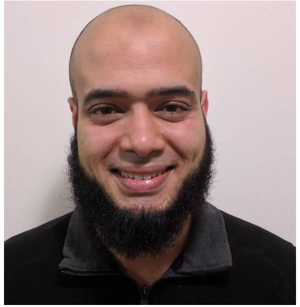
Machine Learning

Bias–Variance tradeoff

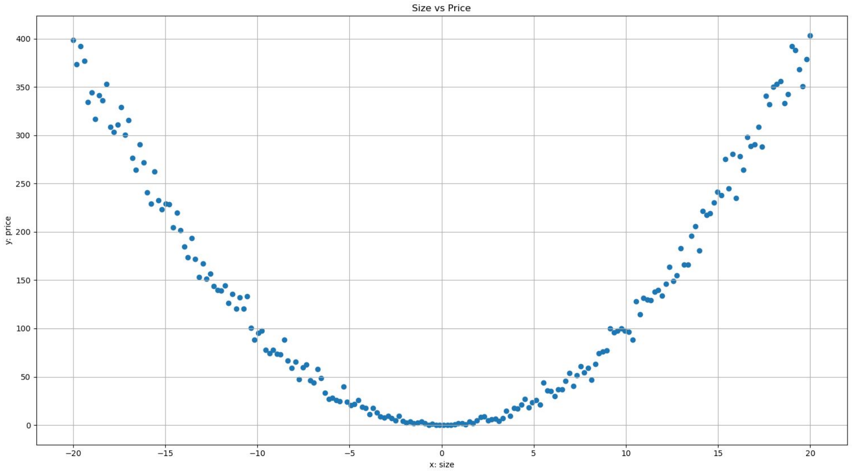
Mostafa S. Ibrahim
Teaching, Training and Coaching for more than a decade!

Artificial Intelligence & Computer Vision Researcher
PhD from Simon Fraser University - Canada
Bachelor / MSc from Cairo University - Egypt
Ex-(Software Engineer / ICPC World Finalist)



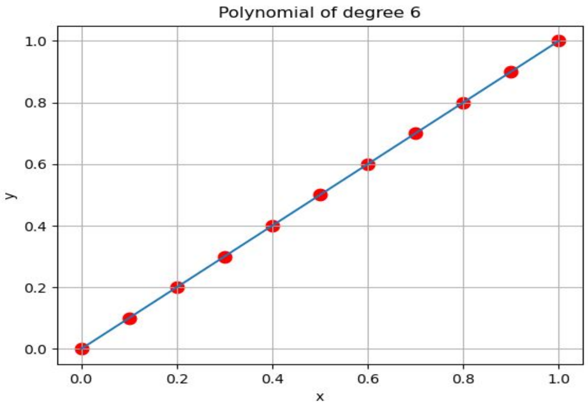
Model Bias

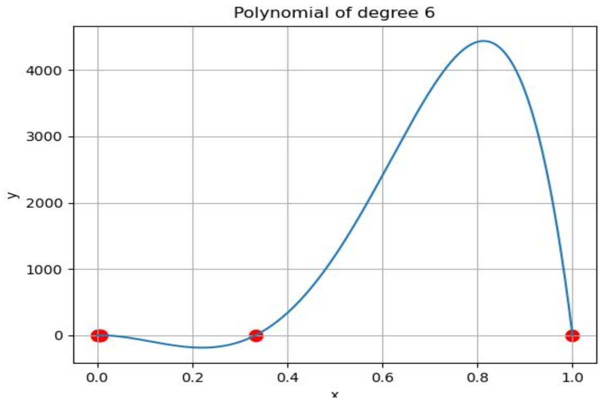
- No matter how you tried to fit a line, it can’t perfectly fit this data!
    - We know this is an example of underfitting
- Why? The line doesn’t have **enough flexibility** to match the data!
    - We call this phenomena **bias**
    - The model is biased toward representing linear input/output relationships
- Models with strong **bias** suffer from **underfitting**

