Steve Lawrence, Andrew D. Back, Ah Chung Tsoi, C. Lee Giles, *On the Distribution of Performance from Multiple Neural Network Trials*, IEEE Transactions on Neural Networks, Volume 8, Number 6, pp. 1507–1517, 1997.

On the Distribution of Performance from Multiple Neural Network Trials

Steve Lawrence,* Andrew D. Back, Ah Chung Tsoi, C. Lee Giles†
{lawrence,giles}@research.nj.nec.com, back@zoo.riken.go.jp,
Ah_Chung_Tsoi@uow.edu.au

Steve Lawrence was with the Department of Electrical and Computer Engineering, University of Queensland. St. Lucia, Australia. He is now with NEC Research, 4 Independence Way, Princeton, NJ 08540.

Andrew D. Back was with the Department of Electrical and Computer Engineering, University of Queensland. St. Lucia, Australia. He is now with the Brain Information Processing Group, Frontier Research Program, RIKEN, The Institute of Physical and Chemical Research, 2-1 Hirosawa, Wako-shi, Saitama 351-01, Japan

Ah Chung Tsoi was with the Department of Electrical and Computer Engineering, University of Queensland. St. Lucia, Australia. He is now with the Faculty of Informatics, University of Wollongong, Northfields Avenue, Wollongong NSW 2522, Australia.

C. Lee Giles is with the NEC Research Institute, 4 Independence Way, Princeton, NJ 08540. He is also with the Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742.

Keywords: neural networks, gradient training, backpropagation, error analysis, convergence, gaussian distribution, probability distributions, statistical methods, box whiskers, kolmogorov-smirnov test, mackey-glass, phoneme classification.

Abstract

The performance of neural network simulations is often reported in terms of the mean and standard deviation of a number of simulations performed with different starting conditions. However, in many cases, the distribution of the individual results does not approximate a Gaussian distribution, may not be symmetric, and may be multimodal. We present the distribution of results for practical problems and show that assuming Gaussian distributions can significantly affect the interpretation of results, especially those of comparison studies. For a controlled task which we consider, we find that the distribution of performance is skewed towards better performance for smoother target functions and skewed towards worse performance

* http://www.neci.nj.nec.com/homepages/lawrence
† http://www.neci.nj.nec.com/homepages/giles.html

for more complex target functions. We propose new guidelines for reporting performance which provide more information about the actual distribution.

1 Introduction

1.1 Performance Measures

It is common in neural network simulations to report results using the mean and standard deviation of a number of trials with different starting conditions [11]. However, in many cases, the distribution of the individual results does not approximate a Gaussian distribution, may not be symmetric, and may be multimodal. For non-Gaussian distributions, more informative methods can be used to present the data (such as plotting the distribution or box-whiskers plots (see section 2.2)).

In this paper, we present a number of experiments which show that the distribution can be non-Gaussian and non-symmetric. Consequently, we propose that, in many cases, results should be treated in a different manner to the common practice of reporting only the mean and standard deviation from a number of trials (a test for normality can be used to determine when the distribution differs significantly from a Gaussian distribution).

1.2 Convergence Properties of Neural Network Training Algorithms

The performance of a neural network simulation is the result of a training process and it is therefore of interest to consider the properties of the training problem. Researchers in computational learning theory have investigated the complexity of neural network learning [9, 5, 15, 19]. Judd [16] showed that even under very restrictive assumptions, the general problem of finding a set of weights consistent with a set of examples is NP-complete. Blum and Rivest [5] have proven that training even a three node network can be NP-complete. These results are for threshold networks. For sigmoid networks, Auer, Herbster and Warmuth [1] have shown that the number of local minima may grow exponentially in the number of parameters. It is conjectured that training may also be NP-complete for sigmoid networks. Therefore, in general, there may be no algorithm capable of finding the optimal set of parameters which has computation time that is bounded by a polynomial in d, the input dimension.

A typical compromise is to use an iterative optimization technique [9] such as backpropagation. In most cases, such techniques are only guaranteed to find a local minimum of the cost function. Backpropagation can fail in very simple cases [6, 17, 18], resulting in a local minimum significantly worse than the global minimum. When the problem and the training algorithm make it hard to find a globally optimal solution, it may be difficult to predict the expected quality of the solutions found. In such cases, there is typically no reason to expect that the distribution of results will always be Gaussian, and therefore the actual distribution is of interest.

A typical method of assessing the performance of a network is to run a number of simulations, each beginning from a different starting point in weight space, and report the mean and standard deviation of the individual

results. This procedure is most suitable when the distribution of the results is Gaussian. For example, if a particular network and training algorithm has a distribution of results which is skewed or multimodal, this will not be observed using the mean and standard deviation. In this case, we propose that a better method of describing the results will provide a more accurate understanding of the true nature of the performance of the network and the algorithm.

The remainder of this paper is organized as follows: in section 2, we examine alternative approaches for describing a distribution. In section 3, we present a number of experiments which show the possible non-Gaussian nature of the distribution of results and indicate how other statistical measures may be better than the traditional mean and standard deviation. Section 4 presents a possible extension to box-whiskers plots. Section 5 provides conclusions and recommendations concerning the presentation of results from multiple neural network trials.

2 Descriptive Statistics

In this section we introduce alternatives to the mean and standard deviation for describing a distribution, and the Kolmogorov-Smirnov test which we use to test a distribution for normality.

2.1 Median and Interquartile Range

The median and the interquartile range (IQR) are simple statistics which are not as sensitive to outliers as the commonly used mean and standard deviation [27]. The median of a probability distribution, $p(x)$, is the value x_m for which smaller and larger values of x are equally probable:

$$\int_{-\infty}^{x_m} p(x)\ dx = \frac{1}{2} = \int_{x_m}^{\infty} p(x)\ dx \tag{1}$$

When given a sample of values from a distribution, the median is estimated as the value x_m which has equal numbers of values above it and below it, i.e. the median is the value in the middle when arranging the distribution in order from the smallest to the largest value. When the number of points is even, it is conventional to estimate the median as the mean of the two central values, i.e. the median is defined as [22]:

$$x_m = \begin{cases} x_{(N+1)/2}, & N \quad \text{odd} \\ \frac{1}{2}\left(x_{N/2} + x_{(N/2)+1}\right), & N \quad \text{even} \end{cases} \tag{2}$$

where the data is in order from the smallest value, x_1, to the largest value, x_N.

If the data is split into two equal groups about the median, then the IQR is the difference between the medians of these groups. The IQR contains 50% of the points. The lower point of the IQR, the median, and the higher point of the IQR are also known as the first, second, and third quartiles respectively (Q1, Q2, and Q3). When comparing the mean and the median, both have advantages and disadvantages. The median is often preferred for distributions with outliers, however the mean takes into account the numerical value of every point whereas the median does not. For example, if a student wishes to average exam results of (5, 90, 94, 92) then the mean

would be more appropriate. However, for neural network result distributions, the distribution of the individual performance results is often of interest, rather than the mean performance from a number of trials. More specifically, the statistic of interest may be the probability that a trial will meet a given performance criterion. The quartiles can provide more information about the distribution of the results and, consequently, the nature of the optimization process.

2.2 Box-Whiskers Plots

Box-whiskers plots [23] use the median and IQR as measures of central tendency and spread rather than the mean and standard deviation. Additionally, they also show the maximum and minimum values of the distribution. The IQR is shown with a box and the median is represented with a bar across the box. Whiskers extend from the ends of the box to the minimum and maximum values. Outliers are sometimes plotted separately (e.g. points greater than 1.5 IQR from the ends of the box may be considered to be outliers and plotted separately [23]).

