Advanced School in Artificial Intelligence

Introduction to Machine Learning

Elena Bellodi

elena.bellodi@unife.it

RegioneEmilia-Romagna

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

FERRARIAE UNIVERSITAS · EX LABORE FRUCTUS · 13 91

Università degli Studi di Ferrara

Outline

- Machine learning (ML) definitions
- Learning paradigms
  - supervised
  - unsupervised
  - semi-supervised
  - reinforcement
- Use of Data in ML
  - training, validation and test set
  - generalization, underfitting and overfitting
  - capacity
  - bias and variance
- Learning protocols

Regione Emilia-Romagna
Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021
Università degli Studi di Ferrara

Generalization, Underfitting and Overfitting

- The central challenge in machine learning is that we must perform well on new, previously unseen inputs
  - **Generalization:** the ability to perform well on new unobserved input

- During the fitting of the model, one of the objectives is to reduce the **training error**, some error measure computed on the training data
- During prediction, the objective is to reduce the **generalization error** or **test error**
- The factors determining how well a machine learning algorithm will perform are its ability to:
  - Make the training error small.
  - Make the gap between training and test error small.
- These two factors correspond to the two central challenges in machine learning: **underfitting** and **overfitting**

Generalization, Underfitting and Overfitting

- **Underfitting** occurs when the model *is not able* to obtain a sufficiently *low error value on the training set*
- **Overfitting** occurs when the *gap between the training error and test error is too large*

- We can control whether a model is more likely to overfit or underfit by altering its **capacity**, the ability to fit a wide variety of target functions.

Model capacity

- Models with **low capacity** may struggle to fit the training set
- Models with **high capacity** can overfit by memorizing properties of the training set that do not serve them well on the test set
- One way to control capacity of a learning algorithm is by **choosing the hypothesis space**, i.e., set of functions that the learning algorithm is working with
  - E.g., the linear regression learning algorithm has the set of all linear functions of its input as the hypothesis space
  - We can generalize to include polynomials in its hypothesis space, which increases model capacity

Model capacity

- A polynomial of degree 1 gives a linear regression model with the prediction

$$Y = aX+b$$

- By introducing x^2 as another features provided to the model, we can learn a model that is quadratic as a function of x

$$Y = a_2X^2 + a_1X + b$$

- We can continue to add more powers of *x* as additional features, e.g., a polynomial of degree 9

$$Y = b + \sum^{9}_{i=1} a_i x^i$$

Model capacity

- Machine Learning algorithms will perform well when **their capacity is appropriate for the true complexity of the task** that they need to perform and the amount of **training data** they are provided with

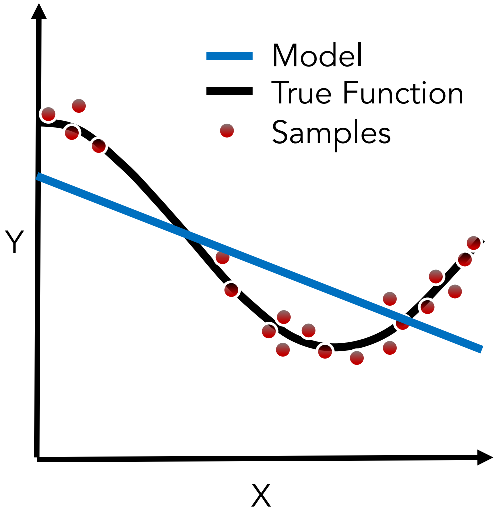
- Models with **insufficient capacity** are unable to solve complex tasks (*underfitting*)

- Models with **high capacity** can solve complex tasks, but when their capacity is *higher* than needed to solve the present task, they may *overfit*
  - A rule of thumb is that, *to avoid overfitting, the number of parameters estimated from the data must be considerably less than the number of data points*.

How well does the model generalize?

