Internal representation in networks of non-monotonic processing units

David B. McCaughan
Department of Computing Science
University of Alberta
dbm@ieee.org

David A. Medler
Center for the Neural Basis of Cognition
Carnegie Mellon University
medler@cnbc.cmu.edu

Michael R.W. Dawson
Department of Psychology
University of Alberta
mike@bcp.psych.ualberta.ca

Abstract

Connectionist networks that use non-monotonic transfer functions tend to adopt highly structured internal representations, revealed as vertical banding in density plots of internal unit activities. Recent work has shown this banding *to be easily analyzed allowing for the extraction of symbolic descriptions of the solution encoded in the network. While the banding phenomenon is well documented, the properties that give rise to this structure have never been formalized. In this paper we detail the geometry that underlies the internal unit activity clustering that banding represents. These results distinguish the operation of non-monotonic units from that of traditional sigmoid devices in terms of the mechanism by which they carve up the input space.*

Introduction

The vast majority of connectionist models make use of an activation function that is monotonic with respect to its net input; that is, the output of a processing unit is directly proportional to its input. The most common examples of this type of unit are those that use a threshold activation function function such as the *sign*:

$$f(\vartheta) = \begin{cases} 1 & \text{if } \vartheta > 0 \\ -1 & \text{otherwise} \end{cases} \quad (1)$$

where ϑ is the net input to the unit, and the sigmoid, for example the logistic:

$$f(\vartheta) = \frac{1}{1 + \exp^{-2\beta\vartheta}}, \quad (2)$$

where β is a gain term.

Recently there has been increasing interest in connectionist models that make use of non-monotonic activation functions such as a Gaussian or sinusoid. These networks tend to adopt remarkably structured internal representations that have been applied to problems of network interpretation and rule extraction. This paper considers the question of what gives rise to this observed structure.

We first introduce the concept of a non-monotonic processing unit and provide an overview this class of network architectures. We then turn to the issue of the internal representations adopted by non-monotonic networks and formalize the geometric properties that underly this structure. These concepts are illustrated with an easily visualized low-dimensional example. Further, we outline a framework for characterizing the classification mechanism of processing units, allowing us to better contrast the mechanism by which the input space is carved up by units using the various activation functions to be discussed. We conclude with a brief discussion of these results and indicate how they relate to current research on non-monotonic processors.

Non-monotonic Processing Units

When considering the behaviour of connectionist processing units there are a number of parameters that are of interest. One fundamental property is the activation function used to map inputs to output.

We can distinguish between processors for which output is proportional to net input, such as those above, from those that are sensitive to only a particular range of values. Ballard refers to the former type of unit as an *integration device*, whereas the latter he terms a *value unit* [1]. In the present context we will refer to any unit for which unit activity is not proportional to its net input as *non-monotonic*.

Dawson and Schopflocher have developed a connectionist architecture based on the concept of a value unit [5]. A *network of value units* is a feed-forward multi-layer perceptron in which processing elements use a Gaussian activation function,

$$f(\vartheta) = e^{-\pi\vartheta^2}. \quad (3)$$

where the net input function ϑ is the Euclidean inner product. The graph of equation 3 is pictured in Figure 1—note that the range of the function is the interval $(0, 1]$, and the maximum value occurs when $\vartheta = 0$.

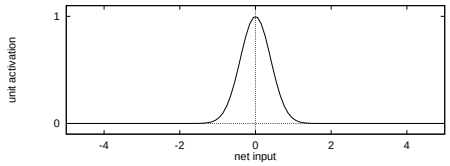

Figure 1. The value unit activation function

This architecture was studied empirically for its suitability as a pattern classification system, and it was demonstrated that for a wide range of problems, the non-monotonic activation function both reduced the time required to learn to solve a given problem and decreased the size of the resulting network. Similar results have been reported for networks that use the periodic activation function,

$$f(\vartheta) = \sin(\vartheta), \tag{4}$$

which is pictured in Figure 2 [7].

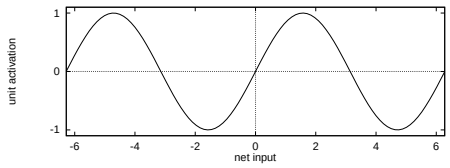

Figure 2. A periodic activation function

Other work in this area has similarly concentrated on empirical comparisons between multi-layer perceptrons whose internal processing units used either a sigmoidal or various non-monotonic activation functions [6, 8]. In any case, little has been done to develop more formal distinctions between these two types of architecture. It is this issue that we turn to now.