2.3 Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov (K-S) test can be used to test the normality of a distribution. The K-S D statistic is [25, 12, 22]:

$$D = \max_{-\infty<x<\infty} |S(x) - N(x)| \tag{3}$$

where $S(x)$ is an estimator of the cumulative distribution function of the distribution to test and $N(x)$ is the cumulative distribution function for (in this case) the normal distribution ($\mathrm{erf}(\frac{x}{\sqrt{2}})/2 + 0.5$ for mean 0 and variance 1). The distribution of the K-S D statistic can be approximated for the null hypothesis that the distributions are the same. We can therefore determine the significance level of a given value of D (as a disproof of the null hypothesis that the distributions are the same). The formula is:

$$P = \text{Prob}(D > \text{observed}) = Q_{KS}\left(\left[\sqrt{N} + 0.12 + 0.11/\sqrt{N}\right] D\right) \tag{4}$$

where N is the number of data points and $Q_{KS}(x)$ is:

$$Q_{KS}(x) = 2\sum_{j=1}^{\infty}(-1)^{j-1}e^{-2j^2x^2} \tag{5}$$

P ranges from 0 to 1 and small values of P indicate that the distributions are significantly different (in this case small values of P indicate that the distribution represented by $S(x)$ is not Gaussian). For the results reported here, $S(x)$ is created from the distribution of results after normalization to zero mean and unit variance.

2.4 Presentation of Results

In order to obtain an indication of the differences in presenting results using box-whiskers plots and mean plus and minus one standard deviation points, four sample distributions were used as shown in figure 1. The

distributions are a) a Gaussian distribution, b) a Cauchy distribution [21], c) a Beta distribution [21], and d) a distribution created from the summation of two Gaussian distributions. The equations for these distributions are (respectively):

$$y = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-x^2}{2}\right) \tag{6}$$

$$y = \frac{0.2/\pi}{0.2^2 + x^2} \tag{7}$$

$$y = \begin{cases} \left(\frac{x}{5} + 0.5\right)^{0.1} \times \left(1 - \left(\frac{x}{5} + 0.5\right)\right)^4, & -2.5 \leq x \leq 2.5 \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

$$y = \frac{0.3}{\sqrt{2\pi}} \exp\left(-4.0(x-2)^2\right) + \frac{0.7}{\sqrt{2\pi}} \exp\left(-4.0(x+2)^2\right) \tag{9}$$

We generated 100 points corresponding to each of the four distributions. Table 1 shows the statistics mentioned earlier calculated from these sets of 100 points, and figure 2 shows box-whiskers plots along with the usual mean plus and minus one standard deviation plots. For distributions b), c), and d), it can be observed that the box-whiskers plots provide more information about the actual distribution. For b), the box reduces in size relative to the whiskers, indicating that the central 50% of points lie within a smaller region. For c) and d) the box moves in the direction that the data is skewed and the median moves towards the bottom of the box (the direction of skew). It is not possible to determine from the box-whiskers plot that d) is multimodal. Hence, box-whiskers plots generally give an indication of significant differences in skew and kurtosis.

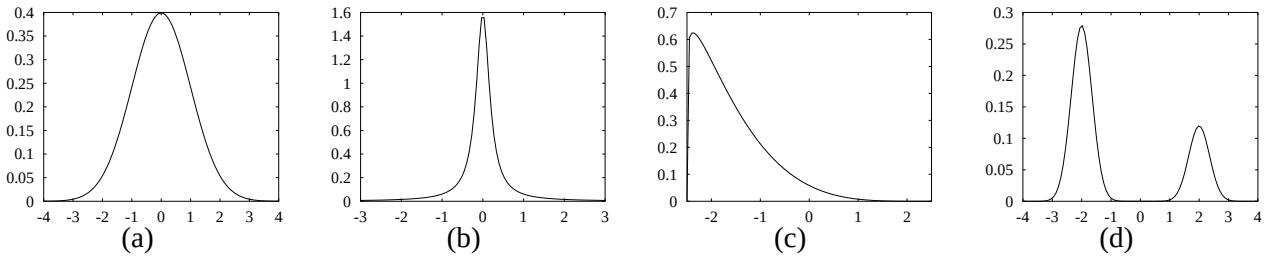

Figure 1. Sample distributions: a) Gaussian, b) Cauchy, c) Beta, and d) summation of two Gaussians.

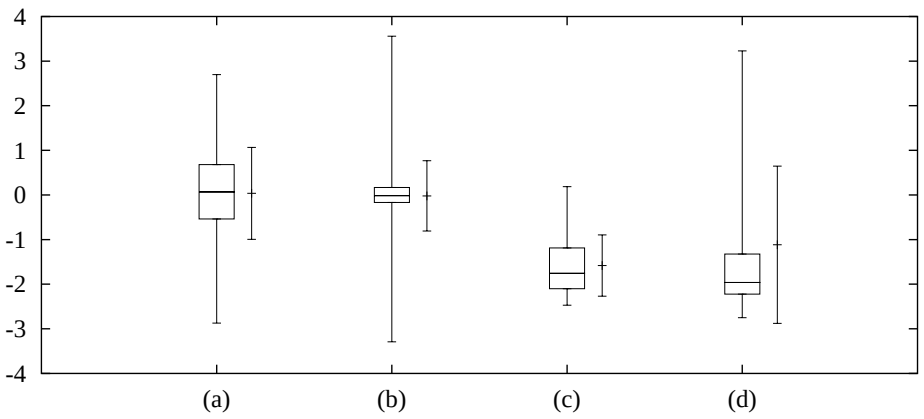

Figure 2. Box-whiskers plots along with the mean and plus/minus one standard deviation for the sample distributions in figure 1. In each case, a box-whiskers plot is shown as a rectangular box with vertical extensors. The box corresponds to the IQR, the bar represents the median, and the whiskers extend to the minimum and maximum values. Immediately on the right of each box plot, the mean and plus/minus one standard deviation is shown as a vertical line.

Distribution	Mean	Median	Std. Dev.	IQR	Q1	Q3	Min	Max	K-S D	K-S P
a	0.036	0.067	1.03	1.22	-0.54	0.68	-2.87	2.70	0.051	0.95
b	-0.020	-0.014	0.79	0.34	-0.17	0.17	-3.29	3.56	0.27	8.5e-7
c	-1.58	-1.75	0.68	0.91	-2.10	-1.19	-2.47	0.18	0.64	4.1e-37
d	-1.11	-1.96	1.76	0.90	-2.22	-1.33	-2.75	3.23	0.69	9.2e-44

Table 1. Statistics for 100 points generated from the distributions in figure 1. Q1 and Q3 are the first and third quartiles (IQR = Q3 - Q1). The K-S values indicate that there is a high likelihood that the points came from a Gaussian distribution only for (a).

3 Empirical Result Distributions

We are primarily interested in the distribution of results for practical problems, and the resulting implications for how results are presented. Therefore, we present the results of a number of experiments using problems that have been commonly used in the neural network literature. In each case, we plot and analyze the distribution of the network error for the training and test data.

