Life Sciences and
Facility Management

Institute of
Computational Life Sciences

Bias-Variance Tradeoff

Machine Learning

Norman Juchler

Learning objectives

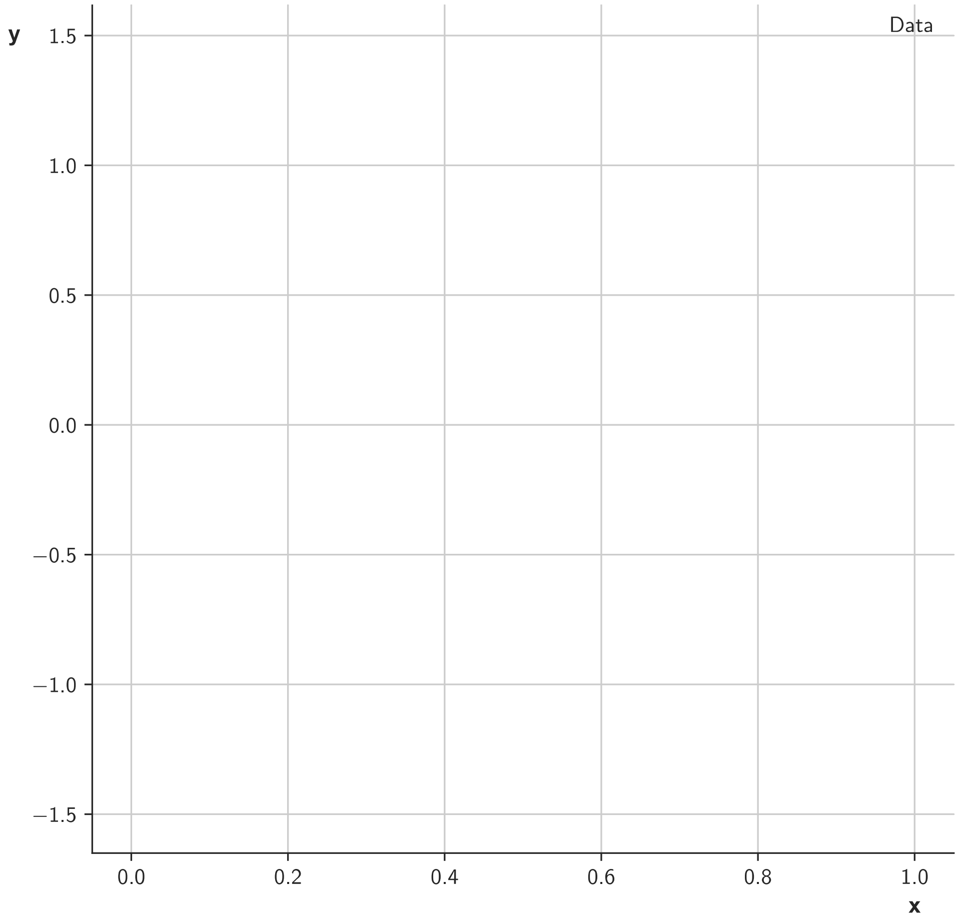
- What are bias and variance?
- How do they contribute to the test error?
- What are the effects of under- and overfitting?
- How to choose optimal model capacity?