Polynomial Degree = 1

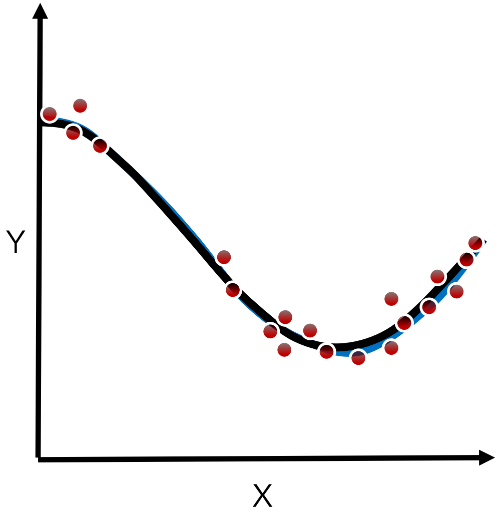
$$Y = aX + b$$



Poor on Training Set
Poor at Predicting

Polynomial Degree = 4

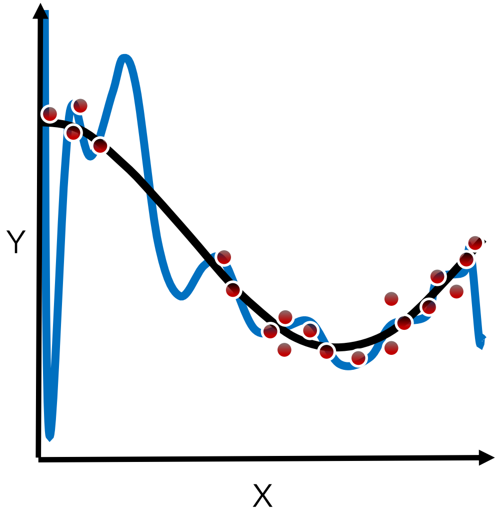
$$Y = aX^4 + bX^3 + cX^2 + dX + e$$



Just Right

Polynomial Degree = 15

$$Y = a_0X^{15} + \cdots + a_{15}$$



Very Good on Training Set
Poor at Predicting

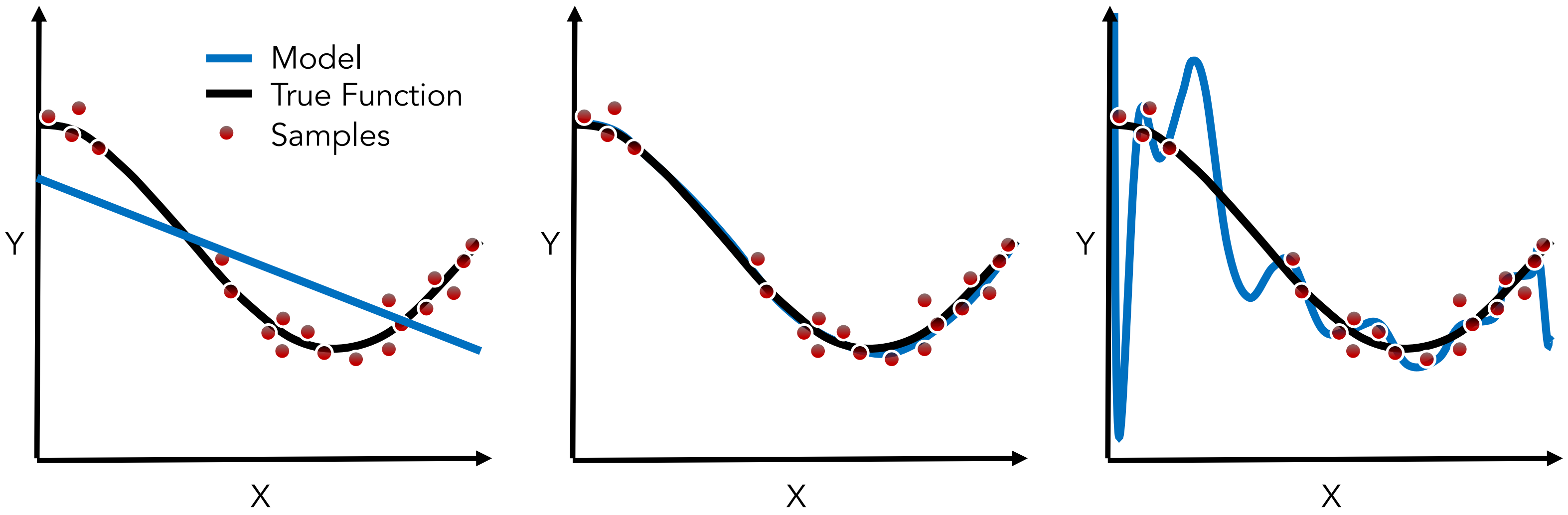
Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Regione Emilia-Romagna

Università degli Studi di Ferrara

How well does the model generalize?

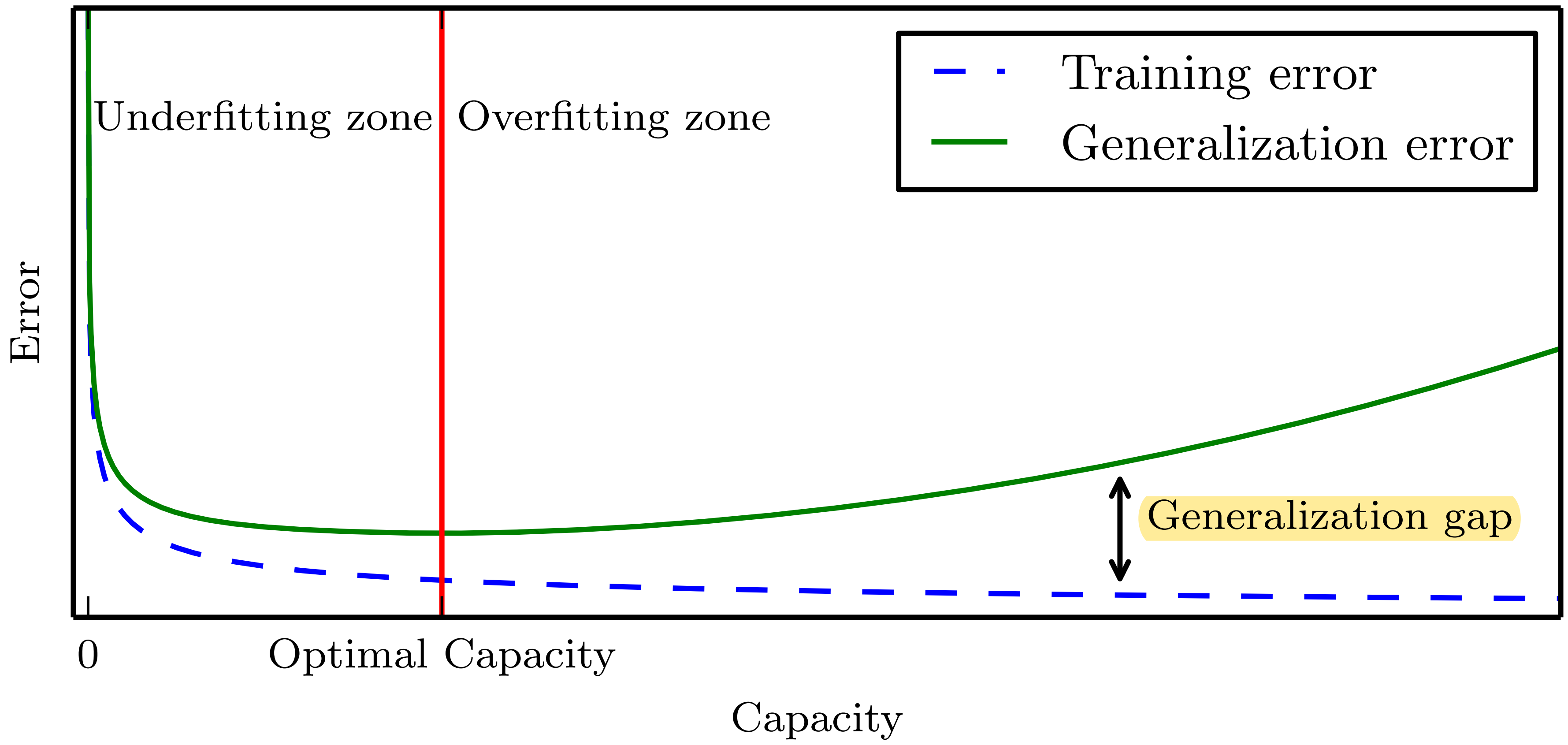
Polynomial Degree = 1	Polynomial Degree = 4	Polynomial Degree = 15
$Y = aX + b$	$Y = aX^4 + bX^3 + cX^2 + dX + e$	$Y = a_0X^{15} + \cdots + a_{15}$