3.1 Training Details

We used standard backpropagation with stochastic update (update after every training point). Except when specified, all networks are MLPs. All inputs were normalized to zero mean and unit variance. The quadratic cost function was used [13]. The learning rate was reduced linearly to zero over the training period from an

initial value of 0.1[1]. Performance is reported in terms of the percentage of examples incorrectly classified (for the classification problem) or normalized mean squared error (NMSE) which is defined as [26]:

Definition 1

$$\text{NMSE} = \frac{\sum_{k=1}^{N_p} \sum_{j=1}^{N_o} (d_{kj} - y_{kj})^2}{\left(\sum_{k=1}^{N_p} \sum_{j=1}^{N_o} \left(d_{kj} - \left(\sum_{k=1}^{N_p} \sum_{j=1}^{N_o} d_{kj} \right) / (N_p N_o) \right)^2 \right) / (N_p N_o)} \tag{10}$$

where d is the desired or target value, y is the predicted value, N_p is the number of patterns, and N_o is the number of outputs. □

3.2 Phoneme Data

These experiments use a database from the ESPRIT ROARS project. The aim of the task is to distinguish between nasal and oral vowels [24]. There are 3600 training patterns, 1800 test patterns, five inputs provided by cochlear spectra, and two outputs. Using 10 hidden nodes and 250,000 iterations per trial, the distribution of results is shown in figure 3. It can be observed that the distributions are skewed towards better performance and are a) not Gaussian and b) not symmetric. Statistics of the distributions are shown in table 2. From the K-S statistics, we observe that the probability of the distributions being Gaussian is very low (note that the underlying distribution of performance is discrete rather than continuous – we are relying on there being sufficient patterns in order to approximate a continuous distribution reasonably well). Figure 4 shows box-whiskers plots and the usual mean and standard deviation plots for the training and test distributions. It can be observed that the median is significantly different to the mean. The fact that the distribution is skewed towards better performance is not evident from the mean and standard deviation values. The extra information contained in the actual distribution and the box-whiskers plots can be important, e.g. the distribution can give an indication of how often a trial meets a given performance criterion and can therefore be useful in determining how many trials are required in a given situation.

3.3 Mackey-Glass Time Series

The Mackey-Glass equation is a time delay differential equation first proposed as a model of white blood cell production [20]:

$$\frac{dx}{dt} = \frac{ax(t-\tau)}{[1 + x^c(t-\tau)]} - bx(t) \tag{11}$$

[1]We have found this to result in similar performance to the "search then converge" learning rate schedules proposed by Darken and Moody [8].

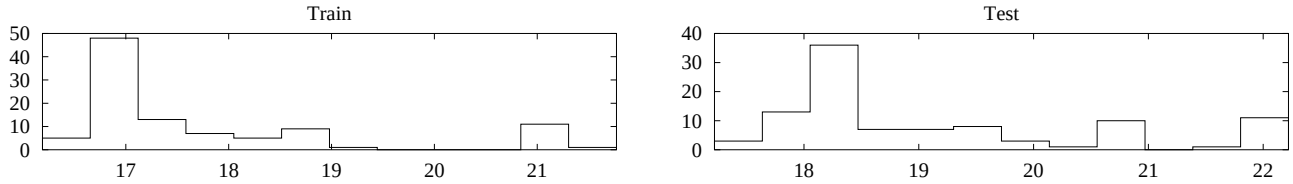

Figure 3. The distribution of classification error for the networks trained on the phoneme problem. The left hand graph shows the training distribution and the right hand graph shows the test distribution. The abscissa corresponds to the percentage of examples incorrectly classified and the ordinate represents the percentage of individual results falling within each section of the histogram. The distribution is created from 200 individual simulations with random starting points. It can be observed that the distribution is skewed towards better performance. Note that the scales change between the graphs.

	Mean	Median	Std. Dev.	IQR	Q1	Q3	Min	Max	K-S D	K-S P
Training error	17.7	17.1	1.42	1.53	16.9	18.4	16.2	21.8	0.22	6.6e-5
Test error	19.2	18.4	1.34	1.45	18.3	19.8	17.2	22.2	0.22	9.6e-5

Table 2. Statistics for the distribution of results for the phoneme task. Q1 and Q3 are the first and third quartiles (IQR = Q3 - Q1). Note that, as would be expected from the distributions, the median is lower than the mean and the minimum and maximum values are not symmetric about either the mean or the median. The K-S P values are very low, indicating that there is a very low probability that the observed samples came from a Gaussian distribution.

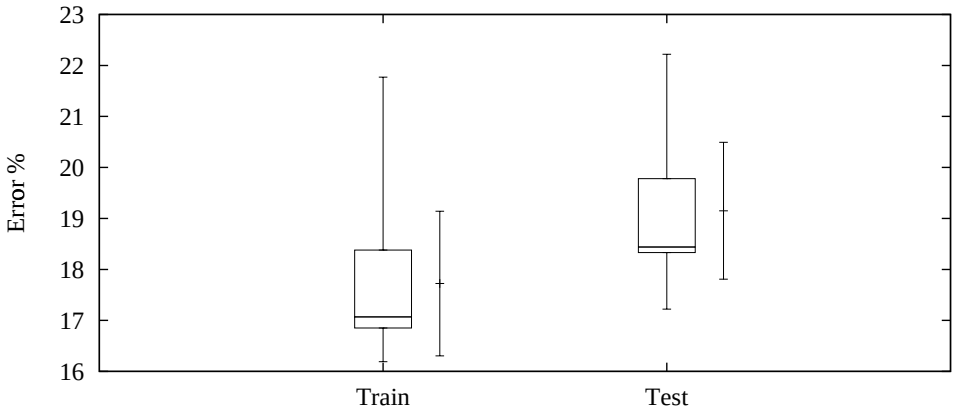

Figure 4. Box-whiskers plots for the phoneme task along with the mean and plus/minus one standard deviation. The box-whiskers plots give an indication that the actual distributions are skewed towards better performance.

where the constants are commonly chosen as $a = 0.2$, $b = 0.1$, and $c = 10$. The delay parameter τ determines the behavior of the system [10]. For $\tau < 4.53$ there is a stable fixed point attractor. For $4.53 < \tau < 13.3$ there is a stable limit cycle attractor. Period doubling begins at $\tau = 13.3$ and continues until $\tau = 16.8$. For $\tau > 16.8$ the system produces a chaotic attractor. For the experiments reported here, $\tau = 30$, the series has been subsampled using $\Delta T = 6$, and the models are trained to predict the value of the series one step ahead.

For the Mackey-Glass problem, the results of a number of architectures are compared: MLP, FIR MLP, and IIR

MLP [3, 4]. The FIR and IIR MLP networks are similar to the standard MLP except each synapse is replaced by FIR and IIR[2] filters respectively. For the FIR MLP, the FIR filters were order 5 (6 taps). For the IIR MLP the IIR filters were order (5, 5). In both cases, filters were only used in the first layer synapses – the second layer synapses contained standard weights. The MLP networks used an input window of 6. Each network had 5 hidden nodes and was trained for 200,000 updates. There were 1,000 training patterns and 1,000 test patterns. The FIR and IIR networks were tested both with and without synaptic gains [2]. It is interesting to observe the difference in the distribution of results in this case. When using synaptic gains an extra parameter is inserted into each synapse which multiplies the weighted sum of the individual filter outputs. Altering a synaptic gain is equivalent to altering all of the weights corresponding to the filter taps. The addition of synaptic gains does not affect the representational power of the networks, however it does affect the error surface and the extra degrees of freedom may make optimization easier [2].

Figure 5 shows the distribution of the normalized mean squared error (NMSE) results. It can be observed that the distribution varies significantly across the various models[3] and that the distributions are often highly skewed and several are multimodal. Figure 6 shows box-whiskers plots and the usual mean and standard deviation plots for these models. Table 3 shows the statistics plotted in figure 6 along with the K-S values D and P, from which it can be observed that all of the distributions have very low probability of being Gaussian except for the MLP case.

