1996

Characterization of a Class of Sigmoid Functions With Applications to Neural Networks

Anil Ravindran Menon
Syracuse University

Kishan Mehrotra
Syracuse University, mehrotra@syr.edu

Chilukuri K. Mohan
Syracuse University, ckmohan@syr.edu

Sanjay Ranka
Syracuse University, ranka@top.cis.syr.edu

Recommended Citation

Menon, Anil Ravindran; Mehrotra, Kishan; Mohan, Chilukuri K.; and Ranka, Sanjay, "Characterization of a Class of Sigmoid Functions With Applications to Neural Networks" (1996). *College of Engineering and Computer Science - Former Departments, Centers, Institutes and Projects*. 49.
https://surface.syr.edu/lcsmith_other/49

Characterization of a class of sigmoid functions with applications to neural networks

Anil Menon
armenon@top.cis.syr.edu

Kishan Mehrotra
kishan@top.cis.syr.edu

Chilukuri K. Mohan
mohan@top.cis.syr.edu

Sanjay Ranka
ranka@top.cis.syr.edu

Syracuse University
Neural Network Group
School of Computer & Information Science, 4-116 CST
Syracuse, NY 13244-4100
Phone: (315) 443-2368
Fax: (315) 443-1122

1 Introduction

Sigmoid functions, whose graphs are "S-shaped" curves, appear in a great variety of contexts, such as the transfer functions in many neural networks.[1] Their ubiquity is no accident; these curves are the among the simplest non-linear curves, striking a graceful balance between linear and non-linear behavior.

Figure 1 shows three sigmoidal functions, and their inverses; the hyperbolic tangent $\tanh(\cdot)$ (graph 'A'), the "logistic" sigmoid $1/(1 + \exp(-x))$ (graph 'B'), and the "algebraic" sigmoid, $x/\sqrt{(1 + x^2)}$ (graph 'C'), with inverses, $\tanh^{-1}(y)$, $\ln y/(1 - y)$, and $y/\sqrt{1 - y^2}$, respectively. In a few cases, sigmoid curves can be described by formulae; this rubric includes power series expansions (e.g., hyperbolic tangent), integral expressions (e.g., error function), composition of simpler functions (e.g., the Gudermannian function), inverses of functions definable by formulae (e.g., the "complexified" Langevin function, a sigmoid defined as the inverse of the function, $1/x - \cot(x)$), differential equations *et cetera*.

Although the level of abstraction in many problems is such that one does not need to work with explicit formulae[2], it is useful to study networks with specific transfer functions for the following reasons:

[1] Other examples of the use of sigmoid functions are the logistic function in population models, the hyperbolic tangent in spin models, the Langevin function in magnetic dipole models, the Gudermannian function in special functions theory, the (cumulative) distribution functions in mathematical statistics, the piecewise approximators in nonlinear approximation theory, the hysteresis curves in certain nonlinear systems etc.

[2] For example, in neural net approximation theory, significant results can be obtained about the existence of realizations within preassigned tolerances, with very few constraints on the nature of the node transfer function; classic results along these lines are found in [5, 7, 11, 20]

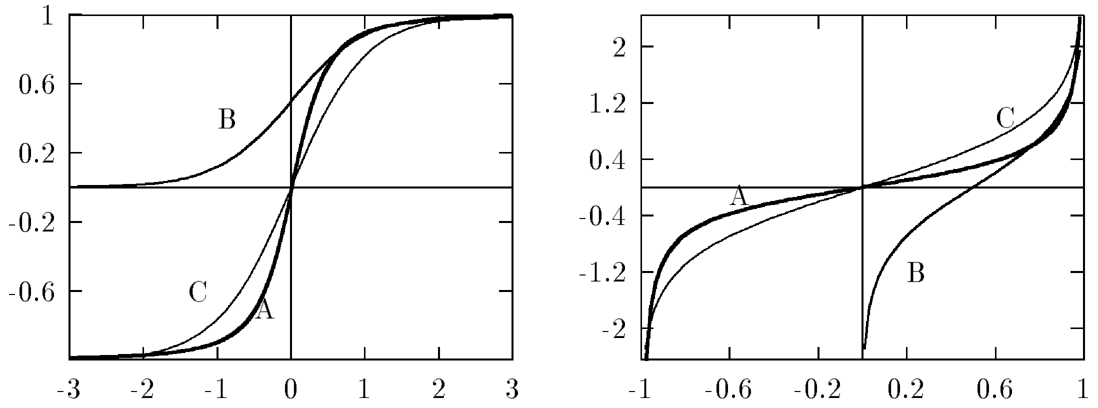

Figure 1: Some sigmoids and their inverses

1. In determining whether a single layered feedforward net is uniquely determined by its corresponding input-output map, Sussmann's elegant proof of uniqueness specifically used the properties of the $\tanh(\cdot)$ function [35]. A later analysis by Sontag obtained the same results with fewer assumptions on the node transfer function, but still requires such functions to be odd, and satisfy certain "independence" properties [1]. With respect to the uniqueness problem, all node transfer functions are *not* equivalent[3].

2. Without tractable analytical forms to work with, many problems relating to sigmoids are resistant to theory. Neural net theory offers many examples. For example, there have been claims in the literature about the advantage (with respect to computability, training times etc.) of certain sigmoidal transfer functions over others in backpropagation networks [8, 17, 33]. Some theoretical support comes from considering the first derivatives (if defined) of the various transfer functions proposed; the first derivatives are partially responsible for controlling the step size in the weight adjustment phase of the back propagation algorithms, which in turn influences the rate of convergence. Explicit expressions for sigmoids are useful in such considerations.

3. The dynamical system describing the continuous Hopfield model raises an intriguing query. If one assumes a $\tanh(\cdot)$ node transfer function, one can show that the Hopfield model is transformable to the Legendre differential equation (see section 6.1); An important question is whether this relationship is *robust* with respect to the choice of the transfer function.

4. The recent study of sigmoidal derivatives by Minai and Williams [26] is another case in point; they derived a connection with Eulerian numbers [15, pp. 252-257] but restricted their inquiry to the very specific logistic sigmoid. Any generalization of their results requires a careful look at sigmoids representable by formulae.

[3] Another example of the non-equivalence of "sigmoids" is offered by Macintyre and Sontag's work on the Vapnik-Chervonenkis (VC) dimension of feedforward networks, which showed that it is finite only for a class of sigmoidal functions they call the exp-RA functions. They showed that analyticity of the transfer function is crucial, and cannot be relaxed by say, making the function C^∞ [23].

5. There are other related issues. For instance, the hyperbolic tangent and logistic sigmoid are essentially equivalent in that, one can be obtained from the other, by simple translation and scaling transformations:

$$\frac{1}{1 + \exp(-x)} - \frac{1}{2} = \frac{1}{2} \tanh(x/2) \tag{1.1}$$

   Many sigmoids have power series expansions which alternate in sign. Many have inverses with hypergeometric series expansions. On the other hand, many sigmoids have no such simple forms, or obvious connections with well known sigmoids. It is natural to ask whether these varied analytical expressions for sigmoids have anything in common. It is difficult to answer such questions without a thorough understanding of the analytical expressions for sigmoid functions.

In view of these considerations, this paper undertakes a study of two classes of sigmoids: the *simple sigmoids*, defined to be odd, asymptotically bounded, completely monotone functions in one variable, and the *Hyperbolic sigmoids*, a proper subset of simple sigmoids and a natural generalization of the hyperbolic tangent. The class of hyperbolic sigmoids includes a surprising number of well known sigmoids. The regular structure of the simple sigmoids often makes a theory tractable, paving the way for more general analysis.