Example problems

Example problem: Regression

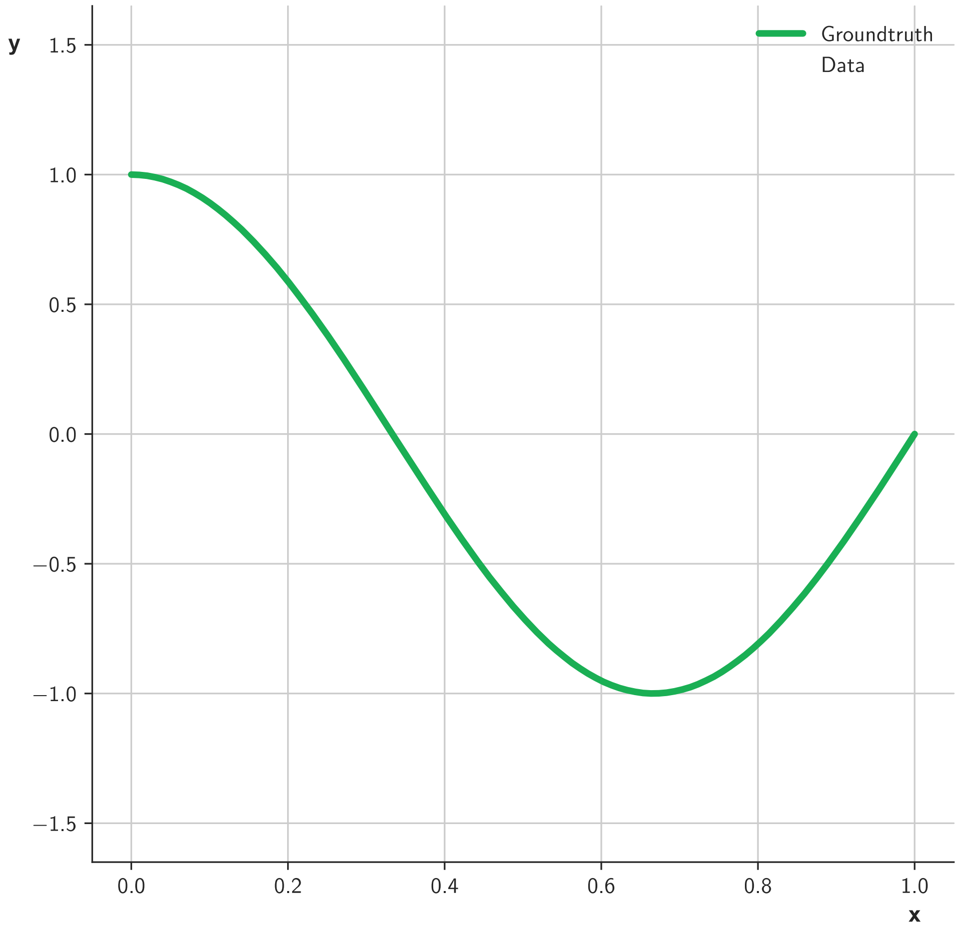
- **Input**:
  - Pairs of values: x, y

- **Goal**:
  - Model the data: For a given x, be able to predict y



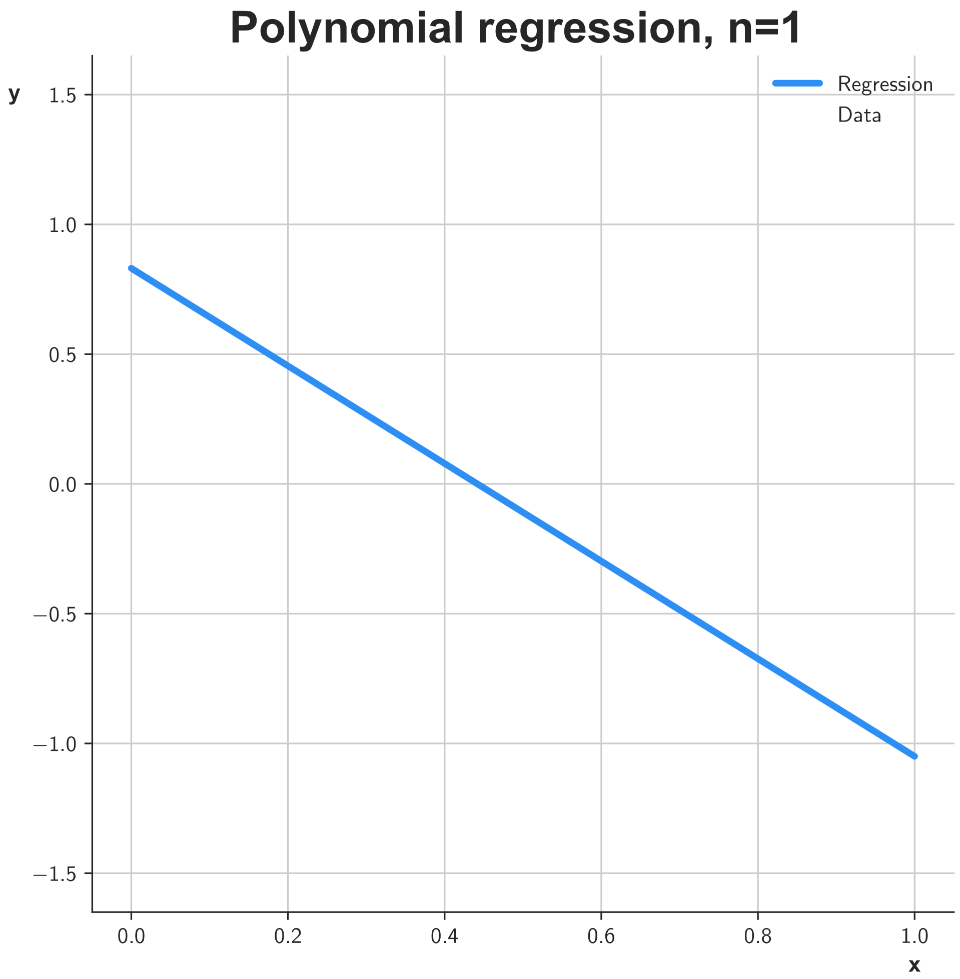
Example problem: Regression

- **Input**:
  - Pairs of values: x, y
- **Goal**:
  - Model the data: For a given x, be able to predict y
- **In this case**:
  - We even know the ground truth: $y = f(x) = \cos(1.5\pi \cdot x)$



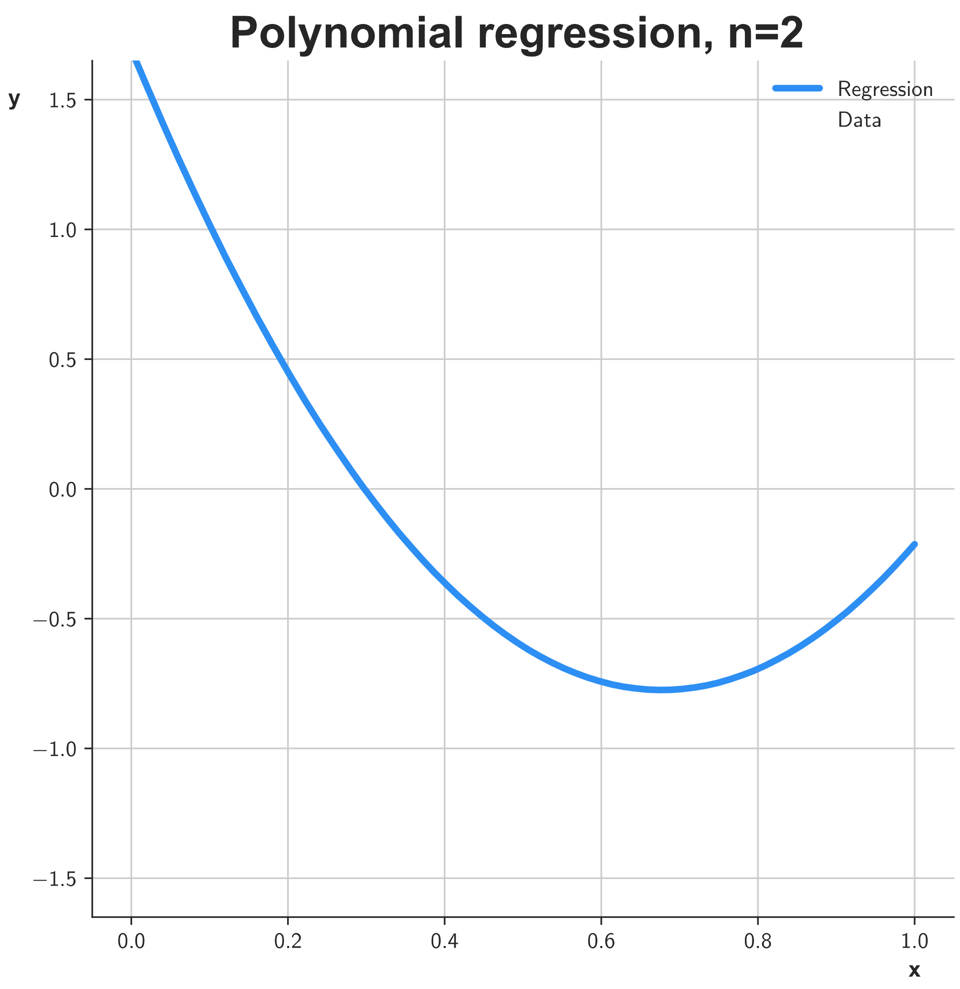
Example problem: Regression

- **Input**:
  - Pairs of values: x, y
- **Goal**:
  - Model the data: For a given x, be able to predict y
- **In this case**:
  - We even know the ground truth: $y = f(x) = \cos(1.5\pi \cdot x)$
- **Solution**:
  - Univariate linear regression
  - Univariate non-linear regression