If normality of the results distributions is assumed, and the actual nature of the distributions is ignored, this would lead to the following problems:

- Significant differences in the distributions of the results for different algorithms would be masked.
- The FIR network with gains would be considered better than the IIR network with gains (because the mean error for these cases is similar and the standard deviation is lower for the FIR gains case). However, in reality, the IIR gains case may be preferred – the median error for the IIR case is almost an order of magnitude lower than that for the FIR gains case. Thus, in this case, if normality of the distribution of results is assumed, then the best performance could be attributed to the wrong network.
- An observer operating under the assumption that the results are approximately normally distributed would be under the impression that a percentage of networks obtained performance better than the mean minus one standard deviation points. However, none of the 100 trials results in such performance for the two IIR test cases – the mean minus one standard deviation is actually lower than the best individual error.

In general, it can be observed that the box-whiskers plots can be more informative than the mean plus standard deviation plots, but are not as informative as the actual distributions.

[2] FIR: Finite Impulse Response, IIR: Infinite Impulse Response.

[3] It is also interesting to note the significantly different distributions for the FIR and IIR MLP networks with and without synaptic gains (recall that the addition of synaptic gains does not alter the computational capabilities of the network).

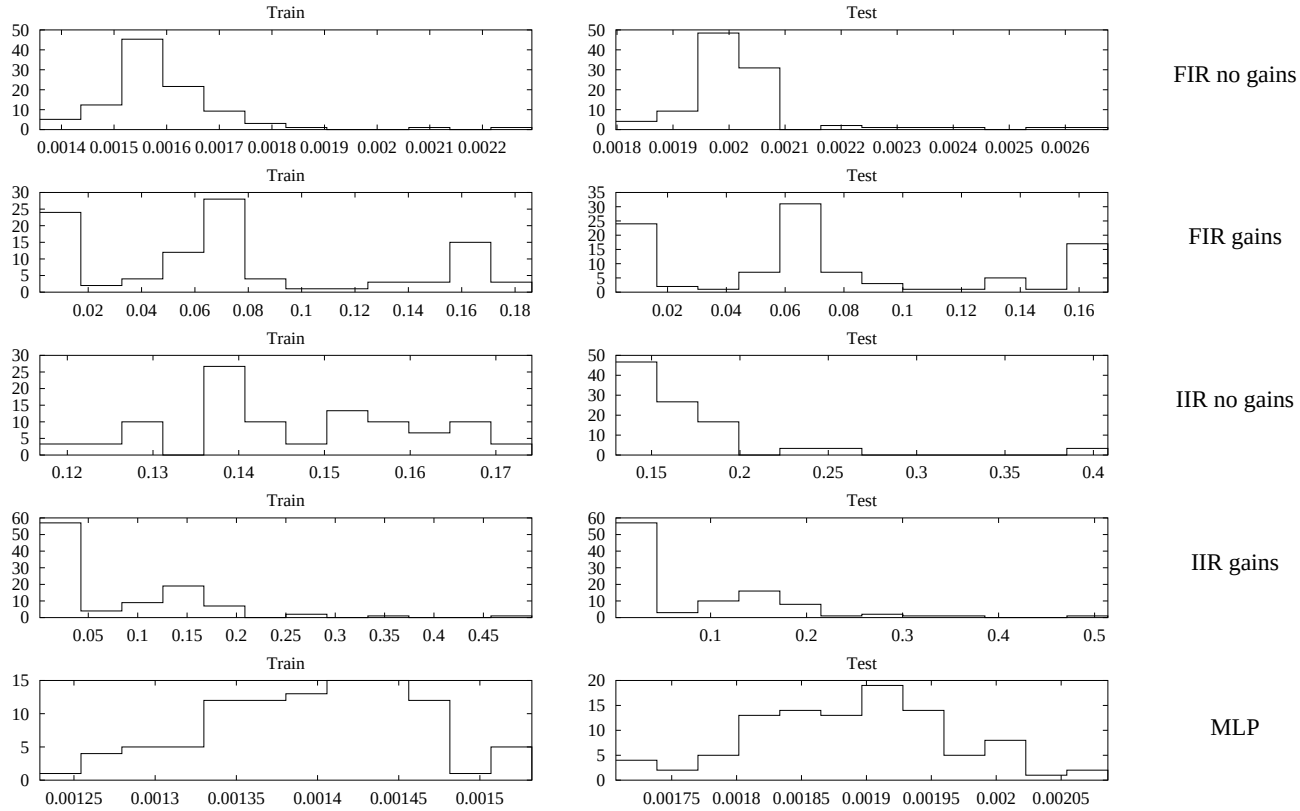

Figure 5. The distribution of the NMSE results for the MLP, FIR MLP, and IIR MLP networks trained on the Mackey-Glass problem (from 100 simulations in each case). The left hand graphs show the distribution of training errors and the right hand graphs show the distribution of test errors. The abscissa corresponds to the mean squared error and the ordinate represents the percentage of individual results falling within each section of the histogram. Note that the scales change from graph to graph.

3.4 Artificial Task

In order to conduct a controlled experiment where we vary the complexity of the target function, we used the following artificial task[4]:

1. An MLP with 5 input nodes, 5 hidden nodes, and 1 output node is initialized with random weights, uniformly selected within a specified range, i.e., w_i in the range $-K$ to K, where w_i are the weights of the network except the biases, and K is a constant. The bias weights are initialized to small random values in the range $(-0.01, 0.01)$. In general, as K is increased, the "complexity" of the function mapping is increased as will be shown later.
2. n_{tr} data points are created by selecting random inputs with zero mean and unit variance and propagating them through the network to find the corresponding outputs. This dataset $\mathcal{S}$ forms the training data for

[4]The task is similar to the procedure used in [7].