Internal Representation

In what follows, we consider the nature of the internal representations as formed by various types of connectionist processing units. We generally develop these results in terms of activation functions that generate outputs in the interval $(0, 1)$; however, most of these concepts generalize to other bounded ranges.

Sigmoid Units

In terms of pattern classification, a sigmoid unit can be thought of as partitioning the input space into two regions. All input patterns for which the net input is below a threshold value produces a low output, often 0 or -1, and all those with a net input that is above a threshold value produces a high output, typically 1. Where the net input is given by the Euclidean inner product (ignoring bias), this partition is defined by a hyperplane through the origin orthogonal to the weight vector as depicted in Figure 3.

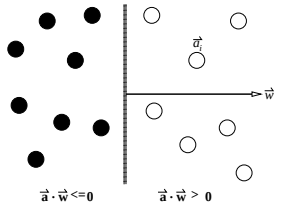

Figure 3. Linear partition induced by sigmoid unit

Banding in Non-monotonic Units

Recent work has studied the highly structured internal representations that the internal units in non-monotonic networks tend to adopt. When the activation values of internal units in trained networks are graphed in density plots (unit activation in response to each pattern in the training set plotted against a random component in order to spread the graph vertically), they tend to cluster into bands. This *banding* has been shown to be easily analyzed allowing for the extraction of symbolic descriptions of the solution encoded in the network [2, 3, 7].

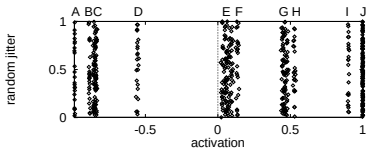

Figure 4. Banding in a non-monotonic unit

While banding has been applied successfully to the problem of extracting symbolic descriptions of network solutions, the properties underlying this phenomenon have not been well understood to this point. The next section turns to this issue, providing a geometric account of the internal structure that is reflected by banding.

The Geometry of Banding

Similar to the decision boundary defined by a sigmoid unit, banding is a consequence of the spatial orientation of the vector of weights into a given processing element with respect to the input space. Since all input vectors that produce the same net input produce the same output, and thus belong to the same band, we initially consider the net input to a unit rather than its transformed output for convenience.

Let $\vec{w}$ be the vector of weights into a unit for which banding is evident and let $\mathcal{A}$ be the set of all input vectors. A band

whose inverse under the activation function corresponds to a net input of ϑ are those that satisfy

$$\{\vec{a} \in \mathcal{A} \mid \vec{a} \cdot \vec{w} = \vartheta\} \tag{5}$$

where net input is computed as the Euclidean inner product. The geometric interpretation of this expression is thus a set of input patterns lying in an affine hyperplane. We can be more specific than this however. Consider the following lemma (E^n denotes the Euclidean n-dimensional space):

Lemma 1 *Given vector $\vec{w} \in E^n$ and constant κ, then $\forall \vec{u} \in E^n$, $\vec{u} \cdot \vec{w} = \kappa \iff \text{proj}_{\vec{w}}\vec{u} = \kappa$*

Proof: The proof of this lemma is a straightforward application of the projection theorem. Let $\vec{w} \neq \vec{0}$ and $\vec{u}$ be vectors such that $\vec{u} \cdot \vec{w} = \kappa$.

$$\vec{u} \cdot \vec{w} = \kappa \tag{6}$$
$$\implies (\text{proj}_{\vec{w}}\vec{u} + (\vec{u} - \text{proj}_{\vec{w}}\vec{u})) \cdot \vec{w} = \kappa \tag{7}$$
$$\implies \text{proj}_{\vec{w}}\vec{u} \cdot \vec{w} = \kappa. \tag{8}$$

The necessity of the lemma follows similarly □.

This is to say that given a reference vector $\vec{w}$, all vectors $\vec{u}$ for which the Euclidean inner product with $\vec{w}$ is equal have the same projection onto $\vec{w}$. Where this reference vector is the weight vector associated with a processing unit, we find that the set of points defined by Equation 5 lie in an affine hyperplane perpendicular to the weight vector, located a distance of $||\text{proj}_{\vec{w}}\vec{u}||$ from the origin. The graphical interpretation of this for E^2 is depicted in Figure 5

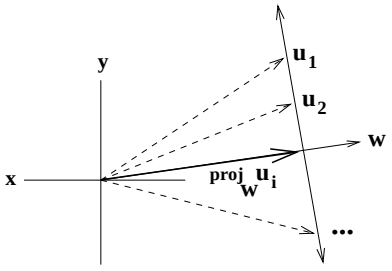

Figure 5. Graphical interpretation of Lemma 1 in E^2

Although this result is derived in terms of net input, we are now in a position to consider what effect a particular activation function might have. For Gaussian activated units, a given band (with the exception of the band at the maximum activation) corresponds to input patterns lying in parallel hyperplanes, perpendicular to the weight vector, equidistant from the origin. This is a result of the symmetric activation function mapping net inputs of equal magnitude but opposite sign to the same output value. Similarly, it should now be clear that for periodic units a band reflects input patterns in an infinite set of parallel hyperplanes perpendicular to the weight vector separated by a distance proportional to the period of the function.

