An empirical study of the test error versus training error in Artificial Neural Networks

FERNANDO MORGADO DIAS[1] and ANA ANTUNES[2]
[1]Departamento de Matemática e Engenharias
and Centro de Ciências Matemáticas - CCM,
Universidade da Madeira, Campus da Penteada, 9000-390 Funchal,
PORTUGAL
http://dme.uma.pt/morgado

[2]Escola Superior de Tecnologia de Setúbal do Instituto Politécnico de Setúbal
Campus do IPS, Estefanilha, 2914-508 Setúbal,
PORTUGAL
[1]morgado@uma.pt [2]aantunes@est.ips.pt

Abstract: - This paper reports an empirical study of the behavior of the test and training errors in different systems. Frequently the test error of Artificial Neural Networks is presented with a monotonic decreasing behavior as a function of the iteration number, while the training error also continuously decreases. The present paper shows examples where such behavior does not hold, with data collected from systems where it is corrupted by either noise or actuation delay. This shows that selecting the best model is not a simple question and points to automatic procedures for the selection of models as the best solution to optimize their capacity, either with the Regularization or Early Stopping techniques.

Key-Words: - Feedforward Neural Networks, Training Error, Test Error, Regularization, Early Stopping, Weight decay

1 Introduction

This paper reports an empirical study of the behavior of the test and training errors in different systems.

It is very common in the literature to present the test error of an Artificial Neural Network (ANN) with a monotonic decreasing behavior as a function of the iteration number, while the training error also continuously decreases. This behavior is, most of the times, illustrated by drawings instead of simulations or data from a real system. Some examples of exceptions can be found in [1] and [2].

The present paper shows examples where such behavior does not hold, with data collected from systems where it is corrupted by either noise or actuation delay.

The behavior of the test error presented points to automatic procedures for the selection of models as the best solution to optimize their capacity, either with the Regularization or Early Stopping techniques.

The models presented here were trained using a non-variable pre-established initial set of weights to enable the comparison of the results without the random effect of a variable set of weights.

2 The test systems

The data presented here is collected from two different systems.

The first one is from a first order system of a cruise control system as shown in equation 1:

$$H(s) = \frac{1}{s+1} \qquad \text{Eq. 1}$$

which is subject to sampling to actuation delay. The second system is a reduced scale prototype kiln affected by measurement noise. Additional details about this system can be found in [6].

Both systems were chosen because the effect of the perturbations provides a behavior which is different from a simulated system.

For both systems training was performed for 10000 iterations using the Levenberg-Marquardt algorithm [3] [4]. After each training epoch, the test sequence was evaluated to allow permanent monitoring of the training and test error.

Usually a train and a test sequence are used. The first one is used to update the weights based in the error obtained at the output and the second is used to test if the ANN is learning the general behavior of the system to be modeled, instead of learning the training sequence.

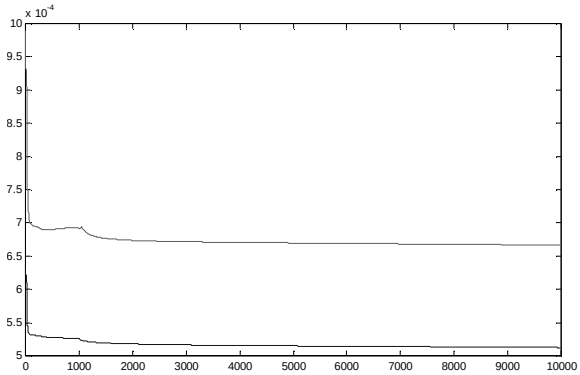

Fig. 1- The test and training error as a function of the iteration for the direct model of the first system with 4 neurons.

Some authors consider using a third sequence to ensure that the ANN is able to generalize the behavior that is desired in different situations or to compare the quality of different networks [5].
Figures 1 to 8, for the first system, and 9 to 23, for the second, represent the plots of both train and test errors for the training of both direct and inverse models of the systems.
In all the figures the training error is always the line with the lower values.