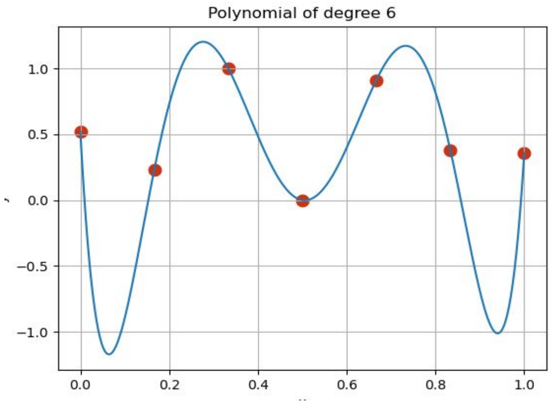


Model Variance

- Assume we generated **several datasets** as below. Then for each one, we fitted a polynomial of 6th degree. You will notice the model is very flexible to be able to adapt itself to different data points / critical points
- The flexibility can vary a lot. The difference between dataset show **variance**
- Models with strong **variance** tend to suffer from **overfitting**
  - Assume the test set is one for all (sinx). All these models will fail to generalize

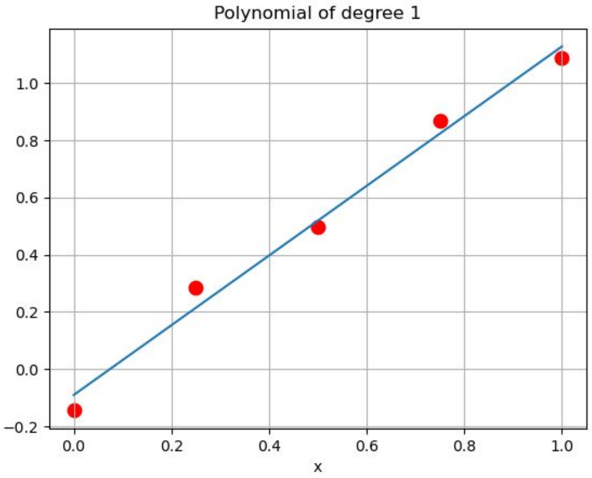


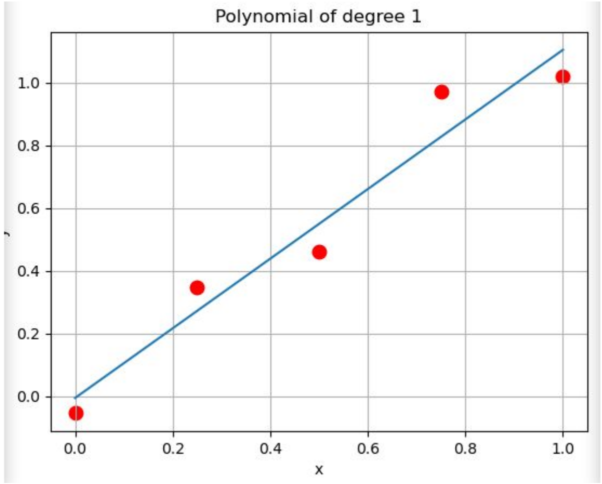


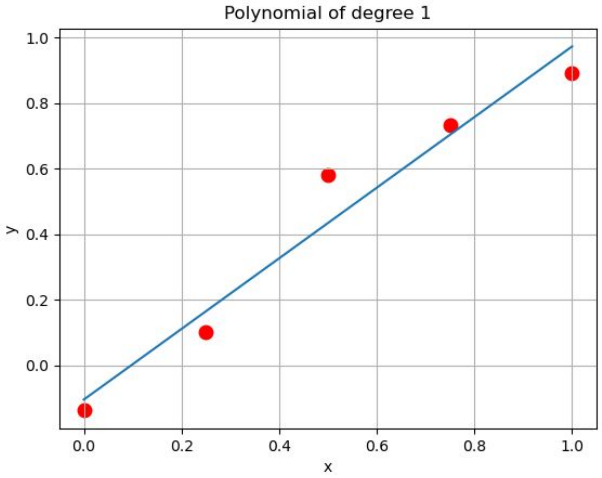


High-Bias Low-Variance

- Assume we generated **several datasets** sampled from: y = x + noise
- We can see the line model change slightly from one dataset to another
- This means the line has **high bias** but low variance

