Lecture 1.4

- Bias, Variance
- Bias and Variance tradeoff

Bias and Variance

- The total error of a model can be decomposed into
  - Reducible errors (Bias and Variance)
  - Irreducible errors (noise)
- The **bias-variance trade-off** is an important concept in statistics and machine learning

Bias

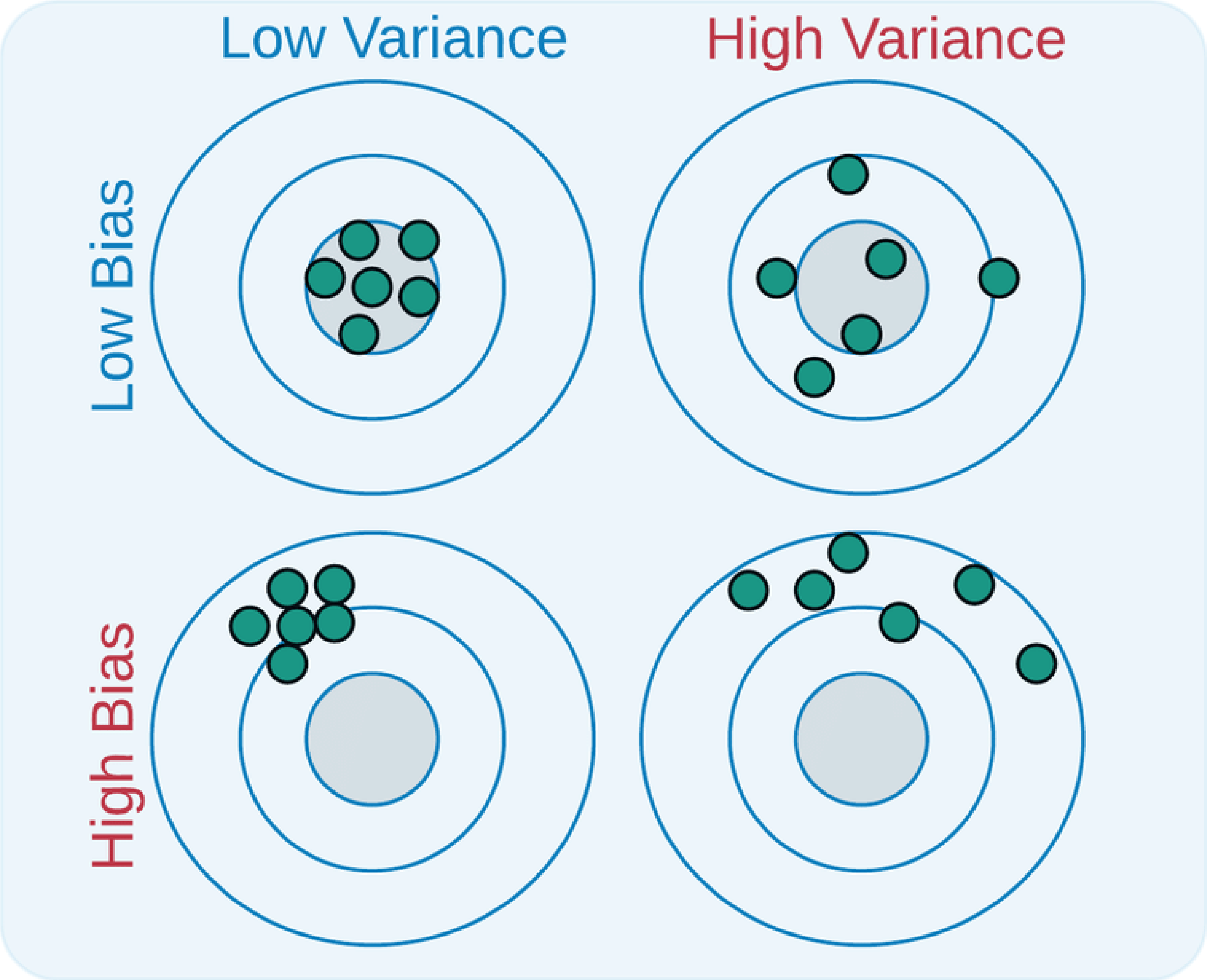
- The inability of a model to accurately capture the true relationship is called **bias**
- Models with high bias are simple and fail to capture the complexity of the data
- Low bias corresponds to a good fit to the training dataset

Variance

- Variance refers to the amount by which the estimate of the true relationship would change on using a different training dataset
- High variance implies that the model does not perform well on previously unseen data (testing data) even if it fits the training data well
- Low variance implies that the model performs well on the testing set

