UNLIMITED

RSRE
MEMORANDUM No. 4341

ROYAL SIGNALS & RADAR ESTABLISHMENT

A SUM RULE SATISFIED BY OPTIMISED FEED-FORWARD LAYERED NETWORKS

Authors: D S Broomhead, D Lowe & A R Webb

PROCUREMENT EXECUTIVE,
MINISTRY OF DEFENCE,
R S R E MALVERN,
WORCS.

RSRE MEMORANDUM No. 4341

CONDITIONS OF RELEASE

0062722 BR-112835

********** DRIC U

********** DRIC Y

Reports quoted are not necessarily available to members of the public or to commercial organisations.

Royal Signals and Radar Establishment
Memorandum 4341

A sum rule satisfied by optimised feed–forward layered networks.

D.S. Broomhead, David Lowe and A.R. Webb

24^{th} January 1989.

Abstract

Take a feed–forward layered network (such as a multilayer perceptron or a radial basis function network) which is to operate as a pattern classifier. The network may have several hidden layers, as many nodes as required and any desired nonlinearities on the hidden units. The transfer functions of the output nodes should be linear. If the network is trained (using any appropriate problem) to minimise the sum squared error over all outputs and patterns such that the output weights have minimum norm, then the output values of the trained network for any subsequent input pattern will sum to a constant.

This note addresses the following interesting observation:

> *Take a feed–forward layered network (such as a multilayer perceptron [1, 2] or a radial basis function network [3]) which is to operate as a pattern classifier. The network may have several hidden layers, as many nodes as required and any desired nonlinearities on the hidden units. The transfer functions of the output nodes should be linear. If the network is trained (using any appropriate problem) to minimise the sum squared error over all outputs and patterns such that the output weights have minimum norm, then the output values of the trained network for* any *subsequent input pattern will sum to a constant.*

A particular example of this behaviour occurs if the target coding scheme for this classifier were chosen as

$$t_k^p = \begin{cases} 1 \; if \; pattern \; p \in \; class \; k \\ 0 \qquad otherwise \end{cases} \tag{1}$$

In this case the output values for any input pattern sum to unity. This case has a special interest since it relates to the idea from pattern recognition that the output of networks operating as classifiers ought to reflect the *likelihood* of a given pattern belonging to a particular class [4, 5]. Unfortunately, to achieve this sum rule it is generally true that some of the output components have to assume negative values, thus invalidating any interpretation in terms of classical probability theory. That such a sum rule should hold *independent of the inputs to the network* is surprising and to our knowledge has not been pointed out previously in the network literature. A possible reason for this is that numerical simulations involving the most common feed–forward network, the multilayer perceptron, do not often use linear output units (except in autoassociative encoding applications [6]). Moreover, even where linear output units are used, training is often terminated after a finite number of iterations, or when the actual outputs are sufficiently close to saturation. A prerequisite for the sum rule to be observed is that the network weights in the final layer have to be optimum, in the sense that a local minimum of the sum squared error is obtained.

The rest of this note is devoted to a proof of the observed sum rule.

The network is trained on P input patterns,

$$|I^p\rangle \in \mathbb{R}^n, \mathrm{p} = 1, 2, \ldots, \mathrm{P}$$

where each pattern is represented by a real valued vector in n dimensions (an extension to complex valued patterns may be made). To each training pattern there corresponds a desired target pattern,

$$|T^p\rangle \in \mathbb{R}^{n'}, \mathrm{p} = 1, 2, \ldots, \mathrm{P}$$

which is a vector in n'–dimensional space where each dimension represents one of the n' possible classes that the input pattern could belong to. The intention is to attempt to produce *actual* outputs from the network, $|O^p\rangle$ which are as close as possible in a least mean squares sense to the *desired* target patterns by choosing an appropriate set of values for the adjustable parameters in the network (i.e. biases and link weights). The output of the network is a weighted sum of the output pattern of the final 'hidden' layer, which in turn is usually a weighted sum of the pattern of the previous hidden layer passed through nonlinear transfer functions. For instance, for a network with a single hidden layer consisting of n_0 nodes, the output of the k–th output node of the network for pattern p may be expressed as

$$O_k = \sum_{j=1}^{n_0} \lambda_{jk} \phi_j^p + \lambda_{0k} \tag{2}$$

where λ_{jk} is the weight value linking the j–th hidden node to the k–th output node, ϕ_j^p is the output of the j–th hidden node for the p–th pattern and λ_{0k} is a constant 'bias' value associated with the k–th output node.