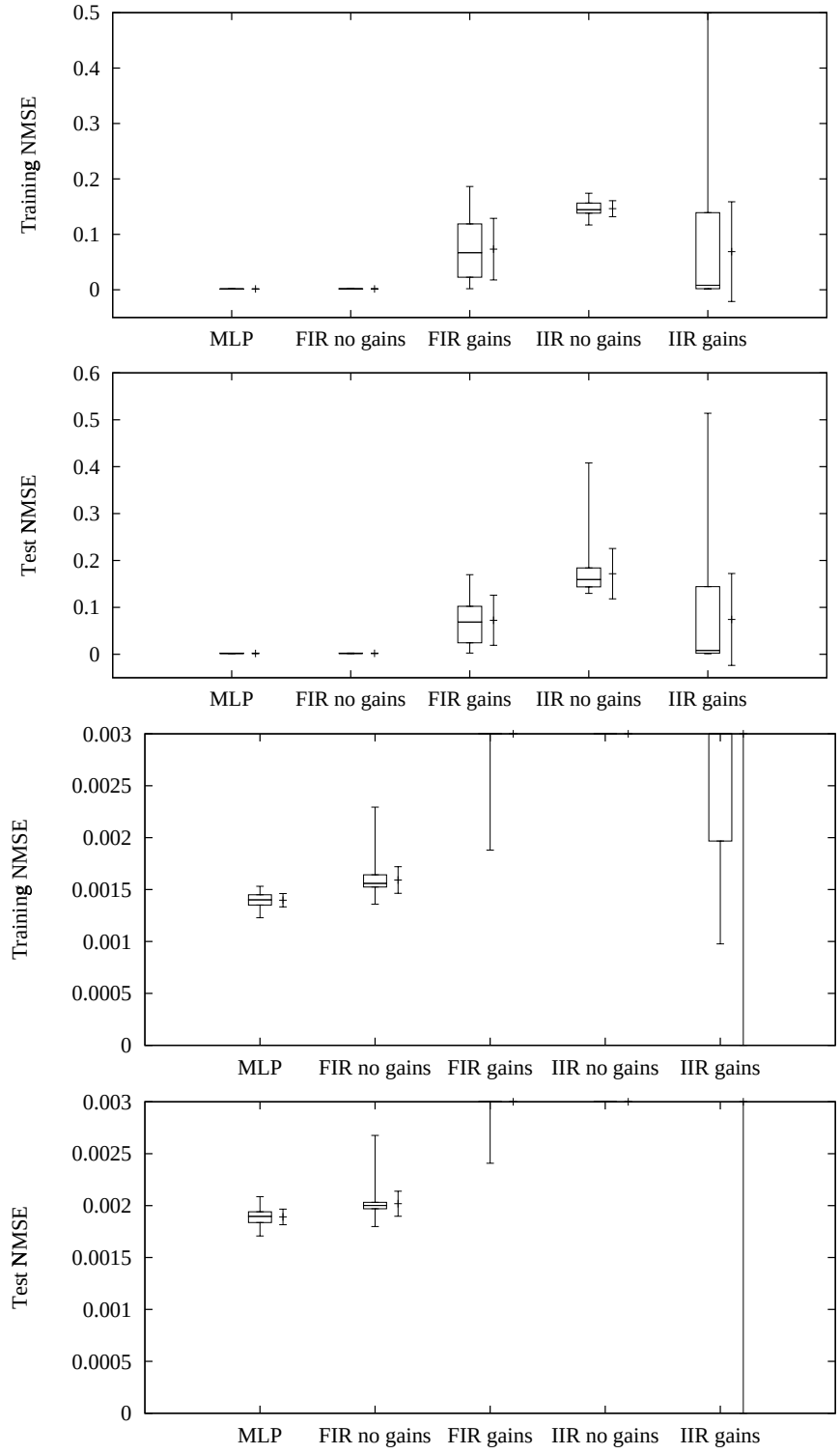

Figure 6. Box-whiskers plots for the Mackey-Glass task along with the mean and plus/minus one standard deviation. The MLP and FIR no gains cases are compressed due to the relatively poor performance of the other cases. Therefore, the plot has been repeated in the lower two graphs with a different scale on the ordinate.

Training normalized mean squared error										
Model	Mean	Median	Std. Dev.	IQR	Q1	Q3	Min	Max	K-S D	K-S P
MLP	0.00140	0.00140	6.41e-5	9.90e-5	0.00135	0.00145	0.00123	0.00153	0.066	0.76
FIR no gains	0.00159	0.00156	0.000128	0.000118	0.00152	0.00164	0.00136	0.00229	0.34	1.3e-10
FIR gains	0.0733	0.0669	0.0556	0.0962	0.0226	0.119	0.00188	0.186	0.18	0.0035
IIR no gains	0.146	0.145	0.0145	0.0182	0.138	0.157	0.117	0.174	0.5	≈ 0
IIR gains	0.0687	0.00822	0.0900	0.137	0.00197	0.139	0.000976	0.499	0.28	2.5e-7

Test normalized mean squared error										
Model	Mean	Median	Std. Dev.	IQR	Q1	Q3	Min	Max	K-S D	K-S P
MLP	0.00189	0.00190	7.55e-5	0.000103	0.00184	0.00194	0.00171	0.00209	0.052	0.95
FIR no gains	0.00202	0.00200	0.000121	6.10e-5	0.00197	0.00203	0.00180	0.00268	0.39	6.8e-14
FIR gains	0.0726	0.0688	0.0533	0.0778	0.0245	0.102	0.00241	0.170	0.17	0.0054
IIR no gains	0.172	0.159	0.0536	0.0404	0.144	0.184	0.130	0.408	0.5	≈ 0
IIR gains	0.0743	0.00820	0.0980	0.142	0.00232	0.144	0.00136	0.514	0.28	2.2e-7

Table 3. Statistics for the distribution of results for the various models on the Mackey-Glass problem. From the K-S values, it can be observed that all of the distributions have very low probability of being Gaussian except for the MLP case.

subsequent simulations. The procedure is repeated to create a test dataset with n_{te} points. n_{tr} is 1,000 and n_{te} is 5,000.

3. The training data set $\mathcal{S}$ is used to train new MLPs. The initial weights of these new networks are set using standard procedures (i.e. they are not equal to the weights in the network used to create the dataset). They are initialized on a node by node basis as uniformly distributed random numbers in the range $(-2.4/F_i, 2.4/F_i)$ where F_i is the fan-in of neuron i [13]. Each network was trained for 200,000 updates.

Figure 7 shows the process graphically.

It is difficult to visualize the function we are trying to approximate. A simple method which gives us an indication is plotted in figure 8 and is created as follows:

```
for each output o
    for each input i
        set all inputs ≠ i equal to 0
        plot output o as input i is varied from -2 to 2
        repeat 4 times
            set all inputs ≠ i to Gaussian random values (μ = 0, σ² = 1)
            plot output o as input i is varied from -2 to 2
```

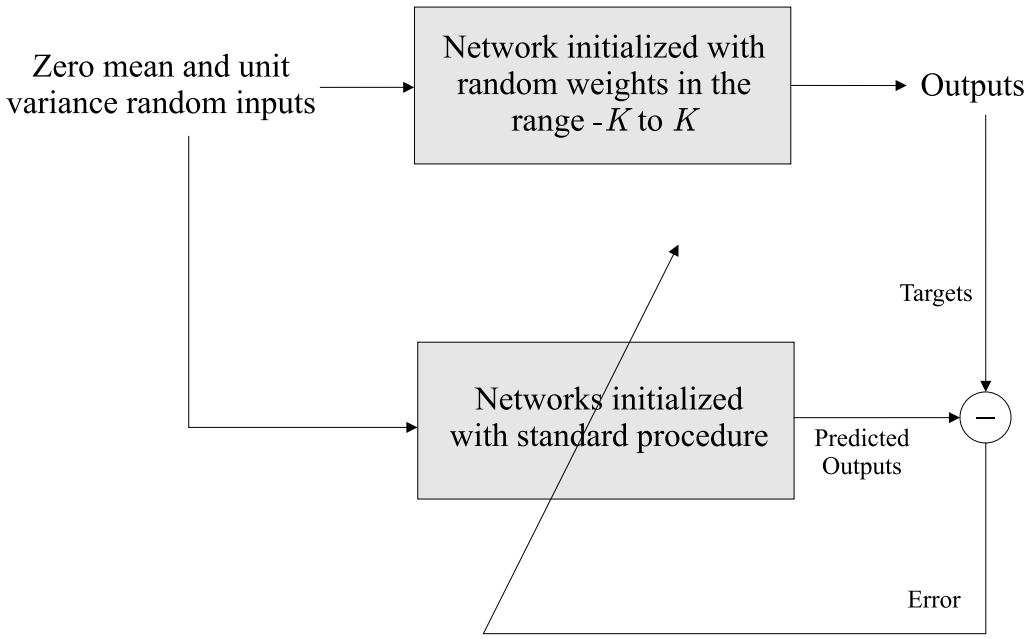

Figure 7. The process of creating the data sets.

Figure 9 shows histograms of the distribution of results for the following four cases: $K = 1, 5, 10, 15$. It can be observed that the distribution of performance is skewed towards better performance for smoother target functions (lower K) and skewed towards worse performance for more complex target functions (higher K), i.e. a general trend can be observed where a higher frequency of the trials resulted in relatively worse performance as K was increased. Note that there is significant multimodality for high K. Figure 10 shows box-whiskers plots and the usual mean and standard deviation plots for these four cases. Table 4 shows the statistics plotted in figure 10 along with the K-S values D and P, from which it can be observed that all of the distributions have very low probability of being Gaussian. Note that the mean minus one standard deviation for $K = 1$ and $K = 5$ is actually lower than the best individual error (from 100 trials).