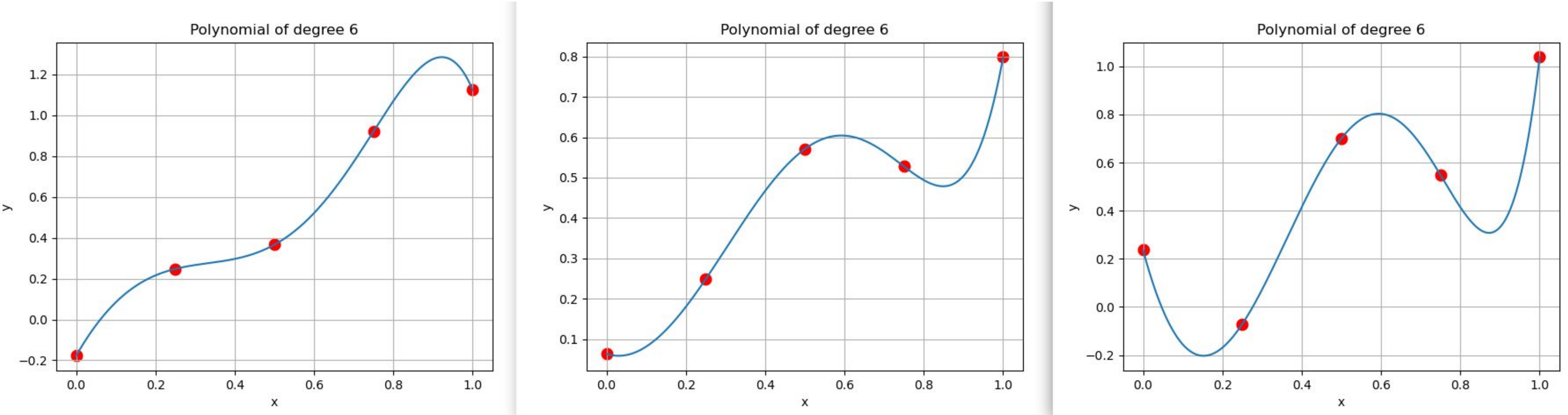






Low-Bias High-Variance

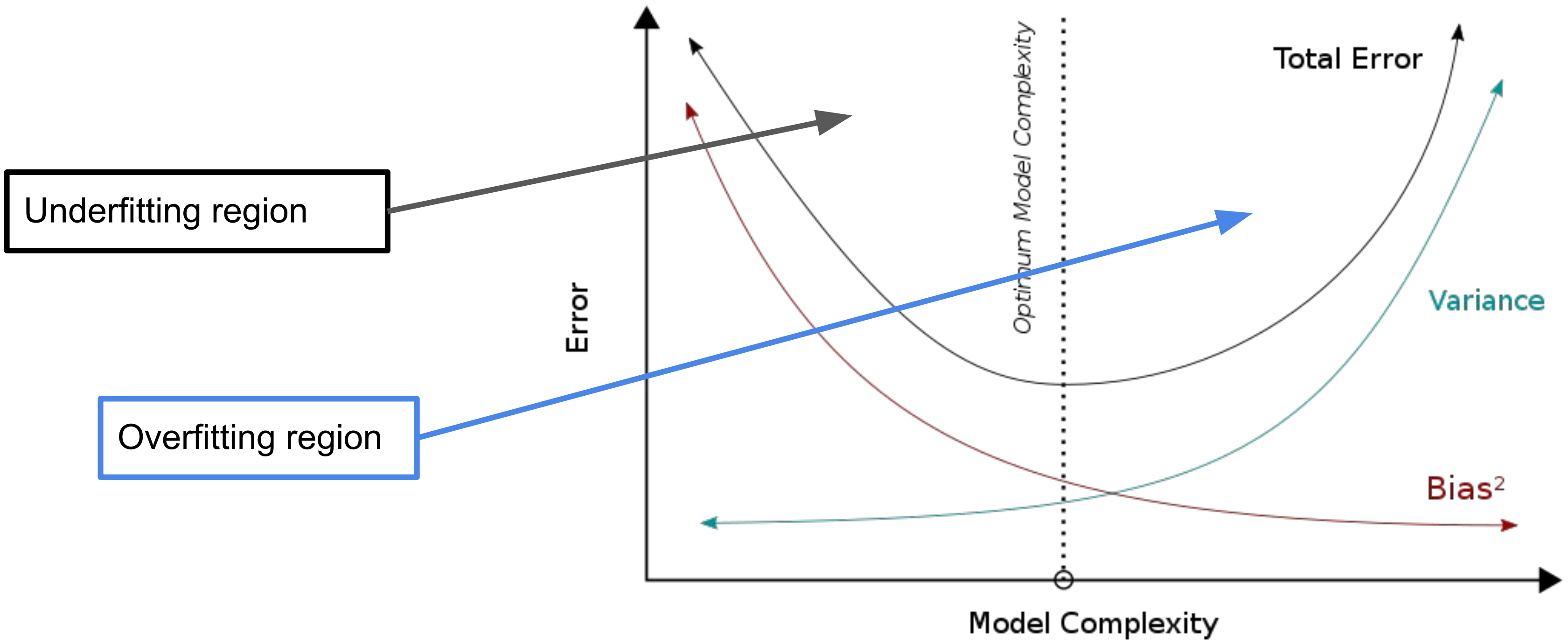
- On the other side, a 6th degree polynomial will vary a lot from one dataset to another, even for simple line data
- This means the 6-degree polynomial has low bias but **high variance**
- The line will provide consistent predictions, while the 6th-poly won't