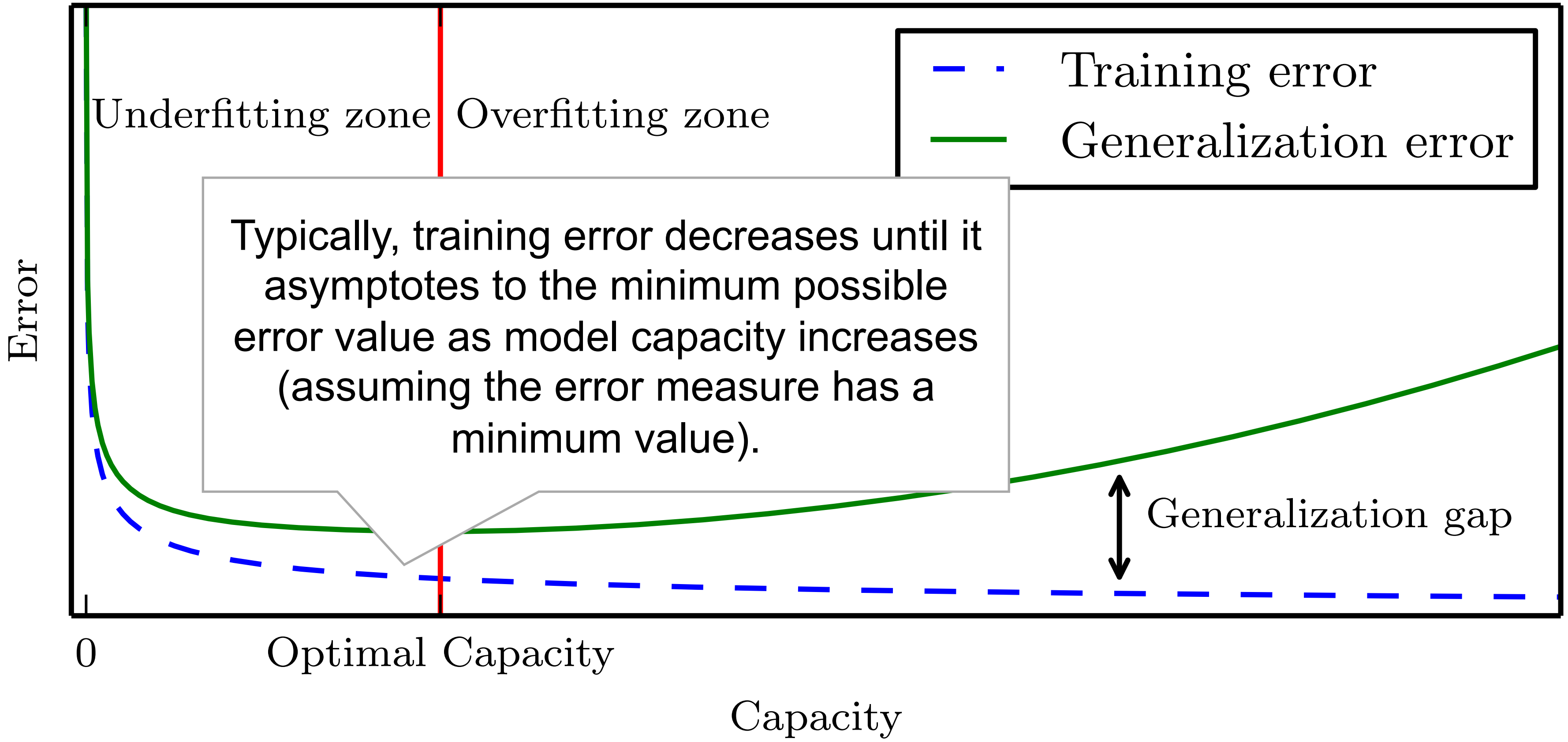
Poor on Training Set Poor at Predicting Underfitting	Just Right	Very Good on Training Set Poor at Predicting Overfitting