As with other descriptive statistics box-whiskers plots can be less informative than the actual distributions. Comparing figures 9 and 10, it can be seen that the histograms of the actual distributions are indeed more informative than the box-whiskers plots. However, as with other descriptive statistics, the box-whiskers plots present a simpler view of the data which is quicker to interpret and compare. Hence, the box-whiskers plots can be seen as being in between the use of the mean and standard deviation and the use of the actual distributions.

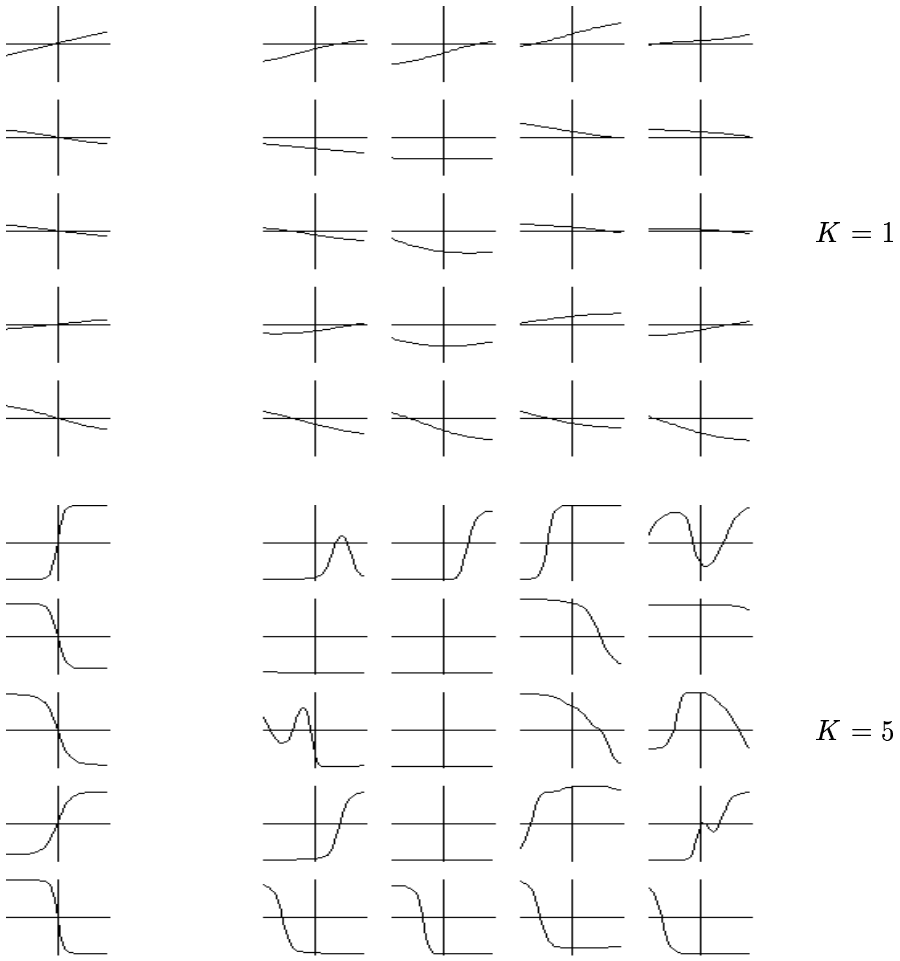

Figure 8. Plots indicating the complexity of the mapping for the MLP with different values of K, the maximum value of the random weights. The plots were created as per the pseudo-code. Each individual plot shows the network output versus one of the inputs. The rows correspond to the five inputs of the networks. The first column corresponds to the case where all other inputs are set to zero, and the remaining columns correspond to the cases where the other inputs are set to random values.

4 Extended Box-Whiskers Plots

Box-whiskers plots provide a better indication of the true distribution than the mean and standard deviation. They are limited however. For example, the bimodal nature of the distribution in figure 1 and the trimodal nature of the distributions in figure 9 for $K = 15$ cannot be identified from the box-whiskers plots.

Box-whiskers plots divide the data into four segments where the division points are the median and the medians of the two segments above and below the original median. Dividing the data into a larger number of segments is one way of increasing the expressive power of the plots. Figure 11 shows the results for the artificial task using division into eight segments instead of division into four (by dividing each of the four segments about the median of the segment). It can be seen that the resulting plot does provide more information. For example, the trimodal nature of the $K = 15$ test NMSE distribution can be identified (from the variation in the size of the

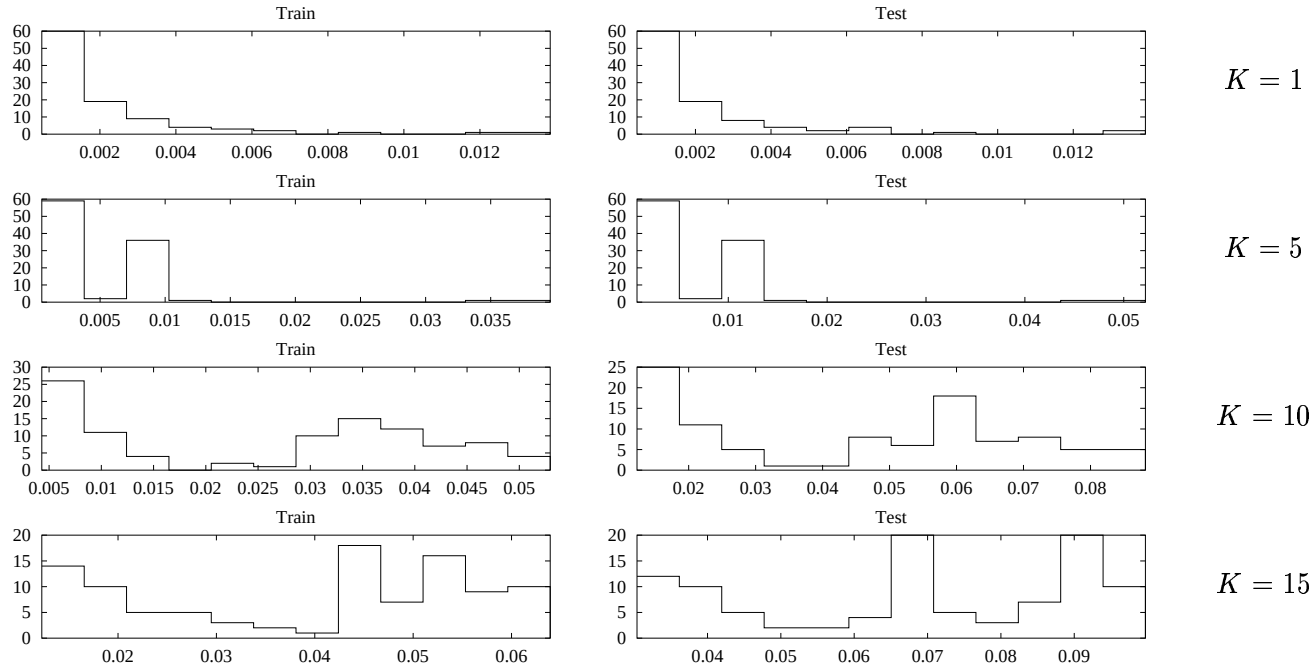

Figure 9. The distribution of errors for the networks trained on the artificial task (from 100 simulations in each case). From top to bottom, the graphs correspond to K values of 1, 5, 10, and 15. The left hand graphs show the distribution of training errors and the right hand graphs shows the distribution of test errors. The abscissa corresponds to the mean squared error and the ordinate represents the percentage of individual results falling within each section of the histogram. Note that the scales change from graph to graph.

segments and the knowledge that a fixed percentage of the samples lie in each segment). However, the trimodal nature of the $K = 15$ training NMSE distribution cannot be detected at this level of division.