The main contributions of the paper are as follows

- Simple and Hyperbolic sigmoids and their inverses are completely characterized in Sections 4 and 5.
- Using series inversion techniques, in Section 5, we obtain the series expansions of hyperbolic sigmoids from those of their inverses. These results extend results of Minai and Williams [26] for the logistic function.
- In section 4, we study the composition of simple sigmoids *via* differentiation, addition, multiplication, and functional composition. These results also completely specify the relationship between Euler's incomplete Beta function and the parameterized sigmoids.
- In Section 6.1 we show that the continuous Hopfield equations belong to the class of non-homogeneous Legendre differential equations if the neural transfer function is a simple sigmoid.
- In Section 6.2 we establish a connection between Fourier transforms and feedforward nets with one summing output and one hidden layer whose nodes contain simple sigmoidal transfer functions.

We do not purport to have discovered a general framework to describe *all* sigmoids; indeed, such a quest is largely meaningless; nor are we arguing for limiting the notion of sigmoids to the classes considered in this paper. Simple sigmoids are rather special sigmoids, but their regular structure often makes a theory tractable, paving the way for more general analysis.

2 Preliminaries

Notation: $\Re$ and $\Re^+$ denote real space, and the set of positive real numbers, respectively. (a, b) and $[a, b]$ denote the open and closed intervals from a to b. If A is a set, then $|A|$ is the cardinality of

A. Given a function f, its domain and range are denoted by $Dom(f)$ and $Ran(f)$, respectively. $f^{(k)}$ refers to the k-th derivative of f (if it exists). Occasionally, we shall use $f'(x)$ in place of $f^{(1)}(x)$. If a function $f(\cdot)$ is k times continuously differentiable on a given interval I, then we write $f \in C^k(I)$. C^∞ functions are called *smooth* functions. The term "Propositions" refers to results cited from external sources.

The concepts of real analytic functions [21, pp. 1-3], absolute monotonic and completely monotonic functions [38, pp. 144-145] and hypergeometric functions [9, pp. 202], are central to what follows; for convenience they are reviewed below.

Definition 2.1 (Real Analyticity) Let $U \subseteq \Re$ be an open set. A function $f : U \rightarrow \Re$ is said to be *real analytic*[4] at $x_0 \in U$, if the function may be represented by a convergent power series on some interval of positive radius centered at x_0, i.e. , $f(x) = \sum_{j=0}^{\infty} a_j(x - x_0)^j$. The function is said to be *real analytic on* $V \subseteq U$, if it real analytic at each $x_0 \in V$. ∎

Definition 2.2 (Monotonicity) A function $f : \Re \rightarrow \Re$ is *absolutely monotonic* in (a, b) if it has non-negative derivatives of all orders there, i.e. , $f \in C^\infty((a, b))$ and,

$$f^{(k)}(x) \geq 0 \qquad a < x < b,\ k = 0, 1, 2 \ldots \tag{2.1}$$

A function $f : \Re \rightarrow \Re$ is *completely monotonic* in (a, b), iff $f(-x)$ is absolutely monotonic in $(-b, -a)$. Equivalently, f is completely monotonic in (a, b) iff $f \in C^\infty((a, b))$ and,

$$(-1)^k f^{(k)}(x) \geq 0 \qquad a < x < b,\ k = 0, 1, 2 \ldots \tag{2.2}$$

A function $f : \Re \rightarrow \Re$ is *completely convex* in (a, b), iff $f \in C^\infty((a, b))$, and for all non-negative k and $x \in (a, b)$, $(-1)^k f^{(k)}(x) \geq 0$. ∎

A fundamental property of absolutely monotone and completely monotone functions is that they are necessarily real analytic on their domains (S. Bernstein's theorem[5] [12, pp. 184]). Additionally, if f is absolutely monotone on an interval $I \subseteq \Re$, then it is non-negative, non-decreasing, convex, and continuous on I.

Definition 2.3 The generalized Gauss hypergeometric (GH) series ${}_pF_q(\alpha_1, \ldots, \alpha_p; \gamma_1, \ldots, \gamma_q; z)$ is defined by,

$${}_pF_q(\alpha_1, \ldots, \alpha_p; \gamma_1, \ldots, \gamma_q; z) = \sum_{k=0}^{\infty} \frac{(\alpha_1)_k(\alpha_2)_k \cdots (\alpha_p)_k}{(\gamma_1)_k(\gamma_2)_k \cdots (\gamma_p)_k} \frac{z^k}{k!} \quad \forall\, i : \gamma_i \neq 0, -1, -2, \cdots \tag{2.3}$$

where $(a)_n = (a)(a + 1)\cdots(a + n - 1)$ is the *rising factorial* or Pochhammer's polynomial in a. By definition, $(a)_0 = 1$. The α_i's are the *numeratorial parameters*, and the γ_i's are referred to as the *denominatorial parameters* of the GH series. ∎

[4] Real analytic functions are also referred to as regular, holomorphic, and monogesic functions.

[5] In full, Bernstein's theorem asserts that given a function $f(x)$, if infinitely many of its derivatives $f^{(n_1)}$, $f^{(n_2)}$, $\cdots$ are of constant sign in the open interval I ($f^{(n_k)}$ is the n_kth derivative of f), and if the sequence $n_1, n_2, \cdots$ does not increase more rapidly than a geometric progression, (i.e. there is a fixed quantity C, such that $\forall k\ n_{k+1}/n_k < C$), then $f(x)$ is analytic on the interval I [12, pp. 184].

In particular, the *classical* GH series[6] in z, ${}_2F_1(\alpha, \beta; \gamma; z)$ is defined by,

$$ {}_2F_1(\alpha, \beta; \gamma; z) \equiv F(\alpha, \beta; \gamma; z) = \sum_{k=0}^{\infty} \frac{(\alpha)_k (\beta)_k}{(\gamma)_k} \frac{z^k}{k!} \tag{2.4} $$

Remark 2.1 The ${}_pF_q$ representation of a hypergeometric series, though a standard one, can be confusing. For example, the series $\sum_{k=0}^{\infty} \frac{z^k}{k!}$ could be viewed as ${}_0F_0(;;z)$, or as ${}_1F_1(1;1;z)$, or as ${}_3F_3(1,1,1;1,1,1;z)$, etc. We shall henceforth use the "minimum" representation, in this case ${}_0F_0(;;z)$. In the case of $\sum_{k=0}^{\infty} \frac{z^k}{k!}$ it is not necessary to have a non-empty list of numeratorial and denominatorial parameters.

Remark 2.2 In general, the parameters α_i's and γ_i's, as well as the variable z, are allowed to be complex; however, we follow common practice and restrict our attention to real values i.e. $\forall i$: $\alpha_i, \gamma_i, z \in \Re$. Even with this restriction, the hypergeometric function is amazingly versatile. Spanier and Oldham list over 170 functions that are representable in terms of the hypergeometric function [32, pp. 149-165]. The hypergeometric function is a periodic table *a la* Mendeleev for mathematical functions; different functions get neatly pegged into various groups[7] by the values of the parameters and the form of the dependent variable.

3 Simple & Hyperbolic Sigmoids

Definition 3.1 (Simple sigmoids) A function $\sigma : \Re \to (-1, 1)$ is said to be a *simple sigmoid* if it satisfies the following conditions:

1. $\sigma(\cdot)$ is a smooth function, i.e., $\sigma(x)$ is C^∞.
2. $\sigma(\cdot)$ is an odd function, i.e., $\sigma(-x) = -\sigma(x)$.
3. $\sigma(\cdot)$ has $y = \pm 1$ as horizontal asymptotes, i.e., $\lim_{x \to \infty} \sigma(x) = 1$.
4. $\sigma(x)/x$ is a completely convex function in $(0, 1)$. ∎

Simple sigmoids are required to be odd smooth functions bound by horizontal aymptotes; constraints impose a degree of standardization on the kinds of sigmoids being considered. The following results clarify the implications of the fourth constraint.

