Report

Andrés Reverón Molina – r0767329

Artificial Neural Networks and Deep Learning

# Supervised learning and generalization

## Perceptron

As we know, due to its simple architecture consisting of a single neuron, a perceptron is only able to learn linearly separable tasks, that is, groups of points that can be separated by and hyperplane in the input space, this is called decision boundary. This makes it a useful tool for basic classification tasks, such as spam detection. On the other hand, a perceptron cannot learn non-linear functions, such as the XOR function.

For this kind of tasks, the perceptron is normally able to achieve a high accuracy score, and it is not computationally expensive at all compared to other network structures. In fact, it is the simplest network consisting of only one layer.

## Backpropagation in feedforward multi-layer networks

The limitation of linearity comes to an end when several layers are introduced. A simple multi-layer network consisting of one hidden layer and one output layers is perfectly capable of being an universal approximator, provided that a sufficient number of hidden units is taken(neurons). Normally, a network with two hidden layers is chosen, paired with an activation function in the output layer, such as tanh or a Radial Basis Function(Gaussian).

A nice advantage of multi-layer networks over polynomial expansion is that they are able to cope better with the curse of dimensionality. This is because the approximation error of the network is virtually independent of the dimionality of the input space, as opposed to polynomial expansion.

The first algorithm that was ever invented for multi-layer perceptrons is called the backpropagation algorithm. It caused a revolution in the area of neural network research. The network optimizes a so-called objective or energy function, the Mean Squared Error(MSE) on the training set(as it is a supervised learning technique). Explained simply, backpropagation is a way of iteratively improving the weights of the network by trying to minimize the value of the error obtained in the previous iteration(epoch). By doing this, the model becomes more reliable, thus improving its ability to generalize.

The algorithm calculates the delta of each unit to see which unit is responsible for most of the error and fine-tune its weight. This is done by, in a sense, sending information back from the output layer, thus effectively backpropagating information.

The backpropagation algorithm is also called generalized delta rule and it is an straightforward way of obtaining an analytic expression for the gradient of the cost function, ever for large networks with a big number of layers.

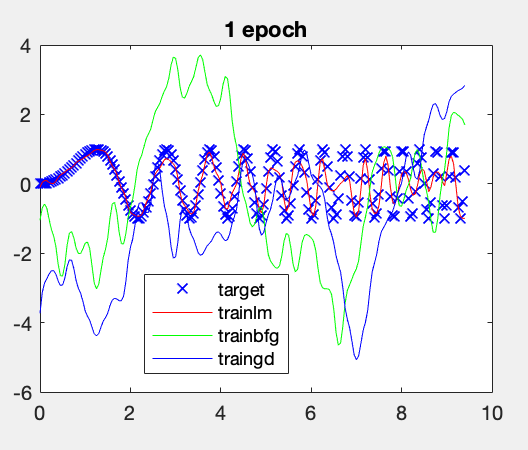
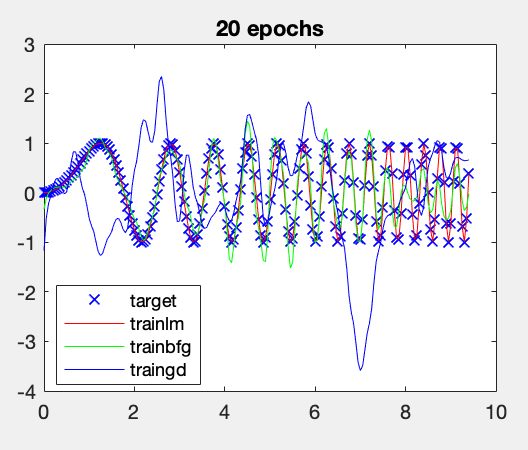
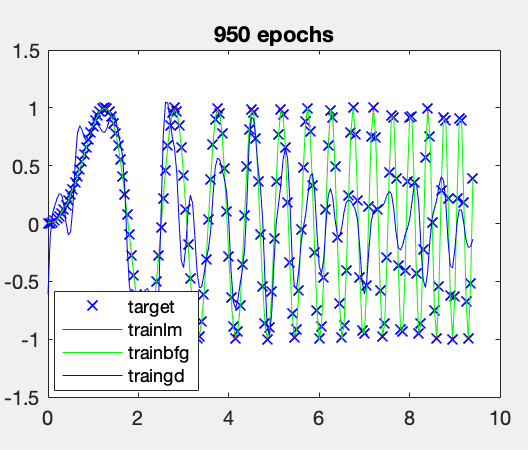
When learning a model, it is important to note that the objective is not to memorize the training data and being able to predict it. The real goal is to approximate the function that is generating this data, in order to be able to generalize this knowledge on new unseen data(the test set). A good model, which is able to generalize, must find a balance in the bias-variance tradeoff. This is basically avoiding overfitting the model with a specific dataset and rendering it unable to generalize. One way of achieveing good generalization without overfitting is via cross-validation.

