Do Backpropagation Trained Neural Networks have Normal Weight Distributions ?

I. Bellido and E. Fiesler

Institut Dalle Molle d'Intelligence Artificielle Perceptive (IDIAP)
Case postale 609, CH-1920 Martigny, Switzerland

Abstract

Although artificial neural networks are employed in an ever growing variety of applications, their inner workings are still viewed as a black box, which is due to the complexity of the non-linear dynamics that govern neural network learning. The key parameters in this learning process are the so called interconnection strengths or weights of the connections between the neurons. Because of the lack of data, mathematical approaches for studying the 'inside' of neural networks have to resort to assumptions like a Normal distribution of the weight values. In order to better understand what goes on inside neural networks, a thorough study of the real probability distribution of the weight values is important. Besides this, knowledge about weight distributions is also a main ingredient for weight reduction schemes enabling the creation of partially connected neural networks and for network capacity calculations. This paper reports on the findings of an extensive empirical study of the distributions of weights in backpropagation neural networks, and tests formally whether the weights of a trained neural network have indeed a Normal distribution.

Keywords: (artificial) neural network, neural computing, neurocomputing, connectionism, interconnection strength distribution, weight distribution, Normal distribution, Gaussian distribution, statistics, neural network complexity, benchmarks, error backpropagation.

1 Introduction

When a large set of numbers is studied, it is usually assumed that its probability distribution is a Normal distribution. This is supported by the *Central Limit theorem*, which states that a large enough set of *independent* random variables tends to be Normally distributed. Although the distribution[1] of the weights in a trained neural network is often assumed to be Normal, the *Central Limit theorem* may not be valid for this case because of the mutual dependencies of the weights due to the learning process. Since the distribution of the weights in a trained neural network is often used in neural network analysis [Hanson-90] and weight reduction methods [Nowlan-91, Banzhaf-90], it is important to examine how weights are distributed by the backpropagation learning process. For example, Hanson and Burr [Hanson-90] observed "Normalness" with large kurtosis (peakedness) in the three large experiments they conducted.

To test whether trained weights are really Normally distributed, a set of widely known benchmark problems with a range of different topologies and complexities, including real-world problems, were selected: six encoder problems [Rumelhart-86], the three Monks problems [Thrun-91], the sonar identification problem [Gorman-88], a genetic sequences identification of the promoters activity problem [Towel-90], and a Finnish vowel phoneme recognition problem [Kohonen-92].

[1] In this publication, the words *distribution* and *probability distribution* are used interchangeably.

2 Statistics

2.1 Comparison

Several statistical tests are available in order to formally test whether a distribution is a Normal distribution. The best known of these is the χ^2 test, which appeared to be too tough for this purpose, most likely due to the non-clustered nature of the weight data. Two goodness-of-fit techniques have been applied: the modified Anderson-Darling test (A^* test) and the D'Agostino's D test [D'Agostino-86]. The value returned by either of these tests is inversely proportional to the goodness-of-fit of the weight distribution to the Normal distribution. This value, together with the number of elements of the data set determines the *level of significance*[2]. For example, a level of significance of 0.005 means that if the hypothesis is rejected, the probability that the distribution is Normal is less than or equal to 0.5%. In this publication, the hypotheses of Normality are rejected with a fixed level of significance. Non-rejected weight distributions are assumed to correspond to Normal distributions.

In order to provide further information, the third and fourth momentums have also been calculated. These statistics give information about the shape of the distribution with respect to the Normal distribution. If the third moment value is smaller than zero, the distribution is skewed to the right, if the value is larger than zero, the distribution is skewed to the left. If the fourth moment (*kurtosis*) value is larger than three, the distribution is more peaked, and if the value is smaller than three, it is less peaked than the Normal distribution.

2.2 Graphical Representation

The main problem in visualizing a weight distribution is how to transform one-dimensional (density) data into a two dimensional weight distribution graph. One solution is to create a histogram of the distribution, dividing the observed weights into several classes and counting the number of weights whose value lies inside those classes. This set of frequencies can be plotted to show the distribution. This method is sensitive to the parameters chosen; small changes in the width of the classes or in the choice of the centers often causes deviations in the obtained graphs. It also requires a large number of classes to obtain an accurate picture.

A second solution is to create as many classes as there are elements in the data set, and to count how many neighbors are inside the interval $[w_i-d, w_i+d]$, where d is a fixed distance and w_i is a weight. After all the elements of the distribution have been processed, a plot can be generated joining the obtained ordinates by straight lines. One problem of this representation is that in areas where the density is supposed to be zero, like large distances between two consecutive weights, the plotted density is not zero. To avoid this problem, the following heuristic can be applied: if the distance between two weight values is larger than $2d$, two zero density values will be introduced between them, one at w_i+d and the other at $w_{i+1}-d$, assuming that the weights are sorted. The selection of d depends on the number of elements available in the data set. If a very small distance is chosen, a very noisy representation (spikes at each weight) is obtained. On the other hand, if a very large distance is chosen, the representation obtained approaches a uniform distribution.

