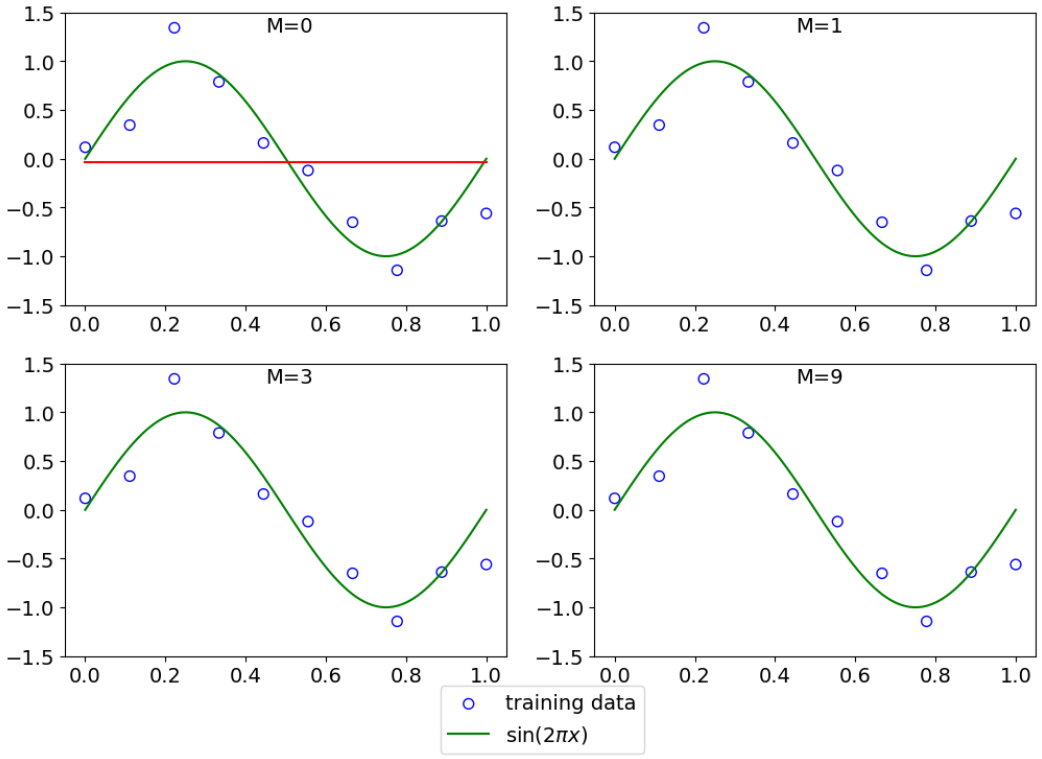
Fitting and model capacity

The following figure shows (noisy) measurements in blue, sampled from a ground truth function shown in green. Consider the use of different polynomial-curve-based models of order M.



1. Complete the sketch (you can draw directly into it) by adding estimations of how the fitted function would look like for the other model orders. An example fit is shown for M=0 as a red curve.
2. Explain under/overfitting on this example.
3. What methods could help to overcome overfitting. Discuss one method in detail.

Inception

Inception blocks have become a fundamental component in many deep neural network architectures. Answer the following questions:

1. Explain the concept of an Inception block. Provide an overview of what an Inception block is, its structure, its key components, and the motivation behind its design in neural networks.

2. What are the advantages of using Inception blocks in deep neural networks? Discuss the benefits of incorporating Inception blocks, such as improved model performance and computational efficiency, in comparison to traditional architectures.
3. What is the role of 1x1 convolutions in an Inception block? Elaborate on the significance of 1x1 convolutions and why this is essential in Inception blocks.

Quantization

Consider quantization as a model compression technique to reduce the overall complexity (compute, memory) of a neural architecture.

1. Discuss uniform and non-uniform quantization with regard to their advantages and disadvantages.
2. Provide and explain a uniform and a non-uniform quantizer using an equation.
3. Optional: For both types of quantization, provide an example demonstrating how to perform calculations using these number formats. You can either describe a specific calculation example, or describe how such computations would be carried out using typical hardware resources.

Computational intensity

A plethora of hardware architecture exists that are tailored to particular forms of neural architectures. Different operations used in neural networks have different requirements on hardware architectures. One way to quantify this is the use of abstract HW metrics, describing the load put on different components of a hardware architecture. Consider computational intensity (FLOPs/Byte) as a key metric here.

1. Discuss the computational intensity of fully-connected and convolutional operations. If possible, provide a quantification (absolute or relative)
2. Explain the use of the roofline model of a given hardware architecture in this context.