To illustrate these concepts, it is instructive to consider a low dimensional problem for which the network solution is easily visualized in terms of the geometry of the input space.

Example: 3-Parity

Even parity is a problem in which a network must learn to output a 1 when an odd number of inputs are 1, and 0 otherwise. In this particular example, we consider 3-bit parity, so the inputs to the network are three binary values. Geometrically, the input space for the 3-parity problem is a cube in 3-dimensional space making it an ideal problem with which to visualize the network solution.

The solution presented here is a typical one arrived at by a feed-forward *network of value units*, using the Gaussian activation function given by Equation 3. Network architecture consisted of 3 input units, 1 internal unit and 1 output unit, and connectivity existed between all units in adjacent layers. After training, a jittered density plot for the single internal unit is generated, and appears in Figure 6. With only 8 input patterns, there are only 8 data points on the jittered density plot, however there are still clearly 3 bands visible. These bands are summarized in Table 1.

Figure 6. Density plot for 3-parity network

Mean activity	Number of patterns	Output
0.05	1	0
0.50	4	1
1.00	3	0

Table 1. Summary of bands in Figure 6

Figure 7 depicts the input space of the 3-parity problem as a cube in which white vertices indicate patterns that are associated with an output of 1, whereas black vertices are those that produce an output of 0 from the network. The weight vector on the inputs to this unit, $\vec{w} = (-0.47, 0.47, -0.47)$, is also shown. This illustrates how the unit orients on the input space such that related points fall into planes perpendicular to $\vec{w}$.

Plane B is the plane orthogonal to the weight vector, resulting in a net input of 0. This value produces maximum activation from the value unit activation function and the unit outputs a 1 in response to all of these patters corresponding to the band with mean activation 1.00. Plane D is sufficiently distant from the origin to produce a large negative net input and thus a low activity in the unit—this is the lone point in

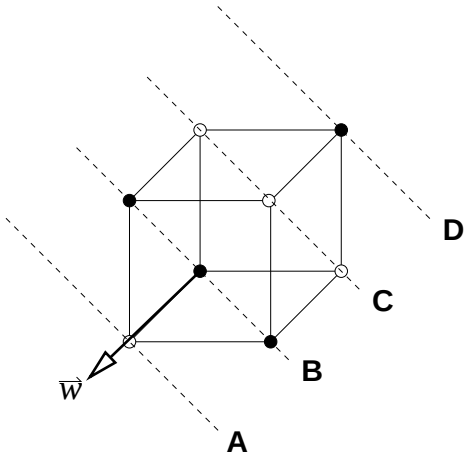

Figure 7. Geometry of 3-parity solution

the band with mean activity of 0.05. Planes A and C are equally distant from the origin, thus produce net inputs of equal magnitude, but opposite sign. Due to the symmetry of the Gaussian activation function, these two planes collapse into the single band with mean activity 0.50.

The correct classification is performed by the output unit with the use of bias to decrease the net input such that the band at 0.5 generates a net input of 0, thus outputting a 1 for these 4 patterns, whereas the patterns producing activities (bands) at 0.05 and 1.0 generate net inputs sufficiently far from 0 to produce near-0 activation in the output unit. This correctly classifies all patterns with a singly unit.

This simple example graphically demonstrates exactly how the geometry of hyperplanes perpendicular to the weight vector into a unit intersecting patterns in the input space reveals itself in the jittered density plots of non-monotonic units. Given this new understanding of internal representation in non-monotonic processing units, we are in a position to develop a framework with which to describe a unit's pattern classification properties in these terms.

A Classification Framework

In this section, we present a framework for describing the manner in which units carve up their input space based on their activation function. This serves to provide a common language with which to visualize and contrast the classification properties of both monotonic and non-monotonic units for which net input is calculated as the Euclidean inner product. This framework is presented primarily to provide the groundwork for future results; however appears here as a natural extension of the results in the preceding section.