Proposition 3.1 : [10, Theorem 3, pp. 222] A function $f : (0, 1) \to \Re$ is absolutely monotone on $(0, 1)$ iff it possesses a power series expansion with non-negative coefficients, converging for $0 < x < 1$. ∎

Lemma 3.1 : A function $f : (0, 1) \to \Re$ is completely monotone on $(0, 1)$ iff it possesses an alternating power series expansion, converging for $0 < x < 1$.

[6] The classical GH series is referred to as the *Gauss function* in the literature [32, pp. 599].

[7] "There must be many universities to-day where 95 per cent, if not 100 per cent, of the functions studied by physics, engineering, and even mathematics students, are covered by this single symbol F(a, b; c; x)." — W. W. Sawyer, cited by Graham *et. al.* [15, pp. 207]

Proof[8]: If f is completely monotone in $(0,1)$, then the power series expansion of f in $(0,1)$ has to be alternating (because, $(-1)^k f^{(k)} \geq 0$). On the other hand, consider an alternating power series $f(x)$ converging for all $0 < x < 1$ and its derivatives:

$$\begin{aligned} f(x) &= a_0 - a_1 x + a_2 x^2 - a_3 x^3 + \cdots \quad a_i \geq 0 (0 < x < 1) \qquad (3.1)\\ (-1)f^{(1)}(x) &= +a_1 - 2a_2 x + 3a_3 x^2 + \cdots \\ f^{(2)}(x) &= 2a_2 - 6a_3 x + \cdots \\ &\ldots\ldots \end{aligned}$$

From real analysis we know that each of $(-1)^n f^{(n)}(x)$ has the same convergence properties as Equation (3.1). Also, the sum of a convergent infinite alternating series is always less than or equal to the first term. This fact, along with the above equations implies that $(-1)^k f^{(k)}(x) \geq 0$ i.e., $f(x)$ is completely monotone on (0, 1). ∎

Corollary 3.1 $\sigma(x)/x$ is a completely convex function in $(0,1)$ iff $\sigma(\sqrt{x})/\sqrt{x}$ is a completely monotone function in $(0,1)$.

Proof: If $\sigma(x)/x$ is completely convex in $(0,1)$, then it has to be analytic in $(0,1)$ [38, 177-179]. Also, $\sigma(x)/x$ is an even function, implying that its power series expansion will consist only of even powers in x, which alternate in sign. From Lemma 3.1, $\sigma(\sqrt{x})/\sqrt{x})$, will hence be completely monotone in $(0,1)$. The same argument suffices for the converse. ∎

If a simple sigmoid is also *strictly* increasing, then a much stronger statement can be made, as demonstrated by the following proposition.

Proposition 3.2 : [21, pp. 9] Let $y = \sigma(x)$ be a strictly increasing simple sigmoid (i.e. $\forall\, x \in \Re$, $\sigma'(x) > 0$). Then:

1. $\eta \equiv \sigma^{-1} : (-1,1) \rightarrow \Re$ exists.
2. $\eta(y)$ is a strictly increasing function, analytic in the interval $(-1,1)$.
3. $\eta'(y) = 1/\sigma'(\eta(y))$, where η' and σ' are the first derivatives of η and σ respectively.
4. $\eta(y)/y$ is absolutely monotone in $(0,1)$. ∎

Remark 3.1 If $\sigma(x)/x$ is completely monotone on $(0,1)$ and σ is invertible then $\eta(y)/y$ is absolutely monotone on $(0,1)$, where η denotes the inverse of σ. The converse is also true, and is an immediate consequence of Lemma 3.1.

Remark 3.2 Since a simple sigmoid has two *horizontal* asymptotes, it implies that its inverse (if it exists) will have two vertical asymptotes (i.e. $\lim_{y \rightarrow \pm 1} \eta(y) \rightarrow \pm\infty$). It will be seen that as they have been defined, sigmoids and their inverses are quite similar; both are odd, increasing, univalent, analytical functions. However, the two differ fundamentally in that sigmoids are aymptotically *bounded*, while their inverses are not.

[8] Lemma 3.1 appears to be "folklore"; we have been unable to find a reference.

Simple sigmoids encompass many of the often used sigmoids described by formulae. The hyperbolic tangent and its close relative, the "exponential" or logistic sigmoid, are often used in many neural network theoretical studies and applications. For example, most of the spin-glass models of the Hopfield net use the hyperbolic tangent.[9] The hyperbolic tangent has, among others, the following properties:

1. It is an odd, strictly increasing analytical function, asymptotically bounded by the lines $y = \pm 1$.
2. Its inverse $\tanh^{-1}(y)$ has a GH expansion given by $yF(1, 1/2; 3/2; y^2)$.
3. The first derivative of $\tanh^{-1}(y)$ is given by $1/(1-y^2) = {}_1F_0(1;;y^2)$, i.e. , the GH expansion of the first derivative of $\tanh^{-1}(y)$ is dependent on only *one* numeratorial parameter.

It can be shown that many other simple sigmoids, such as Elliot's sigmoid [8], the Gudermannian (section 4.2) etc. , also have inverses with classical GH series representations.[10] The function $\tanh^{-1}(y)/y$ satisfies a second order linear homogeneous differential equation, with three *regular* singular points, located at $0, 1$ and ∞. A sigmoid with a similar analytical behavior could be expected to have an inverse that is a solution to some second order Fuchsian equations[11]. Since any second order Fuchsian equation with three singularities can be transformed into the Gauss hypergeometric differential equation, one solution of which is the classical GH series (Klein-Bôcher theorem) [37, pp. 203], it follows that the inverses would have classical series expansions. These considerations motivate the following definition.

Definition 3.2 (Hyperbolic sigmoids) A function $\sigma : \Re \rightarrow (-1, 1)$ is said to be a *hyperbolic sigmoid* function if it satisfies the following conditions:

1. σ is a real analytic, odd, strictly increasing sigmoid, such that $\lim_{x\rightarrow\infty} \sigma(x) = 1$.
2. Let $\eta : (-1, 1) \rightarrow \Re$ denote the inverse of σ, and η' its first derivative. Then,
   (a) $\eta(y)/y$ has a Gauss hypergeometric series expansion in y^2 with *at most three* parameters.
   (b) $\eta'(y)$ has a Gauss hypergeometric series expansion *in* y^2 with *at most one* parameter.

$\blacksquare$

4 Characterization: Inverse hyperbolic sigmoids

The following result is a complete characterization for the inverses of hyperbolic sigmoids. Proofs are presented in the appendix.

[9] Stochastic versions of neural nets often start by replacing a set of deterministic state assignment rules, by probabilistic ones, obtained from some distribution — usually the Gibbsian distribution (e.g. Boltzman machines, Stochastic Hopfield models etc.). Computing expected values for the states of the system then leads to the hyperbolic tangent function. See Hertz et. al. for a typical example [18, pp. 28].

[10] The phenomenon is not unduly surprising. A heuristic argument may be given as follows: If the graphs of two functions "look" the same, their respective differential equations are usually members of the same family.

[11] Fuchsian equations are linear differential equations each of whose singular points are regular [31, pp. 143-168]. $\tanh^{-1}(x)/x$ satisfies such an equation.