This is also the goal of the complexity criteria, which states that the complexity f the model should be as low as possible. It is basically the same principle as in the Occam’s razor. Other methods include pruning and ensemble learning.

### Comparison of various algorithms

There exist multiple methods to learn a feedforward multi-layer network with backpropagation. Three of the most popular and discussed methods are the Levenberg-Marquardt(LM), Quasi-Newton(QN) and Gradient Descent(GD) methods, so for the first exercise, we will compare the performance of these three learning methods on a dataset that constists of points generated by a sine function. This task should not be any problem to a neural network of these characteristics, as they are universal approximators, and thus are capable of approximating non-linear functions, as stated in Hornik’s theorem. The three selected methods were trained on the same dataset, with a varying number of epochs(1, 20 and 950 respectively) to see their impact on each model’s accuracy. The figures shown were plotted using 50 neurons in the hidden layer.

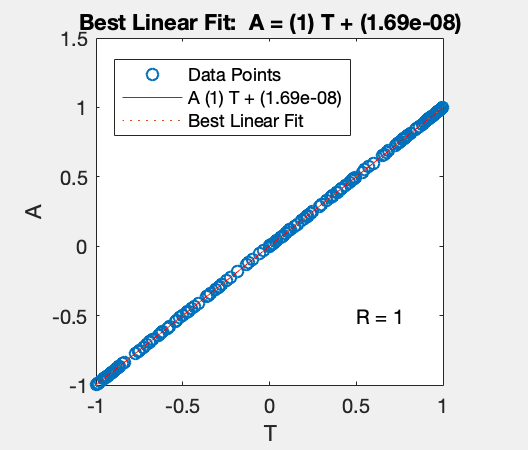
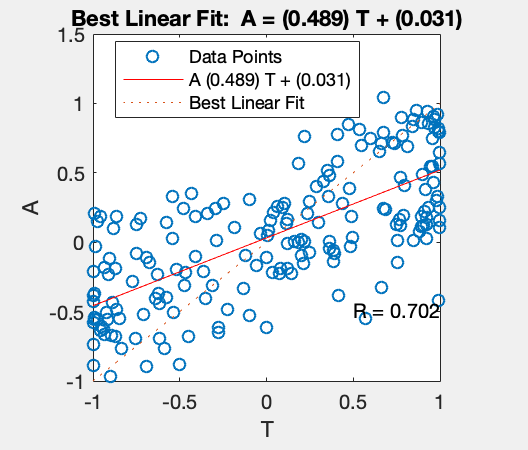
It is important to note that many other learning methods exists, such as Resilient Backpropagation or One Step Secant, but they have been left out of this analysis for the sake of clarity.