3 The Benchmarks

Even though the field of Neural Networks has surged, the number of available benchmarks is surprisingly small. A substantial subset of these is used in this article and listed in table 1.

[2]The probability of rejection of the hypothesis by the test, when the hypothesis is actually true.

Several key characteristics of these problems are also listed: the topology used (the number of layers and the number of units per layer), the number of weights used, and the way in which the inputs and outputs are encoded. The topology notation expresses the number of units per layer from input to output layers (left to right). W is the number of weights of the network (including biases) and P is the *number of input-output pairs used for training.* Biases are considered here as weights of connections from a constant unit, as is customary in backpropagation implementations. The normalization used for the Finnish vowel recognition problem consists of dividing each value of the input pattern by the largest absolute value in the pattern. This preserves the relations between these values.

Benchmark	Data Types		Topology	W	P
Name	Input	Output			
Monks-1	Binary	Binary	17-3-1	58	124
Monks-2	Binary	Binary	17-2-1	39	169
Monks-3	Binary	Binary	17-3-1	58	122
8-bits Encoder	Binary	Binary	8-3-8	59	10
16-bits Encoder	Binary	Binary	16-4-16	148	18
16-bits Encoder	Binary	Binary	16-8-16	280	18
8-bits-Inverse Encoder	Binary	Binary	8-3-8	59	10
16-bits-Inverse Encoder	Binary	Binary	16-4-16	148	18
16-bits-Inverse Encoder	Binary	Binary	16-8-16	280	18
Sonar Signals Classification	Real	Binary	60-15-1	931	104
Gene promoters	Binary	Binary	228-1	229	106
Finnish vowels	Normalized Real	Binary	20-50-5	1305	300

Table 1: Benchmarks used.

4 The Results

The above benchmark problems listed in table 1 have been implemented, their weight distributions have been plotted, and goodness-of-fit techniques have been applied to them. Graphical results show differences in weight distributions depending on the number of weights of the network and the desired training accuracy. For most experiments the same convergence criterion has been used: a maximum error ($\epsilon = \triangle|t - o|$) of 0.1 for all the output neurons (this implies a large accuracy in learning that can lead to poor generalization). Only in the Finnish vowel recognition problem a maximum error (ϵ) of 0.4 is used. Each problem has been evaluated repeatedly with different initial conditions (weight values) to obtain different experiments.

The way the A^* and the D tests have been used (see section 2.1) implies the choice of a level of significance. The level of significance selected for this study is 0.005 for both tests, which implies a very small number of invalid rejections and a very generous acceptance criterion. Table 2 shows the percentage of rejected experiments for each benchmark. It also shows the percentage of experiments that were rejected by both tests. The values in the table for the skewness (third momentum) represent the percentage of cases for which the distribution is deviated to the left or to the right, and the values for kurtosis (fourth momentum) represent the percentage of the experiments with larger or smaller kurtosis (peakedness) as compared to the Normal distribution.

It can be observed from the table that for all the benchmarks but three, the A^* test rejects more than the 60% of the cases, and often more than 95%. The D test rejects a slightly smaller percentage of cases. The differences between the tests used may be caused

	Number of	Rejected by			Skewness		Kurtosis	
	Experim.	A^* test	D test	Both	Left	Right	More	Less
Monks-1	42	95.24	97.62	95.24	57.14	42.86	100.00	0.00
Monks-2	40	62.50	30.00	30.00	52.50	47.50	97.50	2.50
Monks-3	42	14.29	23.81	11.90	59.52	40.48	97.62	2.38
E. 8-3-8	48	97.92	0.00	0.00	0.00	100.00	93.75	6.25
E. 16-4-16	48	33.33	22.92	6.25	2.08	97.92	0.00	100.00
E. 16-8-16	48	97.92	95.83	95.83	12.50	87.50	4.17	95.83
I. E. 8-3-8	48	97.92	12.50	12.50	100.00	0.00	93.75	6.25
I. E. 16-4-16	48	60.42	20.83	18.75	100.00	0.00	2.08	97.92
I. E. 16-8-16	48	97.92	95.83	95.83	37.50	62.50	4.17	95.83
Sonar	44	100.00	100.00	100.00	18.18	81.82	100.00	0.00
Gene promoters	14	7.14	14.29	7.14	78.57	21.43	14.29	85.71
Finnish vowels	10	100.00	100.00	100.00	0.00	100.00	100.00	0.00

In the left column: E. means Encoder problem and I.E. Inverse Encoder problem.