For each processing unit, we will exhaustively classify its input space by identifying 1) the pattern or patterns to which the unit is maximally sensitive, which we will term the *trigger feature* or *trigger plane*; 2) the region to which a unit is not sensitive at all (i.e. inputs producing zero, or near zero, activity), which will be termed the *0-region*; and 3) the regions bounded by 1 and 2 in which a unit must impose some representation in order to perform classification tasks as they produce non-zero activities. This last region will be referred to as the *banding region*, as it will contain the patterns that are candidates for organization into hyperplanes that will appear as banding in density plots of unit activities in non-monotonic networks.

In what follows, A is the set of all potential input patterns to a unit, $\vec{w}$ is the vector of weights on inputs into a given unit, f is the activation function of the unit, ϵ is a threshold activation value below which a unit is considered to be inactive, and σ is an allowed deviation from maximum output above which a unit will still be considered maximally active.

Trigger Feature

The input, or set of inputs, to which a sigmoid unit is maximally sensitive has been referred to as a *trigger feature* [4].

Definition 1 *The **trigger feature** with respect to a given unit's weight vector, $\mathcal{T}_{\vec{w}}$, is the pattern constructed such that the maximum input appears where the corresponding connection weight is positive, and the minimum possible input appears where the weight is negative.*

The geometric interpretation of the trigger feature is a point in space. As it is also common to view sigmoid units as being "on" when the net input exceeds some threshold value, and "off" otherwise. This allows us to think of the input space being partitioned into a *0-region* and *1-region*, agreeing with our existing understanding of the linear partitioning performed by sigmoid units. Both of these conventions are noted in Figure 8 to follow.

As was noted in the case of non-monotonic units, the set of patterns producing any given activity lie in a hyperplane. For the specific instance of the set of patterns producing maximum activity we will simply modify the existing terminology and refer to this set as the *trigger plane* for a unit.

Definition 2 *The **trigger plane** with respect to a given unit's weight vector, $\mathcal{T}_{\vec{w}}$, is the set of patterns satisfying $f(\vec{a}\cdot\vec{w}) = max\{f\}$. This corresponds to the patterns lying in the affine hyperplane in Equation 9.*

$$\mathcal{T}_{\vec{w}} = \{\vec{a} \in A \mid \vec{a}\cdot\vec{w} = f^{-1}(max\{f\})\}. \qquad (9)$$

In practice banding is rarely perfectly clean, as slight error in the spatial orientation of the weight vector is possible, in fact possibly desirable, depending on the nature of the problem. We can accommodate this by introducing a small tolerance σ in maximum activation for which patterns will still be considered to lie on the trigger plane. In this case Equation 9 becomes $\mathcal{T}_{\vec{w}} = \{\vec{a} \in A \mid \vec{a} \cdot \vec{w} = f^{-1}(max\{f\} - \sigma)\}$.

Note that in the case of a periodic activation function with recurring maxima, the unit in question induces an infinite number of trigger planes throughout the input space.

Zero Region

Definition 3 *The **zero region** with respect to a given unit's weight vector, $0_{\vec{w}}$, contains those patterns producing activity below a given inactivity threshold value, i.e. $f(\vec{a} \cdot \vec{w}) < \epsilon$. Thus,*

$$0_{\vec{w}} = \{\vec{a} \in A \mid \vec{a} \cdot \vec{w} < f^{-1}(\epsilon)\} \tag{10}$$

Note that while the sigmoid unit has a single zero region, the symmetry of the Gaussian activation function results in two zero regions located toward the two tails of the curve. The notion of a zero region is only meaningful where there is some interval of the range of the activation function which produces negligible activity in response to values from its domain. Since the periodic activation function does not saturate anywhere in its range, within this framework we would say that such units have no zero region.

Banding Region

Definition 4 *The **banding region** with respect to a given unit's weight vector, $0_{\vec{w}}$, contains those patterns that are candidates for banding. This region is bounded by the zero region and the trigger plane, giving*

$$\mathcal{B}_{\vec{w}} = A - (0_{\vec{w}} \cup \mathcal{T}_{\vec{w}}) \tag{11}$$

Given these definitions, this framework does not admit a banding region for sigmoid units. This suitably mirrors the empirical evidence that demonstrates that these types of units do not typically exhibit banding.

Figures 8, 9 and 10 illustrates the concepts that have been discussed in this section, superimposed on the graphs of the activation functions that we have described in this paper. This is particularly useful for visualizing the relative sophistication with which these units are capable of carving up their input space.

The 3-parity example presented earlier can now be recast in this framework as we would now identify plane B in Figure 7 as the *trigger plane*, plane D lies in the unit's *zero region* so produces negligible activity, and since patterns in planes A and C lie in the *banding region*, they would be expected to appear as non-zero bands in a jittered density plot for this unit.