Example problem: Regression

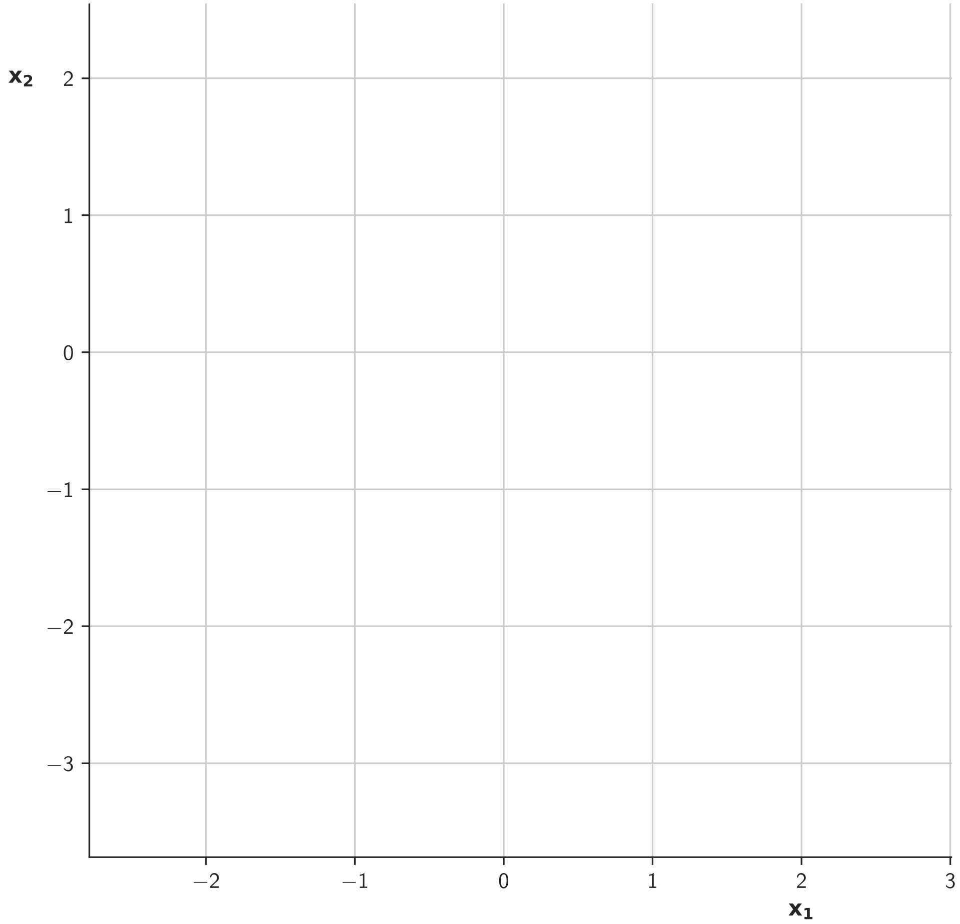
- **Input**:
  - Pairs of values: x, y
- **Goal**:
  - Model the data: For a given x, be able to predict y
- **In this case**:
  - We even know the ground truth: $y = f(x) = \cos(1.5\pi \cdot x)$
- **Solution**:
  - Univariate linear regression
  - Univariate non-linear regression



Example problem: Binary classification

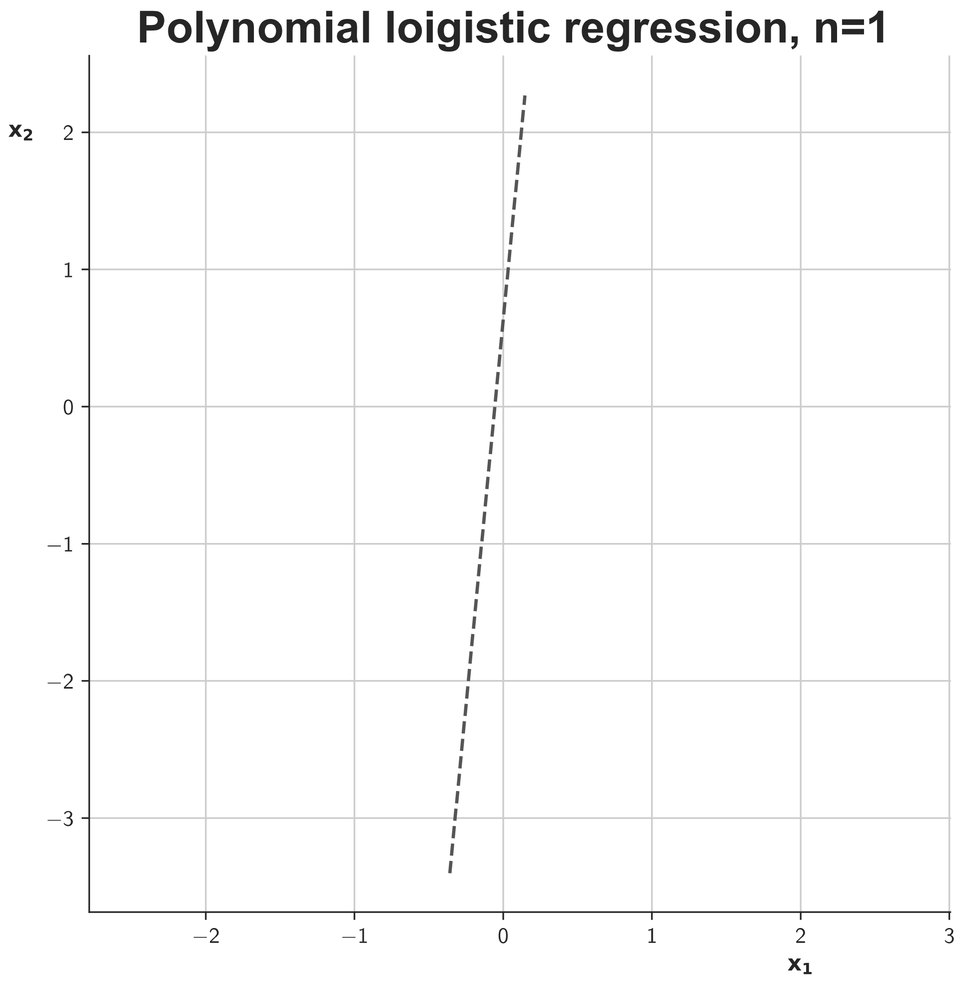
- **Input**:
  - Pairs of values: x_1, x_2
  - Corresponding class labels y

- **Goal**:
  - Predict the class label y for new pairs of values x_1, x_2



Example problem: Binary classification

- **Input**:
  - Pairs of values: x_1, x_2
  - Corresponding class labels y
- **Goal**:
  - Predict the class label y for new pairs of values x_1, x_2
- **Solution**:
  - Binary logistic regression
  - Nonlinear binary logistic regression