Bias–variance tradeoff

- High bias: restricted model due to its assumptions (e.g. line)
  - A high bias might lead to **underfitting** (too simple model)
  - Typically the model will have **low** variance
- High variance: (a lot of freedom)
  - Sensitive to small changes in the data
  - A high variance might lead to **overfitting** (too complex model)
  - Typically the model will have **low** bias
- There is a **conflict** between them
  - People on one extreme are generous (كريم) and the other extreme are miserly (بخيل)
- A good model is somewhere in between these 2 extremes
- Regularization can help us use a complex model, and prevents the model from memorizing the dataset

Bias–variance tradeoff



Bias–variance Decomposition

- Assume we have a dataset from the function y = f(x) + ε (error)
  - var(ε) = σ^2 = Irreducible error
- Whatever model f' we use (e.g. Ridge), the **expected MSE** of point (x0, y0) is defined as 3 terms: irreducible error + bias^2 + variance
  - The expectation is computed based on several datasets (as illustrated in earlier lectures)
  - In other words: 3 error factors that accumulate the **overall** error
  - irreducible error: Noisy data

$$= \mathbb{E}\left[(y_0 - \hat{f}(x_0))^2\right]$$

$$= \sigma^2 + \underbrace{\left(\mathbb{E}[\hat{f}(x_0)] - f(x_0)\right)^2}_{\text{Bias}^2} + \underbrace{\mathbb{E}\left[\left(\hat{f}(x_0) - \mathbb{E}[\hat{f}(x_0)]\right)^2\right]}_{\text{Variance}}$$

Bias–variance Decomposition

- Assume we have 5 datasets. We trained 5 models
- The **expected** performance of your model = The average of all the 5 models
  - E.g. compute average of the 5 house price estimates
  - We call this: E[f^(x)]
- f(x): is the ground truth (real function) for the data
- Refer to Bishop book for further details

$$
\begin{aligned}
&= \mathbb{E}\left[(y_0 - \hat{f}(x_0))^2\right] \\
&= \sigma^2 + \underbrace{\left(\mathbb{E}[\hat{f}(x_0)] - f(x_0)\right)^2}_{\text{Bias}^2} + \underbrace{\mathbb{E}\left[\left(\hat{f}(x_0) - \mathbb{E}[\hat{f}(x_0)]\right)^2\right]}_{\text{Variance}}
\end{aligned}
$$

Bias–Variance Decomposition

- From the equation:
- Bias: the difference between the model's average/expected prediction and the true function
- Variance (general definition): variance is the expectation of the squared deviation of a random variable from its mean
  - Measures how spread out the value are from the mean
  - In our context, variance measures how much the performance of the machine learning model differs when evaluated on different datasets

$$\sigma^2 + \underbrace{\left(\mathbb{E}[\hat{f}(x_0)] - f(x_0)\right)^2}_{\text{Bias}^2} + \underbrace{\mathbb{E}\left[\left(\hat{f}(x_0) - \mathbb{E}[\hat{f}(x_0)]\right)^2\right]}_{\text{Variance}}$$

