Multilayer neural networks and polyhedral dichotomies

C. Kenyon and H. Paugam-Moisy

LIP, URA 1398 CNRS

ENS Lyon, 46 allée d'Italie

F69364 Lyon cedex 07, FRANCE

October 23, 1997

Abstract

We study the number of hidden layers required by a multilayer neural network with threshold units to compute a dichotomy f from $\mathcal{R}^d$ to $\{0,1\}$, defined by a finite set of hyperplanes. We show that this question is far more intricate than computing Boolean functions, although this well-known problem is underlying our research. We present recent advances on the characterization of dichotomies, from $\mathcal{R}^2$ to $\{0,1\}$, which require two hidden layers to be exactly realized.

1 INTRODUCTION

The number of hidden layers is a crucial parameter for the architecture of multilayer neural networks. Early research, in the 60's, addressed the problem of exactly realizing Boolean functions with binary networks or binary multilayer networks. On the one hand, more recent work focused on approximately realizing real functions with multilayer neural networks with one hidden layer [7, 8, 13] or with two hidden units [2]. On the other hand, some authors [1, 14] were interested in finding bounds on the architecture of multilayer networks for exact realization of a finite set of points. Another approach is to search the minimal architecture of multilayer networks for exactly realizing real functions, from $\mathcal{R}^d$ to $\{0,1\}$. Our work, of the latter kind, is a continuation of the effort of [5, 6, 9, 10] towards characterizing the real dichotomies which can be exactly realized with a single hidden layer neural network composed of threshold units. We show how this research is related to geometric algorithms, linear programming and combinatorial optimization.

1.1 Definitions and notations

A finite set of hyperplanes $\{H_i\}_{1\leq i\leq h}$ defines a partition of the d-dimensional space into convex polyhedral open regions (the union of the H_i's being neglected as a subset of measure zero). A *polyhedral dichotomy* is a function $f : \mathcal{R}^d \rightarrow \{0,1\}$, obtained by associating a class, equal to 0 or to 1, to each of those regions. Thus both $f^{-1}(0)$ and $f^{-1}(1)$ are unions of a finite number of convex polyhedral open regions. The h hyperplanes which define the regions are called the *essential hyperplanes* of f. A point P is an *essential point* if it is the intersection of some set of essential hyperplanes.

In this paper, all *multilayer networks* are supposed to be feedforward neural networks of threshold units, fully interconnected from one layer to the next, without skipping interconnections. A network is said to *realize* a function $f : \mathcal{R}^d \rightarrow \{0,1\}$ if, for an input vector x, the network output is equal to $f(x)$, almost everywhere in $\mathcal{R}^d$. The functions realized by our multilayer networks are the polyhedral dichotomies.

1.2 Polyhedral dichotomies and Boolean functions

By definition of threshold units, each unit of the first hidden layer computes a binary function y_j of the real inputs $(x_1, \ldots, x_d)$. Therefore, subsequent layers compute a Boolean function. Since any Boolean function can be written in DNF-form, two hidden layers are sufficient for a multilayer network to realize any polyhedral dichotomy. Two hidden layers are sometimes also necessary, e.g. for realizing the "four-quadrant" dichotomy which generalizes the XOR function [5].

For all j, the j^{th} unit of the first hidden layer can be seen as separating the space by the hyperplane $H_j : \sum_{i=1}^{d} w_{ij} x_i = \theta_j$. Hence the first hidden layer necessarily contains at least one hidden unit for each essential hyperplane of f. Thus each region R can be labelled by a binary number $y = (y_1, \ldots, y_h)$ (see [6]). The j^{th} digit y_j will be denoted by $H_j(R)$.

Usually there are fewer than 2^h regions and not all possible labels actually exist. The *Boolean family* $\mathcal{B}_f$ of a polyhedral dichotomy f is defined to be the set of all Boolean functions on h variables which are equal to f on all the existing labels.

2 EARLY RESULTS

It is straightforward that all polyhedral dichotomies which have at least one linearly separable function in their Boolean family can be realized by a one-hidden-layer network. However the converse is far from true. A counter-example was produced in [6]: adding extra hyperplanes (i.e. extra units on the first hidden layer) can eliminate the need for a second hidden layer (see figure 1). These hyperplanes are called *redundant hyperplanes*. Hence the problem of finding a minimal architecture for realizing dichotomies cannot be reduced to the neural computation of Boolean functions. Finding a generic description of all the polyhedral dichotomies which can be realized exactly by a one-hidden-layer network is a challenging problem.

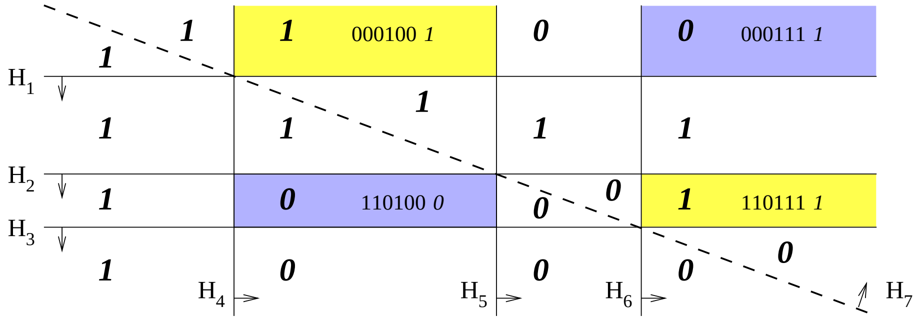


Figure 1: A one-hidden-layer network, with 6 hidden units on the first layer (corresponding to the essential hyperplanes), cannot realize the dichotomy, but a network with a 7^{th} extra unit (associated to the redundant hyperplane H_7, in dotted line) can.

2.1 Geometrical approach

One approach consists of finding geometric configurations which imply that a function is not realizable with a single hidden layer. There are three known such geometric configurations which involve two pairs of regions: the XOR-situation, the XOR-bow-tie and the XOR-at-infinity, as summarized on Figure 2.

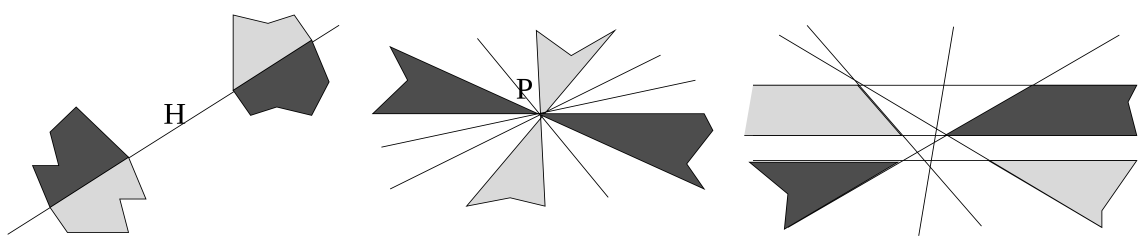


Figure 2: Geometrical representation of XOR-situation, XOR-bow-tie and XOR-at-infinity in the plane (black regions are in class 1, grey regions are in class 0).

Theorem 1 *If a polyhedral dichotomy f, from $\mathcal{R}^d$ to $\{0,1\}$, can be realized by a one-hidden-layer network, then it cannot be in an XOR-situation, nor in an XOR-bow-tie, nor in an XOR-at-infinity.*

The proof can be found in [11, 6] for the XOR-situation, in [17] for the XOR-bow-tie, and in [6] for the XOR-at-infinity. The sketch of these proofs is always the same: the four regions (two in each class) and their respective labellings induce an inconsistency in the system of inequalities associated to a one-hidden-layer solution.

2.2 Network approach

In contrast with our geometrical definitions, note that in [16], Takahashi, Tomita and Kawabata have presented a notion of "cyclicity", in the same context of research, but with a different point of view. They start from the notion of *summability* of boolean functions [15], and n-cyclicity can be viewed as a reinterpretation of n-summability. Given the hyperplanes associated to the hidden units of a *fixed* network (essential hyperplanes, plus redundant hyperplanes), finding the weights which realize the polyhedral dichotomy amounts to solving a system of linear inequalities. The authors of [16] claim that this system has a solution iff there is no "cyclicity", but their notion of cyclicity is only defined with respect to a fixed network. If f is in any of our three cases of XOR, then, no matter what hidden units are added, "cyclicity" occurs. On the other hand, for the example of figure 1, "cyclicity" occurs with six hidden units but not with seven hidden units. Our approach is different since we want a characterization of the polyhedral dichotomies which can be realized by a one-hidden-layer perceptron, independently of the network realizing f. The problem can be addressed in another different way, even less geometric than [16], as presented below.

2.3 Topological approach

Another research direction, implying a function is realizable by a single hidden layer network, is based on a topological approach. The proof uses the universal approximator property of one-hidden-layer networks, applied to intermediate functions obtained constructively adding extra hyperplanes to the essential hyperplanes of f. This direction was explored by Gibson [10], for two dimensions input space. Gibson's result can be reformulated as follows:

Theorem 2 *If a polyhedral dichotomy f is defined on a compact subset of $\mathcal{R}^2$, if f is not in an XOR-situation, and if no three essential hyperplanes (lines) intersect, then f is realizable with a single hidden layer network.*

Unfortunately Gibson's proof is not constructive, and extending it to remove some of the assumptions seems challenging. Both XOR-bow-tie and XOR-at-infinity are excluded by his assumptions of compactness and no multiple intersections. In the next section, we explore the cases, in $\mathcal{R}^2$, which are excluded by Gibson's assumptions. Gibson's present directions of research are turned towards extending the definitions and proofs to go to higher dimensions, where new cases of inconsistency emerge in subspaces of intermediate dimension [12].

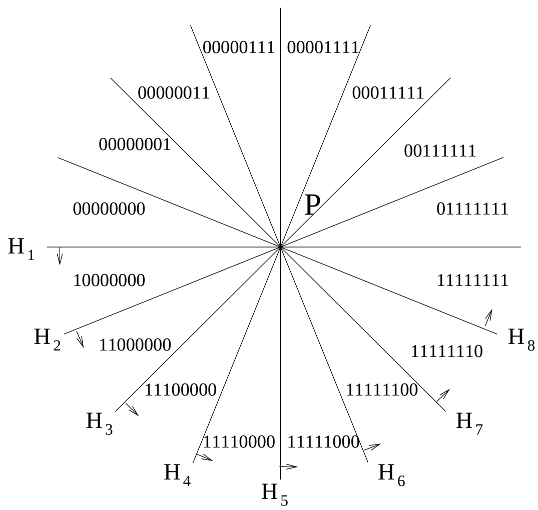
3 RECENT ADVANCES

3.1 Local realization in $\mathcal{R}^2$

The next two theorems proved that, in $\mathcal{R}^2$, the XOR-bow-tie and the XOR-at-infinity are the only restrictions to local realizability. Their proofs can be found in [4, 3].

Theorem 3 *Let f be a polyhedral dichotomy on $\mathcal{R}^2$ and let P be a point of multiple intersection. Let C_P be a neighborhood of P which does not intersect any essential hyperplane other than those going through P. The restriction of f to C_P is realizable by a one-hidden-layer network iff f is not in an XOR-bow-tie at P.*

The proof is in three steps: first, we reorder the hyperplanes in the neighborhood of P, so as to get a nice looking system of inequalities (see figure 3); second, we apply Farkas' lemma; third, we show how an XOR-bow-tie can be deduced.



$$
A = \begin{bmatrix}
\epsilon_1 & 0 & \dots & \dots & 0 & -\epsilon_1 \\
\epsilon_2 & \epsilon_2 & 0 & \ddots & 0 & -\epsilon_2 \\
\vdots & & \ddots & \ddots & \vdots & \vdots \\
\vdots & & & \epsilon_{k-1} & 0 & \vdots \\
\epsilon_k & \dots & \dots & \epsilon_k & \epsilon_k & -\epsilon_k \\
0 & \epsilon_{k+1} & \dots & \dots & \epsilon_{k+1} & -\epsilon_{k+1} \\
0 & 0 & \epsilon_{k+2} & \ddots & \vdots & -\epsilon_{k+2} \\
\vdots & & \ddots & \ddots & \vdots & \vdots \\
\vdots & & & \ddots & \epsilon_{2k-1} & -\epsilon_{2k-1} \\
0 & \dots & \dots & \dots & 0 & -\epsilon_{2k}
\end{bmatrix}
$$

Figure 3: Labels of the regions in the neighborhood of P, and matrix A.

If no two essential hyperplanes are parallel, the case of unbounded regions is exactly the same as a multiple intersection. The case of parallel hyperplanes is more intricate because matrix A is more complex. The proof requires a heavy case-by-case analysis (see [3]).

Theorem 4 *Let f be a polyhedral dichotomy on $\mathcal{R}^2$. Let C_∞ be the complementary region of the convex hull of the essential points of f. The restriction of f to C_∞ is realizable by a one-hidden-layer network iff f is not in an XOR-at-infinity.*

From theorems 3 and 4 we can deduce that a polyhedral dichotomy is locally realizable in $\mathcal{R}^2$ by a one-hidden-layer network iff f has no XOR-bow-tie and no XOR-at-infinity. Unfortunately this result cannot be extended to the global realization of f in $\mathcal{R}^2$ because more intricate distant configurations can involve contradictions in the complete system of inequalities, such as the *critical cycle* which implies that f cannot be realized with one hidden layer. Note that all the geometric configurations which implies that a two-hidden-layer network is required can be defined in $\mathcal{R}^d$ (XOR configurations and critical cycle). The restriction to $\mathcal{R}^2$ is only necessary for advances on converse results.

3.2 Critical cycles

A minimum of twelve regions are required to define the configuration of critical cycle (cf. figure 4). Note that one can augment the figure in such a way that there is no XOR-situation, no XOR-bow-tie, and no XOR-at-infinity. Definition and proof are based on bicolor graph considerations and can be found in [4, 3].

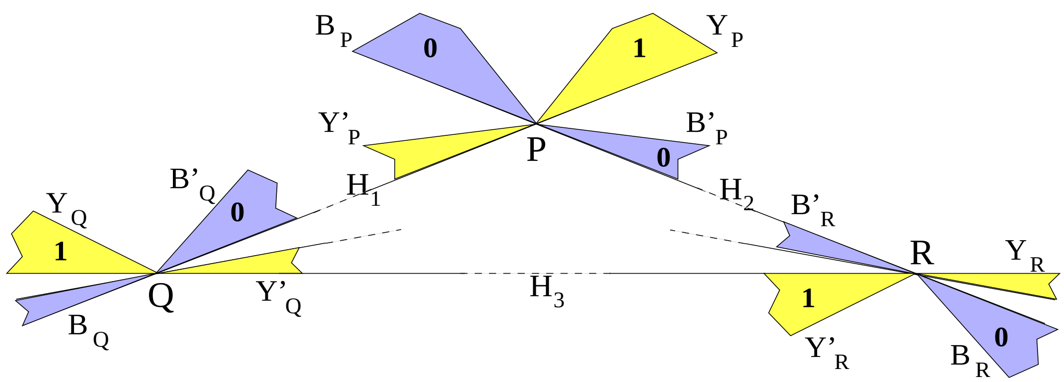


Figure 4: Geometrical configuration of a critical cycle, in the plane.

Theorem 5 *If a polyhedral dichotomy f, from $\mathcal{R}^d$ to $\{0, 1\}$, can be realized by a one-hidden-layer network, then it cannot have a critical cycle.*

4 CONCLUSION AND PERSPECTIVES

This paper reports partial progress towards characterizing functions which can be realized by a one-hidden-layer network, with a particular focus on dimension 2. The principle of using

Farkas lemma for proving local realizability still holds in higher dimensions, but the matrix A becomes more and more complex. In $\mathcal{R}^d$, even for $d = 3$, the labelling of the regions, for instance around a point P of multiple intersection, can become very complex. It seems that none of the work done in dimension 2 can easily be extended to higher dimensions.

Nevertheless, we conjecture that in dimension 2, a function can be realized by a one-hidden-layer network iff it does not have any of the four forbidden types of configurations: XOR-situation, XOR-bow-tie, XOR-at-infinity, and critical cycle. New points of view, such as using planar graphs or looking inside the hypercube of network internal representations are under investigations. We have already solved the problem for cycles of order 2 (which had been proved to be limited to the three XOR configurations, in $\mathcal{R}^2$) but not yet for inconsistencies of higher orders.

References

[1] E. B. Baum. On the capabilities of multilayer perceptrons. *Journal of Complexity*, 4:193–215, 1988.

[2] E. K. Blum and L. K. Li. Approximation theory and feedforward networks. *Neural Networks*, 4(4):511–516, 1991.

[3] G. Brightwell, C. Kenyon, and H. Paugam-Moisy. Multilayer neural networks: one or two hidden layers? Research Report 96-37, LIP, ENS Lyon, 1996.

[4] G. Brightwell, C. Kenyon, and H. Paugam-Moisy. Multilayer neural networks: one or two hidden layers? In M.C. Mozer, M.I. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems 9, Proc. NIPS*96*, pages 148–154. MIT Press, 1997.

[5] M. Cosnard, P. Koiran, and H. Paugam-Moisy. Complexity issues in neural network computations. In I. Simon, editor, *Proc. of LATIN'92*, volume 583 of *LNCS*, pages 530–544. Springer Verlag, 1992.

[6] M. Cosnard, P. Koiran, and H. Paugam-Moisy. A step towards the frontier between one-hidden-layer and two-hidden layer neural networks. In I. Simon, editor, *Proc. of IJCNN'93-Nagoya*, volume 3, pages 2292–2295. Springer Verlag, 1993.

[7] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Math. Control, Signal Systems*, 2:303–314, October 1988.

[8] K. Funahashi. On the approximate realization of continuous mappings by neural networks. *Neural Networks*, 2(3):183–192, 1989.

[9] G. J. Gibson. A combinatorial approach to understanding perceptron decision regions. *IEEE Trans. Neural Networks*, 4:989–992, 1993.

[10] G. J. Gibson. Exact classification with two-layer neural nets. *Journal of Computer and System Science*, 52(2):349–356, 1996.

[11] G. J. Gibson and C. F. N. Cowan. On the decision regions of multilayer perceptrons. *Proceedings of the IEEE*, 78(10):1590–1594, 1990.

[12] C. Z. W. Hassel Sweatman, G. J. Gibson, and B. Mulgrew. Exact classification with two-layer neural nets in n dimensions. *J. of Discrete Applied Mathematics*, 1997. (to appear).

[13] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

[14] S.-C. Huang and Y.-F. Huang. Bounds on the number of hidden neurones in multilayer perceptrons. *IEEE Trans. Neural Networks*, 2:47–55, 1991.

[15] S. Muroga. *Threshold logic and its applications.* John Wiley, 1971.

[16] H. Takahashi, E. Tomita, and T. Kawabata. Separability of internal representations in multilayer perceptrons with application to learning. *Neural Networks*, 6:689–703, 1993.

[17] P. J. Zweitering. *The complexity of multi-layered perceptrons.* PhD thesis, Technische Universiteit Eindhoven, 1994.