In general the nonlinear transfer function of the j–th hidden node takes the form $\phi_j^p = \phi_j(g_j[|I^p\rangle])$ where $g_j[|I^p\rangle]$ is a (scalar) function of the input patterns. The form of ϕ_j and g_j differ depending on the network being used. Such differences are not, however, important to this discussion. Whatever the chosen form of the nonlinear transformation, we are only concerned with the action of the network as described by equation (2). The method of obtaining an appropriate set of weight values is to minimise the error between the desired target patterns and the actual network outputs. The network output for pattern p may be expressed, conveniently, in vector notation as

$$|O^p\rangle = \mathbf{\Lambda}|\phi^p\rangle + |\lambda_0\rangle \tag{3}$$

$\mathbf{\Lambda}$ is an $n' \times n_0$ array of weight values between the n_0 hidden units and the n' output units

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_{11} & \cdots & \lambda_{n_0 1} \\ \vdots & \ddots & \vdots \\ \lambda_{1n'} & \cdots & \lambda_{n_0 n'} \end{bmatrix} \tag{4}$$

$|\lambda_0\rangle$ is the n'–dimensional column vector of bias values associated with the output units and $|\phi^p\rangle$ is the n_0–dimensional column vector of output values of the hidden units.

For P patterns the output matrix may be expressed as

$$\mathbf{O} = |\lambda_0\rangle\langle 1| + \mathbf{\Lambda H} \tag{5}$$

where $\langle 1|$ is a P–dimensional row vector of 1's, $\mathbf{H}$ is the $n_0 \times P$ array of hidden unit outputs

$$\mathbf{H} = \begin{bmatrix} \phi_1^1 & \cdots & \phi_1^P \\ \vdots & \ddots & \vdots \\ \phi_{n_0}^1 & \cdots & \phi_{n_0}^P \end{bmatrix} \tag{6}$$

and $\mathbf{O}$ is a matrix having the output vectors $|O^p\rangle$ as columns.

The adjustable parameters (contained explicitly in $|\lambda_0\rangle$ and $\mathbf{\Lambda}$ and implicitly in $\mathbf{H}$) are determined by minimising the sum squared error at the network output

$$E = \sum_{p=1}^{P} || \ |T^p\rangle - |O^p\rangle \ ||^2 \tag{7}$$

where $||\ldots||$ denotes the Euclidean vector norm. Differentiating this expression with respect to the bias weights, $|\lambda_0\rangle$ and equating to zero gives the optimum choice of values for the bias vector as

$$|\lambda_0\rangle = |\overline{T}\rangle - \mathbf{\Lambda}|m^H\rangle \tag{8}$$

where $|\overline{T}\rangle = \mathbf{T}|1\rangle/P$ is the mean target vector over all training patterns, $\mathbf{T}$ is a matrix having target vectors $|T^p\rangle$ as columns, and $|m^H\rangle$ is the mean pattern vector over all the patterns in the training set produced at the output of the final hidden layer of nodes.

It is clear from equation (8) that the rôle of the bias vector is to compensate for the difference between the mean of the desired target vectors and the mean of the actual outputs over the training patterns. Inserting this expression for the bias into the error expression gives

$$E = ||\hat{\mathbf{T}} - \mathbf{\Lambda}\hat{\mathbf{H}}||_F^2 \tag{9}$$

where $||\ldots||_F$ is the Frobenius matrix norm, $\hat{\mathbf{T}} = \mathbf{T} - |\overline{T}\rangle\langle 1|$ is the mean–shifted matrix of targets and $\hat{\mathbf{H}} = \mathbf{H} - |m^H\rangle\langle 1|$ is the mean–shifted matrix of outputs at the final hidden layer.