5 Conclusions

Publications commonly report the performance of neural networks using the mean and standard deviation of a number of simulations with different starting conditions. Other papers recommend reporting confidence intervals using Gaussian or t-distributions (for a number of runs less than 100) and testing the significance of comparisons using the t-test [11]. However, these assume symmetric distributions. The distribution of results for neural network simulations can vary widely depending on the network architecture, the data, and the training algorithm. Comparisons based on the mean and standard deviation of simulation results can therefore be misleading if the observer assumes the distributions are Gaussian. Alternative means of presenting results can be more informative. For example, it is possible to obtain an indication of how often a particular network and algorithm will produce an acceptable result. In a practical situation, the distribution of results can affect the desirable number of trials, e.g. if the results of multiple trials do not vary greatly, then it may be reasonable to use a smaller number of trials.

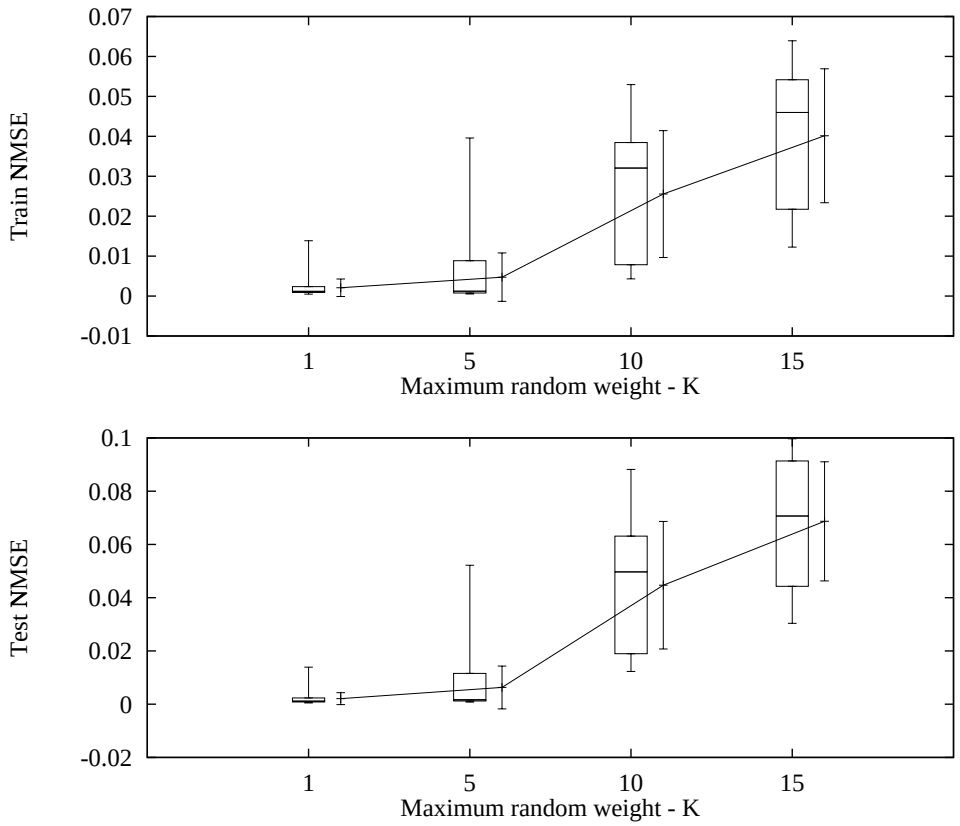

Figure 10. Box-whiskers plots for the artificial task along with the mean and plus/minus one standard deviation.

The possible non-normality of performance distributions adds an element of risk to the training process. The expected or average performance may not be the most important measure of success, e.g. the likelihood that a particular trial meets a given performance criterion depends on the actual distribution of results. In some cases, it may be desirable to trade lower average performance for an increased certainty of obtaining a given performance level. Depending on the distribution of results, a portfolio approach [14] may be used to reduce, for example, the expected time for training to reach a given target performance (in a portfolio approach, multiple networks are trained in parallel and the fraction of time spent training each network can be adjusted in order to optimize a given training criterion, such as the time required for 95% of trials to reach a given target performance).

Our recommendations are:

1. Plot the distribution of results for visual inspection. Distributions can be significantly multimodal and neither the mean plus standard deviation nor box-whisker plots show the complete picture.

2. Use the median, interquartile range, minimum and maximum values as well as the mean and standard deviation for interpreting results. When plotting results, use box-whiskers plots [23].

3. In certain cases it may be possible to approximate a normal distribution by removing outliers. For the case where a relatively small number of trials result in comparatively poor convergence, the practice of removing

Training normalized mean squared error										
K	Mean	Median	Std. Dev.	IQR	Q1	Q3	Min	Max	K-S D	K-S P
1	0.00209	0.00119	0.00220	0.00148	0.000887	0.00237	0.000471	0.0139	0.25	8.3e-6
5	0.00473	0.00121	0.00605	0.00812	0.000751	0.00887	0.000527	0.0396	0.26	1.9e-6
10	0.0255	0.0321	0.01589	0.0306	0.00783	0.0384	0.00430	0.0530	0.17	0.0045
15	0.0401	0.0460	0.0168	0.0324	0.0217	0.0542	0.0123	0.0639	0.19	0.0015

Test normalized mean squared error										
K	Mean	Median	Std. Dev.	IQR	Q1	Q3	Min	Max	K-S D	K-S P
1	0.00209	0.00115	0.00226	0.00151	0.000843	0.00235	0.000453	0.0139	0.25	6.3e-6
5	0.00629	0.00167	0.00804	0.0104	0.00118	0.0115	0.000766	0.0522	0.25	3.3e-6
10	0.0447	0.0497	0.0239	0.0442	0.0190	0.0631	0.01229	0.0882	0.17	0.0071
15	0.0687	0.0707	0.0224	0.0471	0.0443	0.0914	0.03038	0.0997	0.12	0.088

Table 4. Statistics for the distribution of results for the artificial task. Note that as K increases, the median goes from being below the mean to being above the mean.