Theorem 4.1 (Inverses) Let $y = \sigma(x)$ be a hyperbolic sigmoid, and let $\eta : (-1,1) \to \Re$ be its inverse. Then, either

$$\eta(y) = yF(\alpha, \frac{1}{2}; \frac{3}{2}; y^2) = y \sum_{k=0}^{\infty} \frac{(\alpha)_k}{(2k+1)} \frac{y^{2k}}{k!} \qquad \alpha \geq 1 \tag{4.1}$$

or

$$\eta(y) = yF(\alpha, -; -; y^2) = \frac{y}{(1 - y^2)^{\alpha}} \qquad \alpha > 0 \tag{4.2}$$

where, by $F(\alpha, -; -; y^2)$, we mean $F(\alpha, \beta; \beta; y^2)$ $(\beta \in \Re)$. ∎

Notation: Each inverse hyperbolic sigmoid is denoted by η_α and is characterized by a single parameter α.

Corollary 4.1 The set of hyperbolic sigmoids is a proper subset of the set of simple sigmoids.

A proof for Corollary 4.1 may be given along the following lines. If σ is a hyperbolic sigmoid, then it is simple on the interval $(-1,1)$: For, from Theorem 4.1, the series representation for its inverse in $(-1,1)$ has non-negative coefficients, and this implies $\eta(y)/y)$ is absolutely monotone (Proposition 3.1). Hence $\sigma(x)/x$ is completely monotone, and therfore simple. (Lemma 3.1 and Remark 3.1). The converse is *not* true. Simple sigmoids need not be hyperbolic. The error function $\text{erf}(\cdot)$ is simple, but one can use Carlitz's study of the function to show that it does *not* have an inverse representable by a classical hypergeometric series [4]. It follows that $\text{erf}(\cdot)$ is not a hyperbolic sigmoid, and hence the set of hyperbolic sigmoids is a proper subset of the set of simple sigmoids. ∎

For specific values of its parameters, the hypergeometric function often reduces to other well known special functions. When inverse hyperbolic sigmoids are characterized by Equation (4.1), there is an intimate connection with Euler's *incomplete Beta* function.

Proposition 4.1 : [32, pp. 573] Let α, β and γ be such that, $\beta = \gamma - 1$. Then,

$$F(\alpha, \gamma - 1; \gamma; z) = \frac{(\gamma - 1)B(\gamma - 1; 1 - \alpha; z)}{z^{\gamma - 1}} \tag{4.3}$$

where $B(v; u; z)$ is the *incomplete beta* function, defined by $\int_0^z t^{v-1}(1-t)^{u-1}dt$, where $0 \leq z < 1$. In particular, $\frac{1}{2} B(1/2; 1 - a; z^2) = \int_0^{\tanh^{-1}(z)} \cosh^{2(a-1)}(t)\, dt$. ∎

Spanier and Oldham give a detailed description of the many properties of this important special function [32, pp. 573-580]. The following corollary is an immediate consequence of Theorem 4.1 and Proposition 4.1. It gives the connection between inverse hyperbolic sigmoids, and Euler's incomplete Beta function.

Corollary 4.2 If $\eta_\alpha(y) = yF(\alpha, 1/2; 3/2; y^2)$, then $\eta_\alpha(y) = \frac{1}{2}B(1/2; 1 - \alpha; y^2)$. ∎

The relationship between hyperbolic sigmoids and the incomplete Beta function, also makes explicit the relationship between $\tanh^{-1}(\cdot)$, and inverse hyperbolic sigmoids of form $yF(\alpha; 1/2; 3/2; y^2)$. Other consequences include:

1. The availability of good approximations for small values of y and $(1 - y)$.
2. Rapidly converging series expansions for y close to 1.
3. Connections with other indefinite integrals of powers of trigonometric or hyperbolic functions.
4. Connections with statistics *via* the function $I_y(p, q)$ [32, pp. 573-580].

When inverse hyperbolic sigmoids are characterized by Equation (4.2), we can use the identity,

$$\cosh(\tanh^{-1}(y)) = \frac{1}{\sqrt{1 - y^2}} \tag{4.4}$$

to show that,

$$y \cosh^{2a}(\tanh^{-1}(y)) = \frac{y}{(1 - y^2)^a} \tag{4.5}$$

The fundamental role played by the hyperbolic tangent is once again evident. Here, it relates the two types of hyperbolic sigmoids defined by Equations 4.1 and 4.2.

4.1 New Inverses from Old

Theorem 4.1 makes it possible to generate new inverse hyperbolic sigmoids from others. The key idea is that if $yF(\alpha, 1/2; 3/2, y^2)$ is an inverse hyperbolic sigmoid, then so is $yF(\alpha + 1, 1/2; 3/2; y^2)$. A similar statement may be made for inverse hyperbolic sigmoids of the form $yF(\alpha, -; -; z^2)$. GH functions such as $F(\alpha, \beta; \gamma; z)$, and $F(\alpha + 1, \beta; \gamma; z)$ are said to be *contiguous*, and there exist several differential identities between them [9, pp. 102-104]. Lemma 4.1 is a straightforward consequence of three such identities.

Lemma 4.1 If $\eta_\alpha : (-1, 1) \to \Re$ is an inverse hyperbolic sigmoid, then the functions $\eta_{\alpha+1}$ and $\eta_{\alpha-1}$ defined by:

$$\eta_{\alpha+1}(y) \equiv \frac{y^{2(1-\alpha)}}{2\alpha} \frac{d}{dy} (y^{2\alpha-1} \eta(y)) \qquad \alpha \geq 1 \tag{4.6}$$

$$\eta_{\alpha-1}(y) \equiv \frac{-y^{2\alpha-1}}{(2\alpha - 3)(1 - y^2)^{\alpha-2}} \frac{d}{dy} \left\{ \frac{(1 - y^2)^{\alpha-1}}{y^{2(\alpha-1)}} \eta(y) \right\} \qquad \alpha \geq 2 \tag{4.7}$$

are also inverse hyperbolic sigmoids. Also, there exist functions $K_1(\alpha, z)$, $K_2(\alpha, z)$ and $K_3(\alpha, z)$ such that, following relation holds:

$$K_1(\alpha, z)\, \eta_{\alpha-1}(y) + K_2(\alpha, z)\, \eta_\alpha(y) + K_3(\alpha, z)\, \eta_{\alpha+1}(y) = 0 \tag{4.8}$$

Proof: Equation (4.6) that defines $\eta_{\alpha+1}(y)$ results from the following identity:

$$(\alpha)_n\, z^{\alpha-1}\, F(\alpha + n, \beta; \gamma; z) = \frac{d^n}{dz^n} [z^{\alpha+n-1}\, F(\alpha, \beta; \gamma; z)] \tag{4.9}$$

In the following we will use $F(\theta)$ as an abbreviation for $F(\theta;\beta;\gamma;z)$. Equation (4.7) follows from the identity:

$$(\gamma - \alpha)_n\, z^{\gamma - \alpha - 1}(1 - z)^{\alpha + \beta - \gamma - n}\, F(\alpha - n) = \frac{d^n}{dz^n}[z^{\gamma - \alpha + n - 1}(1 - z)^{\alpha + \beta - \gamma}F(\alpha)] \tag{4.10}$$

Equation (4.8), relating $\eta_{\alpha - 1}(y)$, $\eta_\alpha(y)$ and $\eta_{\alpha - 1}(y)$ is a consequence of the identity:

$$(\gamma - \alpha)F(\alpha - 1) + (2\alpha - \gamma - \alpha z + \beta z)F(\alpha) + \alpha(z - 1)F(\alpha + 1) = 0 \;\blacksquare \tag{4.11}$$

Inverse hyperbolic sigmoids come in two flavors; one form has three parameters (Equation (4.1)), while the other has two "missing" parameters (Equation (4.2)). Subject to a minor condition, the latter form is always obtainable from the former:

Lemma 4.2 Let $\eta_\alpha = yF(\alpha, 1/2; 3/2; y^2)$, where $\alpha > 1$. Then the function $\eta_{\alpha - 1}$ defined by:

$$\eta_{\alpha - 1}(y) \equiv y(1 - y^2)\frac{d}{dy}\eta_\alpha(y) = yF(\alpha - 1, -; -; y^2) \tag{4.12}$$

is an inverse hyperbolic sigmoid, with parameter $\alpha - 1$. $\blacksquare$

For inverse hyperbolic sigmoids with "missing" parameters, there is a very simple composition rule;