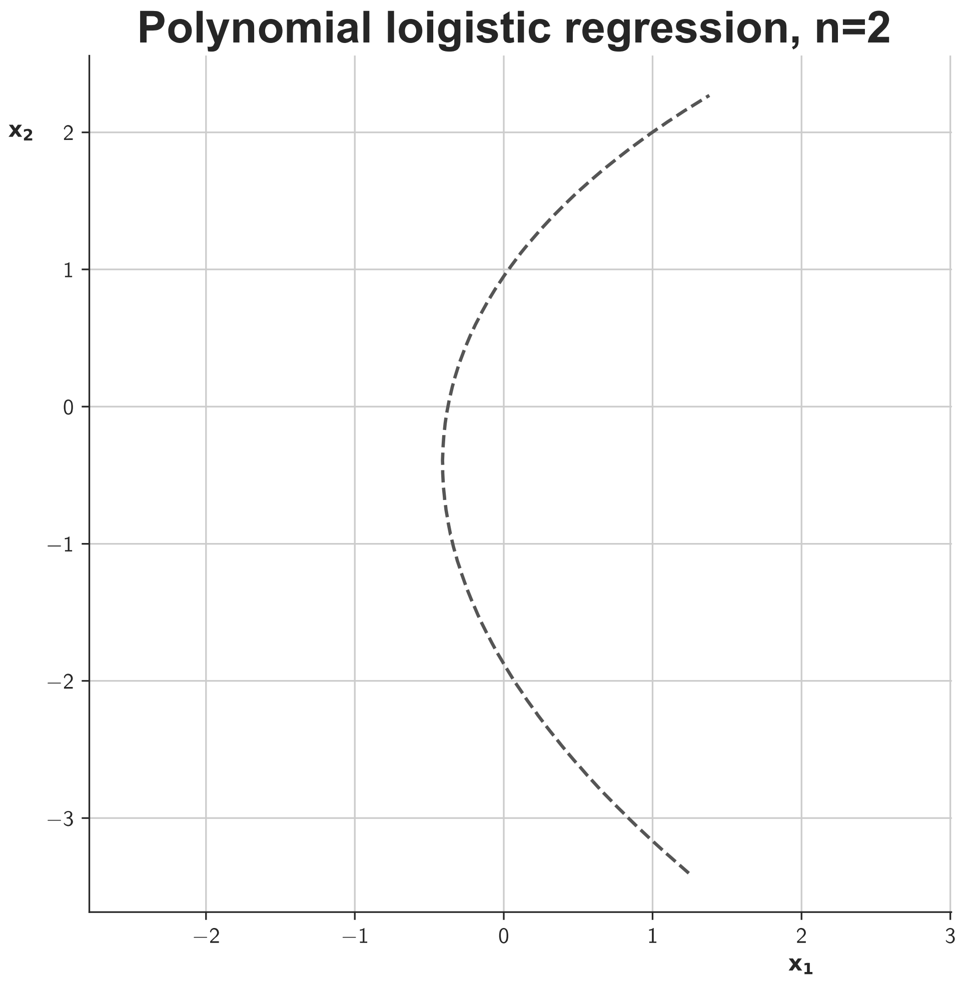


Example problem: Binary classification

- **Input**:
  - Pairs of values: x_1, x_2
  - Corresponding class labels y

- **Goal**:
  - Predict the class label y for new pairs of values x_1, x_2

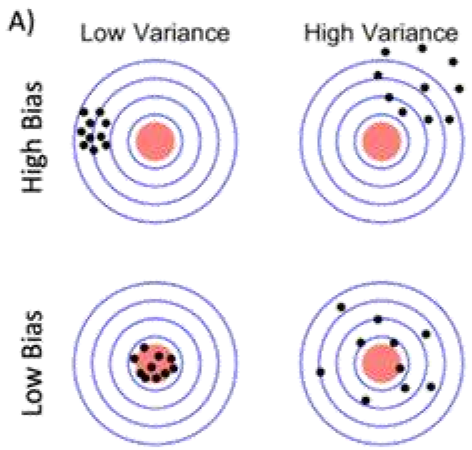
- **Solution**:
  - Binary logistic regression
  - Nonlinear binary logistic regression



Bias and variance

Bias and variance in statistics

- The **bias** is a systematic error that leads to a consistent deviation of the estimate from the true value.
- The **variance** is a measure of dispersion (or variability) in a set of data.



Bias and variance in machine learning

- The **bias** is a property of the model that causes it to miss relevant relations between the features and the target output, e.g. because of wrong assumptions in the model.
- The **variance** measures how susceptible the model is to small changes in the training data. A model with high variance will change a lot if the training data is changed slightly.

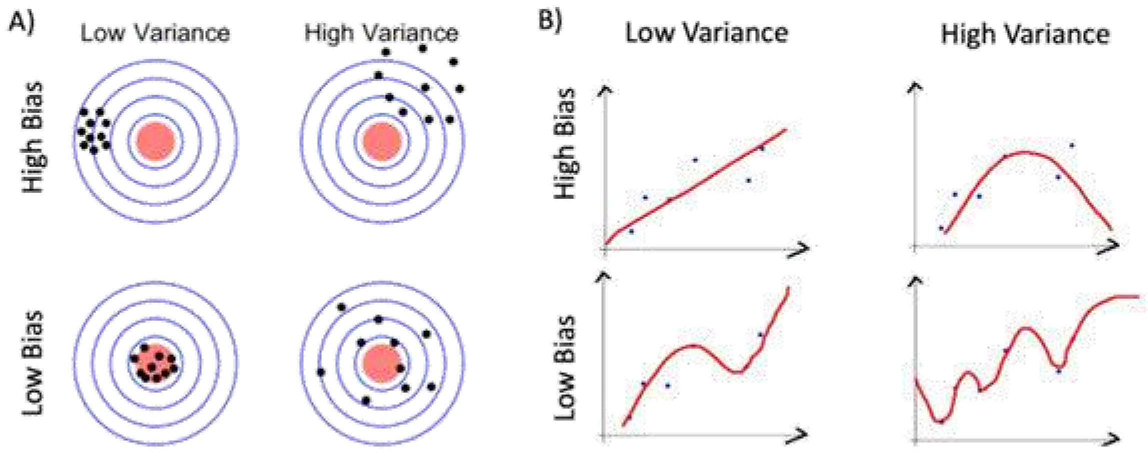


Figure 4: A: Statistical properties of bias and variance. B: Example of regression models for different combinations of bias and variance (assuming that the red curve in the lower left is the true data generating process).

What is the bias-variance tradeoff?

- Bias:
  - Measures how much the model’s predictions differ from the true outcomes (on average).
  - High bias means the model is too simple or makes too strong assumptions, possibly missing important patterns (underfitting).

- Variance:
  - Measures how much the model’s predictions fluctuate for different datasets.
  - High variance means the model is too complex and fits noise or random fluctuations in the training data (overfitting).