Overfitting and Underfitting

- Overfitting occurs when a model captures the noise along with the data pattern
  - A model that fits the the training data well but fails to do so on the testing set is an overfit to the data.
  - Overfitted models have low bias and high variance.
- Underfitting occurs when a model fails to even capture the pattern of the data
  - Such models have high bias and low variance


Low Variance
High Variance
Low Bias
High Bias

The Bias-Variance Trade-off

- This is a way to make sure that the model is neither overfitted nor underfitted
- Ideally, a model should have low bias so that it can accurately model the true relationship and low variance so that it can produce consistent results and perform well on testing data
- This is called a trade-off as it is challenging to find a model for which both the bias and variance are low

Total Error

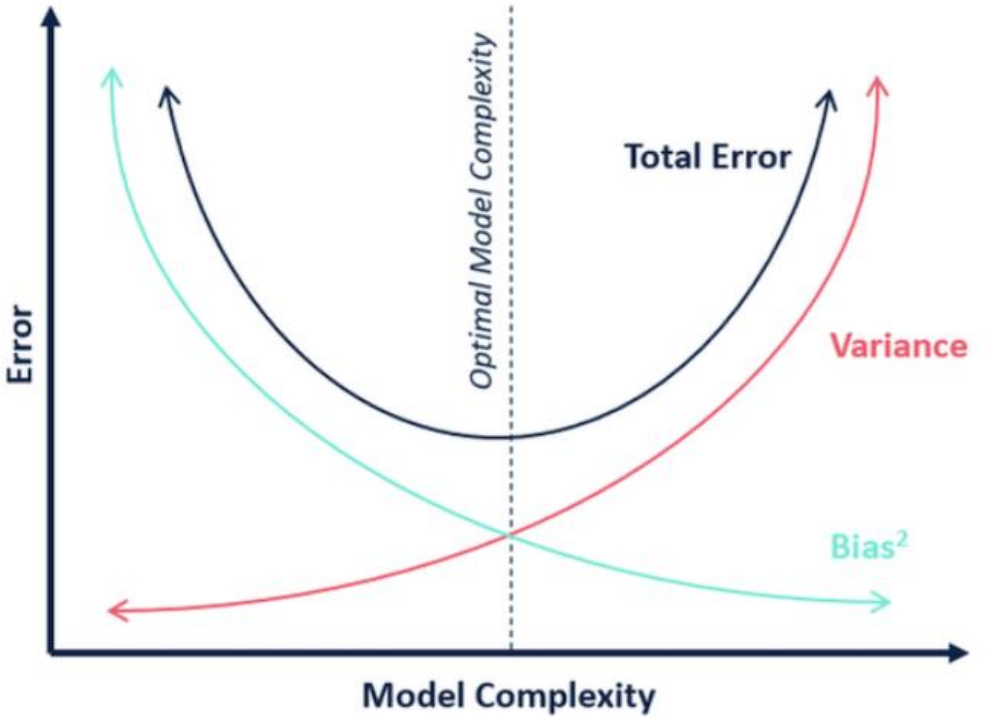
$$Err(x) = \mathrm{Bias}^2 + \mathrm{Variance} + \mathrm{Irreducible\ Error}$$

OR

$$\mathbb{E}[(y - \hat{f}(x))^2] = \mathrm{bias}[\hat{f}(x)]^2 + \mathrm{var}(\hat{f}(x)) + \sigma_\epsilon^2$$

- This equation suggests that we need to find a model that simultaneously achieves low bias and low variance
- Variance is a non-negative term and bias squared is also non negative which implies the total error can never go below the irreducible error

- **This graph suggests that as we increase model complexity(flexibility), the bias initially decreases faster than variance increases, consequently, the total error decreases**
- **However, at one point, increase the model complexity has little effect on the bias but the variance increases significantly, consequently , the total error also increases**



Derivation of Total Error 1

- We have independent variables **x** that affect the value of a dependent variable **y**
- Function **f** denotes the **true relationship** between **x** and **y**
- In real life problems it is very hard to know this relationship
- **y** is given by this formula along with some **noise** which is represented by the random variable **ϵ** with zero mean and variance σ_ϵ^2

$$y = f(x) + \epsilon$$

Where,

$$\mathbb{E}[\epsilon] = 0, \operatorname{var}(\epsilon) = \mathbb{E}[\epsilon^2] = \sigma_\epsilon^2$$

Derivation of Total Error 2

- Now, when we try to model the underlying real-life problem, we try to find a function $\hat{f}$ that can accurately predict the true relationship f
- The goal is to bring the prediction as close as possible to the actual value $(y \approx \hat{f}(x))$ to minimize the error

$$\mathbb{E}[(y - \hat{f}(x))^2] = \mathrm{bias}[\hat{f}(x)]^2 + \mathrm{var}(\hat{f}(x)) + \sigma_\epsilon^2$$