Lemma 4.3 If $\eta_\alpha(y) = y/(1 - y^2)^\alpha$ and $\eta_{\alpha'}(y) = y/(1 - y^2)^{\alpha'}$ are two inverse hyperbolic sigmoids with $\alpha, \alpha' > 0$, then the function $(\eta_\alpha(y)\eta_{\alpha'}(y))/y$ is also an inverse hyperbolic sigmoid with parameter $(\alpha + \alpha')$. $\blacksquare$

In general, the set of inverse hyperbolic sigmoids is *not* closed under multiplication or addition. But if η_α and $\eta_{\alpha'}$ are inverses of two hyperbolic sigmoids then their sum would also be an inverse hyperbolic sigmoid η_μ for some $\mu \in \Re$, i.e., $\eta_\alpha + \eta_{\alpha'} = K\eta_\mu$, for some K, if and only if

$$\eta_\alpha + \eta_{\alpha'} = K\eta_\mu \Rightarrow (\alpha)^n + (\alpha')^n = K(\mu)^n \quad \forall n \geq 1 \tag{4.13}$$

which in turn, is possible[12] if and only if $\alpha = \alpha'$, or $\alpha = 0$, or $\alpha' = 0$.

The definition of hyperbolic sigmoids implies that their inverses have GH expansions in y^2. Theorem 4.2 relaxes this requirement by only requiring GH expansions in some odd, injective C^1 function $g(y)$. A proof is provided in Appendix I.

Theorem 4.2 Let $\sigma : \Re \to (-1, 1)$ be a real analytic, odd, strictly increasing sigmoid, such that its inverse $\eta : (-1, 1) \to \Re$ has a GH series expansion in some injective, odd, increasing C^1 function $g(\cdot)$, with at most three parameters, convergent in $(-1, 1)$. Also let η' have a GH series expansion *in* $g(\cdot)$, with at most one parameter. Then, either

$$\eta(y) = g(y)F(\alpha, \frac{1}{2}; \frac{3}{2}; (g(y))^2) = g(y)\sum_{k=0}^{\infty} \frac{(\alpha)_k}{2k + 1}\frac{(g(y))^{2k}}{k!}, \qquad \text{for } \alpha \geq 1, \tag{4.14}$$

$$\text{or} \qquad \eta(y) = g(y)F(\alpha, -; -; (g(y))^2) = \frac{g(y)}{(1 - (g(y))^2)^\alpha}, \qquad \text{for } \alpha > 0 \tag{4.15}$$

provided $\lim_{y \to 1} \dfrac{g'(y)}{(1 - y^2)^\alpha} \to \infty$, where $g'(\cdot)$ is the first derivative of $g(\cdot)$. $\blacksquare$

[12] Equation (4.13), with $K = 1$, provides an amusing application for Fermat's last theorem; if we accept that for all $n > 2$, there cannot exist positive integers a, b and c satisfying the identity $a^n + b^n = c^n$, then we may conclude that the sum of inverse hyperbolic sigmoids with different integral parameters cannot be an inverse hyperbolic sigmoid with an integral parameter.

In the case $g(y) = y$, we obtain the characterization for inverse hyperbolic sigmoids. Another interesting special case is when $g(y) = \eta(y)$, where $\eta(y)$ is an inverse hyperbolic sigmoid (since $\eta(y)$ is an injective, smooth, odd, increasing function the conditions of the theorem are satisfied). The elementary composition rules presented here allows the generation of an infinite variety of inverse hyperbolic sigmoids[13]. The next section presents some examples.

4.2 Examples

Any function of the form $y/(1 - y^2)^\alpha$, where $\alpha > 0$, is the inverse of a hyperbolic sigmoid. For example, for $\alpha = 2$, the function $y/\sqrt{1 - y^2}$ is the inverse of the hyperbolic sigmoid $x/\sqrt{1 + x^2}$.

Of all inverse hyperbolic sigmoids of the form $yF(\alpha, 1/2; 3/2; y^2)$, the function $\tanh(\cdot)$ is noteworthy; firstly, it corresponds to the case $\alpha = 1$, secondly, all inverse hyperbolic sigmoids with integral values of α may be generated from $\tanh(x)$ by a process of differentiation (Lemma 4.1), and thirdly, it is a function often encountered in neural nets [19]. As was mentioned in the Introduction, the logistic function may be thought of as a translated and scaled version of the hyperbolic tangent.

There is a good example of the hypergeometric composition described in Theorem 4.2. Since $\tan(\beta y)$ is an odd, injective, smooth, increasing function of y (for some constant $\beta > 0$), from Theorem 4.2, one may conclude that for positive α the function, $\tan(\beta y)F(\alpha, 1/2; 3/2; \tan^2(\beta y))$ is the inverse of some real analytic, odd, strictly increasing sigmoid. It turns out that the inverse Gudermannian function[14], may be obtained from this function, by choosing $\alpha = 1$ as follows:

$$\begin{aligned} \mathrm{gd}^{-1}(y) &= \ln(\sec(y) + \tan(y)) \text{ for } -\frac{\pi}{2} < y < \frac{\pi}{2} \\ &= 2\tan(y/2)\, F(1, 1/2; 3/2; \tan^2(y/2)) \end{aligned}$$

Many such examples could be generated.[15]

5 Characterization: Hyperbolic Sigmoids

It is often desirable and necessary to work with sigmoids themselves, rather than their inverses. In this section, we obtain power series expansions of sigmoids.

If $x = \eta(y)$ is an inverse hyperbolic sigmoid, then $\sigma \equiv \eta^{-1}$ must have a Maclaurin series expansion of the following form: $y = \sigma(x) = x \sum_{k=0}^{\infty} \dfrac{b_{2k+1}}{(2k+1)!} x^{2k}$. We are interested in determining the coefficients $\{b_{2l+1}\}_{l=0}^{\infty}$ associated with the inverse hyperbolic sigmoids: $\dfrac{y}{(1 - y^2)^\alpha}$ and $yF(\alpha, 1/2; 3/2; y^2)$.

5.1 Hyperbolic Sigmoids of the First Kind

When an inverse hyperbolic sigmoid is of the form $y/(1 - y^2)^\alpha$, a remarkably explicit form for the coefficients $\{b_{2l+1}\}_0^\infty$ may be given:

[13] An intriguing case is Elliot's piecewise rational sigmoid [8], defined as $\sigma(x) = y/(1 + |x|)$. Although its inverse $\eta(y) = y/(1 - |y|)$ does not fit in an obvious way into the framework developed in the last few sections, it is fairly simple to relax the conditions placed on $g(y)$, in Theorem 4.2, so as to include this sigmoid as well.

[14] The inverse Gudermannian function finds use in relating circular and hyperbolic functions, without the use of complex functions.

[15] In particular, [32, pp. 149-165], [16, pp. 196-198] are minelodes of such functions and expansions.