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Generalization and Capacity

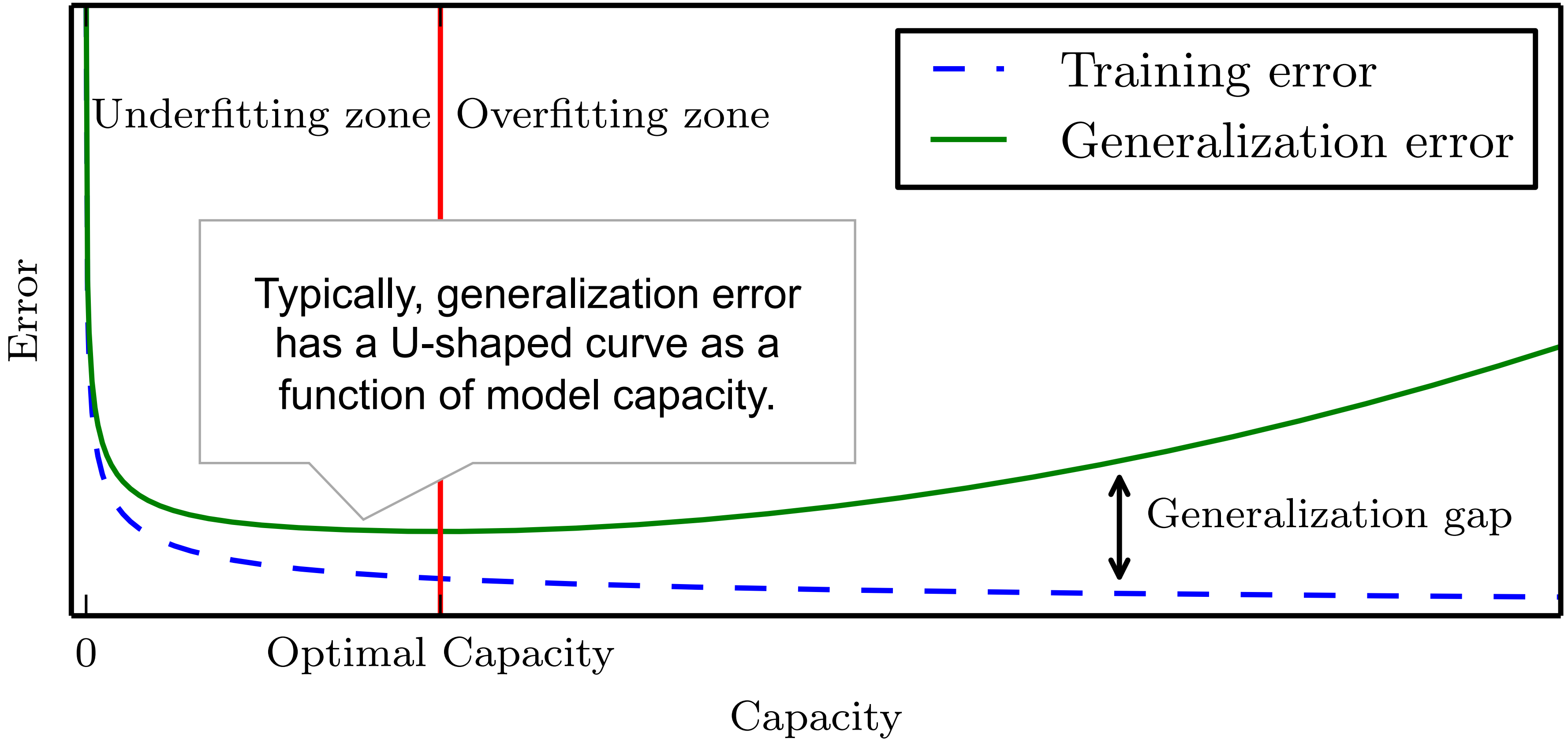


Generalization and Capacity

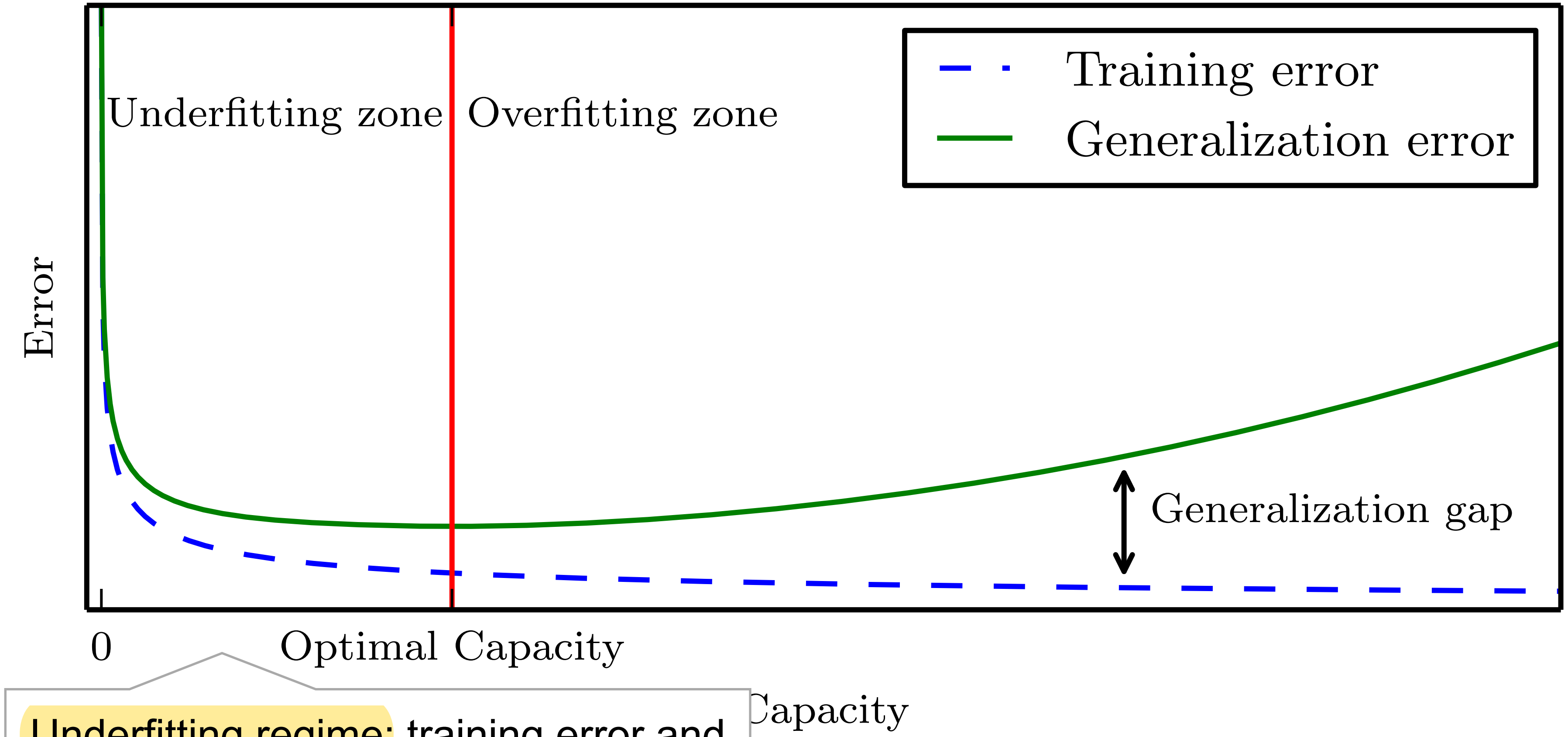


Typically, training error decreases until it asymptotes to the minimum possible error value as model capacity increases (assuming the error measure has a minimum value).