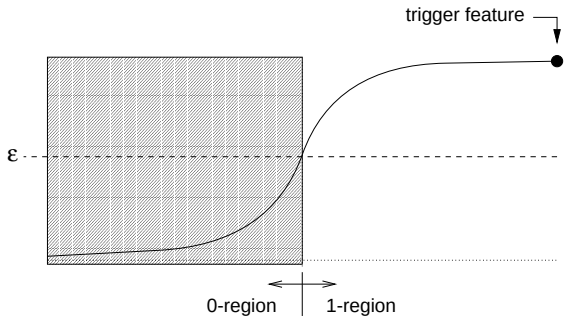

Figure 8. Classification by a sigmoid activation function

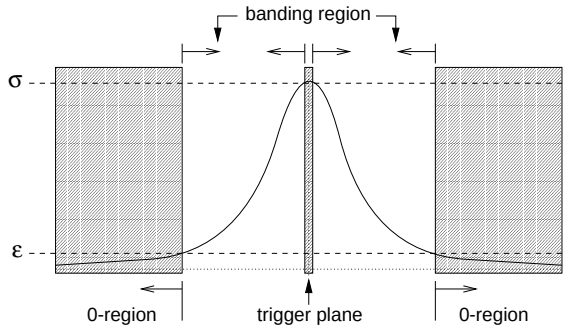

Figure 9. Classification by a Gaussian activation function

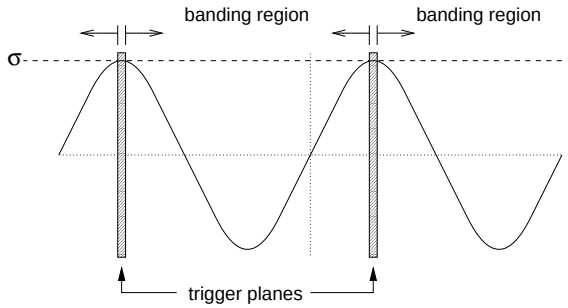

Figure 10. Classification by a periodic activation function

Conclusions

This paper has shown banding to be the result of a non-monotonic processing unit orienting itself to define parallel sets of affine hyperplanes that intersect points in the input space. In contrast, a sigmoid or threshold device partitions the input space into two linearly separable regions. This un-

derlines a fundamental difference in the way these two types of processors compute on their input.

A geometric framework was introduced with which to describe the classification properties of non-monotonic processing units, contrasting this with that of the well known sigmoid or threshold devices. A unit's response to its input space can be completely characterized by its *trigger feature*, *banding region* and *zero region*, if one exists. In this way we are able to better distinguish the characteristics of the partitioning of the pattern space induced by units using various monotonic and non-monotonic activation functions.

The material introduced in this paper form the basis for current work establishing the computation complexity of non-monotonic networks, and placing them in context with the already well understood sigmoid networks.

References

[1] D. H. Ballard. Cortical connections and parallel processing: Structure and function. *The Behavioral and Brain Sciences*, 9:67–120, 1986. 304i

[2] I. S. Berkeley, M. R. Dawson, D. A. Medler, D. P. Schopflocher, and L. Hornsby. Density plots of hidden value unit activations reveal interpretable bands. *Connection Science*, 7(2):167–186, 1995. 304ii

[3] M. R. Dawson, I. S. Berkeley, D. A. Medler, and D. P. Schopflocher. Density plots of hidden value unit activations reveal reveal interpretable bands and microbands. In *Proceedings of the Machine Learning Workshop at AI/GV/VI'94*, pages (iii) 1–9, 1994. 304ii

[4] M. R. Dawson, S. C. Kremer, and T. N. Gannon. Identifying the trigger feature for hidden units in a PDP model of the early visual pathway. In R. Elio, editor, *Proceedings of the Tenth Canadian Conference on Artificial Intelligence*, pages 115–119, 1994. 304iv

[5] M. R. Dawson and D. P. Schopflocher. Modifying the generalized delta rule to train networks of non-monotonic processors for pattern classification. *Connection Science*, 4(1):19–31, 1992. 304i

[6] K. Hara and K. Nakayama. Comparison of activation functions in multilayer neural network for pattern classification. In *Proceedings of the IEEE International Conference on Neural Networks*, pages 2997–3002, 1994. 304ii

[7] D. B. McCaughan. On the properties of periodic perceptrons. In *Proceedings of the 1997 International Conference on Neural Networks (ICNN'97)*, pages 188–193, 1997. 304ii, 304ii

[8] M. Van Alstyne. Remaking the neural net: A perceptron logic unit. In *International Neural Network Society 1988 First Annual Meeting*, 1988. 304ii