Theorem 5.1 (Hyperbolic sigmoids - I) If the inverse sigmoid is given by $y/(1 - y^2)^\alpha$, $\alpha > 0$, then in some neighborhood of the origin, we have the valid expansion $\sigma(x) = x \sum_{k=0}^{\infty} \frac{b_{2k+1}}{(2k+1)!} x^{2k}$ where,

$$b_{2k+1} = (-1)^k (2k+1)! \binom{(2k+1)\alpha}{k} \tag{5.1}$$

Proof : (see Appendix I) ∎

5.2 Hyperbolic Sigmoids of the Second Kind

When an inverse hyperbolic sigmoid is of the form $x = yF(\alpha, 1/2; 3/2; y^2)$, the problem is much harder. The Lagrange inversion formula leads to an intractable expression. Kamber's formulae, as presented by Goodman, can be used to give explicit expressions for the coefficients [14, Theorem 7, pp. 56-57]. Unfortunately, the resulting expressions involve determinants, and are of little computational value. The method of repeated differentiation is more successful. The starting point for this line of attack is the observation that if $x = \eta(y)$ is an inverse hyperbolic sigmoid, then:

$$\frac{dx}{dy} = \frac{d}{dy}\eta(y) = \eta'(y) = \frac{1}{(1 - y^2)^\alpha} \tag{5.2}$$

From Theorem 3.2, we see that for $y = \sigma(x)$,

$$\frac{dy}{dx} = \frac{d}{dx}\sigma(x) = \frac{1}{\eta'(y)} = (1 - y^2)^\alpha \tag{5.3}$$

By virtue of Equation (5.3), we can compute the higher derivatives of $\sigma(\cdot)$ and hence compute $b_{2k+1} = \left.\frac{d^{2k+1}\sigma(x)}{dx^{2k+1}}\right|_{x=0}$. Note that $\frac{dy}{dx}$ is expressed in terms of y; this necessitates the use of the chain rule. For example, to calculate the second derivative:

$$\frac{d^2y}{dx^2} = \left(\frac{d}{dy}(1 - y^2)^\alpha\right)\frac{dy}{dx} = (1 - y^2)\left(\frac{d}{dy}(1 - y^2)^\alpha\right) \tag{5.4}$$

The following theorem presents an efficient way to implement this procedure.

Theorem 5.2 (Hyperbolic sigmoids — II A) Let the inverse hyperbolic sigmoid be $\eta_\alpha = yF(\alpha, 1/2; 3/2; y^2)$, and $\sigma \equiv \eta_\alpha^{-1}$. Let $D \equiv \frac{d}{dx}$. Then,

$$D^n(y) = D^n(\sigma(x)) = G_{n-1}(y)(1 - y^2)^{n\alpha} \tag{5.5}$$

where $G_n : (-1, 1) \to \Re$ is a function satisfying the recursion

$$\begin{aligned} G_0(y) &= 1, \\ G_n(y) &= \frac{d}{dy}G_{n-1}(y) - \frac{2yn\alpha}{1 - y^2}G_{n-1}(y) \qquad n \geq 1 \end{aligned} \tag{5.6}$$

In particular, $b_{2k} = 0$, and $b_{2k+1} = D^{2k+1}(\sigma(x)) = G_{2k}(0)$.

Proof: Theorem 5.2 is easily proved by an induction argument on n. ∎

While the procedure implicit in Theorem 5.2 is efficient, it does involve the computation of the derivative of $G_n(y)$. Equation (5.6) is a partial difference equation with variable coefficients. Therefore there is little hope of solving it in any generality and obtaining a closed form expression. Even more sophisticated methods such as Truesdell's generating function technique and Weisner's group theoretic approach (see [25]), do not give any special insight into the nature of the polynomials $G_n(y)$.[16] The next theorem offers a somewhat different approach to the method of repeated derivatives.

Theorem 5.3 (Hyperbolic sigmoids - II B) Let $\sigma(x) = \sum_{k=0}^{\infty} \frac{b_{2k+1}}{(2k+1)!} x^{2k}$ be an expansion for a hyperbolic sigmoid, whose inverse is of the form $yF(\alpha, 1/2; 3/2; y^2)$, valid in some neighborhood of the origin. Then $b_{2k} = 0$, and $b_{2k+1} = C(2k+1, k)$, where the sequence $C(n,k)$ satisfies:

$$\begin{aligned} C(1,0) &= 1 \\ C(n,k) &= 0 \qquad \forall\, k \geq n,\, k < 0 \\ C(n+1,k) &= (2k-n+1)C(n,k) - 2(n\alpha - k + 1)C(n,k-1) \quad n \geq 1 \end{aligned} \tag{5.7}$$

n and k are natural numbers, $D^n(\sigma(x))$, the nth derivative of σ, is given by:

$$D^n(y) = D^n(\sigma(x)) = \sum_{k=0}^{n-1} C(n,k) y^{2k-n+1}(1-y^2)^{n\alpha - k}; \text{ for } n \geq 1 \tag{5.8}$$

Proof: See Appendix I. ∎

The recursive system described by Equation (5.7) does not involve any differentiation. The desired value b_{2k+1} may be obtained by computing the value of $C(2k+1, k)$. Equation (5.8) gives information about the shapes of the derivatives of the hyperbolic sigmoid. From Equation (5.7),

$$b_1 = 1 \qquad\qquad b_3 = -2\alpha, \tag{5.9}$$

$$b_5 = 4\alpha(7\alpha - 3) \qquad b_7 = -8\alpha(127\alpha^2 - 123\alpha + 30) \tag{5.10}$$

Theorem 5.3 may be viewed as a generalization of the work of Minai and Williams on the derivatives of the logistic sigmoid [26]. They obtained relations similar to Equation (5.7)[17]. In general, Equation (5.7) is a partial difference equation with variable coefficients, and the system does not appear to be related to any well known sets of numbers. A closed form solution for the numbers $C(n,k)$ appears to be intractable.

6 Applications

In this section, we present two applications. The first shows that if the neural network transfer function is a hyperbolic sigmoid, then the dynamical equations describing the Hopfield neural network

[16] Equation (5.6) is a differential-difference system of the *ascending* type; it can then be shown that the polynomials $\{G_n(y)\}_{n=1}^{\infty}$ satisfy Truesdell's F-equation. Unfortunately, the resulting generating function for $G_n(y)$ is too complicated for any practical use.

[17] Interestingly, in the case of the logistic sigmoid, these relations happened to be the recursions corresponding to the Eulerian numbers [15, pp. 253-257]; in other words, the coefficients arising in the computation of higher order derivatives of the logistic sigmoid turn out to be the Eulerian numbers.

[19] can be transformed into a set of non-homogeneous associated Legendre differential equations. Some conclusions regarding the behavior of the Hopfield model, as the outputs saturate (i.e. output → ±1) can then be drawn.

The second application derives an interesting connection between Fourier transforms and 1-hidden layer feedforward nets (1-HL nets). Subject to an additional minor constraint, we show that the use of 1-HL nets with simple sigmoidal transfer functions for function approximation is tantamount to assuming that the function being approximated is the product of two functions; one the derivative of a bounded non-negative function, and the other satisfying some linear n-th order differential equation, where n is the number of nodes in the hidden layer.

6.1 Continuous Hopfield nets & Legendre Differential Equations

The continuous Hopfield network model [19] with N neurons is described by the following dynamics:

$$\frac{du_i}{dt} + g_i u_i = \sum_j T_{ij} v_j + I_i = E_i = -\frac{\partial E}{\partial v_i} \quad \forall\, i \in \{1, \ldots, N\} \tag{6.1}$$

where u_i and v_i are the net input and net output of the i^{th} neuron, respectively, I_i is a constant external excitation, and E is the so called "energy" of the network, given by:

$$E = -\frac{1}{2}\sum_{i,j} T_{ij} v_i v_j = \sum_i v_i \frac{\partial E}{\partial v_i} = -\sum_i v_i E_i \tag{6.2}$$

Assume $-1 < v_i < 1$. Let $v_i = \sigma(u_i)$, where $\sigma(\cdot)$ is a hyperbolic sigmoid. Let $\eta \equiv \sigma^{-1}$, or, $u_i = \eta(v_i)$. There are two cases to consider.