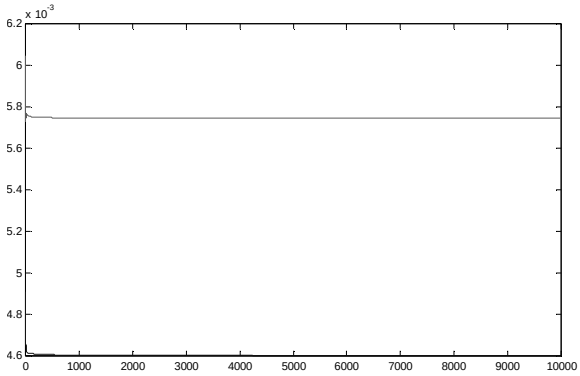

Fig. 2- The test and training error as a function of the iteration for the inverse model of the first system with 4 neurons.

3 Discussion

As stated in the introduction, the literature presents frequently both the training and test errors with a monotonic behavior. It can be seen from the two examples shown here that such behavior is not always found in real systems.

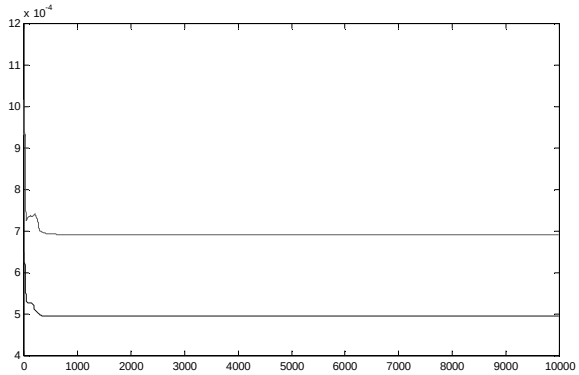

Fig. 3- The test and training error as a function of the iteration for the direct model of the first system with 6 neurons.

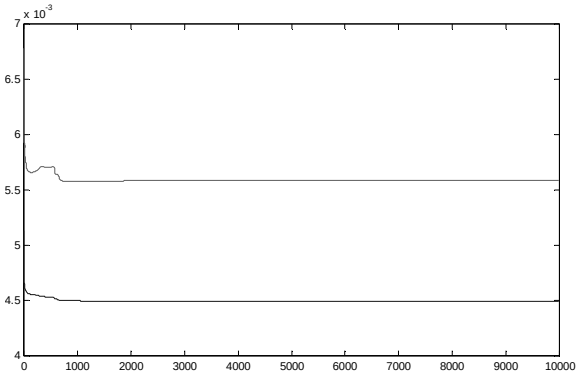

Fig. 4- The test and training error as a function of the iteration for the inverse model of the first system with 6 neurons.

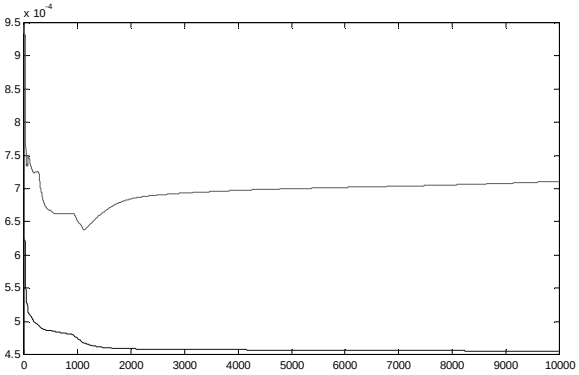

Fig. 5- The test and training error as a function of the iteration for the direct model of the first system with 8 neurons.

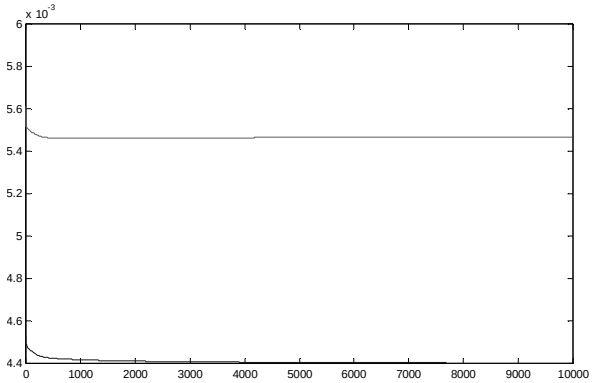

Fig. 6- The test and training error as a function of the iteration for the inverse model of the first system with 8 neurons.