- The Tradeoff:
  - Low bias leads to high variance, and vice versa.
  - Goal: find a model that balances bias and variance to minimize overall prediction error!

Prediction error

- Important: The prediction we can estimate only on the **test data**

$$\epsilon_{test} = \underbrace{\left(\epsilon_{train} - \epsilon_0\right)}_{\text{Avoidable Bias}} + \underbrace{\left(\epsilon_{test} - \epsilon_{train}\right)}_{\text{Variance}} + \epsilon_0$$

- **Bias** resulting from erroneous assumptions in the learning algorithm;
- **Variance** resulting from capturing effects in the model that are actually only noise;
- **Irreducible error**, or optimal error rate, resulting from noise in the measurement, denoted by ϵ_0.

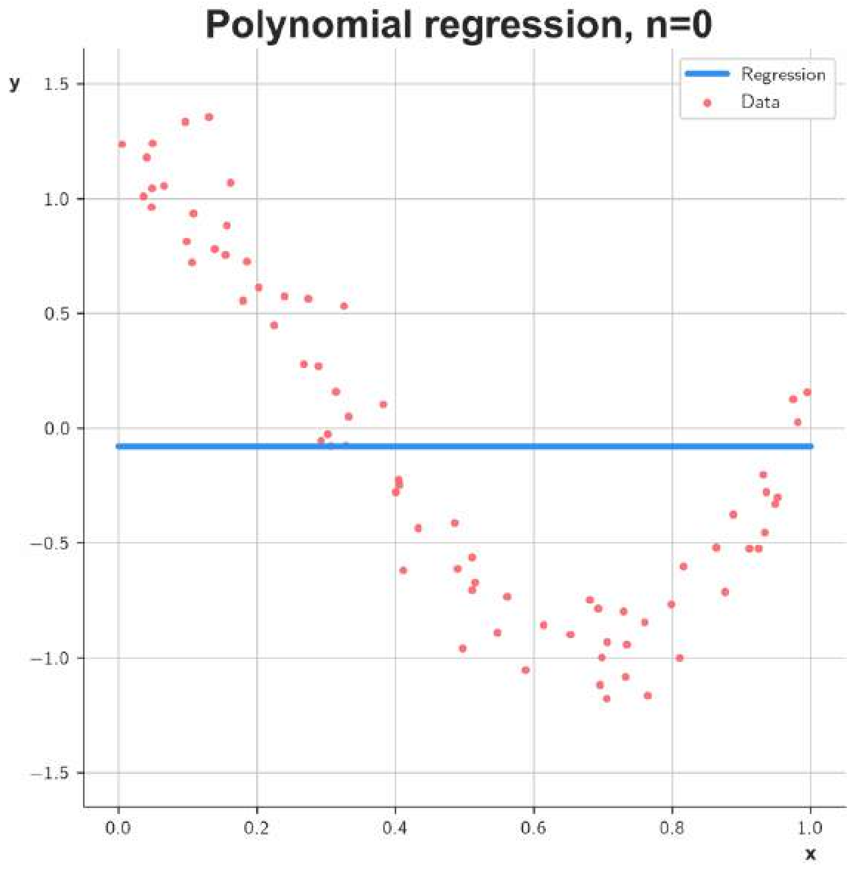
Model complexity

Example problem: Regression

- The model complexity is linked to the degree of the polynomial
- Model:

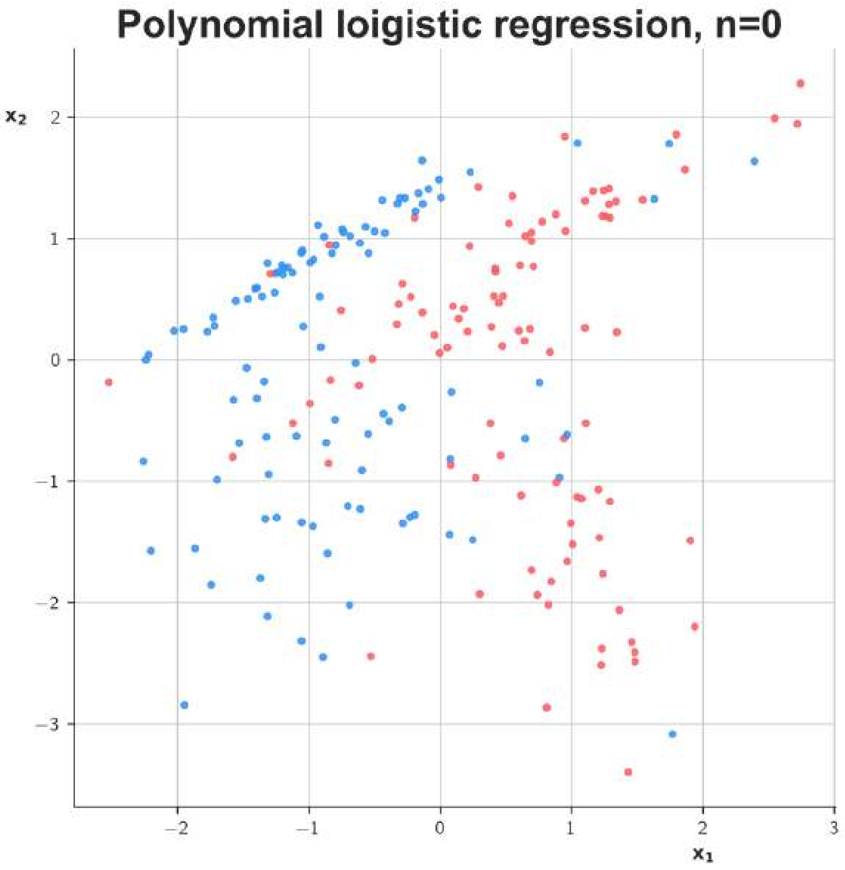
$$y = f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots + \beta_d x^d + \epsilon$$

- y: The dependent variable (target or outcome).
- x: The independent variable (feature).
- $\beta_0, \beta_1, \ldots, \beta_d$: The **coefficients** (parameters) of the model.
- d: The **degree** of the polynomial.
- ϵ: The **error term**, representing noise or deviations from the true relationship.