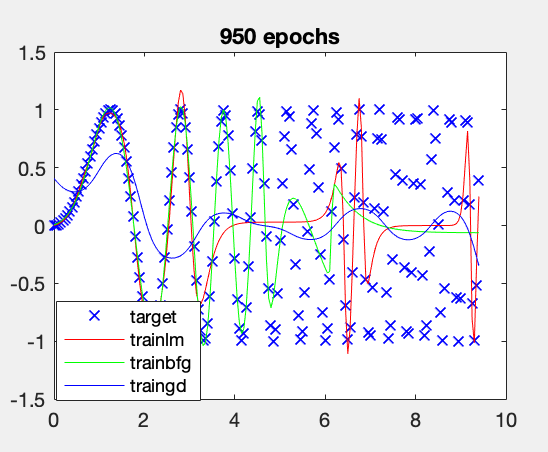
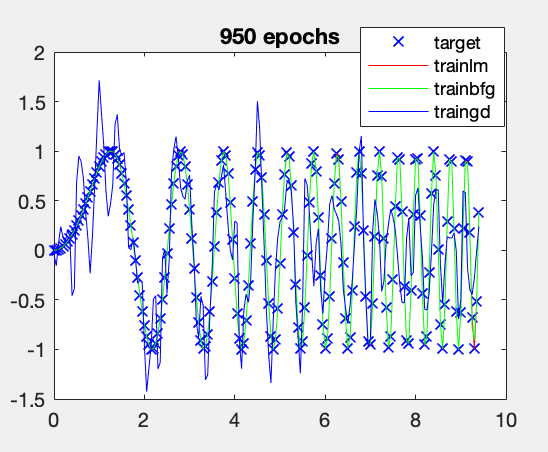
As we can see in the figure, LM achieves by far the best approximation, and it does so quickly, at the cost of more memory usage. Under the same conditions(initial weights, bias, number of epochs and neurons), a pure GD approach gives really disappointing results. This is expected as the LM method if working towards finding the best error reduction, not just the biggest one at a given point.

Given some more epochs, the QN approach also achieves good results, but, again, we can observe that, no matter the number of iterations over the training data, the GD algorithm produces subpar results and it is not able to approximate the input function.

Overall, out of these 3 methods, LM achieves the best results, and also has the best training time. These results are confirmed by performing a regression analysis of the difference between targets and outputs. For example, here is the comparison of the linear fit achieved by LM and GD with 950 epochs:

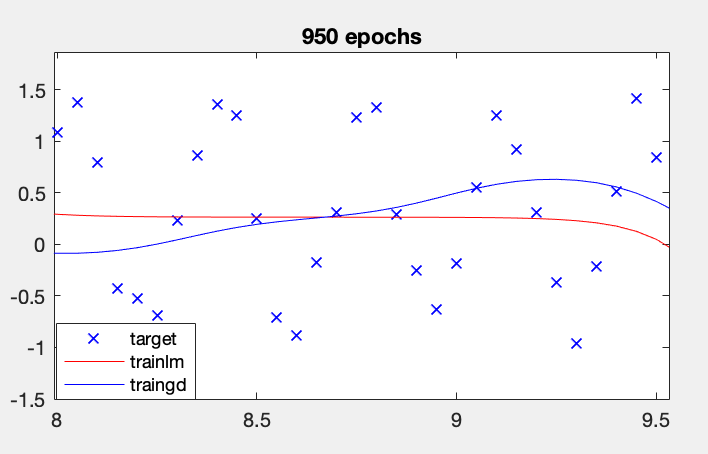
 

The effect of the number of neurons in the hidden layer was also studied. Lowering the number of neurons to 10 translates into significantly worse predictions, while increasing the number of neurons up to 100 exponentially increases the computation cost without affecting the quality of the predictions in a significant way. Although they improve slightly in the case of GD. A good balance between prediction accuracy and computational cost lies around 35 neurons.

### Learning from noisy data

Although these methods work fairly well with the task of approximating a sine function, the results vary tremendously when noise is introduced in the data, and the number of training points in increased slightly. In the following figure, the number of training points has been increase from 3pi to 10 pi, and a random noise factor has been included.

****

We have zoomed to a random location of the plot in order to be able to appreciate the target function points. The difference compared to the previous plots in highly noticeable. Both algorithms are unable to learn the underlying function from the data, thus showing that these methods are not really robust against noise. This is a clear case of overfitting: the models are not able to ignore the noise and thus their predictions do not approximate the function generating the noise.