The examples show, in many situations, that both for direct and inverse models, while the training error is always monotonic, the test error finds frequently hills and vales.

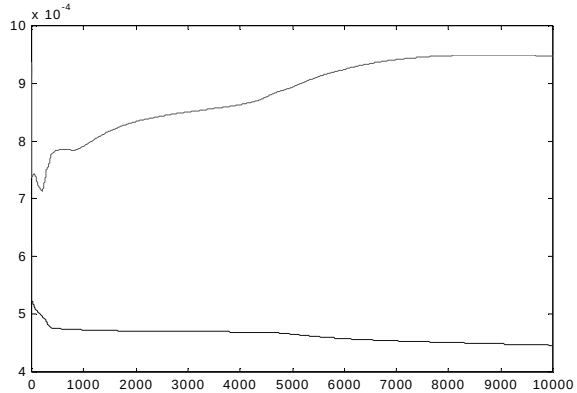

Fig. 7- The test and training error as a function of the iteration for the direct model of the first system with 10 neurons.

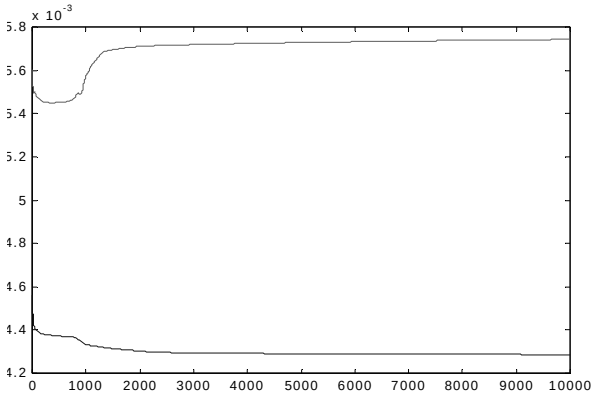

Fig. 8- The test and training error as a function of the iteration for the inverse model of the first system with 10 neurons.

Figures 1 to 4 of the first model and all the figures of the direct model of the second system show an initial phase of learning, followed by a very flat zone where learning is very slow for the test and training error.

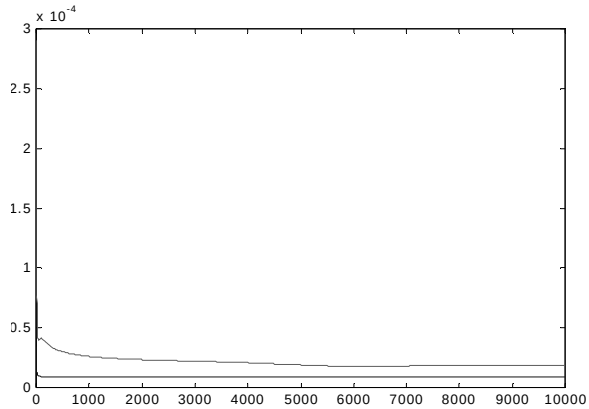

Fig. 9- The test and training error as a function of the iteration for the direct model of the second system with 4 neurons.

In the rest of cases it is possible to find a very different behavior for the train and test error. While the training error continues to decrease with iterations, the test error shows frequently hills and vales.

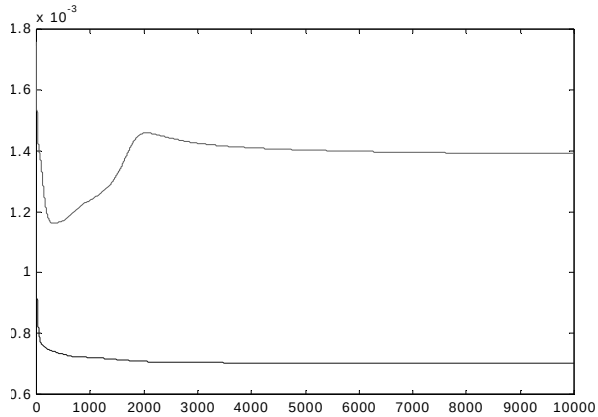

Fig. 10- The test and training error as a function of the iteration for the inverse model of the second system with 4 neurons.

The different tests shown in the figures correspond to the two systems referred above, modeled with different number of neurons. While usually a larger set of weights allows obtaining a lower training error, the existence of more degrees of freedom enable the test error to be composed of much more ups and downs then it is the case with less parameters.

