Ultrametric Structure in Autoencoder Error Surfaces

Marcus Gallagher, Tom Downs and Ian Wood
Department of Computer Science and Electrical Engineering
University of Queensland
St Lucia Qld 4072, Australia
{*marcusg,td,wood*}*@csee.uq.edu.au*

Abstract

We use sampling methods to analyse the "apparent minima" of the error surfaces of feedforward neural networks learning encoder problems. First and second-order statistics of a sample of these points of attraction are shown to provide qualitative statistical information about the structure of the error surface, allowing a simple description of this structure. Following methods previously used in the analysis of other complex configuration spaces (such as spin glass models and several combinatorial optimization problems), the third-order statistics of the points of attraction are examined and found to be arranged in a highly *ultrametric* way, using the normal Euclidean distance measure. The implications of this result are discussed.

1 Introduction

1.1 Learning as Error Surface Optimization

The error surface is essentially a cost function of the type that arises in general multivariate optimization problems. Given a vector of weights in a feedforward neural network or Multi-Layer Perceptron (MLP), the task of learning a set of training patterns is to find the weight vector $\mathbf{w}^* = (w_1, \ldots, w_n)$ which minimizes some given cost or error function $E(\mathbf{w})$

$$\mathbf{w}^* = \arg\min \mathrm{E}(\mathbf{w}).$$

Finding $\mathbf{w}^*$ can be viewed as searching an error surface sitting above weight space for a minimum point, the height of which is determined by $E(\mathbf{w})$. Formulating the training problem in this way is quite general: the network is simply treated as a black box (mapping input vectors to output vectors) with N adjustable parameters, allowing for arbitrary error surfaces. Normally specific choices are made regarding the cost function, the model (in this case an MLP of fixed topology) and the training set to be used. Such choices limit the kinds of possible error surfaces formed and this is one important reason for exploring error surface structure.

From a more practical viewpoint, any training algorithm which is able to utilize information about the structure of the error surface, either explicitly or

implicitly (through knowledge or manipulation of the network topology, error function or training set) has the potential to outperform a "blind" algorithm: "any algorithm performs only as well as the knowledge concerning the cost function put into the cost algorithm" [10] (see [7] for an example).

1.2 Exploring the Structure of the Error Surface

In general, the very high dimensionality of the error surface makes its investigation a difficult task. For problems involving small networks or small training sets (e.g, XOR), the error surface can be studied by visualization methods such as plotting different two dimensional "slices" of the surface, or analytically [3]. These methods become impractical for larger networks/training sets.

Despite this, some limited results concerning error surfaces exist. For example, if the training patterns are linearly separable, then an MLP error surface has a unique minimum under some loose assumptions [2]. It is also well known that error surfaces contain a large degree of redundancy, due to symmetry [1]. Permuting hidden units in the same layer (including all ingoing and outgoing connected weights), as well as flipping the sign of each weight connected to a hidden unit (for an odd activation function), leaves the input-output function of the network unchanged, meaning that for every point on the error surface there are $M!2^M$ equivalent points, where M is the number of hidden layer units.

An interesting feature of the error surface concerns its "local minima". Since it is impractical to locate minima precisely on a continuous surface, it is normal practice to conduct repeated runs of backpropagation with a small learning rate, to obtain points which are "close" to minima after significant numbers of training epochs. Note that it is often suggested that MLP error surfaces contain a number of wide, flat areas rather than true critical points of the cost function - we make no distinctions between them here. To make this clear we refer to points collected at the *end* of training runs as apparent minima (AM).

2 Statistical Properties of Error Surfaces

2.1 Encoder Networks

The (auto)encoder problem is a simple problem which is commonly applied to neural network architectures. For an N-input/output encoder, $log_2(N)$ hidden units are required to perform a binary encoding. When the hidden units use intermediate activation levels, an $N-2-N$ encoder can be constructed that can solve the encoder problem for arbitrarily large N [4], although backpropagation has great difficulty in finding such a solution for $N > 8$ [5]. The encoder problem is convenient because it can be scaled to any desired size, and the difficulty of the problem can be somewhat controlled.

In this paper we examine the error surfaces of encoder networks for $N = 4$ and $N = 8$. In each case the number of hidden units is varied between 1 and N. All experiments were conducted using an MLP with a single hidden layer, bias inputs for the hidden and output layers and tanh() activation functions on

hidden and output units. In the training set, desired output values of $[\pm 0.9]$ were used, to avoid units being forced to saturation. Standard backpropagation was used with no momentum and learning rate $\eta = 0.1$. At the end of each training run (30000 epochs), all weight vectors were transformed to lie within a unique wedge of weight space [1], to remove the permutation and sign-flip symmetries of the error surface. Data samples consisted of 1000 points.