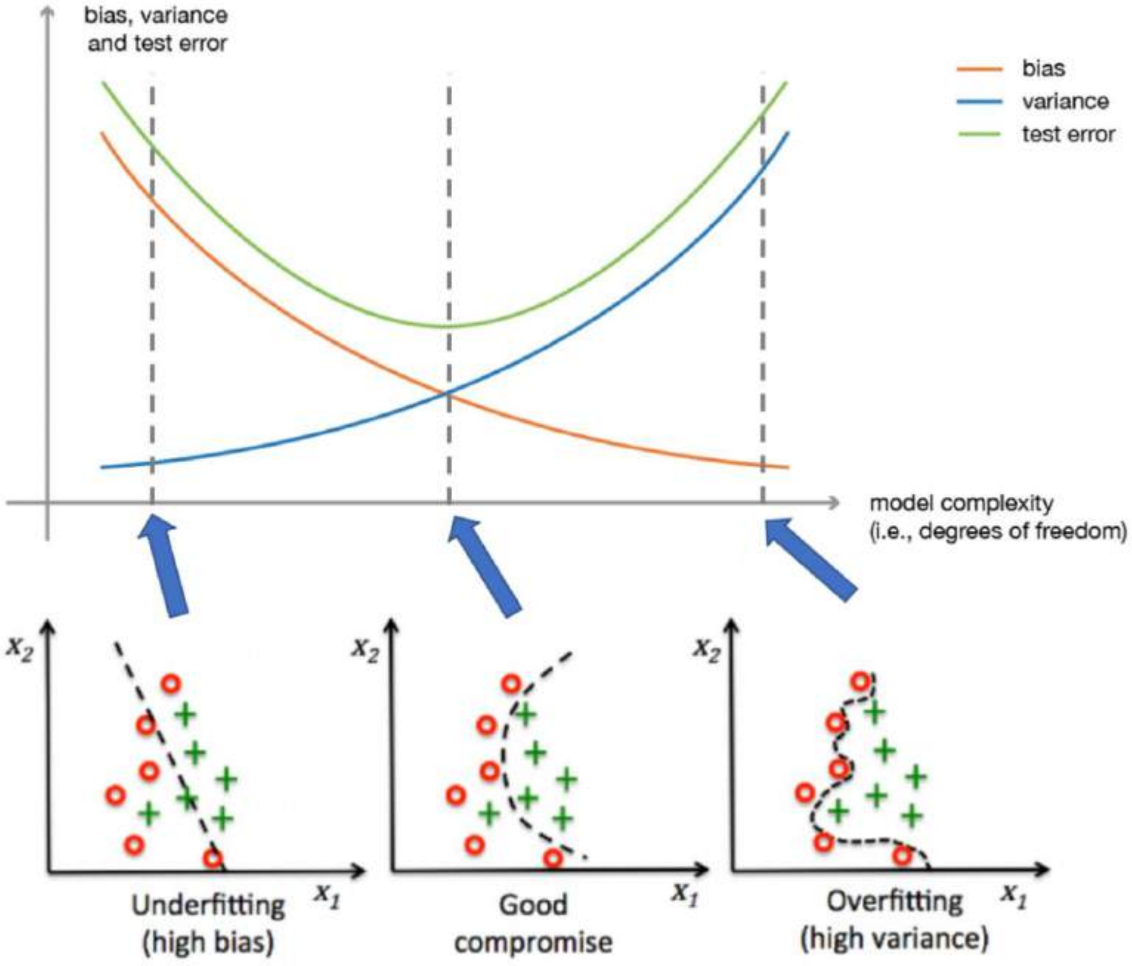
Example problem: Binary classification

- The model complexity is linked to the degree of the polynomial
- Model:
  - ...please be patient... ☺
  - Under the hood, there's also a polynomial function with parameters β_i

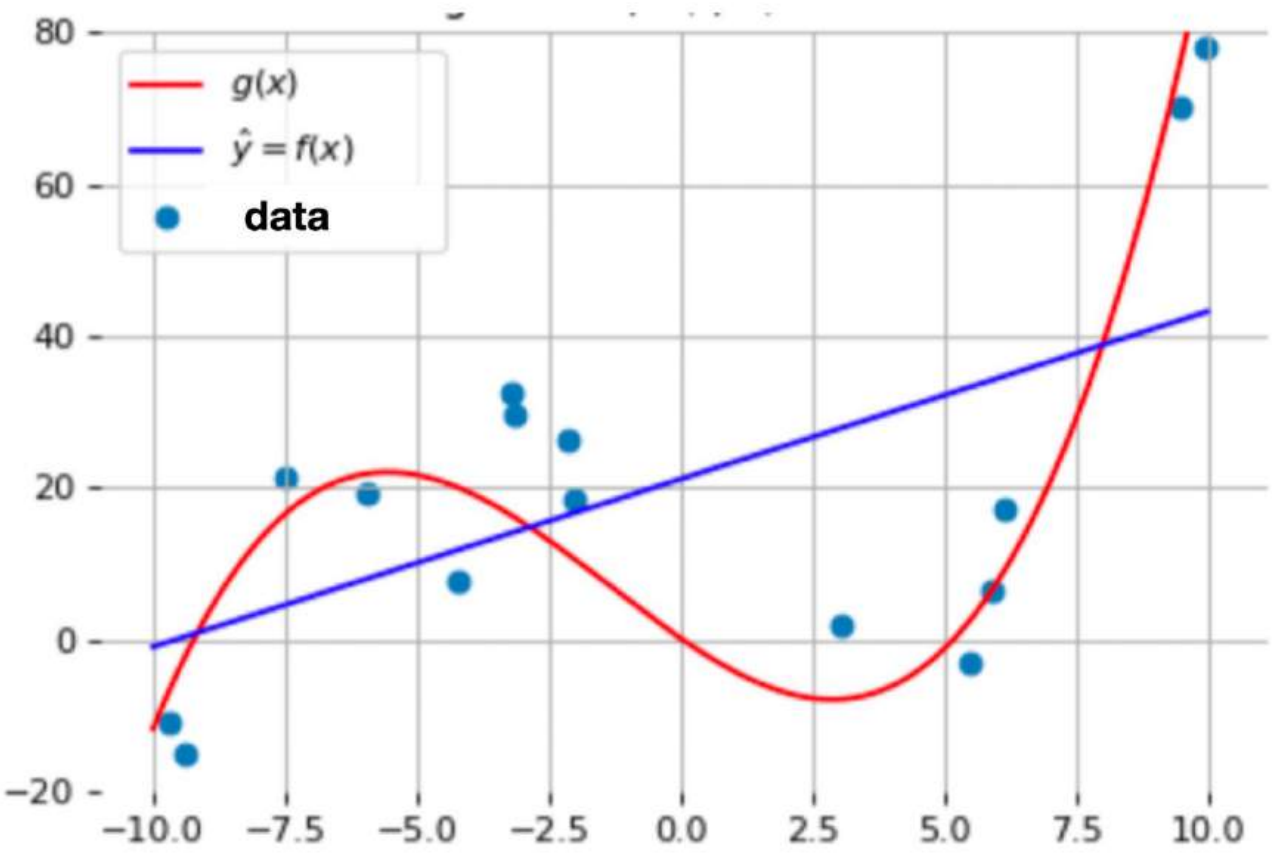


Model complexity vs. bias-variance tradeoff

We need to find the sweet spot in the middle!