For the second system more figures are shown because, with a larger number of parameters it presents curves which are less common.

4 Regularization and Early Stopping

The oscillations found in the test error of the systems chosen as an example lead to the necessity of choosing carefully the length of the training stage or to apply another solution to obtain the best quality for the models.

Two possible solutions that can be used to cope with these problems are Early Stopping and Regularization techniques.

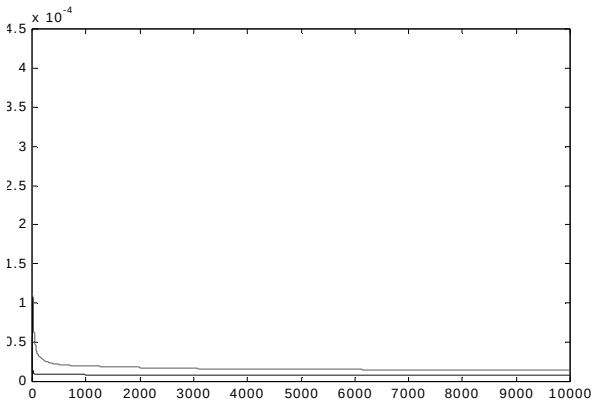

Fig. 11- The test and training error as a function of the iteration for the direct model of the second system with 6 neurons.

4.1 Regularization

For the training algorithms that are based on derivatives the first parameters to be updated are the ones with larger influence in the criteria to be minimized, while in a second phase other less important parameters are updated.

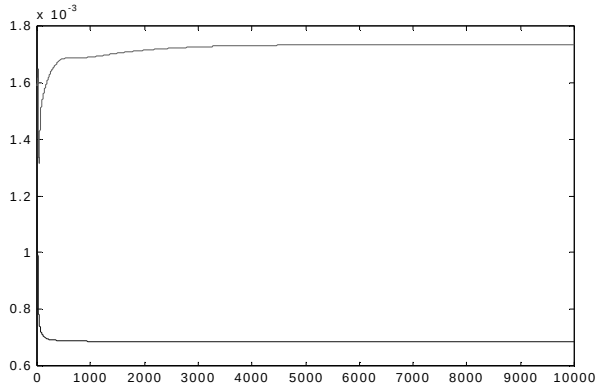

Fig. 12- The test and training error as a function of the iteration for the inverse model of the second system with 6 neurons.

These last parameters to be updated are the ones responsible for the overtraining problem by learning characteristics of the training signal and the noise. The overtraining, i.e. excessive training, situation results in a network that over fits the training sequence but is not capable of the same performance with a test sequence, because the ANN has learned details of the training sequence instead of the general behavior.

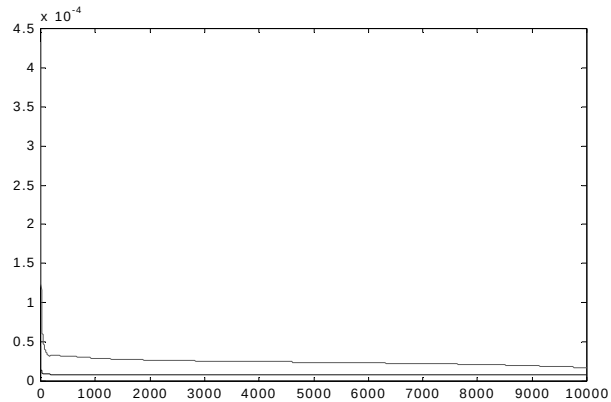

Fig. 13- The test and training error as a function of the iteration for the direct model of the second system with 8 neurons.

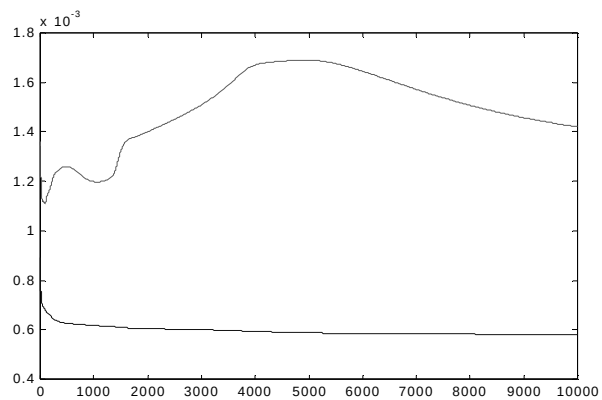

Fig. 14- The test and training error as a function of the iteration for the inverse model of the second system with 8 neurons.