Generalization and Capacity

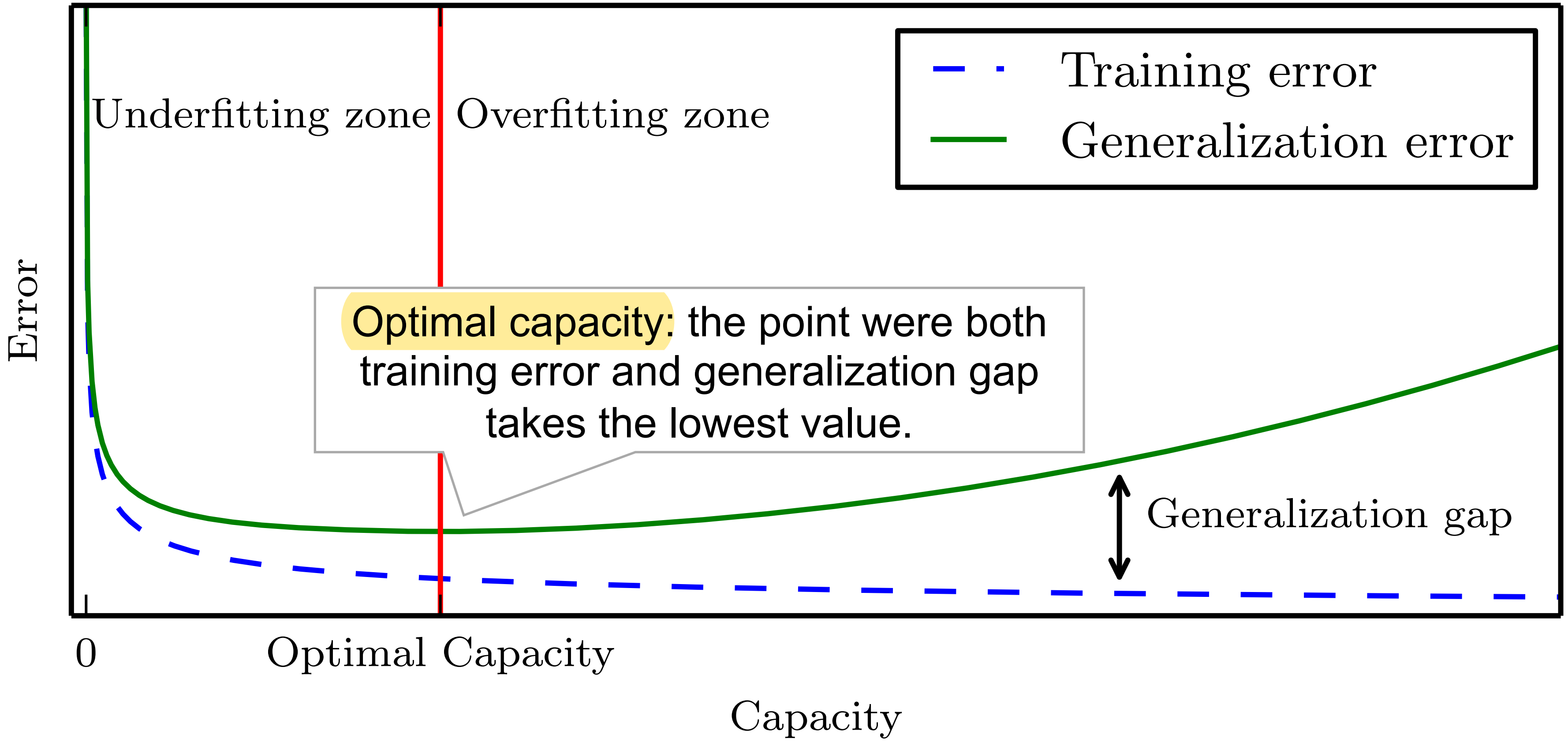


Generalization and Capacity



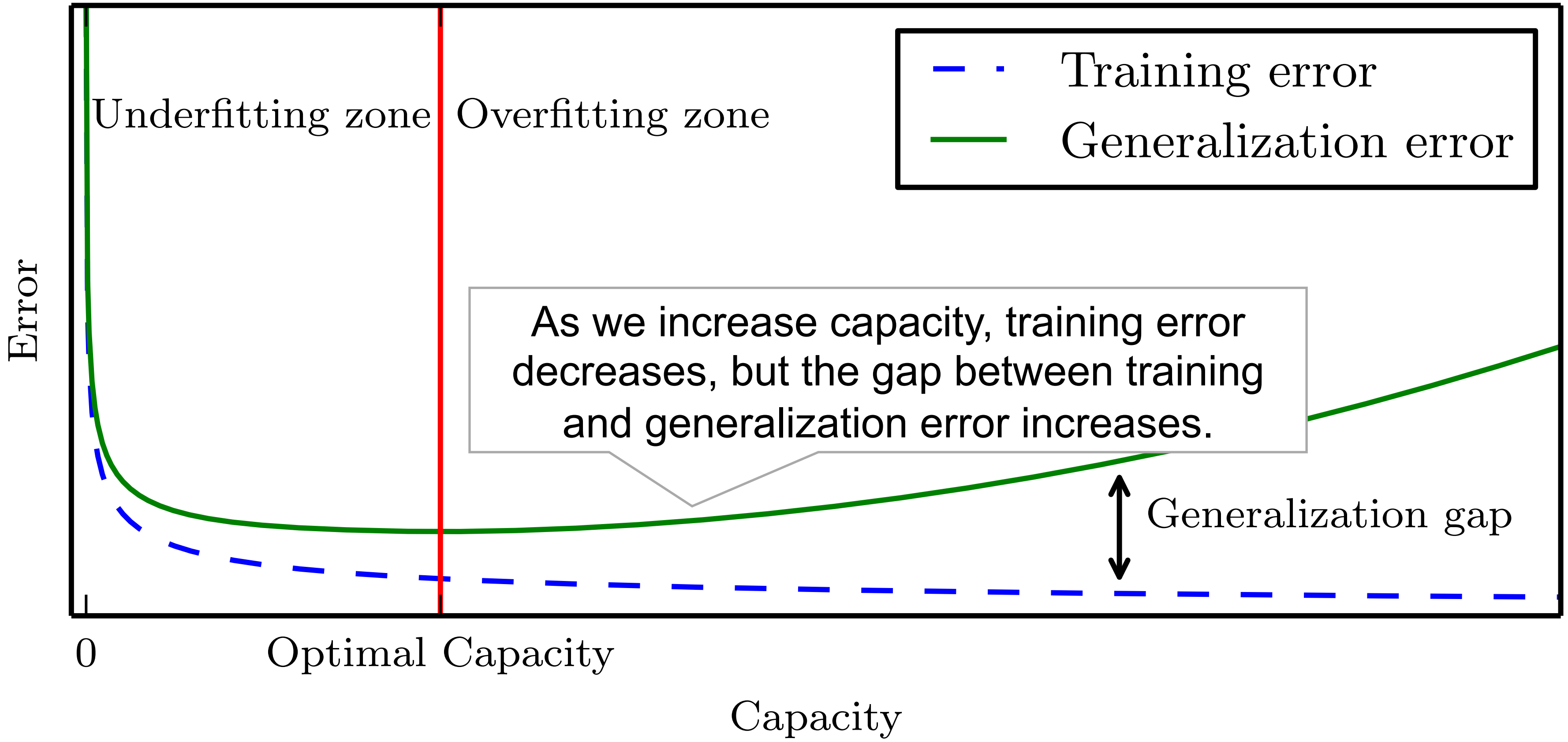
Underfitting regime: training error and generalization error are both high.