2.2 Error Histograms and Pairwise Distances

Given a sample of AM and their corresponding error values, the cumulative frequency distribution of the different error values in the sample can be examined (results for several encoder networks are shown in Fig. 1). The first prominent feature for many of the networks was the step-like nature of the curves, indicating that a small number of error values often dominate a sample. Secondly, the curves shift upwards and to the left as the number of hidden units increases, indicating an increasing chance of an AM being an (increasingly) good solution ($E \approx 0$). To examine how the AM are distributed on the error surface, we cal-

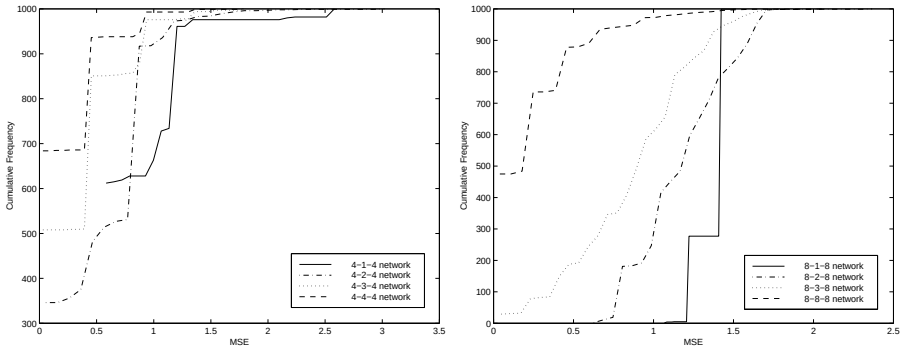

Figure 1: Error distributions for AM samples.

culate the probability distribution $P(q)$ for the distances between two randomly selected points, where $q = d(\mathbf{w_a}, \mathbf{w_b})$ is the Euclidean distance between points $\mathbf{w_a}$ and $\mathbf{w_b}$. Fig. 2 shows typical $P(q)$ histograms for samples of AM. Many of these distributions were skewed to the left (esp. for the 4-1-4, 8-1-8, 8-2-8 and 8-3-8 networks). This suggests that a degree of clustering is present in the AM. Further examination is required to determine the nature of the clustering (e.g, the number of clusters). One way of visualizing the AM is using Principal Component Analysis (see next section).

3 Ultrametricity

It is known that under certain conditions, the spin glass models of statistical physics, combinatorial optimization problems and other systems exhibit the

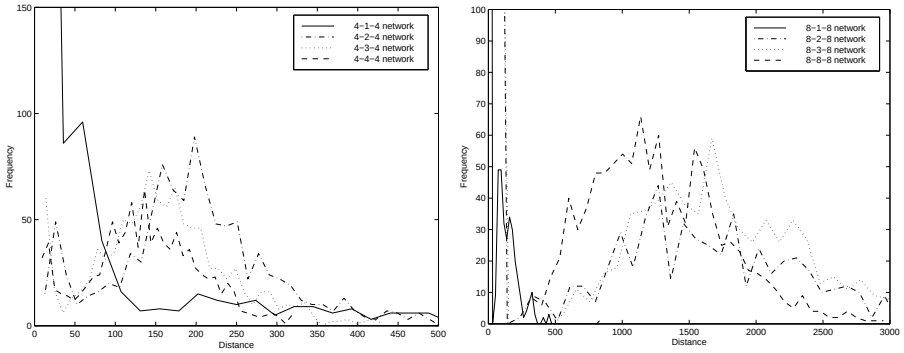

Figure 2: Distribution of distances between AM.

phenomenon of ultrametricity (see [8] for a review). In general, a distance in a metric space obeys the triangular inequality

$$d(\mathbf{w_a}, \mathbf{w_c}) \leq d(\mathbf{w_a}, \mathbf{w_b}) + d(\mathbf{w_b}, \mathbf{w_c})$$

whereas an ultrametric space is endowed with an (ultrametric) distance measure satisfying a stronger inequality

$$d(\mathbf{w_a}, \mathbf{w_c}) \leq Max\{d(\mathbf{w_a}, \mathbf{w_b}), d(\mathbf{w_b}, \mathbf{w_c})\}.$$

For any three points in such a space, the two of them that are nearer to each other are equidistant from the third. This means that all triangles in an ultrametric space must be either equilateral, or isosceles with a small base (third side shorter than the two equal ones). It is known that configuration spaces with ultrametrically distributed minima are quasi-fractal, and empirical evidence suggests that simulated annealing can work well in such spaces [6],[9]. Other algorithms might also be developed to make use of this structural information.

Given a sample of points in a configuration space, the degree of ultrametricity can be estimated using a correlation function of distances between sample points in the configuration space [9], which uses the two longest sides of a sample of triangles randomly generated from the data points. Having no knowledge of the distribution of the data, we use the distribution-free rank correlation coefficient S

$$S = 1 - \frac{6 \sum_{i=1}^{k} d_i^2}{k(k^2 - 1)}$$

where k denotes the number of points in the sample, and d_i is the difference between the ranks of the ith pair of longest sides. Fig. 3 shows values of S for several of the AM samples as a function of the number of epochs. While the random starting points return a small value of S, as training progresses the distribution of the points on the error surface becomes highly ultrametric.