One way to avoid this second phase in training is called regularization and it consists of changing the criteria to be minimized according to:

$$W(\theta) = V(\theta) + \delta\|\theta\|^2 \qquad \text{Eq. 2}$$

where δ, the weight decay is a small value and is the original criterion to be minimized.

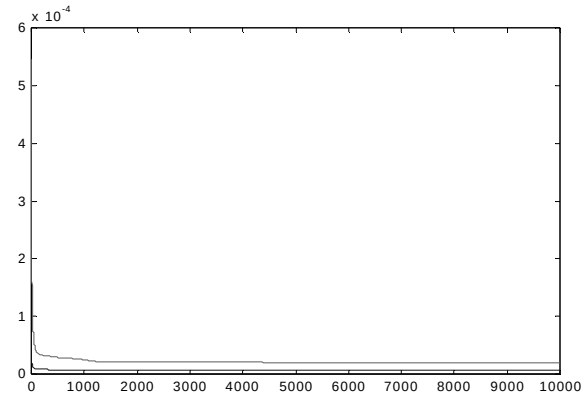

Fig. 15- The test and training error as a function of the iteration for the direct model of the second system with 10 neurons.

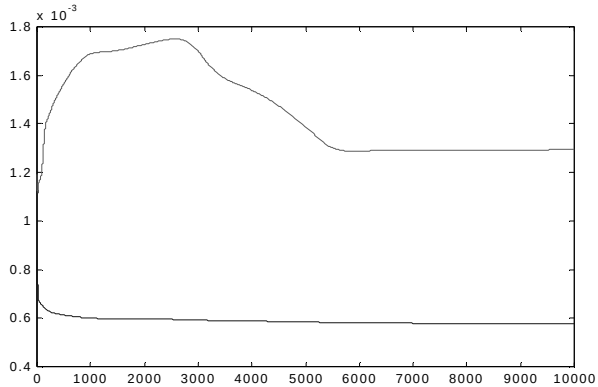

Fig. 16- The test and training error as a function of the iteration for the inverse model of the second system with 10 neurons.

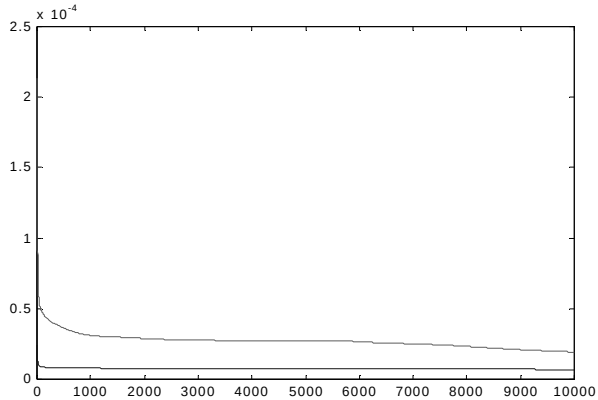

Fig. 17- The test and training error as a function of the iteration for the direct model of the second system with 15 neurons.

The idea is to eliminate the so called second phase in learning where parameters with small influence are updated by introducing a trend towards zero in the parameters.

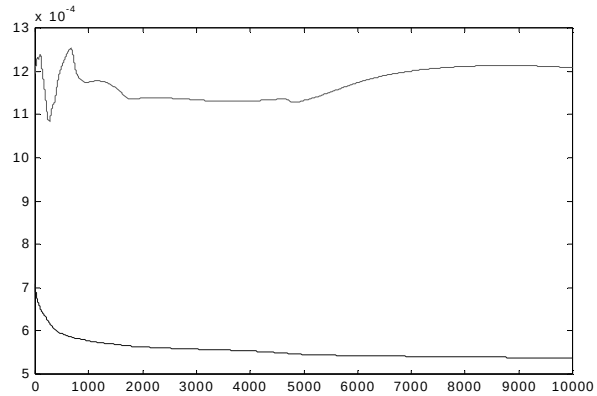

Fig. 18- The test and training error as a function of the iteration for the inverse model of the second system with 15 neurons.