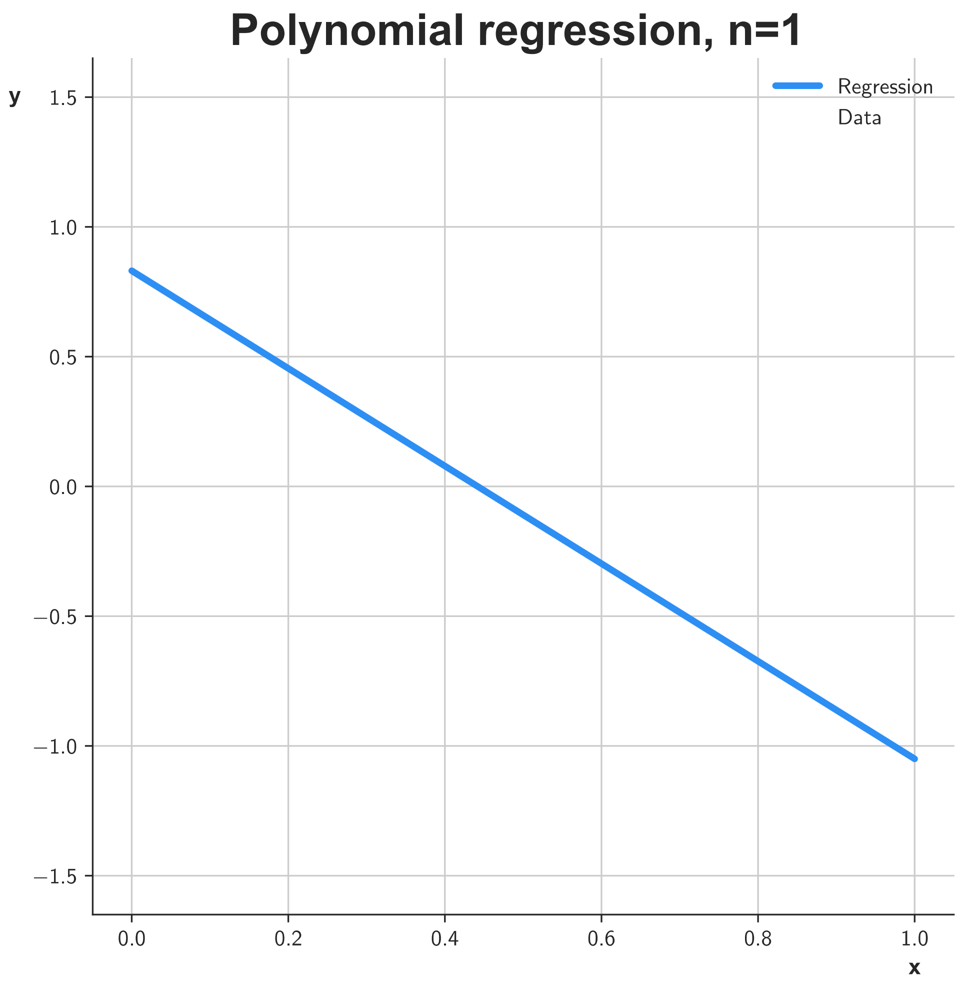


Question: Which model should we choose?



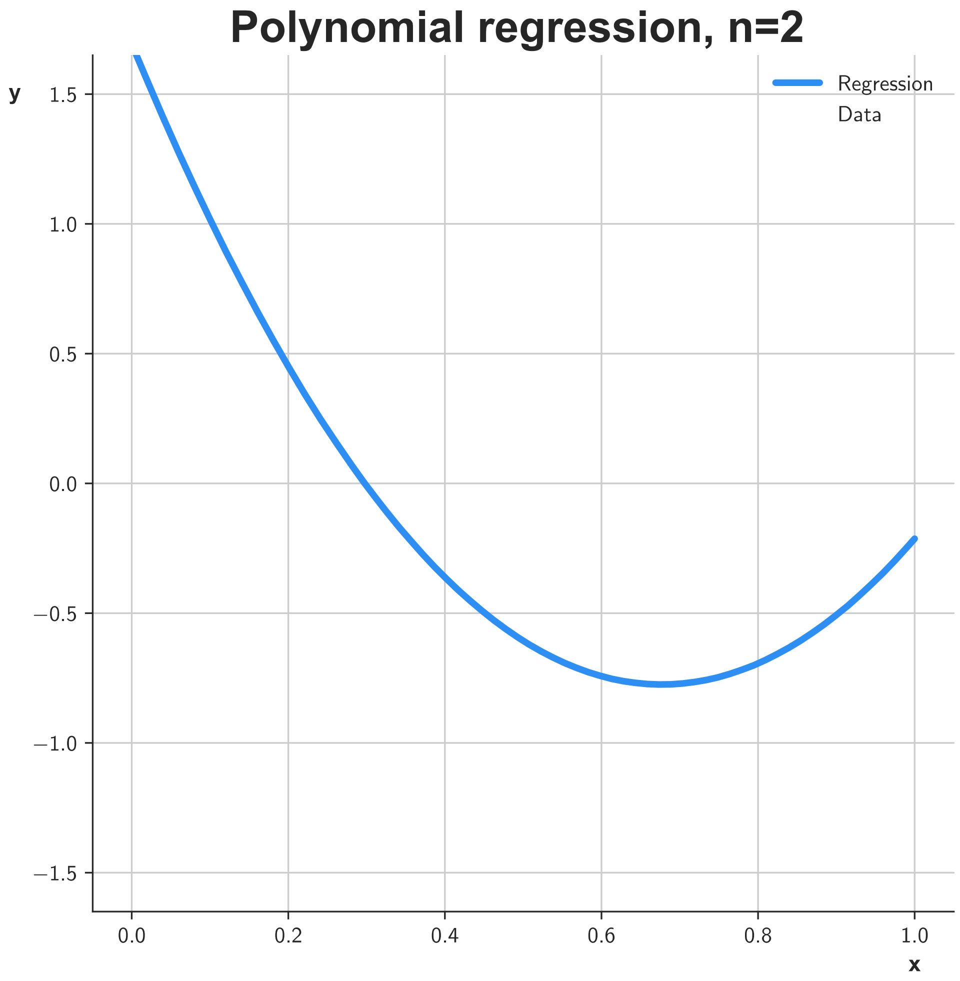
Underfitting

- If a model has too few parameters it might not be able to capture all the critical aspects of the data.
- Result: It will underfit and therefore underperform on new data.
- Choosing a more complex model will reduce the error (on the training set).