Generalization and Capacity

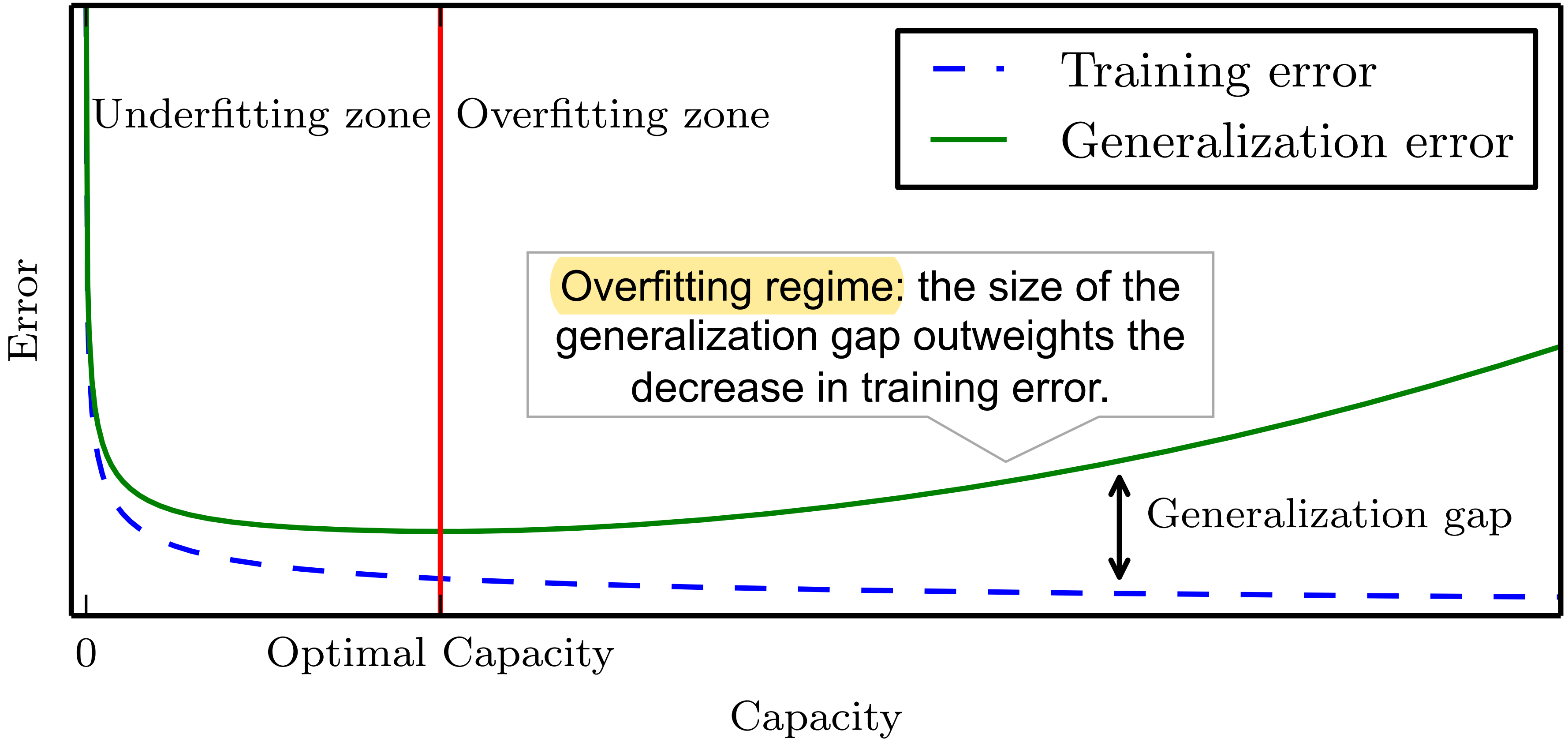


Optimal capacity: the point were both training error and generalization gap takes the lowest value.