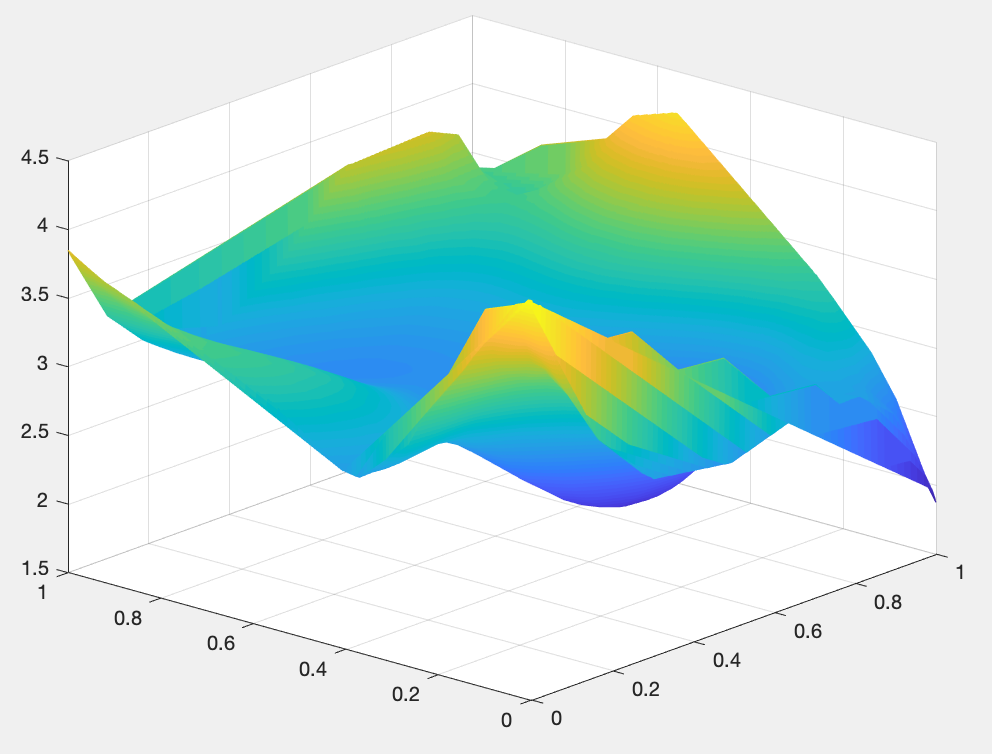
### Approximating an unknown function

Now, we will try to learn an unknown function from data. For this task, a custom dataset was built based on my student number. From an exising dataset T, the custom dataset Tnew was built such as:

Tnew = (d1T1 + d2T2 + d3T3 + d4T4 + d5T5) / (d1 + d2 + d3 + d4 + d5)

Where d1, d2, d3, d4 and d5 are the largest 5 digits from the student number in descending order.

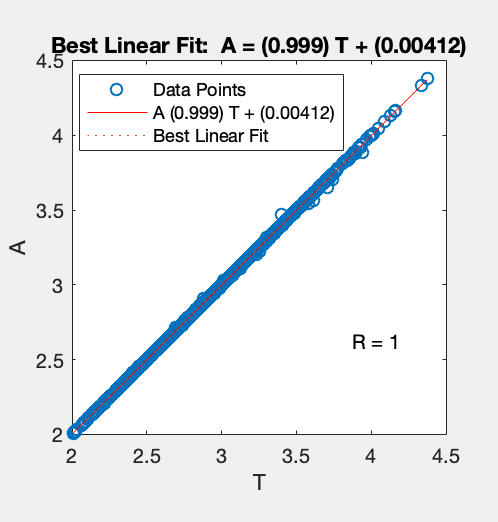
The dataset consists of 3 independent samples of 1000 points each, built with the *datasample* function. The resulting training dataset can be plotted using the mesh function and can be seen in this figure:



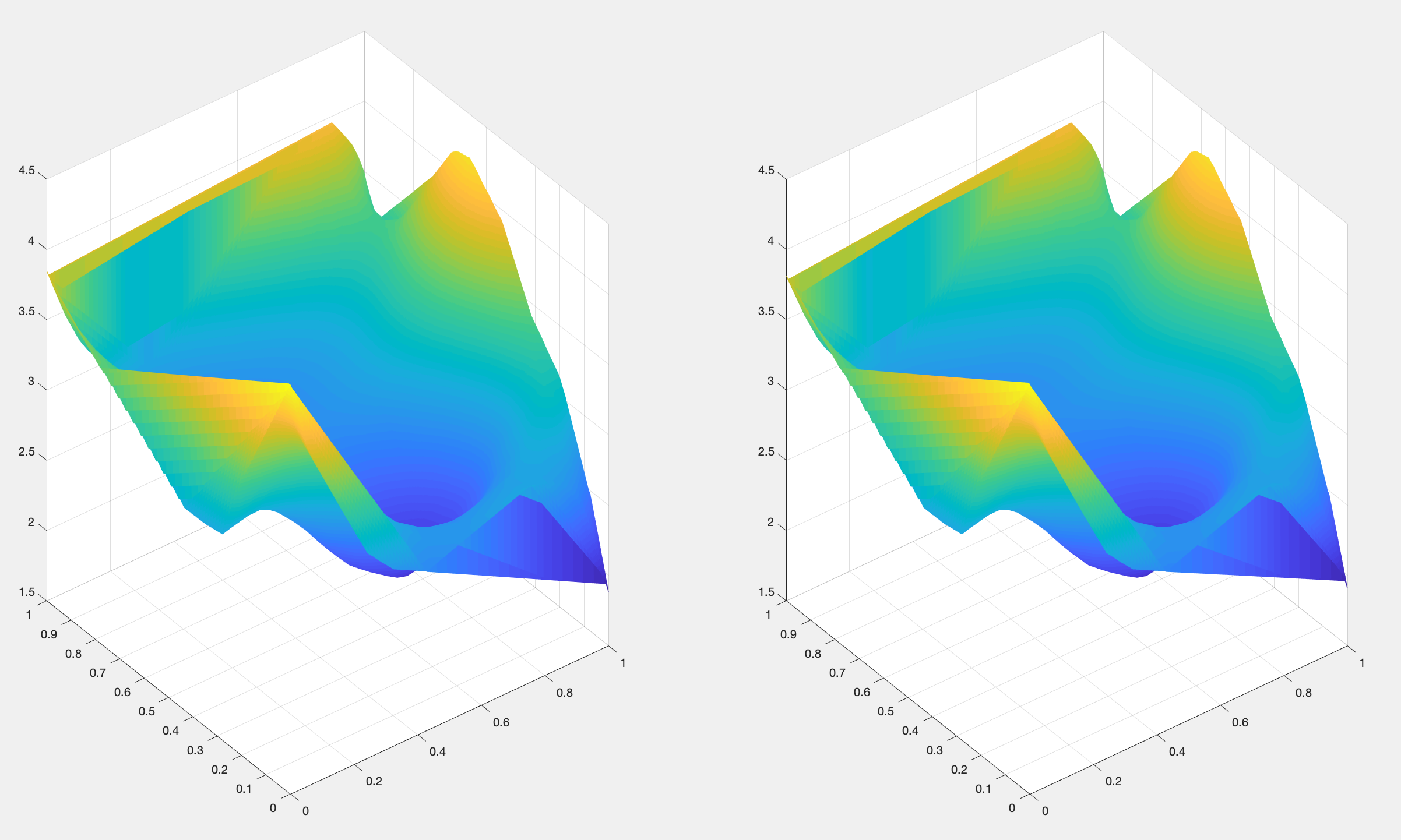
Once the dataset is prepared, we need to find the best possible neural network for the task at hand. In order to find the optimal combination, an algorithm was designed. This algorithm iterates over 6 different learning functions and different number of neurons for the network, and outputs the combination that gives the best R score with the lowest number of neurons possible. We noticed that increasing the number of hidden layers to anything more than 2 hidden layers had a big impact on computational complexity, without improving accuracy significantly, so all tests were performed with 1 and 2 hidden layers.

As these algorithms are non-deterministic, the test was run multiple times and the resulting choice was done via majority vote. As expected based on the previous exercise, the best algorithm for the task was Levenberg-Marquardt, which achieved an R-score of 0.999 with just 15 neurons in two hidden layers.

On the test dataset, the network scores fairly well, which indicates that it is able to generalize the learned concept to new datapoints correctly. Calculating the mean squared error of the predictions a value of almost 0.



Here we can see the actual test points vs the prediction of the model.