Case I: $\eta(v_i) = v_i F(\alpha, 1/2; 3/2; v_i^2)$. In this case,

$$\frac{d\eta}{dv_i} = \frac{1}{(1 - v_i^2)^\alpha} \tag{6.3}$$

and substituting Equation (6.3) in Equation (6.1), we get:

$$\frac{1}{1 - v_i^2}\frac{dv_i}{dt} + g_i u_i = E_i \tag{6.4}$$

The following sequence of operations are applied to Equation (6.4):

1. Substitute $y_i = \frac{dv_i}{dt}$, and differentiate with respect to v_i,
2. multiply throughout by $(1 - v_i^2)^{\alpha+1}$, and
3. differentiate once more with respect to v_i.

Equation (6.4) is then transformed into:

$$(1 - v_i^2)\frac{d^2 y_i}{dv_i^2} - 2(1 - \alpha) v_i \frac{dy_i}{dv_i} + 2\alpha y_i = Q_i \tag{6.5}$$

where $Q_i = \frac{d}{dv_i}[(1 - v_i^2)^{\alpha+1}\frac{dE_i}{dv_i}] + 2g_iv_i$. Finally, put $y_i = z_i(1 - 1\, v_i^2)^{\alpha/2}$ in Equation (6.5) yielding,

$$(1 - v_i^2)\frac{d^2z_i}{dv_i^2} - 2v_i\frac{dz_i}{dv_i} + [\alpha(\alpha + 1) - \frac{\alpha^2}{1 - v_i^2}]z_i = R_i \tag{6.6}$$

where $R_i = (1 - v_i^2)^{-\alpha/2}\, Q_i$. Recall that the associated Legendre differential equation is of the form [32, pp. 594-597],

$$(1 - x^2)\frac{d^2f}{dx^2} - 2x\frac{df}{dx} + \left[\nu(\nu + 1) - \frac{\mu^2}{1 - x^2}\right]f = 0 \tag{6.7}$$

It is clear that the left hand side in equation(6.6), is the associated Legendre differential equation with parameters $n = -\alpha$ (Equation (6.5) requires us to choose $\mu = -\alpha$, rather than $+\alpha$), and $\nu = \alpha$. In other words, the continuous Hopfield model with a neural transfer function given by $\eta(v_i) = v_iF(\alpha, 1/2; 3/2; v_i^2)$, is reducible to the non-homogeneous associated Legendre differential equation with parameters $\mu = -\alpha$ and $\nu = \alpha$.

Case II: $\eta(v_i) = v_iF(\alpha, -; -; v_i^2)$. An analogous approach leads to the very same conclusion, as in Case I, i.e., it is possible to transform the continuous Hopfield equation with the above transfer function to a non-homogeneous associated Legendre equation. However, the right hand side of the transformed equation is complicated and we do not consider this case further.

We emphasize that the link between the continuous Hopfield equation and the Legendre differential equation is not accidental, given that it can be established for all hyperbolic sigmoidal transfer functions. For $u_i = \tanh^{-1}(v_i)$, $\alpha = 1$, and the above equations have a rather elementary form.

An immediate application of the above transformation is in studying the saturation behavior of the Hopfield neural net. By saturation, we mean that the outputs of the neurons tend to ±1. This usually occurs when the network is heading towards a critical point (local or global) [19]. Saturation implies that as anode output $v_i \rightarrow \pm 1$, the quantity $R_i \rightarrow 0$. In other words, we may study the saturation behavior of the continuous Hopfield model by considering the homogeneous version of Equation (6.6) viz.,

$$(1 - v_i^2)\frac{d^2z_i}{dv_i^2} - 2v_i\frac{dz_i}{dv_i} + [\alpha(\alpha + 1) - \frac{\alpha^2}{1 - v_i^2}]z_i = 0 \tag{6.8}$$

From the theory of associated Legendre equations, it is seen that Equation (6.8) has a solution in terms of the associated Legendre functions, $P_\nu^{(\mu)}(x)$, and $Q_\nu^{(\mu)}(x)$ [9, pp. 121-179]. Here, $\mu = -\alpha$, $\nu = \alpha$, and $x \equiv v_i$, and we have:

$$\begin{aligned} z_i &= c_1P_\alpha^{(-\alpha)}(v_i) + c_2Q_\alpha^{(-\alpha)}(v_i) \\ \frac{y_i}{(1 - v_i^2)^{\alpha/2}} &= c_1P_\alpha^{(-\alpha)}(v_i) + c_2Q_\alpha^{(-\alpha)}(v_i) \\ \frac{1}{(1 - v_i^2)^{\alpha/2}}\frac{dv_i}{dt} &= c_1P_\alpha^{(-\alpha)}(v_i) + c_2Q_\alpha^{(-\alpha)}(v_i) \end{aligned} \tag{6.9}$$

Neglecting the effect of g_i, as is common practice, we obtain from Equation (6.4):

$$\frac{dv_i}{dt} \approx (1 - v_i^2)^\alpha\, E_i \tag{6.10}$$

Equation (6.9), in conjunction with Equation (6.10), implies:

$$E_i = \frac{\partial E}{\partial v_i} = (1 - v_i^2)^{\alpha/2}[c_1 P_\alpha^{(-\alpha)}(v_i) + c_2 Q_\alpha^{(-\alpha)}(v_i)] \tag{6.11}$$

Equation (6.11) in conjunction with Equation (6.2) implies that the overall energy at saturation may be written as follows:

$$E = \sum_i v_i E_i = \sum_i v_i (1 - v_i^2)^{\alpha/2}[c_1 P_\alpha^{(-\alpha)}(v_i) + c_2 Q_\alpha^{(-\alpha)}(v_i)] \tag{6.12}$$

E_i does not depend on E_j for $i \neq j$. Thus, to a crude first approximation, the Hopfield network "dissociates" at saturation, into independent units, and the quadratic energy function may be written as a linear sum of non-linear univalent functions, given by Equation (6.11) and Equation (6.12).

We wish to stress the *possibilities* revealed by dealing with the Hopfield equation in a general context. For example, in Equation (6.6),

$$(1 - v_i^2)\frac{d^2 z_i}{dv_i^2} - 2v_i\frac{dz_i}{dv_i} + [\alpha(\alpha + 1) - \frac{\alpha^2}{1 - v_i^2}]z_i = R_i \tag{6.13}$$

where $R_i = (1 - v_i^2)^{-\alpha/2} Q_i$, and $\frac{d}{dv_i}[(1 - v_i^2)^{\alpha+1}\frac{dE_i}{dv_i}] + 2g_i v_i$, consider the case when $Q_i = K$ is a constant. Then the above equation reduces to the non-homogeneous equation,

$$(1 - v_i^2)\frac{d^2 z_i}{dv_i^2} - 2v_i\frac{dz_i}{dv_i} + [\alpha(\alpha + 1) - \frac{\alpha^2}{1 - v_i^2}]z_i = K(1 - v_i^2)^{-\alpha/2} \tag{6.14}$$

which may be solved using the special function $s_{\alpha,1}^{-\alpha}$, defined and described by Babister [2, pp. 256-264]. Recall that Equation (6.14) first arose in the context of solving for Poisson's equation in spherical polar co-ordinates [2, pp. 362-363].

The fact that the connection between Legendre differential equations and the Hopfield equation holds for such a wide variety of sigmoids, and is not just an accidental consequence of a particular sigmoid, strongly indicates that further exploration is warranted.