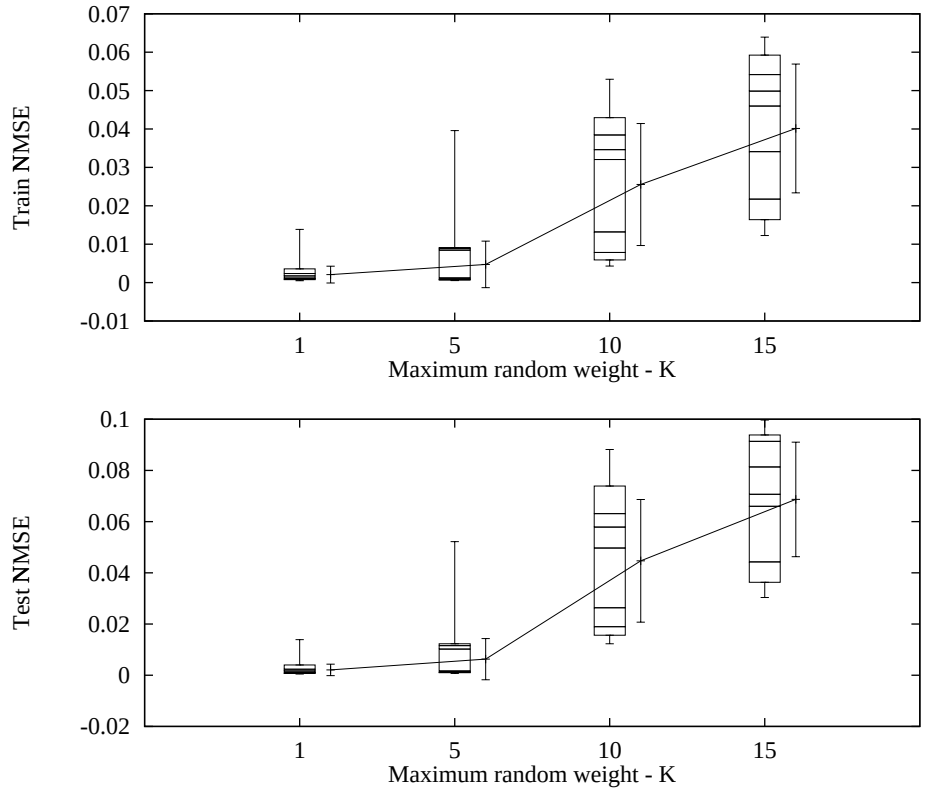

Figure 11. Extended box-whiskers plots for the artificial task along with the mean and plus/minus one standard deviation. In comparison with figure 10, these plots provide a better indication of the actual distribution. For example, the trimodal nature of the $K = 15$ test NMSE plot can be seen here (from the variation in the size of the segments and the knowledge that a fixed percentage of the samples lie in each segment).

those trials from the statistics and reporting the percentage of "failed" trials may be reasonable. Note that we do not know *a priori* what the distribution of results will be for a given application. Therefore, any strategy for identifying and removing outliers should ideally be chosen based on observation of the actual distribution of results.

It may sometimes be difficult to perform enough simulations in order to accurately characterize the distribution of performance within a reasonable time. Therefore it may not always be possible to follow the recommendations given in this paper. In the case of a limited number of runs, the distribution and box-whiskers plots approximate the true distribution, and the number of runs should be clearly stated in order to aid interpretation of the results. For the case of only five simulations, box-whiskers plots reduce to showing all of the five data points. This can still be helpful, e.g. one or more runs may be identifiable as outliers even with such few data points and even one "failed" run can significantly alter the mean and standard deviation thereby presenting a very different picture to a box-whiskers plot.

Acknowledgments

We would like to thank the anonymous reviewers for their useful comments and suggestions. This work has been partially supported by the Australian Research Council (ACT/ADB) and the Australian Telecommunications and Electronics Research Board (SL).

References

[1] P. Auer, M. Herbster, and M.K. Warmuth. Exponentially many local minima for single neurons. In D. Touretzky, M. Mozer, and M. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8. MIT Press, 1996.

[2] A.D. Back. *New Techniques for Nonlinear System Identification: A Rapprochement Between Neural Networks and Linear Systems.* PhD thesis, Department of Electrical Engineering, University of Queensland, 1992.

[3] A.D. Back and A.C. Tsoi. FIR and IIR synapses, a new neural network architecture for time series modeling. *Neural Computation*, 3(3):375–385, 1991.

[4] A.D. Back, E. Wan, Steve Lawrence, and A.C. Tsoi. A unifying view of some training algorithms for multilayer perceptrons with FIR filter synapses. In J. Vlontzos, J. Hwang, and E. Wilson, editors, *Neural Networks for Signal Processing 4*, pages 146–154. IEEE Press, 1995.

[5] A.L. Blum and R.L. Rivest. Training a 3-node neural network is NP-complete. *Neural Networks*, 5(1):117–127, 1992.

[6] M.L. Brady, R. Raghavan, and J. Slawny. Back propagation fails to separate where perceptrons succeed. *IEEE Transactions on Circuits and Systems*, 36:665–674, 1989.

[7] Roger Crane, Charles Fefferman, Scott Markel, and John Pearson. Characterizing neural network error surfaces with a sequential quadratic programming algorithm. In *Machines That Learn*, Snowbird, 1995.

[8] C. Darken and J.E. Moody. Note on learning rate schedules for stochastic optimization. In R.P. Lippmann, J.E. Moody, and D. S. Touretzky, editors, *Advances in Neural Information Processing Systems*, volume 3, pages 832–838. Morgan Kaufmann, San Mateo, CA, 1991.

[9] András Faragó and Gábor Lugosi. Strong universal consistency of neural network classifiers. *IEEE Transactions on Information Theory*, 39(4):1146–1151, 1993.

[10] J. D. Farmer. Chaotic attractors of an infinite-dimensional dynamical system. *Physica*, 4D:366, 1982.

[11] Arthur Flexer. Statistical evaluation of neural network experiments: Minimum requirements and current practice. Technical Report OEFAI-TR-95-16, The Austrian Research Institute for Artificial Intelligence, Schottengasse 3, A-1010 Vienna, Austria, 1995.

[12] T.G. Gonzalez, S. Sahni, and W.R. Franta. An efficient algorithm for the Kolmogorov-Smirnov and Lilliefors tests. *ACM Transactions on Mathematical Software*, 3(1):60–64, 1977.

[13] S. Haykin. *Neural Networks, A Comprehensive Foundation.* Macmillan, New York, NY, 1994.

[14] B.A. Huberman, R.M. Lukose, and T. Hogg. An economics approach to hard computational problems. *Science*, 275:51–54, 1997.

[15] J.S. Judd. On the complexity of loading shallow neural networks. *Journal of Complexity*, 4:177–192, 1988.

[16] J.S. Judd. *Neural Network Design and the Complexity of Learning.* MIT Press, Cambridge, Massachusetts, 1990.

[17] Steve Lawrence, C. Lee Giles, and A.C. Tsoi. What size neural network gives optimal generalization? Convergence properties of backpropagation. Technical Report UMIACS-TR-96-22 and CS-TR-3617, Institute for Advanced Computer Studies, University of Maryland, College Park MD 20742, April 1996.

[18] Steve Lawrence, C. Lee Giles, and A.C. Tsoi. Lessons in neural network training: Overfitting may be harder than expected. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence, AAAI-97*, pages 540–545. AAAI Press, Menlo Park, California, 1997.

[19] J.H. Lin and J.S. Vitter. Complexity results on learning by neural nets. *Machine Learning*, 6:211–230, 1991.

[20] M.C. Mackey and L. Glass. Oscillation and chaos in physiological control systems. *Science*, 197:287, 1977.

[21] A. Papoulis. *Probability, Random Variables, and Stochastic Processes.* McGraw-Hill, 3rd edition, 1991.

[22] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery. *Numerical Recipes.* Cambridge University Press, Cambridge, second edition, 1992.

[23] J.W. Tukey. *Exploratory Data Analysis.* Addison-Wesley, Reading, MA, 1977.

[24] M. Verleysen, J.L. Voz, P. Thissen, and J.D. Legat. A statistical neural network for high-dimensional vector classification. In *Proceedings of the IEEE International Conference on Neural Networks, ICNN 95*, Perth, Western Australia, 1995. IEEE.

[25] R. von Mises. *Mathematical Theory of Probability and Statistics.* Academic Press, New York, 1964.

[26] A.S. Weigend and N.A. Gershenfeld. Results of the time series prediction competition at the Santa Fe Institute. In *IEEE International Conference on Neural Networks*, pages 1786–1793. IEEE Press, Piscataway, NJ, 1993.

[27] N.A. Weiss and M.J. Hassett. *Introductory Statistics.* Addison-Wesley, Reading, Massachusetts, second edition, 1987.