Generalization and Capacity



Generalization and Capacity



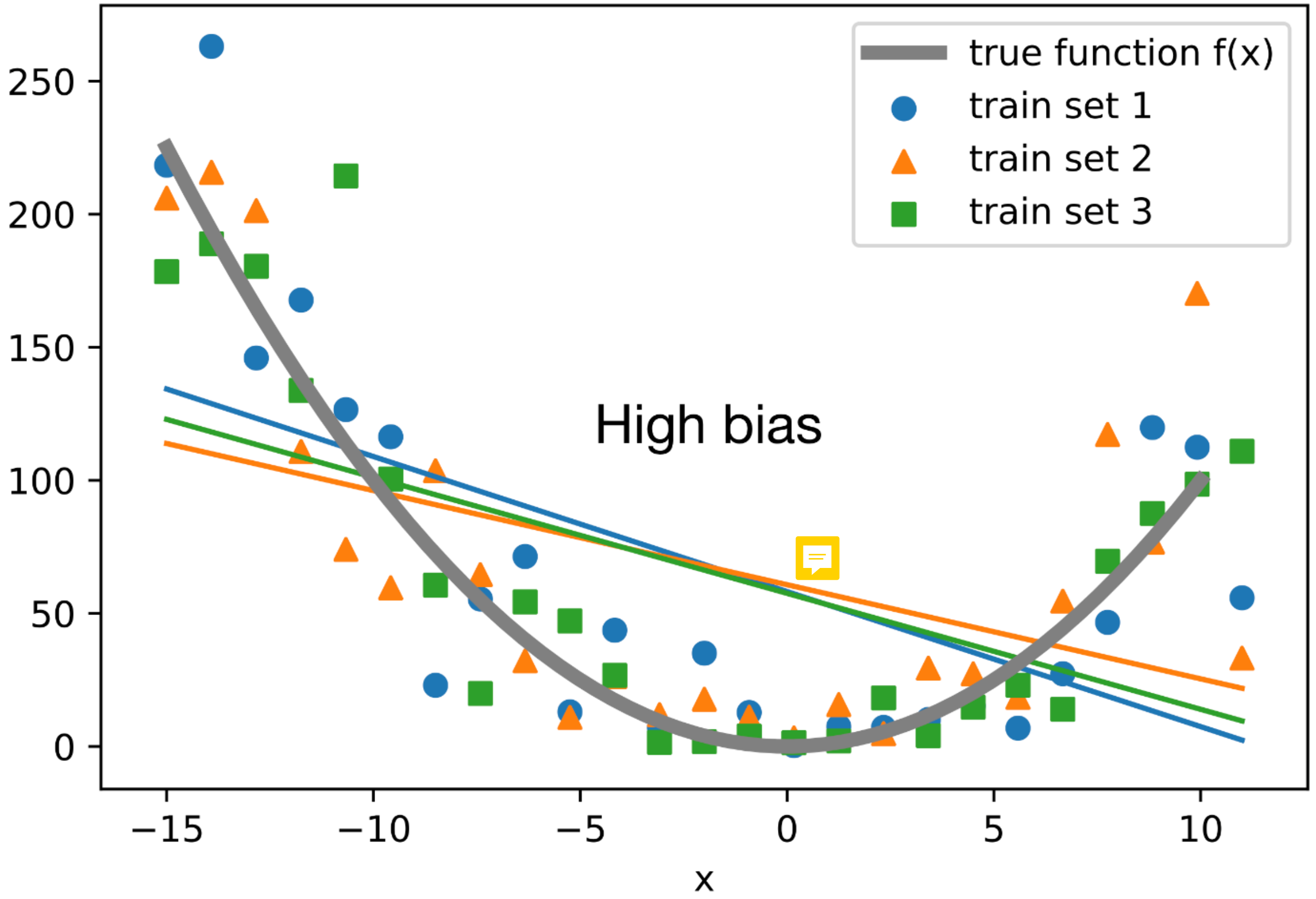
Bias and Variance: Bias

- The difference between
  - the predicted target value given by the model $\hat{Y}$ *=g(x)*
  - and the true values *Y = f(x)*

$$bias(\hat{Y}) = \mathrm{E}(\hat{Y}) - Y = \mathrm{E}(\hat{Y} - Y)$$

- $\hat{Y}$ is **unbiased** if its bias is 0, which means that $\mathrm{E}(\hat{Y}) = Y$
- Bias indicates how far are the values predicted by the model from the real values
- If the **average predicted values are far from the actual values then the bias is high**.
- **High bias** → the model is too simple (cannot capture complexity of the model) → **underfitting**

Bias



Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Bias and Variance: Variance

- The estimate that the **target function will change if different training data was used**

$$variance(\hat{Y}) = \mathrm{E}[(\hat{Y} - \mathrm{E}(\hat{Y}))^2]$$

- Ideally, it should not change too much from one training dataset to the next, meaning that the algorithm is good at picking out the hidden underlying mapping between the inputs and the output variables.

- The variance tells us how dispersed the expected values are compared to the actual value, if the **predicted values are far from their average then the variance is high**.

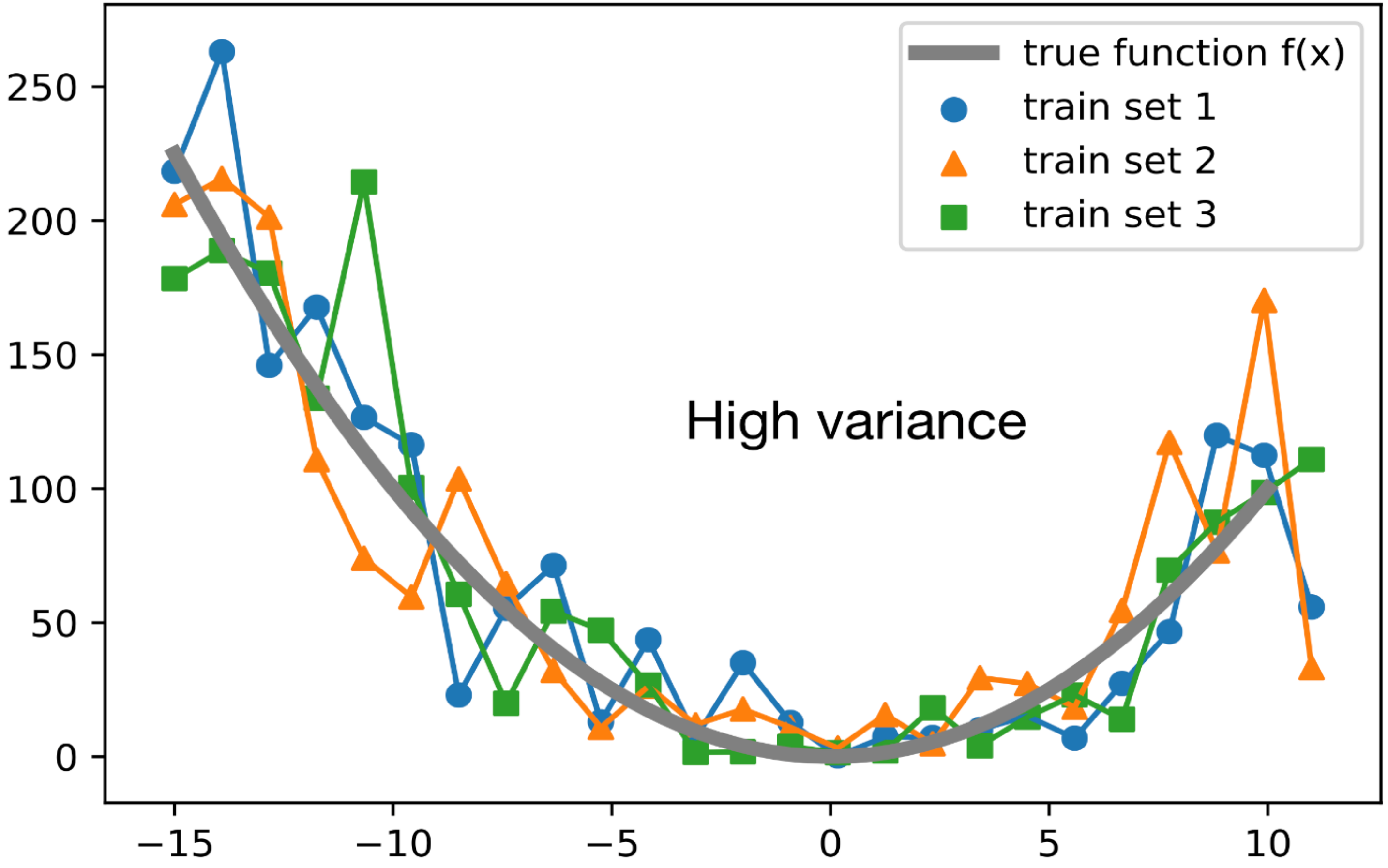
Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Bias and Variance: Variance

- **Low Variance**: Suggests *small changes* to the estimate *of the target function* with changes to the training dataset.