The difficulty is to determine the appropriate value of δ for performing regularization.

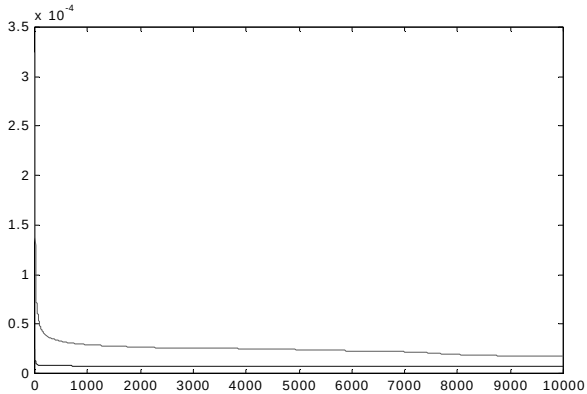

Fig. 19- The test and training error as a function of the iteration for the direct model of the second system with 20 neurons.

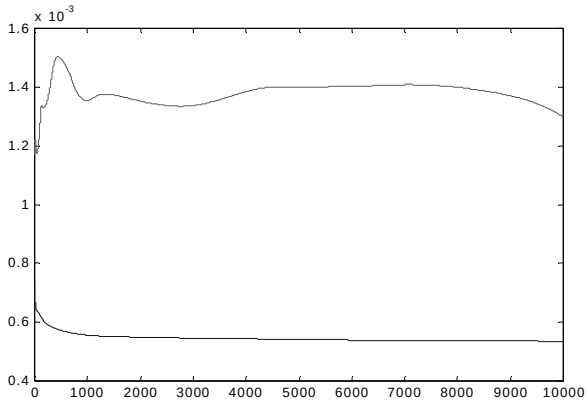

Fig. 20- The test and training error as a function of the iteration for the inverse model of the second system with 20 neurons.

4.2 Early Stopping

Another way to avoid the overtraining, called early stopping, which is quite intuitive, consists in stopping training before the second phase of training starts but after the first one is concluded so that the characteristics of the system are learned.

Clearly the difficulty here is to find the exact number of iterations for performing the training.

Both solutions have been proved to be formally equivalent in [2]. Nevertheless it is important to take into account the difficulties to determine the regularization parameter for explicit regularization or the number of iterations to use for early stopping. One example of comparison of both techniques in an automated procedure can be found in [7].

5 Conclusion

Two systems were presented where the behavior of the test and training error is not the monotonic decreasing usually pointed out in the literature.

The objective is to show that for systems subject to noise or actuation delay it is quite common to find a behavior different from simulated systems.

The differences found in the test error for ANNs with more parameters are also relevant since this larger set of adjustable weights can lead to a better network, but only if the test error is carefully evaluated.

These examples suggest that the choice of the models' characteristics must be careful to avoid getting a worst model after a larger training period. This urges for the use of the Early Stopping and Regularization techniques that can be very helpful both for manual and automatic model selection.

References:

[1] M. Nørgaard, "System Identification and Control with Neural Networks", PhD Thesis, Department of Automation, Technical University of Denmark, 1996.

[2] Sjöberg, J. & Ljung, L., "Overtraining, regularization, and searching for minimum with application to neural nets", International Journal of Control 62(6), 1391–1407, 1995.

[3] K. Levenberg, "A method for the solution of certain problems in least squares", Quart. Appl. Math., 2:164—168, 1944.

[4] D. Marquardt, "An algorithm for least -squares estimation of nonlinear parameters", SIAM J. Appl. Math., 11:431—441, 1963.

[5] Lutz Prechelt, "Automatic early stopping using cross validation: Quantifying the Criteria", Neural Networks, 1998.

[6] F. Morgado Dias and A. Mota, "Comparison between Different Control Strategies using Neural Networks", Mediterranean Control Conference, Dubrovnik, 2001.

[7] F. Morgado Dias, A. Antunes and A. Mota, "Automating the Construction of Neural Models for Control Purposes using Genetic Algorithms", 28th Annual Conference of the IEEE Industrial Electronics Society, Sevilha, 2002.