Table 2: Results obtained for each benchmark.

by the differences in the power of each test. The A^* test is one of the most powerful tests for Normality, which implies that the rejections are actually valid with large probability. For the sonar identification and Finnish vowel problems the results are unanimous: the distribution of the weights is not Normal. Following the A^* results, only one benchmark, the gene promoters problem, passes this Normality test for more than 90% of the cases. This is coherent with the D test result. It has to be noted that the gene promoters problem is the only problem solved with a network without hidden layers.

Figure 1 shows an example of the distribution of weights in a monks-2 problem compared to its corresponding Normal distribution. Two large maxima can be observed in this case, which was reported as non-Normal by both the A^* and the D tests. Figure 2 corresponds to a distribution of weights for the sonar signal identification problem. The kurtosis given by the fourth momentum test is clearly seen. The third figure corresponds to the gene promoters problem, which yielded minimal rejection ratios.

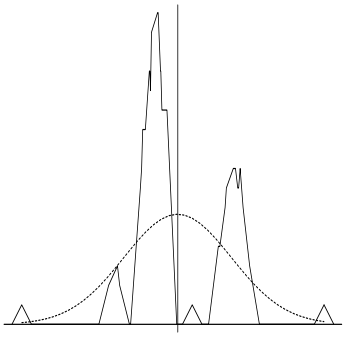

Figure 1: Weight distribution and corresponding Normal distribution for the Monks-2 problem.

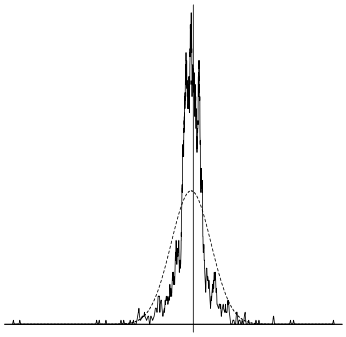

Figure 2: Weight distribution and corresponding Normal distribution for the Sonar problem.

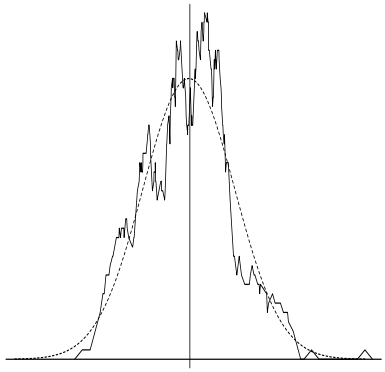

Figure 3: Weight distribution and corresponding Normal distribution for the Gene promoters problem.

5 Conclusions

Even while using a very small probability of invalid rejections (0.005), the majority of the 480 weight distributions were labeled not Normal by either of the statistical tests used. The results also show a strong problem dependency for the weight distributions.

References

[Banzhaf-90] W. Banzhaf, T. Ishii, S. Nara, and T. Nakayama, "A Sparsely Connected Asymmetric Neural Network and its Possible Application to the Processing of Transient Spatio-Temporal Signals", *Proceedings of the International Neural Network Conference (INNC) 90*, volume 2, pages 1005–1008, Kluwer Academic Publishers, 1990.

[D'Agostino-86] R. B. D'Agostino, "Tests for the Normal Distribution", in R. B. D'Agostino and M. A. Stephens Editors *Goodness of Fit Techniques*, chapter 9, Marcel Decker Inc., 1986.

[Gorman-88] R. P. Gorman and T. J. Sejnowsky, "Analysis of Hidden Units in a Layered Network Trained to Classify Sonar Targets", *Neural Networks*, volume 1, pages 75–89, 1988.

[Hanson-90] S. J. Hanson and D. J. Burr, "What Connectionists Models Learn: Learning and Representation in Connectionists Networks", *Behavioral and Brain Sciences*, volume 13, number 3, pages 471–518, 1990.

[Kohonen-92] T. Kohonen, J. Kangas, J. Laaksonen, and K. Torkkola, "LVQ_PAK: A program package for the correct application of Learning Vector Quantization algorithms", *Proceedings of the International Joint Conference on Neural Networks*, volume I, pages 725–730, Baltimore, June 1992.

[Nowlan-91] S. J. Nowlan and G. E. Hinton, "Simplifying Neural Networks by Soft Weight-Sharing", *Neural Computation*, volume 4, number 4, July 1992.

[Rumelhart-86] D. E. Rumelhart and J. L. McClelland, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume 1, chapter 8, pages 318-362, MIT Press, Cambridge, MA, 1986.

[Towel-90] G. G. Towel, J. W. Shavlik, and M. O. Noordewier, "Refinement of Approximate Correct Domain Theories by Knowledge-Based Neural Networks", *Proceedings of the AAAI'90*, pages 861–866, AAAI Press, 1990.

[Thrun-91] S. B. Thrun et al., "The MONK's Problems: A Performance Comparison of Different Learning Algorithms", Carnegie Mellon University Technical Report CMU-CS-91-197, December 1991.