One solution for $\mathbf{A}$ which minimises the above error and gives a solution with minimum (Frobenius) norm is given by

$$\mathbf{A} = \hat{\mathbf{T}}\hat{\mathbf{H}}^{+} \tag{10}$$

where $\hat{\mathbf{H}}^{+}$ is the $P \times n_0$ Moore–Penrose pseudo–inverse of $\hat{\mathbf{H}}$ [9].

Thus using these optimum values of the weight matrix and bias vector, the expression for the general output $|O\rangle$ of the trained network for an input pattern giving rise to the pattern vector $|h\rangle$ at the output of the final hidden layer is

$$|O\rangle = |\overline{T}\rangle + \hat{\mathbf{T}}\hat{\mathbf{H}}^{+}\left(|h\rangle - |m^{H}\rangle\right) \tag{11}$$

We can sum the outputs by multiplying on the left by $\langle 1|$, thus

$$\text{Sum over outputs } = \langle 1|O\rangle = \langle 1|\overline{T}\rangle + \langle 1|\hat{\mathbf{T}}\hat{\mathbf{H}}^{+}\left(|h\rangle - |m^{H}\rangle\right) \tag{12}$$

However, consider the bra vector $\langle 1|\hat{\mathbf{T}}$ in the situation when the sum of each column of $\mathbf{T}$ is a constant ($= t$ say);

$$\begin{aligned}
\langle 1|\hat{\mathbf{T}} &= \langle 1|\mathbf{T} - \langle 1|\overline{T}\rangle\langle 1| \\
&= \langle 1|\mathbf{T} - \frac{\langle 1|\mathbf{T}|1\rangle\langle 1|}{P} \\
&= t\langle 1| - t\frac{\langle 1|1\rangle\langle 1|}{P} \\
&= \langle 0|
\end{aligned} \tag{13}$$

since $\langle 1|1\rangle = P$.

Therefore, the sum over the outputs of the trained, optimised network is given by

$$\begin{aligned}
&\text{Sum over outputs } = \\
&\langle 1|O\rangle = \langle 1|\overline{T}\rangle \\
&\qquad = \text{Sum of mean target vector.}
\end{aligned} \tag{14}$$

This completes the proof of the original observation. In particular, for a one–from–n' coding scheme where the components of the target vector belong to $\{1, 0\}$, the sum of the outputs of the network for *any* input sum to unity.

However, it is now apparent that a more general result holds:

Theorem.

Consider a network having linear output units. Let the weights associated with the connections to these units be determined by linear minimum norm least squares optimisation. Then if there exists an arbitrary linear constraint of the form

$$\langle u|T^p\rangle = \langle u|\overline{T}\rangle \ \forall\ p = 1, 2, \ldots, P$$

with $\langle u|$ a constant vector, then the general output of the network $|O\rangle$ satisfies:

$$\langle u|O\rangle = \langle u|\overline{T}\rangle$$

Proof

The general output of the network is given by equation 11:

$$|O\rangle = |\overline{T}\rangle + \hat{\mathbf{T}}\hat{\mathbf{H}}^{+}\left(|h\rangle - |m^H\rangle\right)$$

Therefore

$$\langle u|O\rangle = \langle u|\overline{T}\rangle + \langle u\hat{\mathbf{T}}\hat{\mathbf{H}}^{+}\left(|h\rangle - |m^H\rangle\right)$$

But

$$\langle u|\hat{\mathbf{T}} = \langle u|\mathbf{T} - \langle u|\overline{T}\rangle\langle 1|$$

By hypothesis, $\langle u|\mathbf{T} = \langle u|\overline{T}\rangle\langle 1|$. Therefore

$$\langle u|O\rangle = \langle u|\overline{T}\rangle \quad \blacksquare$$

Remark: If the set of target vectors satisfy several linear constraints simultaneously, then so will the general network outputs.