- $\mathbb{E}[(y - \hat{f}(x))^2]$ is called Mean Squared Error, commonly known as MSE
- This is defined as the average squared difference of a prediction $\hat{f}(x)$ from its true value y.
- Bias is defined as the difference of the average value of prediction from the true relationship function f(x)

$$\mathrm{bias}[\hat{f}(x)] = \mathbb{E}[\hat{f}(x)] - f(x)$$

- Variance is defined as the expectation of the squared deviation of $\hat{f}(x)$ from its expected value $\mathbb{E}[\hat{f}(x)]$

$$\mathrm{var}(\hat{f}(x)) = \mathbb{E}[(\hat{f}(x) - \mathbb{E}[\hat{f}(x)])^2]$$

Derivation of Total Error 3

$$
\begin{aligned}
\mathbb{E}[(y-\hat{f}(x))^2] &= \mathbb{E}[(f(x)+\epsilon-\hat{f}(x))^2] \\
&= \mathbb{E}[(f(x)-\hat{f}(x))^2] + \mathbb{E}[\epsilon^2] + 2\mathbb{E}[(f(x)-\hat{f}(x))\epsilon] \\
&= \mathbb{E}[(f(x)-\hat{f}(x))^2] + \underbrace{\mathbb{E}[\epsilon^2]}_{=\sigma_\epsilon^2} + 2\mathbb{E}[(f(x)-\hat{f}(x))]\underbrace{\mathbb{E}[\epsilon]}_{=0} \\
&= \mathbb{E}[(f(x)-\hat{f}(x))^2] + \sigma_\epsilon^2
\end{aligned}
$$

Derivation of Total Error 4

- Now, by further expanding the term on the RHS, $\mathbb{E}[(f(x) - \hat{f}(x))^2]$

$$
\begin{aligned}
\mathbb{E}[(f(x)-\hat{f}(x))^2] &= \mathbb{E}\left[\left((f(x)-\mathbb{E}[\hat{f}(x)]) - (\hat{f}(x)-\mathbb{E}[\hat{f}(x)])\right)^2\right] \\
&= \mathbb{E}\left[\left(\mathbb{E}[\hat{f}(x)] - f(x)\right)^2\right] + \mathbb{E}\left[\left(\hat{f}(x)-\mathbb{E}[\hat{f}(x)]\right)^2\right] \\
&\quad - 2\mathbb{E}\left[\left(f(x)-\mathbb{E}[\hat{f}(x)]\right)\left(\hat{f}(x)-\mathbb{E}[\hat{f}(x)]\right)\right] \\
&= \underbrace{(\mathbb{E}[\hat{f}(x)] - f(x))^2}_{=\text{bias}[\hat{f}(x)]} + \underbrace{\mathbb{E}\left[\left(\hat{f}(x)-\mathbb{E}[\hat{f}(x)]\right)^2\right]}_{=\text{var}(\hat{f}(x))} \\
&\quad - 2\left(f(x)-\mathbb{E}[\hat{f}(x)]\right)\mathbb{E}\left[\left(\hat{f}(x)-\mathbb{E}[\hat{f}(x)]\right)\right] \\
&= \text{bias}[\hat{f}(x)]^2 + \text{var}(\hat{f}(x)) \\
&\quad - 2\left(f(x)-\mathbb{E}[\hat{f}(x)]\right)\left(\mathbb{E}[\hat{f}(x)]-\mathbb{E}[\hat{f}(x)]\right) \\
&= \text{bias}[\hat{f}(x)]^2 + \text{var}(\hat{f}(x))
\end{aligned}
$$

Derivation of Total Error 5

- $\mathbb{E}[\hat{f}(x)] - f(x)$ is a constant since we subtract $f(x)$, a constant , from $\mathbb{E}[\hat{f}(x)]$ which is also a constant
- So, $\mathbb{E}[(\mathbb{E}[\hat{f}(x)] - f(x))^2] = (\mathbb{E}[\hat{f}(x)] - f(x))^2$
- Further expanding using the linearity property of expectation we get the value of $\mathbb{E}[(f(x) - \hat{f}(x))^2]$
- Plugging this value back into the equation for $\mathbb{E}[(y - \hat{f}(x))^2]$, we arrive on our final equation

$$\mathbb{E}[(y - \hat{f}(x))^2] = \mathrm{bias}[\hat{f}(x)]^2 + \mathrm{var}(\hat{f}(x)) + \sigma_\epsilon^2$$