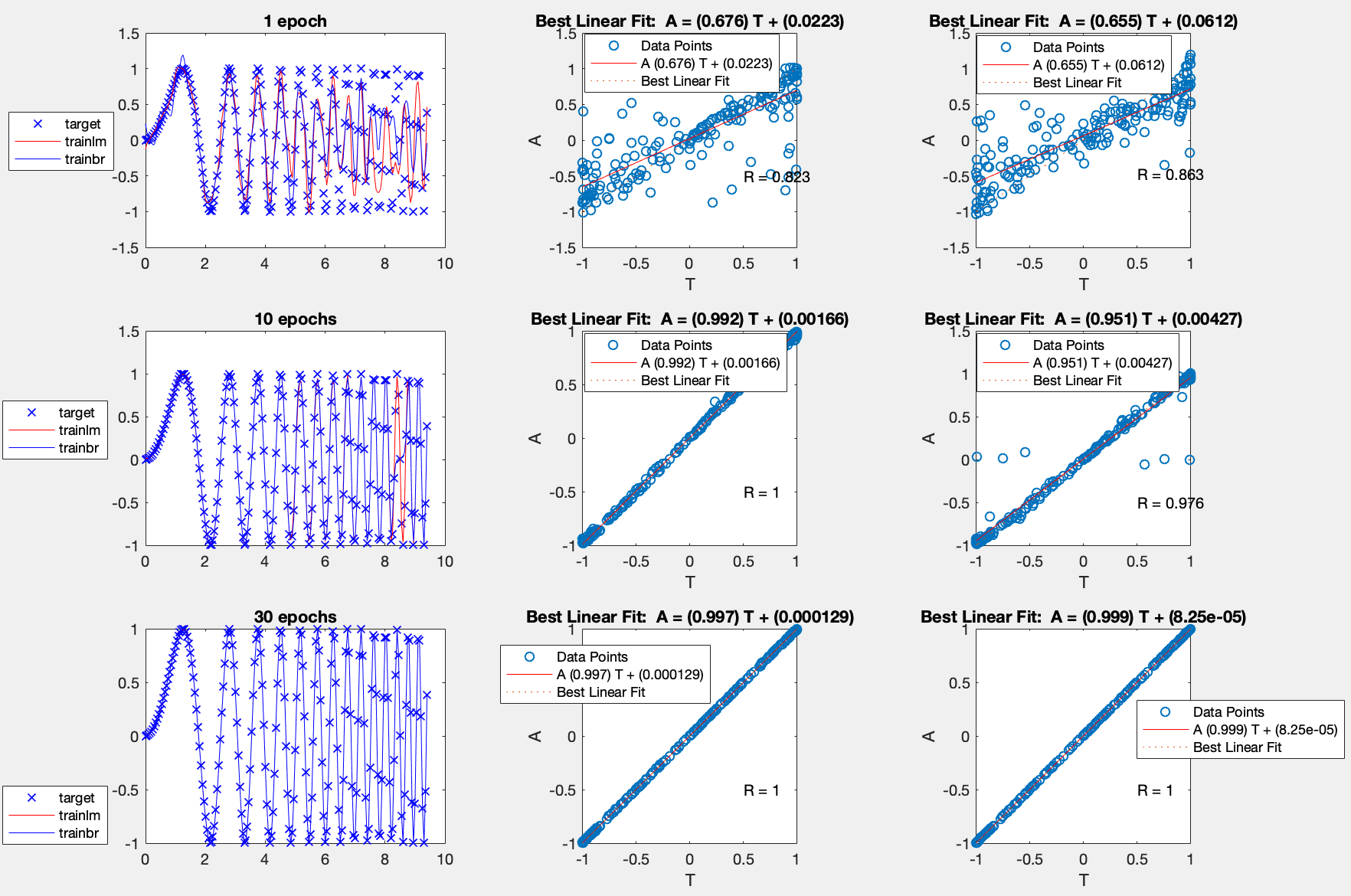


As it can be seen, the figures are almost identical.

### Bayesian Learning Regularization

For the last exercise, we will repeat the analysis performed in the first part of the practical session, but now the Bayesian Learning Regularization(*trainbr*) algorithm will we studied. This algorithm updates the weight and bias values according to the previously studied Levenberg-Marquardt learning algorithm. The main difference is that *trainbr* minimizes a combination of squared errors and weights, and then determines the correct combination so as to produce a network that generalizes well. The process is called Bayesian regularization.

Given this definition, one would expect this algorithm to perform better on test data, as it has better generalization properties. Surprisingly, that is not the case, and *trainlm* performed better in every test. On this figure we can see how *trainlm* is consistently better, no matter the number of epochs.



# Recurrent neural networks

## Hopfield networks

## Long short-term memory networks

# Deep feature learning

# Generative models