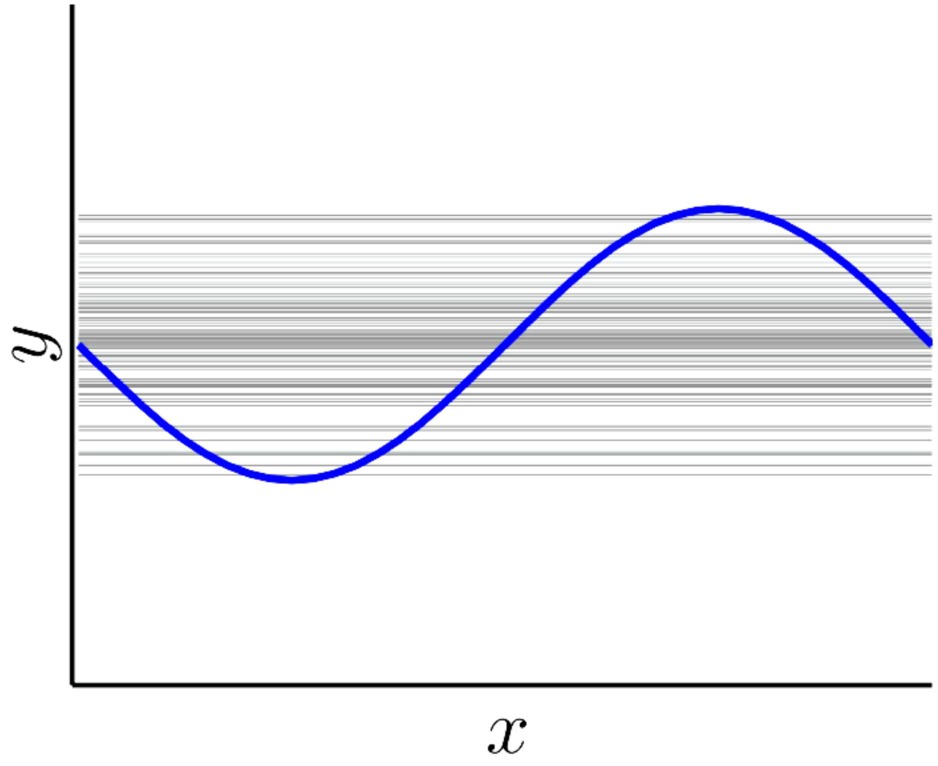


Linear Model

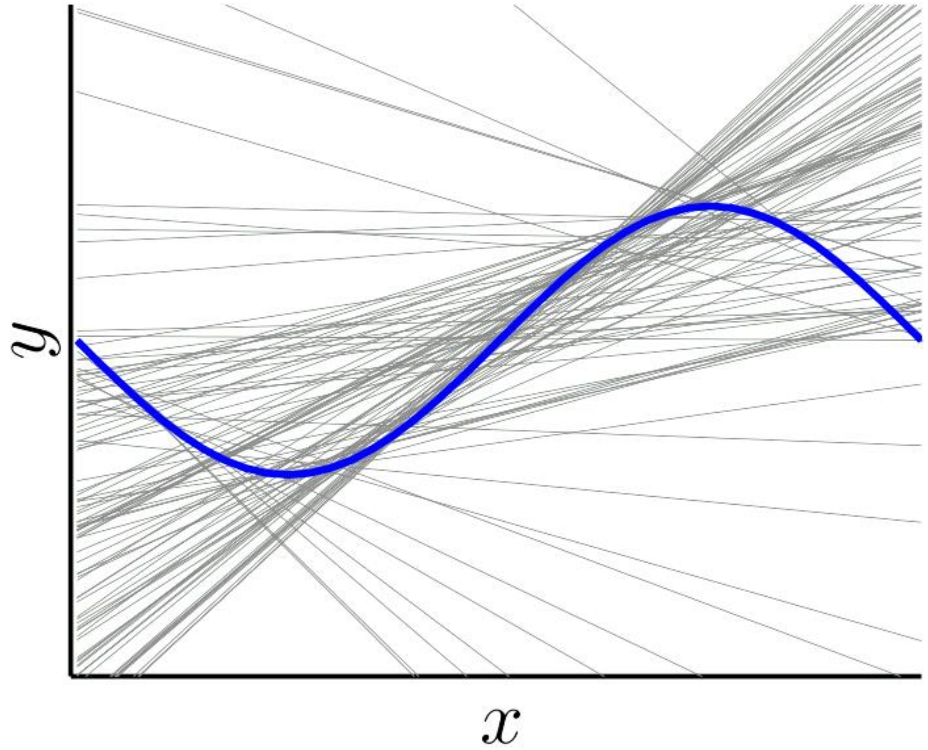


Repeating the Model Building with Different Data Sets

$y = b$

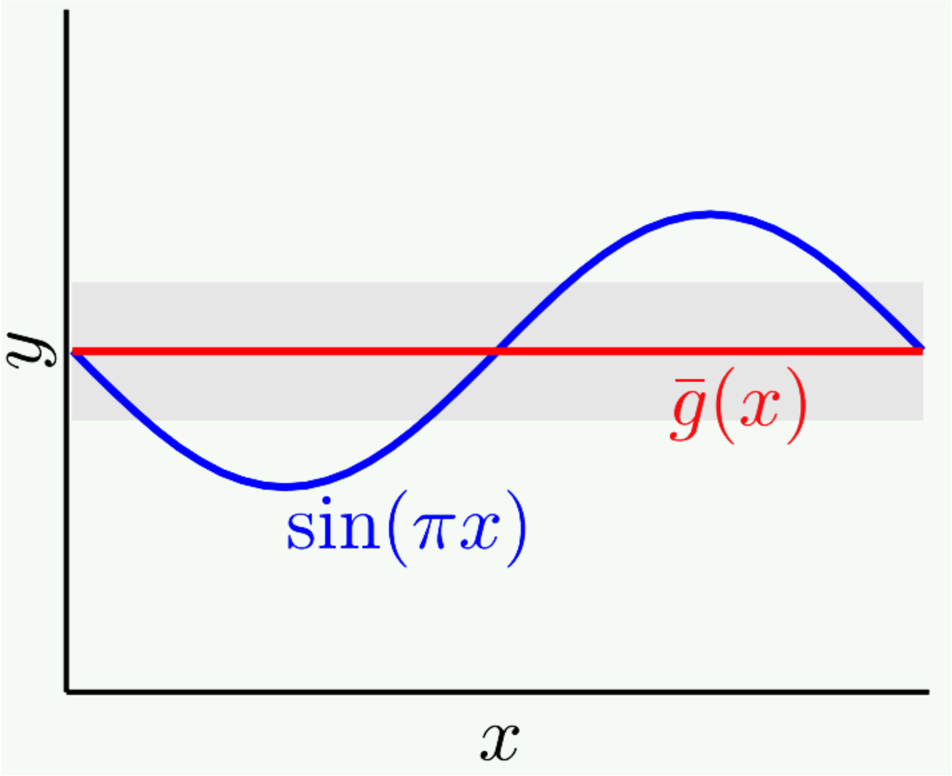


$y = b + mx$



Example – cont.

$y = b$

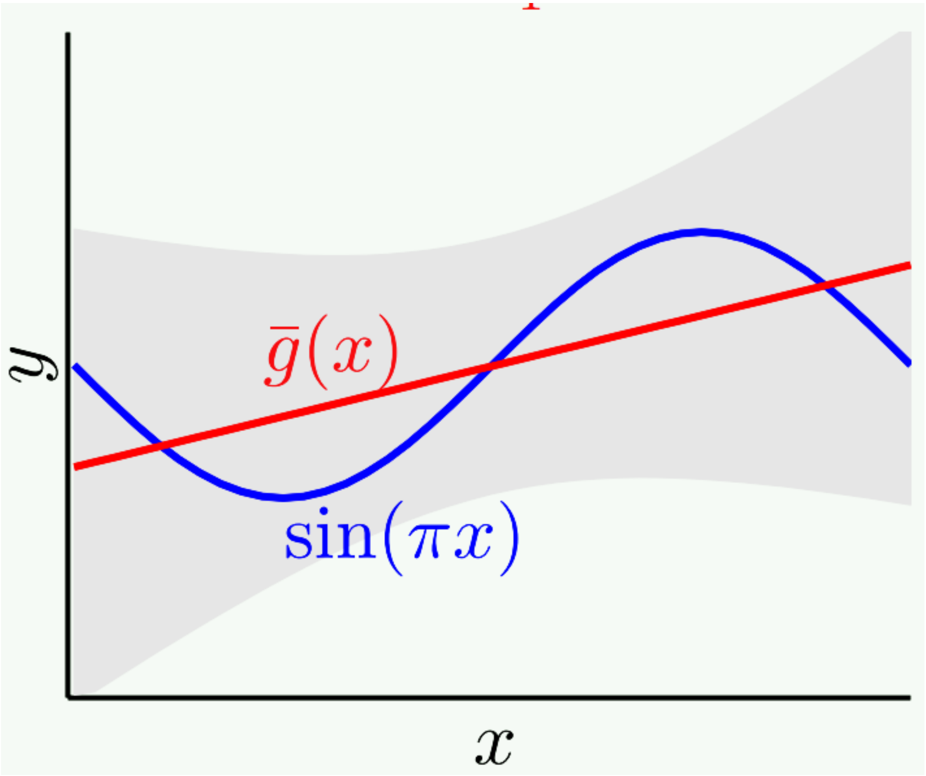


Bias $=?$

Variance = ?

Model too simple ⇔ High bias/low variance

$y = b + mx$



Bias = ?

Variance = ?

Model too complex ⇔ Low bias/high variance

Summary

- The bias and variance can not be computed in practice:
  - You need to know the true target function f(x)
  - You need to know the x probability distribution function(PDF) to find expected value.
- It’s a conceptual tool that helps with developing models.
- Techniques that are used to help with finding optimum points on bias-variance trade-off plots include
  - regularization
  - validation.

References

- Learning From Data by Yaser S. Abu-Mostafa, Malik Magdon-Ismail, and Lin
- An introduction to Statistical Learning with Applications in R, by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani