Chapter 0

Clifford Networks

Justin Pearson

This chapter introduces a class of feed-forward neural networks which have Clifford valued weight and activation values. Clifford algebras generalise the Complex and Quaternion algebras to higher-dimensions, thus Clifford networks are natural generalisations of complex valued networks. In this chapter the back-propagation algorithm is derived for Clifford valued networks and an approximation theorem is proved. Because Clifford algebras are a generalisation of the Complex and Quaternion algebras the approximation results also shows that Complex networks are universal functional approximators.

1 Introduction

The complex numbers can be motivated either algebraically as providing a field where the equation $x^2 = -1$ can be solved or geometrically as providing an algebra of two-dimensional space. The geometric interpretation is used in engineering an signal processing to provide an algebraic framework that allows reasoning about frequency and phase information.

Clifford algebras are born out of geometry. A Clifford algebra includes the normal vector algebra but provides an algebraic framework where symmetries and geometric transformation s can be reasoned about in an algebraic framework.

In this chapter Clifford networks are presented which are a natural extension of Complex valued networks. An approximation theorem is proved for Clifford networks which shows that they are universal approximators in the sense of (Hornick, Stinchcombe & White 1989). This proof specialises and gives a proof of the approximating power of Complex valued feed-forward networks.

2 Clifford Algebras

In this chapter we will only consider Clifford algebras defined over real valued vector spaces. Clifford algebras arise from geometric motivations which are sketched below, algebraically they are algebras compatible with quadratic forms. Quadratic forms allow inner products and norms to be studied in non-euclidian spaces. The relationship is studied below.

In this section: first a direct construction of Clifford algebras over real vector spaces is given; then this is generalised to Clifford algebras generated from Quadratic forms over vector spaces; this is then related back to the geometrical motivation given below; finally a classification is given of all Clifford algebras over real vectors spaces showing that many matrix algebras over the Complex numbers, the Real Numbers and the Quaternions are in fact Clifford algebras. As pointed out before Clifford algebras and hence networks generalise Complex and Quaternion Valued networks.

2.1 Geometric Motivation for Clifford Algebras

A Clifford Algebra is an extension of a vector algebra. It has two operations: an addition, $+$, corresponding to normal vector addition and Clifford Multiplication which generalises the dot product and the cross product of vectors. As an example the construction of the Clifford Algebra Corresponding to 2-dimensional Euclidian Space is considered. A vector, $\vec{v}$, can be be represented as the sum of two unit

length bases elements, $\vec{e_1}$ and $\vec{e_2}$ thus $\vec{v} = x\vec{e_1} + y\vec{e_2}$ then the scaler product of a vector $\vec{v}$ with its self, $\vec{v} \cdot \vec{v}$, is equal to the square of the length that is $x^2 + y^2$. If the vector is formally multiplied by its self then $(x\vec{e_1} + y\vec{e_2})(x\vec{e_1} + y\vec{e_2})$ becomes:

$$x^2\vec{e_1}^2 + y^2\vec{e_2}^2 + xy(\vec{e_1}\vec{e_2} + \vec{e_2}\vec{e_1})$$

In Euclidian space the length of a unit length basis element $\vec{e_i}$ is equal to 1, this gives the first equation true in a Clifford algebra of a euclidian space, that is $\vec{e_i}^2 = 1$. The elements $\vec{e_1}\vec{e_2}$ and $\vec{e_2}\vec{e_1}$ can be thought of as directed areas and are referred as bivectors. For an arbitrary pair of vectors, $\vec{v_1}$ and $\vec{v_2}$ the bivectors $\vec{v_1}\vec{v_2}$ and $\vec{v_2}\vec{v_1}$ represent directed areas. Then the second main equation of in a Clifford algebra (true in all Clifford Algebras) is:

$$\vec{e_1}\vec{e_2} = -\vec{e_2}\vec{e_1}$$

hence:

$$(x\vec{e_1} + y\vec{e_2})(x\vec{e_1} + y\vec{e_2}) = x^2 + y^2$$

giving the the normal scaler product of two vectors $\vec{v_1}$ and $\vec{v_2}$. The symmetric inner product is defined as:

$$\vec{v_1} \cdot \vec{v_2} = \frac{1}{2}\left(\vec{v_1}\vec{v_2} + \vec{v_2}\vec{v_1}\right)$$

by cancelling out the terms $\vec{e_1}\vec{e_2} + \vec{e_2}\vec{e_1} = 0$ the expression reduces to:

$$\frac{1}{2}(2x_1x_2 + y_1y_2) = x_1x_2 + y_1y_2$$

which is the normal scaler product of a vector.

The non-symmetrical version of the scaler product, the outer product is defined as:

$$\vec{v_1} \wedge \vec{v_2} = \frac{1}{2}(\vec{v_1}\vec{v_2} - \vec{v_2}\vec{v_1})$$

corresponds to the cross product of two vectors in three dimensions, but instead of giving a new vector gives a sum of directed areas, thus

giving a dimension independent definition of the cross product which is intrinsic to the vector space, for example:

$$(x_1\vec{e_1} + y_1\vec{e_2} + z_1\vec{e_3}) \wedge (x_2\vec{e_1} + y_2\vec{e_2} + z_2\vec{e_3})$$

which is equal to:

$$(x_1y_2 - y_1x_2)\vec{e_1}\vec{e_2} + (z_1x_2 - x_1z_2))\vec{e_3}\vec{e_1} + (y_1z_2 - z_1y_2)\vec{e_2}\vec{e_3}$$

Using the above two rules, the product of any two vectors $\vec{v_1}$ and $\vec{v_2}$ can be written as:

$$\vec{v_1}\vec{v_2} = \vec{v_1} \cdot \vec{v_2} + \vec{v_1} \wedge \vec{v_2}$$

Clifford algebras have applications in non-Euclidian spaces, that is vector spaces where $e_i^2 = -1$ for some elements e_i. The Lorentzian inner product used in special relativity where:

$$(x_1\vec{e_1} + y_1\vec{e_2} + z_1\vec{e_3} + t_1\vec{e_4}) \cdot (x_1\vec{e_1} + y_1\vec{e_2} + z_1\vec{e_3} + t_1\vec{e_4})$$

is defined to be:

$$x_1x_2 + y_1y_2 + z_1z_2 - t_1t_2$$

has an associated Clifford Algebra where $e_1^2 + e_2^2 + e_3^2 = 1$ and $e_4^2 = -1$.

Clifford algebras have been used extensively in physics to aid calculations (Chisholm & Common 1986, Hestenes 1986)

2.2 A Direct Construction

A $p + q$ dimensional real vector space will be denoted as $\mathcal{R}^{p+q}$ (the reason for the notation $p + q$ will become clear later on). A Clifford algebra $\mathcal{R}_{p,q}$ is two things, a 2^{p+q} vector space constructed from $\mathcal{R}^{p+q}$ and a set of algebraic rules defining multiplication and addition of vectors (when constructing Clifford algebras from quadratic form theory, these rules come out naturally).

The vector space $\mathcal{R}^{p+q}$ has a basis of the form:

$$e_1, e_2, e_3, \ldots e_{p+q}$$

From this construct a $2^{n=p+q}$ dimensional vector space with basis elements:

$$\{e_A = e_{(h_1 \ldots h_r)} | A = (h_1, \ldots, h_r) \in \mathcal{P}(\mathcal{N}), 1 \leq h_1 < \ldots < h_r \leq n\}.$$

(where $\mathcal{P}(\mathcal{N})$ represents the set of subsets of the set $\{1, \ldots, n\}$). For example the vector space over $\mathcal{R}^{1,2}$ would have the basis,

$$e_\emptyset,\ e_{(1)},\ e_{(2)}\ e_{(1,2)}, e_{(1,3)}, e_{(2,3)}, e_{(1,2,3)}$$

For notational convenience when no confusion can arise, $e_{(h_1, \ldots h_r)}$ will be denoted as $e_{h_1 h_2 \ldots h_r}$ and $e_\emptyset = e_0$ or since e_0 acts as the unit of the algebra it is often dropped when writing out elements a Clifford Algebra.

An element of the Clifford algebra is written as a formal sum:

$$x = \sum_A x_A e_A$$

with each $x_A \in \mathcal{R}$. In what follows a summation with a capital letter near the beginning of the alphabet denotes a sum over the basis elements of a Clifford algebra.

Addition of two elements of the algebra is defined as for vectors:

$$x + y = \sum_A (x_A + y_A) e_A$$

Multiplication is slightly more complicated. It is done formally element by element as in expanding brackets subject to the following algebraic rules:

$$e_i^2 = 1 \quad , \quad i = 1, \ldots, p \tag{1}$$

$$e_i^2 = -1 \quad , \quad i = p+1, \ldots, p+q \tag{2}$$

$$e_i e_j = -e_j e_i \quad , \quad i \neq j \tag{3}$$

with $1 \leq h_1 < \ldots h_r \leq n$, $e_{h_1} \cdot e_{h_2} \cdots e_{h_r} = e_{h_1 \ldots h_r}$. (the geometric motivation of these rules are as in section 2.1). This can be expressed more compactly in the following way,

$$e_A e_B = \kappa_{A,B} e_{A \Delta B},$$

where:

$$\kappa_{A,B} = (-1)^{\#((A \cap B) \setminus P)} (-1)^{p(A,B)} \tag{4}$$

P stands for the set $1, \ldots p$, and $\#X$ represents the number of elements in the set X,

$$p(A,B) = \sum_{j \in B} p'(A,j), \;\; p'(A,j) = \#\{i \in A | i > j\}, \tag{5}$$

and the sets A, B and $A \Delta B$(the set difference of A and B) are ordered in the prescribed way.

For example in $\mathcal{R}_{1,1}$ given $x = 3 + 4e_1 + e_2$ and $y = e_2 + 2e_{12}$ then

$$\begin{aligned} xy &= (3 + 4e_1 + e_2)(e_2 + 2e_{12}) \\ &= 3e_2 + 6e_{12} + 4e_1 e_2 + 8e_1 e_{12} + e_2^2 + 2e_2 e_{12} \end{aligned}$$

By using the reduction rules $e_1 e_{12} = e_1 e_1 e_2 = e_1^2 e_2 = e_2$ and $e_2 e_{12} = -e_2 e_{21} = -e_2^2 e_1 = e_1$. So the product xy is equal to $xy = -1 + 2e_1 + 11e_2 + 10e_{12}$ In general a Clifford algebra is associative but non-commutative.

2.3 Quadratic Forms

This section is intended to show that Clifford algebras arise naturally as mathematical structures and in particular how the rules (1 - 3) arise.

An orthogonal space is a real linear vector space X together with a symmetric inner product $(\cdot|\cdot)$ from X^2 to $\mathcal{R}$ which is linear in each

component. That is, the following equations are satisfied for all vectors x,y and z and all reals α and β, first is symmetric in x and y that is $(x, y) = (y, x)$ and further:

$$(\alpha x + \beta y, z) = \alpha(x, z) + \beta(y, z)$$
$$(x, \alpha y + \beta z) = \alpha(x, y) + \beta(x, z)$$

These equations abstract the standard inner product on euclidian vector spaces. A quadratic form Q on X, can be constructed from an inner product by defining:

$$Q(x) = (x|x)$$

The inner product is recoverable from the quadratic form by the following equation:

$$(x|y) = \frac{Q(x) + Q(y) - Q(x - y)}{2} \tag{6}$$

Thus the quadratic form uniquely determines the inner product and vice versa.

A Clifford algebra for a vector space X with respect to a quadratic form Q is the universal real associative algebra A which has the vector space X embedded in, such that for each element x of X, the following is true (see (Porteous 1995, Blaine Lawson Jr. & Michelsohn 1989) for details of the universal construction):

$$x^2 = -Q(x) \tag{7}$$

(where x^2 is carried out in the algebra A).

There is a theorem in quadratic form theory due to Sylvester (for a proof see (Lam 1973) or (Bromwich 1986, *Introduction to Quadratic Forms* 2000)) that given a real vector space and a quadratic form, it is possible by change of basis to represent the quadratic form as:

$$Q(x) = -x_1^2 - x_2^2 - \ldots x_p^2 + x_{p+1}^2 + \ldots x_{p+q}^2$$

where p, q is called the signature of the Quadratic form, and $p + q$ is equal to the dimension of X.

Thus to get the equations (1 - 3), we apply the equation (7) to the basis elements of X:

$$e_i^2 = -Q(e_i) = +1 \qquad \text{For } 0 < i \leq q \tag{8}$$
$$e_i^2 = -Q(e_i) = -1 \quad \text{For } q < i < p + q \tag{9}$$
$$\tag{10}$$

The anti-commutative law (equation (3)) can be derived from the following proposition.

Proposition 1 *For all $x, y \in X$ then in A,*

$$(x|y) = \frac{-(xy + yx)}{2} \tag{11}$$

Proof: From the formula (6):

$$2(x|y) = Q(x)+Q(y)+Q(x-y) = -x^2-y^2+(x-y)^2 = -xy-yx$$

In particular for distinct basis elements e_i and e_j we have:

$(e_i\|e_j) = 0$	Because e_i and e_j are orthogonal
$-2(e_i\|e_j) = e_ie_j + e_je_i$	From the equation (11)

which implies that $e_ie_j + e_je_i = 0$.

2.4 Some Familiar Clifford Algebras

We now show that the Clifford algebras arising from vector spaces over the reals include the Complex numbers the Quaternions and various matrix algebras; a complete classification can be found in (Porteous 1981, Blaine Lawson Jr. & Michelsohn 1989, Porteous 1995). It is easy to show that the complex numbers are simply the Clifford algebra $\mathcal{R}_{0,1}$.

The quaternion algebra is generated from the basis elements $1, i, j, k$ with the relations $i^2 = j^2 = k^2 = -1$ and

$$ij = k = -ji \quad , \quad jk - i = -kj \quad , ki = j = -ik$$

The quaternions are isomorphic to $\mathcal{R}_{0,2}$ with the following isomorphisms,

$$e_0 \cong 1 \ , \ e_1 \cong i \ , \ e_2 \cong j \ , \ e_{12} \cong k$$

Some perhaps less familiar examples include the Dirac algebra $\mathcal{R}_{4,1}$ and the Pauli algebra $\mathcal{R}_{3,3}$.

The algebra ${}^2\mathcal{R}$ (often denoted as $\mathcal{R} \oplus \mathcal{R}$) is defined over ordered pairs (x_1, x_2) with addition and multiplication defined as:

$$(x_1, x_2) + (y_1, y_2) = (x_1 + y_1, x_2 + y_2)$$
$$(x_1, x_2) * (y_1, y_2) = (x_1 * y_1, x_2 * y_2)$$

This algebra is isomorphic to the algebra $\mathcal{R}_{1,0}$ under the isomorphism ϕ given by:

$$\phi(\alpha + \beta e_1) \mapsto (\alpha + \beta, \alpha - \beta)$$

with

$$\phi^{-1}(a, b) \mapsto \frac{a+b}{2} + \frac{a-b}{2} e_1$$

The algebra $\mathcal{R}_{1,0}$ is called the algebra of hyperbolic complex numbers and can be used in relativity calculations in physics.

$R_{1,1}$ is isomorphic to the set of 2 by 2 real valued matrices. This can seen by the following identification,

$$e_1 \cong \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad e_2 \cong \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$
$$e_{12} \cong \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \quad e_0 \cong \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

This is a basis for $\mathcal{R}(2)$ and defines the isomorphism of algebras. This can be generalised to show that $\mathcal{R}_{n,n} \cong \mathcal{R}(2^n) \cong \text{End}(\mathcal{R}^n)$.

2.5 A partial Classification of Clifford Algebras

Every Clifford algebra is either isomorphic to a matrix algebra of $\mathcal{R}$, $\mathcal{C}$, $\mathcal{H}$ or a direct product of such matrix algebras. Not all Clifford algebras are distinct, and there is the so called periodicity theorem which relates higher dimensional Clifford algebras to low dimensional algebras. This section is a short guide to the relationships between Clifford algebras.

All proofs in this section are omitted and can be found in Chapter 13 of (Porteous 1981). So far the following relationships have been demonstrated:

$$\mathcal{R}_{0,0} \cong \mathcal{R}\ ,\ \mathcal{R}_{0,1} \cong \mathbf{C}\ ,\ \mathcal{R}_{0,2} \cong \mathcal{H}\ ,\ \mathcal{R}_{n,n} \cong \mathcal{R}(2^n)\ ,\ \mathcal{R}_{1,1} \cong {}^2\mathcal{R}$$

In fact all Clifford algebras defined over the real numbers can be constructed in some way from $\mathcal{R}$,$\mathcal{C}$ or $\mathcal{H}$. Table 1 extends this information. The reader interested in the explicit construction of the table should again consult (Porteous 1981) or (Blaine Lawson Jr. & Michelsohn 1989). The most useful fact (again for a proof see (Porteous 1981)) is perhaps that $\mathcal{R}_{p+1,q} \cong \mathcal{R}_{q+1,p}$.

To complete the table to arbitrary algebras it is enough to know that

$$\mathcal{R}_{p,q+8} \cong \mathcal{R}_{p,q} \otimes_{\mathcal{R}} \mathcal{R}(16) \cong \mathcal{R}_{p,q}(16).$$

where $\otimes_{\mathcal{R}}$ denotes the real tensor product of two algebras. This is the so called periodicity theorem a proof can be found in (Porteous 1981) or (Blaine Lawson Jr. & Michelsohn 1989)

3 Clifford Back-Propagation

Now the material on Clifford Algebras has been setup, we derive the back-propagation equations for a feed-forward Clifford network with Clifford valued weights and synapses. The derivation of a Clifford valued feed-forward network is similar to the derivation of the

Table 1. A table of Clifford algebras up to dimension 256, p goes horizontally and q vertically

$\mathcal{R}_{p,q}$	0	1	2	3	4	5	6	7	8
0	$\mathcal{R}$	$\mathcal{C}$	$\mathcal{H}$	${}^2\mathcal{H}$	$\mathcal{H}(2)$	$\mathcal{C}(4)$	$\mathcal{R}(8)$	${}^2\mathcal{R}(8)$	$\mathcal{R}(16)$
1	${}^2\mathcal{R}$	$\mathcal{R}(2)$	$\mathcal{C}(2)$	$\mathcal{H}(2)$	${}^2\mathcal{H}(2)$	$\mathcal{H}(4)$	$\mathcal{C}(8)$	$\mathcal{R}(16)$	${}^2\mathcal{R}(16)$
2	$\mathcal{R}(2)$	${}^2\mathcal{R}(2)$	$\mathcal{R}(4)$	$\mathcal{C}(4)$	$\mathcal{H}(4)$	${}^2\mathcal{H}(4)$	$\mathcal{H}(8)$	$\mathcal{C}(16)$	$\mathcal{R}(32)$
3	$\mathcal{C}(2)$	$\mathcal{R}(4)$	${}^2\mathcal{R}(4)$	$\mathcal{R}(8)$	$\mathcal{C}(8)$	$\mathcal{H}(8)$	${}^2\mathcal{H}(8)$	$\mathcal{H}(16)$	$\mathcal{C}(32)$
4	$\mathcal{H}(2)$	$\mathcal{C}(4)$	$\mathcal{R}(8)$	${}^2\mathcal{R}(8)$	$\mathcal{R}(16)$	$\mathcal{C}(16)$	$\mathcal{H}(16)$	${}^2\mathcal{H}(16)$	$\mathcal{H}(32)$
5	${}^2\mathcal{H}(2)$	$\mathcal{H}(4)$	$\mathcal{C}(8)$	$\mathcal{R}(16)$	${}^2\mathcal{R}(16)$	$\mathcal{R}(32)$	$\mathcal{C}(32)$	$\mathcal{H}(32)$	${}^2\mathcal{H}(32)$
6	$\mathcal{H}(4)$	${}^2\mathcal{H}(4)$	$\mathcal{H}(8)$	$\mathcal{C}(16)$	$\mathcal{R}(32)$	${}^2\mathcal{R}(32)$	$\mathcal{R}(64)$	$\mathcal{C}(64)$	$\mathcal{H}(64)$
7	$\mathcal{C}(8)$	$\mathcal{H}(8)$	${}^2\mathcal{H}(8)$	$\mathcal{H}(16)$	$\mathcal{C}(32)$	$\mathcal{R}(64)$	${}^2\mathcal{R}(64)$	$\mathcal{R}(128)$	$\mathcal{C}(128)$
8	$\mathcal{R}(16)$	$\mathcal{C}(16)$	$\mathcal{H}(16)$	${}^2\mathcal{H}(16)$	$\mathcal{H}(32)$	$\mathcal{C}(64)$	$\mathcal{R}(128)$	${}^2\mathcal{R}(128)$	$\mathcal{R}(256)$

learning rule for a complex valued network as in (Georgiou & Koutsougeras 1992) or elsewhere in this book.

In what follows the norm on an arbitrary Clifford number, $\|\cdot\|$ will be used, where,

$$\|x\| = \left(\sum_A [x]_A^2\right)^{\frac{1}{2}} \tag{12}$$

where $[x]_A$ represents the A'th part of the Clifford number x.

A feed-forward Clifford network with n inputs and m outputs will have a transfer function,

$$\Psi : (\mathcal{R}_{p,q})^n \to (\mathcal{R}_{p,q})^m$$

Where $(\mathcal{R}_{p,q})^n$ is the n-dimensional left module over the Clifford algebra $\mathcal{R}_{p,q}$.

To implement Clifford back-propagation an error measure E is defined which measures how well the network models a data set X. The basic form is the same,

$$E = \frac{1}{2}\sum_{\mathbf{x}\in X} \|\Psi - \Phi\|^2$$

where X is a set of training vectors.

It is convenient from the point of view of the derivation to define $\|\cdot\|$ as:

$$\|\mathbf{x}\|^2 = \sum_{i=1}^{k} |(x)_i|^2$$

where $(x)_i$ is a Clifford number representing the i'th part of $\mathbf{x}$ in the m-dimensional Clifford module over $\mathcal{R}_{p,q}$.

Assume that each node in the network has the same Clifford valued activation function $f : \mathcal{R}_{p,q} \to \mathcal{R}_{p,q}$. The output o_j of the j'th neuron can be written as,

$$o_j = f(net_j) = \sum_A u_A^j e_A$$

With u_A^j a function from $\mathcal{R}_{p,q}$ to $\mathcal{R}$ and

$$net_j = \sum_l \omega_{lj} o_l$$

where l sums over all the inputs to neuron j.

It is important to notice since $\mathcal{R}_{p,q}$ is in general non-commutative the order of multiplication in the above equation is important, although it will be shown later (in Section 4) that networks with left weight multiplication are equivalent in expressive power to networks with right multiplication.

In the real case E depends on the number of weights in the network. In the Clifford case E depends not only on all the weights but on the components of each of the weights. Again define $\lambda = \|\Psi - \Phi\|^2$. Then:

$$\frac{\partial E}{\partial [\omega_{ij}]_A} = \frac{1}{2} \sum_{\mathbf{x} \in X} \frac{\partial \lambda}{\partial [\omega_{ij}]_A}$$

and using the chain rule,

$$\frac{\partial \lambda}{\partial [\omega_{ij}]_A} = \sum_B \left(\frac{\partial \lambda}{\partial u_B^j} \left(\sum_C \frac{\partial u_B^j}{\partial [net_j]_C} \frac{\partial [net_j]_C}{\partial [\omega_{ij}]_A} \right) \right)$$

The partial derivative

$$\frac{\partial [net_j]_C}{\partial [\omega_{ij}]_A}$$

needs a bit of care. Using equation 3:

$$\frac{\partial [net_j]_C}{\partial [\omega_{ij}]_A} = \sum_l \frac{\partial [\omega_{lj} x_l]_C}{\partial [\omega_{ij}]_A} = \frac{\partial [\omega_{ij} o_i]_C}{\partial [\omega_{ij}]_A}$$

Then using the fact that $\omega_{ij} o_i = \sum_{D,E} [\omega_{ij}]_D [o_i]_E e_D e_E$ and

$$\frac{\partial [\omega_{ij} o_i]_C}{\partial [\omega_{ij}]_A} = \frac{\partial \left(\sum_{D,E} [\omega_{ij}]_D [o_i]_E \kappa_{A,E} \right)}{\partial [\omega_{ij}]_A}$$

with κ defined as in (4) and D, E summing over all the elements such that $e_D e_E = \pm e_C$. Since the denominator of the partial derivative only refers to $[\omega_{ij}]_A$ the partial derivative will equal:

$$\frac{\partial [\omega_{ij} o_i]_C}{\partial [\omega_{ij}]_A} = \frac{\partial [\omega_{ij}]_A [o_i]_E \kappa_{A,E}}{\partial [\omega_{ij}]_A} = \kappa_{A,E} [o_i]_E \text{ with } e_A e_E = \pm e_C \tag{13}$$

For example in the algebra $\mathcal{R}_{2,0}$ the table of derivatives would look like,

$\frac{\partial [x]_B}{\partial [\omega_{jl}]_A}$	$B=0$	1	2	12
$A=0$	$[x_{jl}]_0$	$[x_{jl}]_1$	$[x_{jl}]_2$	$[x_{jl}]_{12}$
1	$[x_{jl}]_1$	$[x_{jl}]_0$	$[x_{jl}]_{12}$	$[x_{jl}]_2$
2	$[x_{jl}]_2$	$-[x_{jl}]_{12}$	$[x_{jl}]_0$	$-[x_{jl}]_1$
12	$-[x_{jl}]_{12}$	$[x_{jl}]_2$	$-[x_{jl}]_1$	$[x_{jl}]_0$

The error derivative is now quite easy to calculate. If j is an output neuron then,

$$\frac{\partial \lambda}{\partial u_A^j} = \frac{\partial}{\partial u_A^j} \|\Psi - \Phi\|$$

$$\frac{\partial}{\partial u_A^j} |o_j - \Phi_j^2|^2 = 2[o_j - \Phi_j]_A$$

If j is not an output unit then the chain rule has to be used again.

$$\frac{\partial \lambda}{\partial u_A^j} = \sum_k \frac{\partial \lambda}{\partial u_A^k} \left(\sum_{B,C} \frac{\partial u_B^k}{\partial [net_k]^C} \frac{\partial [net_k]^C}{\partial u_A^j} \right)$$

with k running over the neurons that receive input from neuron j.

The term

$$\frac{\partial [net_k]_C}{\partial [u_j]_A}$$

is calculated in a similar manner to (13),

$$\frac{\partial [net_k]_C}{\partial u_A^j} = \kappa_{A,E} [\omega_{jk}]_D$$

where $\kappa_{A,E} e_A e_E = e_C$. The derivatives:

$$\frac{\partial u_B^k}{\partial [x_k]_C}$$

play the same rôle as $f'(net_j)$ does in the real-valued case and depends on the activation function used; this will be discussed in the next section.

Bringing this all together we have,

$$\frac{\partial E}{\partial [\omega_{ij}]_A} = \frac{1}{2} \sum_{x \in X} \sum_B \lambda_j^B \left(\sum_C \frac{\partial u_B^j}{\partial [net_j]_C} \kappa_{A,E} [o_k]_E \right)$$

with $e_A e_E = \pm e_C$ and

$$\lambda_j^B = \frac{\partial \|\Psi - \Phi\|^2}{\partial u_B^j} = 2[o_j - \Phi_j]_B$$

if j is an output neuron. If j is not an output unit then the chain rule has to be used again:

$$\lambda_j^B = \sum_k \lambda_k^B \left(\sum_{B,C} \frac{\partial u_B^k}{\partial [net_k]^C} \kappa_{A,D} [\omega_{jk}]_D \right)$$

with k running over the neurons that receive input from neuron j and $e_A e_D = e_C$.

The choice of activation function for a Clifford network as with a complex network requires some care. The normal sigmoid function $(1 + e^{-x})^{-1}$ can not be used. In (Georgiou & Koutsougeras 1992) a simple complex activation is proposed:

$$f(z) = \frac{z}{c + \frac{1}{r}|z|}$$

This function extended to the Clifford Domain is a suitable activation function and has been used successfully in experiments (Pearson & Bisset 1992, Pearson & Bisset 1994, Rahman, Howells & Fairhurst 2001, Pearson 1995) and applications.

4 Approximation Results

Now we have derived the back-propagation rules for Clifford networks this proves that Clifford networks are universal approximators: that is they any compact continuous function can be approximated arbitrarily close with a feed-forward Clifford network, these results generalise the results in (Hornick et al. 1989, Cybenko 1989) from real valued networks to Clifford valued networks.

There are essentially two ways of analysing feed-forward networks. The first views a feed-forward network as a pattern classifier and uses statistical techniques to assess the performance of a network; see (MacKay 1992). The second treats a feed-forward network essentially as a function approximator, that is, given a network with n inputs and m outputs and a set of weight values ω_{ij} the network can be seen to be computing a function:

$$\Phi_\omega : \mathcal{R}^n \to \mathcal{R}^m$$

In the Clifford case the real numbers $\mathcal{R}$ are replaced by an arbitrary Clifford algebra $\mathcal{R}_{p,q}$. The sort of question then asked is how well can

a given class of networks approximate classes of functions? Various theorems have been proved (Cybenko 1989, Hornick et al. 1989, Ito 1991, Kůrková 1991) which show that feed-forward networks with one hidden layer are sufficient to approximate continuous functions. Further results by Sontag (Sontag 1992) show that in certain problems two hidden layers are required; this is because the function trying to be approximated is too discontinuous to be approximated by a single hidden layer network.

This section first extends Cybenko's (Cybenko 1989) proof, that real valued networks with a single hidden layer can approximate any bounded continuous function with compact support, to networks over an arbitrary Euclidean Clifford algebra (that is algebras with signature $0, q$), these include the Complex numbers $\mathcal{R}_{0,1}$ and the Quaternions $\mathcal{R}_{0,2}$. Before the proof can be made some background material has to be provided in Clifford modules (section 4.1) and Clifford analysis.

4.1 Clifford modules and Clifford Analysis

This section deals with a generalisation of vector spaces, the theory of Modules over rings: specifically Clifford modules. This generalisation is necessary in order to state the relevant approximation theorems. Various theorems are stated which are generalisations of traditional theorems such as the Hahn-Banach theorem and the Riesz representation theorem (Rudin 1966, Rudin 1973); all proofs are omitted, but these can be found in (Brackx, Delenghe & Sommen 1982).

From now on the convention adopted in (Brackx et al. 1982) is used, where a Euclidean Clifford algebra is referred as an $\mathcal{A}$ algebra. A module is a generalisation of a vector space, where the set of coefficients come from some ring instead of a field, thus modules have a different geometrical structure from vector spaces.

Definition 1 *A unitary left* $\mathcal{A}$-module $X_{(l)}$ *is an Abelian group* $X_{(l)}, +$ *and an operation* $(\lambda, f) \rightarrow \lambda f$ *from* $\mathcal{A} \times X_{(l)}$ *into* $X_{(l)}$ *s.t.*

for all $\lambda, \mu \in \mathcal{A}$ and $f, g \in X_{(l)}$ the following hold:

$$(\lambda + \mu)f = \lambda f + \mu f$$

$$(\lambda\mu)f = \lambda(\mu f)$$

$$\lambda(f + g) = \lambda f + \lambda g$$

$$e_0 f = f$$

We have already met an example of a Clifford Module in Section 3, the space $\mathcal{R}^n_{p,q}$.

Definition 2 *Let $X_{(l)}$ be a unitary left $\mathcal{A}$-module, then a function $p : X_{(l)} \to \mathcal{R}$ is said to be a* proper semi-norm *if there exists a constant $C_0 \geq 0$ s.t. for all $\lambda \in \mathcal{A}$ and $f, g \in X_{(l)}$ the following conditions are satisfied:*

$$p(f + g) \leq p(f) + p(g)$$

$$p(\lambda f) \leq C_0|\lambda|p(f)$$

$$p(\lambda f) = |\lambda|p(f) \ \textit{if} \ \lambda \in \mathcal{R}$$

$$\textit{If} \ p(f) = 0 \ \textit{then} \ \ f = 0$$

Definition 3 *Given a module $X_{(l)}$ the* algebraic dual $X^{*alg}_{(l)}$ *is defined to be the set of left $\mathcal{A}$-linear functionals from $X_{(l)}$ into $\mathcal{A}$. That is the set of functionals $T : X_{(l)} \to \mathcal{A}$ s.t.*

$$T(\lambda f + g) = \lambda T(f) + T(g)$$

$f, g \in X_{(l)}$ and $\lambda \in \mathcal{A}$.

Definition 4 *The set of bounded T functionals with respect to a semi-norm p is denoted $X^*_{(l)} \subset X^{*alg}_{(l)}$. Explicitly for all functionals T and for all $f \in X_{(l)}$:*

$$|T(f)| \leq Cp(f)$$

for some real constant C.

The following theorem is a a corollary to a Hahn-Banach type theorem for Clifford modules for details and proof see sections 2.10-2.11 in (Brackx et al. 1982).

Theorem 1 *Let $X_{(l)}$ be a unitary left $\mathcal{A}$-module provided with a semi norm p and let $Y_{(l)}$ be a submodule of $X_{(l)}$. Then $Y_{(l)}$ is dense in $X_{(l)}$ if and only if for each $T \in X^*_{(l)}$ such that $T|Y_{(l)} = 0$[1] we have $T = 0$ on $X_{(l)}$.*

Now a useful class of function spaces is introduced.

Definition 5 *The space $C^0(\mathcal{K}; \mathcal{A})$. Let $\mathcal{K}$ be a compact subset of $\mathcal{R}^r$ ($r \geq 1$). Then $C^0(\mathcal{K}; \mathcal{A})$ stands for the unitary bi-$\mathcal{A}$-module of $\mathcal{A}$-valued continuous functions on $\mathcal{K}$.*

This can be thought of as a product of classical real valued functions i.e.:

$$C^0(\mathcal{K}; \mathcal{A}) = \Pi_A C^0(\mathcal{A}; \mathcal{R}) e_A \tag{14}$$

where A runs over all the basis elements in the Clifford algebra in question. A norm can be defined for each $f \in C^0(\mathcal{K}; \mathcal{A})$:

$$||f|| = \sup_{x \in \mathcal{K}} |f(x)|$$

This norm is equivalent to the product norm taken from (14).

Definition 6 *Given an open set $\Omega \subset \mathcal{R}^n$ and a sequence $(\mu_B)_B$ of real valued measures on Ω. Then for any open set in Ω an $\mathcal{A}$ valued measure can be defined:*

$$\mu(I) = \sum_B \mu_B(I) e_B$$

Definition 7 *An $\mathcal{A}$-valued function:*

$$f = \sum_A f_a e_A$$

is said to be μ-integrable in Ω if for all A and B ranging over the basis elements of $\mathcal{A}$ each f_A is μ_B integrable.

[1] T restricted to $Y_{(l)}$ equal to zero

Definition 8 *For any μ-integrable function f define:*

$$\int_{\Omega} f(x)d\mu = \sum_{A,B} e_A e_B \int_{\Omega} f_A(x)d\mu_B$$

A Riesz representation type theorem can be obtained.
Theorem 2 *Let T be a bounded $\mathcal{A}$ valued function in $C^0_{(l)}(\mathcal{K};\mathcal{A})$. Then there exists a unique $\mathcal{A}$ valued measure μ with support contained in $\mathcal{K}$ such that for all $f \in C^0_{(l)}(\mathcal{K};\mathcal{A})$:*

$$T(f) = \int_{\mathcal{K}} f(x)d\mu$$

For a proof again see (Brackx et al. 1982).

4.2 The Approximation result

Now all the machinery has been set up and the approximation result can be proved. A feed-forward network with one output neuron and N inputs units and K hidden units computes a function:

$$\Phi(\mathbf{x}) = \sum_{j=1}^{K} \alpha_j f(\sum_{i=1}^{N} y_{ij}x_i + \theta_j)$$

with f the activation function x_i the i'th input, y_{ij} weight values for the connection between the input layer and the hidden layer and α_j the weights from the hidden layer to the output node.

The function $\Phi(\mathbf{x})$ can be seen as a function from $\mathcal{R}^{N2^n}$ (where 2^n is the dimension of $\mathcal{A}$) to $\mathcal{A}$ and hence a member of $C^0_{(l)}(\mathcal{R}^{N2^n};\mathcal{A})$. This is why the material of the last section was relevant. The next definition is important. What is shown is that all activation functions satisfying the definition, when used in feed-forward networks, are universal approximators. Then to complete the proof all that is needed to show is that the activation functions considered in Section 3 satisfy the definition.

Definition 9 *An activation function f (considered as a function from $\mathcal{R}^{N2^n}$ to $\mathcal{A}$) is said to be* discriminating *if for any given Clifford valued measure μ with support I^{N2^n} if:*

$$\int_{I^{N2^n}} f(\sum_{i=1}^{N} y_i x_i + \theta) d\mu(x) = 0$$

for all $y_i, \theta \in \mathcal{R}_{0,n}$ implies $\mu(x) = 0$.

The following theorem is the heart of the approximation result.

Theorem 3 *Let f be any continuous discriminating functions. Then finite sums of the form:*

$$\Phi(x) = \sum_{j=1}^{K} \alpha_j f(\sum_{i=1}^{N} y_{ij} x_i + \theta_j) \tag{15}$$

are dense in $C^0_{(l)}(I^{N2^n}; \mathcal{A})$

Proof: The proof is essentially a modification of Cybenko's Theorem 1 in (Cybenko 1989) using the theory of Clifford modules in the last section.

Let S be the function space generated by sums of the form (15). Assume that the closure of S is not all of $C^0_{(l)}(I^{N2^n}; \mathcal{A})$; denote the closure of S by R. By the Hahn-Banach type theorem 1 there is a bounded linear functional T on $C^0_{(l)}(I^{N2^n}; \mathcal{A})$, with $T \neq 0$ but $T(R) = T(S) = 0$. By Theorem 2 this bounded linear functional is of the form:

$$T(h) = \int_{I^N 2^n} h(x) \mu(x)$$

for some measure μ and $h \in C^0_{(l)}(I^{k2^n}, \mathcal{A})$. In particular since $f \in C^0_{(l)}(I^{k2^n}, \mathcal{A})$ is in R, for any y_i:

$$T(f) = \int_{k2^n} f(\sum_{i=1}^{N} y_i x_i + \theta) d\mu(x) = 0$$

Since f is discriminating this implies $\mu = 0$ contradicting our assumption hence S must be dense in $C^0_{(l)}(I^{N2^n}; \mathcal{A})$.

So to prove that the class of feed-forward networks considered in Chapter 3 are universal approximators, we have to show that functions of the form:

$$f(x) = \frac{x}{1 + |x|}$$

are discriminating.

Theorem 4

$$f(x) = \frac{x}{1 + |x|}$$

is discriminatory.

Proof: A function $f(x)$ is discriminatory if:

$$\int_{N2^n} f(\sum_{i=1}^{N} y_i x_i + \theta) d\mu(x) = 0$$

for all y_i implies that $\mu(x) = 0$. This is equivalent to saying that:

$$\int_{N2^n} f(\sum_{i=1}^{N} y_i x_i + \theta) d\mu(x) = \sum_{A,B} e_A e_B \int_{N2^n} f_A(\sum_{i=1}^{N} y_i x_i + \theta) d\mu_B(x) = 0$$

for all y_i.

Define $\gamma_A(x) : I^{N2^n} \to \mathcal{R}$ to be the limit of:

$$\gamma_A(x) = \lim_{\lambda \to \infty} f_A(\lambda x)$$

(where λx is a Clifford multiplication, with λ a real number). So

$$f_A(\lambda x) = \frac{[\lambda z]_A}{1 + |\lambda z|} = \frac{\lambda [z]_A}{1 + \lambda |z|}$$

So

$$\gamma_A(z) = \begin{cases} 1 & \text{if } [z]_A > 0 \\ 0 & \text{if } [z]_A = 0 \\ -1 & \text{if } [z]_A < 0 \end{cases}$$

In our case:

$$\gamma_A(\sum_{i=1}^{N} y_i x_i + \theta) = \begin{cases} 1 & \text{if } [\sum_{i=1}^{N} y_i x_i + \theta]_A > 0 \\ 0 & \text{if } [\sum_{i=1}^{N} y_i x_i + \theta]_A = 0 \\ -1 & \text{if } [\sum_{i=1}^{N} y_i x_i + \theta]_A < 0 \end{cases}$$

The sets defined by $[\sum_{i=1}^{N} y_i x_i + \theta]_A = 0$ are hyper-planes, since $[\sum_{i=1}^{N} y_i x_i + \theta]_A$ is just a set of linear equations in the components of x_i.

The rest of the proof is almost verbatim from Lemma 1 of Cybenko (Cybenko 1989). So let $\Pi_{y,\theta}^{A} \subset I^{2^n}$ be the hyper-plane defined by:

$$\left\{ x \middle| \left[\sum_{i=1}^{N} y_i x_i + \theta \right]_A = 0 \right\}$$

and let $H_{y,\theta}^{A}$ be the half space defined by:

$$\{x | [\sum_{i=1}^{N} y_i x_i + \theta]_A > 0\}$$

Then by the Lebesgue bounded convergence theorem we have:

$$0 = \int_{I^{N2^n}} f_A(\lambda x) d\mu_B(x) = \int_{I^{N2^n}} \gamma_A(x) d\mu_B(x) = \mu(H_{y,\theta}^{A})$$

Now if μ_B were always a positive measure the result would be trivial, but since μ_B is an arbitrary measure the result is harder (since positive bits of μ might cancel out negative bits of μ_B).

Fix the y_i's and define:

$$F_A(h) = \int_{I^{N2^n}} h([\sum_{i=1}^{K} y_i x_i]_A)$$

for some bounded μ_B measurable function $h : \mathcal{R} \to \mathcal{R}$. F_A is a bounded functional on $L^\infty(\mathcal{R})$.

Let h be the indicator function on the interval $[\theta_A, \infty)$, then:

$$F(h) = \int_{k2^n} h([\sum_{i=1}^{K} y_i x_i]_A) = \mu_B(\Pi^A_{y,\theta}) + \mu(H^A_{y,\theta})$$

Similarly $F(h) = 0$. If h is the indicator of any open interval, by linearity $F(h) = 0$ and hence for any simple function. Since the simple functions are dense in $L^\infty(\mathcal{R})$, $F = 0$.

In particular given the two functions $s(x) = \sin(x)$, $c(x) = \cos(x)$:

$$F_A(s(x) + ic(x)) = \int_{I^{k2^n}} s([\sum_{k=1}^{K} y_k x_k]_A) + ic([\sum_{k=1}^{K} y_k x_k]_A) d\mu_B =$$

$$\int_{I^{k2^n}} \exp(i[\sum_{k=1}^{K} y_k x_k]_A) d\mu_B = 0 \quad (16)$$

for all y_k. Therefore the Fourier transform of μ_B is zero, hence μ_B must be zero and hence f is discriminatory.

One important thing to point out with this proof is that the order of weight multiplication is irrelevant; the whole proof could be repeated with networks where multiplication was done on the right. Thus it does not matter theoretically which sort of nets (left or right weight multiplication) is used for a particular problem. Practically not much is known, but in all the examples the author has tried, the performance of the net does not seem to be affected by the order of weight multiplication.

5 Conclusion and related work

This chapter has been largely theoretical, but it has been shown that it is possible to derive a back-propagation algorithm for Clifford valued feed-forward networks. Such networks are a natural extension of Complex valued networks by virtue of Clifford algebras being the natural geometric extension of the Complex numbers. Due to space limitations no experimental results have been presented, applications of Clifford networks can be found in (Pearson 1995, Rahman et al. 2001). Other work on Clifford valued neural networks has been done with self-organising networks with Clifford Algebras applied to motion modelling systems (Bayro-Corrochano, Buchholz & Sommer 1996*b*, Bayro-Corrochano, Buchholz & Sommer 1996*a*, Bayro-Corrochano 1996).

Author's address

Justin Pearson: Uppsala University Department of Information Technology Box 337 SE-751 05 Uppsala Sweden.

References

Bayro-Corrochano, E. 1996, Clifford selforganizing neural network, clifford wavelet network., *in* 'Proc. 14th IASTED Int. Conf. Applied Informatics. Innsbruck, Austria,', pp. 271–274.

Bayro-Corrochano, E., Buchholz, S. & Sommer, G. 1996*a*, A new selforganizing neural network using geometric algebra, *in* 'Proc. Int. Neural Network Society 1996 Annual Meeting: World Congress on Neural Networks, WCNN'96, CA, San Diego, USA', pp. 245–249.

Bayro-Corrochano, E., Buchholz, S. & Sommer, G. 1996*b*, Selforganizing clifford networks, *in* 'Proceedings of ICNN', Vol. 1, pp. 120–125.

Blaine Lawson Jr., H. & Michelsohn, M.-L. 1989, *Spin Geometry*, Princeton University Press, Princeton, New Jersey.

Brackx, F., Delenghe, R. & Sommen, F. 1982, *Clifford Analysis*, Research notes in mathematics; 76, Pitman Advanced Publishing Program.

Bromwich, T. I. 1986, *Quadratic Forms and Their Classification by Means of Invariant-Factors*, Vol. 3 of *Cambridge Tracts in Mathematics and Mathematical Physics*, Cambridge University Press.

Chisholm, J. & Common, A., eds 1986, *Clifford algebras and their applications in mathematical physics*, Vol. 183 of *NATO ASI series, C:Mathematical and physical sciences*.

Cybenko, G. 1989, 'Approximation by superpositions of a sigmoidal function', *Mathematics of Control Signals and Systems* pp. 303–314.

Georgiou, G. M. & Koutsougeras, C. 1992, 'Complex domain backpropagation', *IEEE Transactions on Circuits and Systems* pp. 330–334.

Hestenes, D. 1986, *New foundations for Classical Mechanics*, Reidel.

Hornick, K., Stinchcombe & White 1989, 'Multilayer feedforward networks are universal approximators', *Neural Networks* **2**, 359–366.

Introduction to Quadratic Forms 2000, Springer Verlag.

Ito, Y. 1991, 'Representation of functions by superpositions of step or sigmoid functions and their applications to neural network theory', *Neural Networks* **4**, 385–394.

Kůrková, V. 1991, 'Kolmogorov's theorem is relevant', *Neural Computation* **3**(4), 617–622.

Lam, T. 1973, *The algebraic theory of quadratic forms*, Mathematics lecture note series, W.A. Benjamin, Reading, Mass.

MacKay, D. J. 1992, Bayesian Methods for Adaptive Models, PhD thesis, California Institute of Technolgy, Pasadena, California.

Pearson, J. & Bisset, D. 1992, Back Propagation in a Clifford Algebra, *in* 'ICANN Brighton'.

Pearson, J. & Bisset, D. 1994, Neural Networks in the Clifford Domian, *in* 'IEEE94 symposium on Neural Networks Orlando'.

Pearson, J. K. 1995, Clifford Networks, PhD thesis, University of Kent (Electronic engineering).

Porteous, I. 1981, *Topological Geometry*, Cambridge University Press.

Porteous, I. 1995, *Clifford Algebras and the Classical Groups*, Cambridge Studies in Advanced Mathematics, Cambridge University Press.

Rahman, A., Howells, W. & Fairhurst, M. 2001, 'A multi-expert framework for character recognition: A novel application of clifford networks', *IEEE Neural Networks* **12**(1).

Rudin, W. 1966, *Real and complex analysis*, McGraw-Hill series in higher mathematics, McGraw-Hill.

Rudin, W. 1973, *Functional analysis*, McGraw-Hill series in higher mathematics, McGraw-Hill.

Sontag, E. 1992, 'Feedback stablisation using two hidden layer nets', *IEEE Transactions on Neural Networks* pp. 981–990.