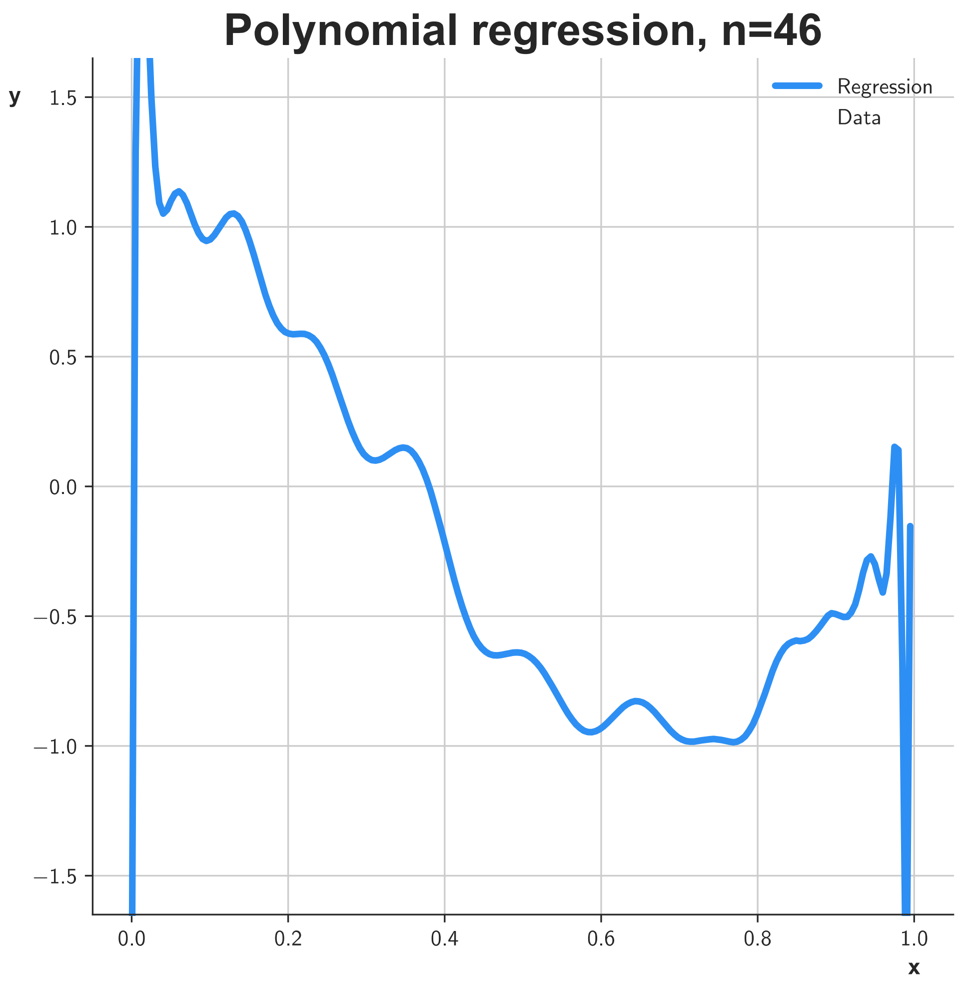
Fitting

- If a model has too few parameters it might not be able to capture all the critical aspects of the data.
- Result: It will underfit and therefore underperform on new data.
- Choosing a more complex model will reduce the error (on the training set).



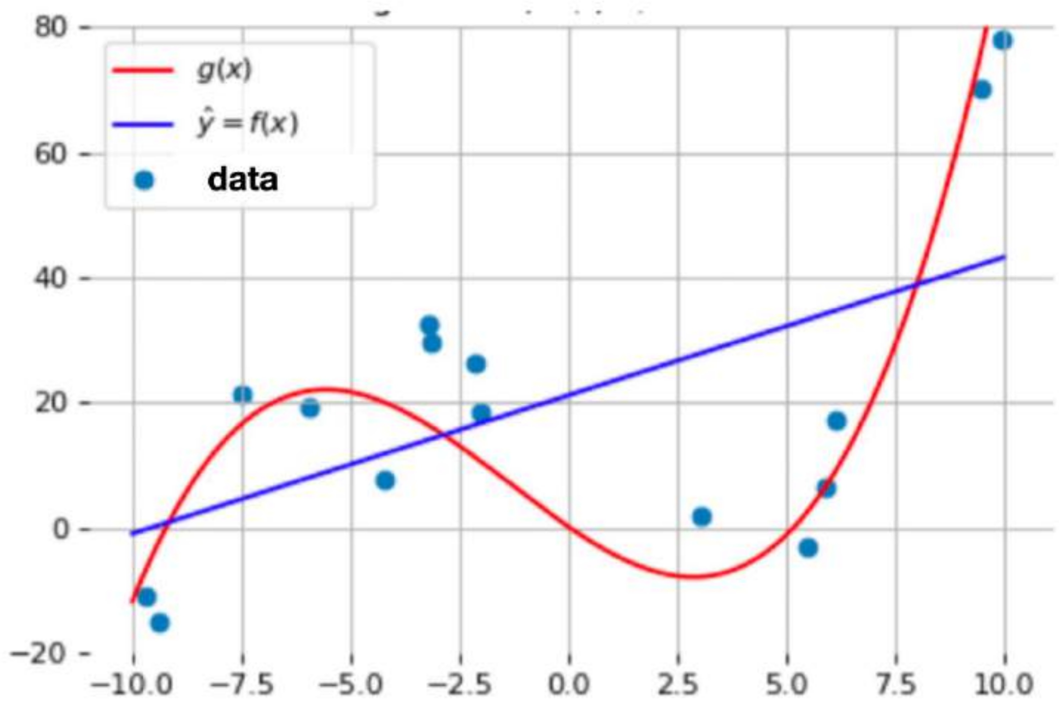
Overfitting

- Overfitting occurs when a model captures variations in its parameters that are, in fact, only noise.
- Result: the model will underperform on new data where this noise is not present.
- Overfitting is difficult to detect since we can only see it on the test set, and training performance even increases the more we overfit.



Question: Which model should we choose?

- Under- and overfitting are a natural consequence of modeling
- We can only aim to balance the degree of under- and overfitting
-



Task

Recap: Expected value

- **Definition:** The **Expected Value (E[X])** of a random variable X represents the average or mean outcome if an experiment is repeated infinitely many times.

- **Discrete case**:

$$E[X] = \sum_{i=1}^{n} p(x_i) \cdot x_i$$

x_i: Possible outcomes

$p(x_i)$: Probability of outcome x_i

- **Sample mean**:
  - We can try to estimate the expected value for a specific dataset using the sample mean

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

$\bar{X}$: Sample mean

X_i: Observed values in the sample

n: Number of observations

Task

- Study this tutorial: https://mlu-explain.github.io/bias-variance/
- Can you link the topics discussed in these lecture slides with this tutorial?