Bias–Variance in practice

- In practice, we use the train/val error (or accuracy) as a **proxy** for the tradeoff
- Cross-validation mean/variance is indicator to bias-variance
- Low train error + low validation error ⇒ **Perfect** fit (low bias, low variance)
  - Assuming no data leakage or experimentation errors
- Low train error + high validation error ⇒ **Overfitting** (low bias, high variance)
- High train error ⇒ **Underfitting** (high bias)
  - We typically don't validate then!
  - If there is a gap between train and validation errors, this also indicates high variance
- The terms "low" and "high" are relative to data complexity and model complexity
  - A 3rd degree polynomial can be: underfit, perfect fit or overfit
  - We can see the line as a model with high bias albeit one which can fit the data perfectly

Question!

- Assume that our model has 17% train error and 20% test error in a classification task that involves classifying 50 species for 100 animals
  - There are more than **11,000 bird species**
- Is this a high error?

- Imagine we asked you to learn these 50 x 100 cases for a month
- Then we tested you with 1000 images
- What is your expected error?

- A Human error in this task can be very high
- In theory, understanding the model error should be relevant to human error
- Side note: this is one task where ML can easily beat human performance!

Bagging and Resampling techniques

- Assume we have a low-bias high variance model
- Then we train this model on multiple subsets of our datasets
- In testing, we calculate the **average** result of **all** the models
- This **averaged model** reduces the **variance**!
- This means that we may be able to achieve a low-bias model that also **reduces** the variance

K-fold and the trade-off

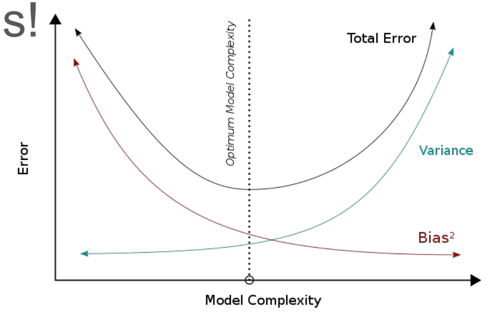
- We already mentioned good k-fold values (e.g. 5 and 10)
- In theory:
  - Lower K has cheaper computations, less variance and more bias
  - Higher K has more computations, more variance, and lower bias
- Assume we computed models mean and standard deviation.
- Which model to select?
  - Logically, the model with minimum error
  - Another **empirical** rule on the classical ML days: “one-standard error”:
    choose the **simplest** model within **one standard error** of the minimum error
  - Why: hopefully reduce overfitting by balancing between model complexity and performance

Hyperparameter λ and bias-variance tradeoff

- A higher λ value
  - increases the amount of regularization
  - more shrinkage of the coefficients
  - **lower variance but potentially higher bias**
- A lower λ value
  - reduces the amount of regularization
    - allowing the model to fit the training data more closely
  - **lower bias but higher variance**
  - when the dataset is larger, cleaner, or when the underlying relationships are complex and require a more flexible model
- Practically; use cross-validation

Deep learning and Overfitting

- Intentionally, I explained these concepts in a way that is similar to the popular old machine learning books and mainstream ideas/posts
- Deep Learning actually challenges some of what we learned in classical ML
- For example, we found that deeper models (which are more complex) decrease the test error, which contradicts our expectations!
    - Does the test error really follow a U-shape? It seems not
    - Is the test error enough for bias-variance analysis? Also no



- In general, deep learning often behave **differently** than traditional ML
    - Large amount of data + smart regularization (dropout) + Architectural Design (e.g. structural constraints from CNN) are key factors

Relevant resources

- Tradeoff: Article, Article, Article, Article
- Concerns relevant to deep learning
  - Article
  - Paper: A Modern Take on the Bias-Variance Tradeoff in Neural Networks
  - Double Descent
- Bayes error rate (~irreducible error)
- Linear regression - inductive bias

"Acquire knowledge and impart it to the people."

"Seek knowledge from the Cradle to the Grave."