To partially visualize this structure present in the AM, we use Principal Component Analysis and plot the AM in the first three principal components

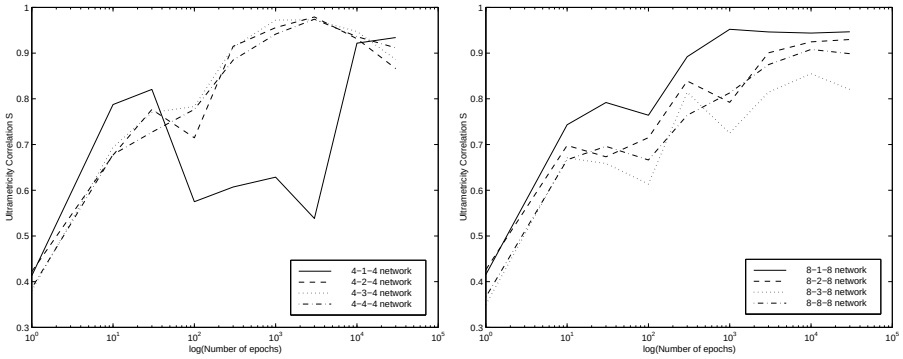

Figure 3: Degree of ultrametricity in samples of AM as a function of the number of epochs.

of the sampled data. Fig. 4 shows an example of such a plot for a 4-1-4 network, with points collected after 30000 epochs ($S = 0.93$). In this case 83% of the distance information on the error surface is captured by the first three principal components.

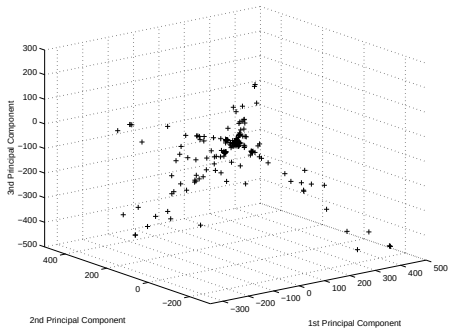

Figure 4: A visualization of how the AM are organized in weight space, using the first three principal components, for the 4-1-4 encoder.

4 Discussion

From the above, a simple description of the error surfaces of autoencoders is possible.

The fact that AM often corresponded to only a few error values suggests that they are either very tightly clustered (into a small number of clusters), or that these particular values are present in many places about the error surface. When the pairwise distribution is dominated by distances near zero, the former is true (esp. in the 4-1-4, 8-1-8, and 8-2-8 encoders). However, for many of the other encoders, a wide range of distances between points is shown, meaning AM are scattered over the error surface. A staircase-like error surface is consistent with these observations.

The high degree of ultrametric structure detected in AM is quite unexpected. Previously [8],[9], the configuration spaces which have displayed this structure were defined on discrete spaces (together with an appropriate distance measure), and are thus quite different to our Euclidean setting. Fig. 3 shows that the degree of ultrametricity increases rapidly from the start of training runs with the number of epochs. This indicates that the paths followed by smooth gradient descent are ultrametrically distributed, not just the AM obtained after a large number of epochs.

5 Summary

This paper shows how the structure of the error surface can be explored and its statistical properties measured. The results show that ultrametricity is embedded in the error surface of encoder networks. Examining the effectiveness of simulated annealing for searching such a surface [6], and designing an algorithm which can utilize other available information about the error surface (e.g, gradient, higher-order statistical information) are interesting areas for future work.

References

[1] A. M. Chen, H. Lu, and R. Hecht-Nielsen. On the geometry of feedforward neural network error surfaces. *Neural Computation*, 5(6):910–927, 1993.

[2] P. Frasconi, M. Gori, and A. Tesi. Successes and failures of backpropagation: a theoretical investigation. In O. Omidvar and C. L. Wilson, editors, *Progress in Neural Networks*. Ablex Publishing, 1993.

[3] Leonard G. C. Hamey. The structure of neural network error surfaces. In *Proc. Sixth Australian Conference on Neural Networks*, pages 197–200, Sydney, 1995.

[4] L. Kruglyak. How to solve the N bit encoder problem with just two hidden units. *Neural Computation*, 2(4):399–401, 1990.

[5] Raymond Lister. Visualizing weight dynamics in the N-2N encoder. In *IEEE International Conference on Neural Networks*, volume 2, pages 684–689, 1993.

[6] Raymond Lister. Fractal strategies for neural network scaling. In Michael A. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*, pages 403–405. MIT Press, 1995.

[7] R. Parisi, E. Di Claudio, and G. Orlandi. A generalized learning paradigm exploiting the structure of feedforward neural networks. *IEEE Transactions on Neural Networks*, 7(6):1450–1459, 1996.

[8] R. Rammal, G. Toulouse, and M. A. Virasoro. Ultrametricity for physicists. *Reviews of Modern Physics*, 58(3):765–788, 1986.

[9] Sara A. Solla, Gregory B. Sorkin, and Steve R. White. Configuration space analysis for optimization problems. In E. Bienenstock et. al., editor, *Disordered Systems and Biological Organization, NATO ASI Series*, volume F20, pages 283–293, 1986.

[10] David H. Wolpert and William G. Macready. No free lunch theorems for search. Technical Report SFI-TR-95-02-010, Santa Fe Institute, February 1995.