Acknowledgements.

We would like to thank John Bridle for many thought-provoking discussions and in particular for providing a geometrical argument pointing out the confines of the sum rule discussed in this note.

References

[1] Rumelhart, D.E., Hinton, G.E., and Williams, R.J., (1986), Learning internal representation by error propagation. in *Parallel distributed processing: Explorations in the microstructure of cognition*, (Vols, 1 and 2), Cambridge, MA:MIT Press.

[2] Lippmann, Richard P., (1987), An Introduction to Computing with Neural Nets, *IEEE ASSP Magazine, April, 1987*, 4–22.

[3] Broomhead, D.S., Lowe, David,(1988). Multi–variable Functional Interpolation and Adaptive Networks, *Complex Systems*, **2**, No. 3, 269–303.

[4] Eric B. Baum, (1988), "Supervised Learning of Probability Distributions by Neural Networks", Neural Information Processing Systems, Denver, Colorado, 1987, pp 52–61, (American Institute of Physics, NY, 1988).

[5] John. S. Bridle, (1988), "Speech recognition:Statistical and Neural Information Processing Approaches", To appear in *Proceedings IEEE Conf. on Neural Information Processing Systems, Denver, Colorado, 1988*, (American Institute of Physics, NY, 1989).

[6] Bourlard, H., Kamp, Y., (1988), Auto–association by Multilayer Perceptrons and Singular Value Decomposition, *Biological Cybernetics*, **59**, 291–294.

[7] M.J.D. Powell, (1985), "Radial basis functions for multivariable interpolation: A review", *IMA conference on "Algorithms for the Approximation of Functions and Data", RMCS Shrivenham.*

[8] M.J.D. Powell, (1987), "Radial basis function approximations to polynomials", DAMPT preprint 1987 NA/6, (presented at the 1987 Dundee biennial Numerical Analysis Conference).

[9] Golub, G., Kahan, W.,(1965). Calculating the Singular Values and Pseudo–Inverse of a Matrix, *Journal SIAM Numerical Analysis, Series B*, **2**(2), 205–224.

DOCUMENT CONTROL SHEET

Overall security classification of sheet ..... UNCLASSIFIED .....

(As far as possible this sheet should contain only unclassified information. If it is necessary to enter classified information, the box concerned must be marked to indicate the classification, eg (R), (C) or (S))

1. DRIC Reference (if known)	2. Originator's Reference	3. Agency Reference	4. Report Security Classification
	MEMO 4341		UNCLASSIFIED

5. Originator's Code (if known)	6. Originator (Corporate Author) Name and Location
7784000	ROYAL SIGNALS & RADAR ESTABLISHMENT ST ANDREWS ROAD, GREAT MALVERN WORCESTERSHIRE WR14 3PS

5a. Sponsoring Agency's Code (if known)	6a. Sponsoring Agency (Contract Authority) Name and Location

7. Title

A SUM RULE SATISFIED BY OPTIMISED FEED-FORWARD LAYERED NETWORKS

7a. Title in Foreign Language (in the case of Translations)

7b. Presented at (for Conference Papers): Title, Place and Date of Conference

8. Author 1: Surname, Initials	9a. Author 2	9b. Authors 3, 4 . . .	10. Date	pp. ref.
BROOMHEAD D S	LOWE D	WEBB A R	1989.01	6

11. Contract Number	12. Period	13. Project	14. Other Reference

15. Distribution Statement

UNLIMITED

Descriptors (or Keywords)

Continue on separate piece of paper

Abstract

Take a feed-forward layered network (such as a multilayer perceptron or a radial basis function network) which is to operate as a pattern classifier. The network may have several hidden layers, as many nodes as required and any desired nonlinearities on the hidden units. The transfer functions of the output nodes should be linear. If the network is trained (using any appropriate problem) to minimise the sum squared error over all outputs and patterns such that the output weights have minimum norm, then the output values of the trained network for any subsequent input pattern will sum to a constant.