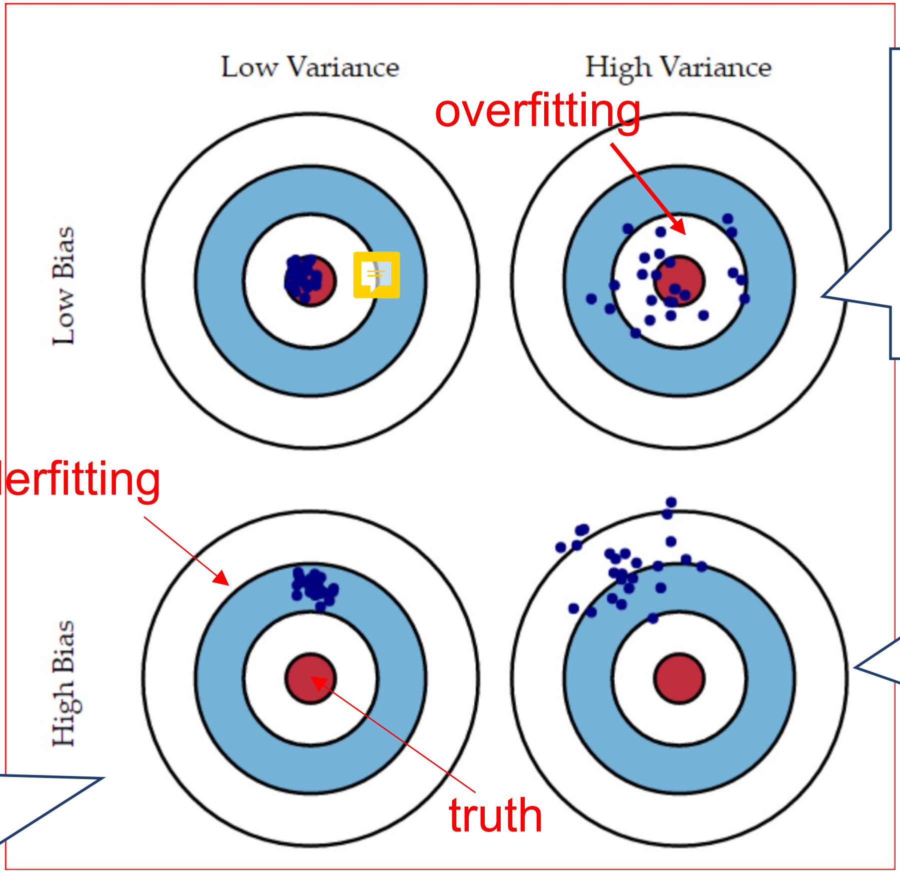
- **High Variance**: Suggests *large changes* to the estimate of the target function with changes to the training dataset.
  - the model also depicts random noise present in the training data → **overfitting**

Variance



Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Bias and Variance



Low Bias High variance: Models are somewhat accurate but inconsistent on average. A small change in the data can cause a large error (high-complexity models)

High Bias Low Variance: Models are consistent but inaccurate on average (low-complexity models)

High Bias High Variance: Models are inaccurate and also inconsistent on average

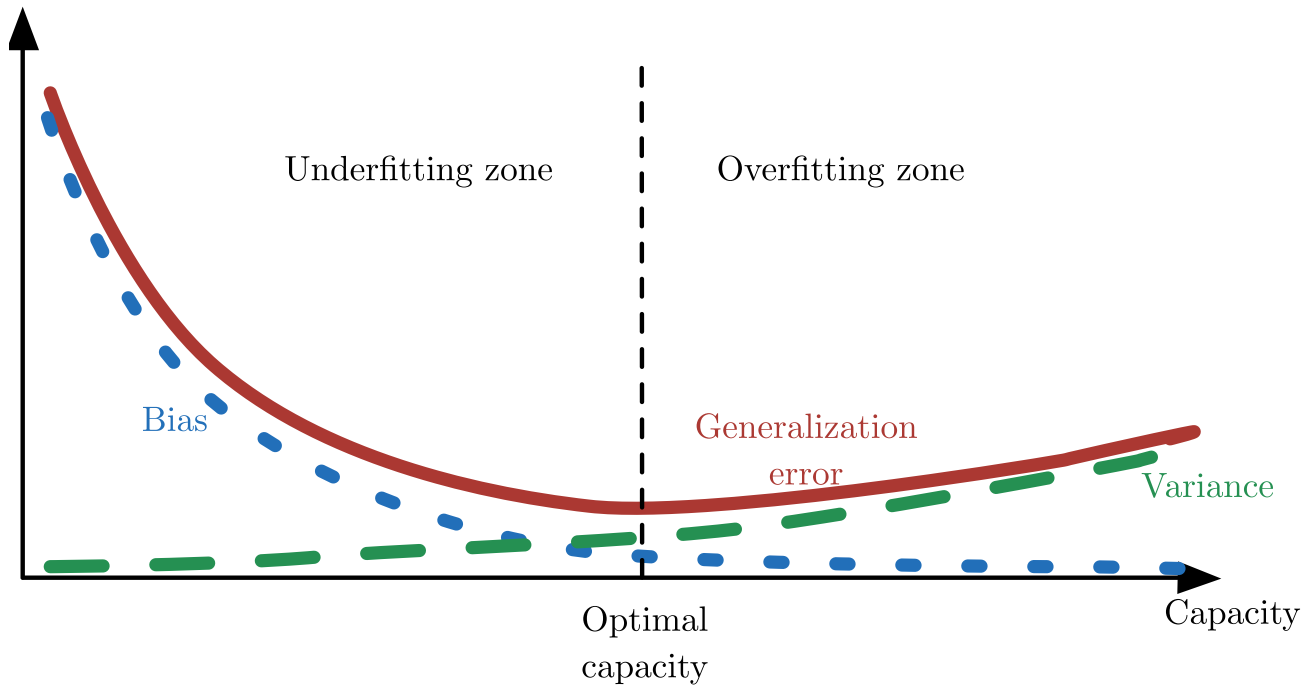
Image from: “An Introduction to Statistical Learning”
by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Regione Emilia-Romagna

Università degli Studi di Ferrara

Bias-Variance Tradeoff



Total error = bias + variance + irreducible error

The generalization error is the sum of a reducible error, that is measured and optimized during the learning of the model, and *irreducible* error, that cannot be removed (noise in our data).