6.2 Fourier transforms & Feedforward nets

There have been many different attempts to describe the behavior of feedforward networks such as the group theoretic analysis of the Perceptron, proposed by Minsky and Papert [27], the space partition (*via* hyperplanes) interpretation discussed by Lippman [22] (and many others), the *metric synthesis* viewpoint introduced by Pao and Sobajic [29], the statistical interpretation emphasized by White [36], *et cetera*. In 1988, Gallant and White showed that a 1-HL feedforward net with "monotone cosine" squashing at the hidden layer, and a summing output node, embeds as a special case a "Fourier network" that yields a Fourier series approximation to a given function as its output [13]. We present a related construction in this section; it is shown that a one hidden layer (1-HL) nets with simple sigmoidal *convex* transfer functions (at the hidden layer), and a single summing output, can be thought of as performing trigonometric approximation (regression) [34, Chap. 4]. Specifically, the inverse Fourier transform of the function (to be learned) is approximated as a linear combination of weighted sinusoids.

The result is a consequence of a connection between a class of simple sigmoids and Fourier transforms, that facilitates a novel interpretation of 1-HL feedforward nets. Polya's theorem is a starting point [30].

Proposition 6.1 (Polya's theorem) : [12] A real valued and continuous function $f(x)$ defined for all real x and satisfying the following properties:

1. $f(0) = 1$,
2. $f(x) = f(-x)$,
3. $f(x)$ is convex for $x > 0$,
4. $\lim_{x \to \infty} f(x) = 0$,

is always a characteristic function (Fourier transform) of an absolutely continuous distribution function[18], i.e., $f(x) = \mathcal{F}(h(t); x) = \int_{-\infty}^{\infty} e^{ixt}h(t)dt$. Furthermore, the density $h(t)$ is an *even* function, and is continuous everywhere except possibly at $t = 0$. ▮

The following result connects simple sigmoids with Fourier transforms.

Theorem 6.1 Let $\sigma(x)$ be a simple sigmoid. If $\sigma(x)/x$ is a convex function, then it is the Fourier transform of an absolutely continuous distribution function i.e.,

$$\frac{\sigma(x)}{x} = \mathcal{F}(h(t); x) = \int_{-\infty}^{\infty} e^{ixt}h(t)dt \tag{6.15}$$

Proof: It suffices to prove that $\sigma(x)/x$ satisfy the conditions of Polya's theorem. $\sigma(x)$ being simple is bounded, and hence $\lim_{x \to \infty} \sigma(x)/x = 0$. Also, $\sigma(-x)/-x = -\sigma(x)/-x = \sigma(x)/x$. Since $\sigma(x)$ is completely monotone in $(0, 1)$, it follows that $\lim_{x \to 0} \sigma(x)/x = K$ (some positive constant). There is no loss of generality in assuming $K = 1$, since one can always scale $\sigma(\cdot)$ appropriately. Finally, the convexity of $\sigma(x)/x$ ensures that all of the conditions of Polya's theorem are satisfied and the conclusion follows. ▮

Remark 6.1 Polya's theorem is a sufficient but *not* necessary condition for $f(x)$ to be the Fourier transform of some function $h(t)$. Hence, Theorem 6.1 is also only a sufficient condition for a simple sigmoid to be a Fourier transform. A case in point is the function $\tanh(x)$ which is not convex, but is still a Fourier transform [28, pp. 42, item # 240], i.e.,

$$\frac{\tanh(x)}{x} = \mathcal{F}(\log(\frac{1}{\pi}\coth(\pi t)); x) \tag{6.16}$$

In other words, the conclusions we draw in the next few paragraphs may be valid for some non-convex simple sigmoids as well.

Remark 6.2 In Equation (6.15) $h(t)$ is an *even* function. Hence the transform is a Fourier cosine transform. The sine component vanishes during the course of an integration.

Consider a 1-HL net, with k input nodes, n hidden layer nodes with *convex* simple sigmoidal transfer functions $\sigma(\cdot)$, and one summing output node. Let w_{ij} denote the weight of the connection between the ith node in the hidden layer and jth node in the input layer; similarly, let c_i denote the weight

[18] Recall that an absolutely continuous function $F(x)$ is a distribution function if it can be written in the form $F(x) = \int_{-\infty}^{x} h(t)dt$, where $h(t)$ is called the density of $F(x)$.

of the connection between the ith hidden layer node and the output node. Then the output O may be expressed as,

$$O = \sum_{i=1}^{n} c_i \, y_i = \sum_{i=1}^{n} c_i \, \sigma(u_i) = \sum_{i=1}^{n} c_i \, \sigma(\sum_{j=1}^{k} w_{ij} x_j + \theta_i) \tag{6.17}$$

where u_i and θ_i are the input and bias for the ith hidden node, respectively. Since $\sigma(\cdot)$ is a convex simple sigmoid, using Lemma 6.1, Equation (6.17) may be rewritten as,

$$O(t) = \sum_{i=1}^{n} c_i \, y_i = \sum_{i=1}^{n} c_i \, u_i \mathcal{F}(h(t); u_i) \tag{6.18}$$

where $\mathcal{F}(h(t); u_i)$ denotes the fact that $\mathcal{F}(h(t); x)$ is to be evaluated at the point $x = u_i = \sum_{j=1}^{k} w_{ij} x_j + \theta_i$. Using the well known property of Fourier transforms, that if $f(x) = \mathcal{F}(h(t); x)$, then $x f(x) = -i\mathcal{F}(h'(t); x) = \mathcal{F}(-ih'(t); x)$, where $h'(\cdot)$ is the first derivative of $h(\cdot)$, and $i = \sqrt{-1}$ [6, pp. 100], Equation (6.18) may be rewritten[19] as,

$$O(t) = \sum_{i=1}^{n} c_i \, \mathcal{F}(-ih'(t); u_i) \tag{6.19}$$

Equation (6.19) can be recognized as being analogous to the Heaviside expansion formula in Laplace transform theory[20], which allows the reconstruction of a time varying function using information relating to its spectral components. Equation (6.19) suggests that 1-HL nets with convex simple sigmoidal transfer functions can be thought of as implementing a spectral reconstruction of the output using the weighted inputs $u_i's$ to evaluate the associated pole coefficients (residues) of the Heaviside expansion.

In particular, it can be demonstrated that the results of Gallant and White [13] are implied by Equation (6.19). In what follows, we shall use $\mathcal{F}_s(h; x)$ and $\mathcal{F}_c(h; x)$ to indicate the Fourier sine and cosine transforms of $h(t)$.

Since $h(t)$, the continuous distribution function corresponding to $\sigma(x)/x$ is an even function (from Polya's theorem), it follows that $\sigma(x) = x\mathcal{F}(h(t); x) = x\mathcal{F}_c(h(t); x)$. Using the property of Fourier transforms that $x\mathcal{F}_c(g(t); x) = \mathcal{F}_s(-g'(t); x)$ [6, pp. 104], we may conclude that $\sigma(x) = \mathcal{F}_s(-h'(t); x)$.

Let $u_i = u + r_i$, where r_i are appropriate functions of the x_i's (since the u_i's are functions of the inputs x_i's).

$$O(u) = \sum_{i=1}^{n} c_i \, \mathcal{F}_s(-h'(t); u + r_i) \tag{6.20}$$

From the frequency shifting property of Fourier transforms [6, pp. 104], viz. ,

$$\frac{1}{2}\mathcal{F}_s(f(t); x + a) = \mathcal{F}_s(f(t)\cos(at); x) + \mathcal{F}_c(f(t)\sin(at); x) \tag{6.21}$$

[19] In Equation (6.19), the i term in $\mathcal{F}(-ih'(t); u_i)$ converts the Fourier cosine transform representation of $\sigma(x)/x$ (see remark 6.2) into a Fourier sine transform.

[20] For convenience we restate a simple version of the formula: If the Laplace transform of a function $h(t)$, is given by $f(x)$, i.e. $f(x) = \mathcal{L}(h(t); x) = \int_0^\infty h(t) \exp(-xt)\, dt$, and $f(x)$ has only first order poles at $x_1, x_2 \cdots x_n$, then $h(t) = \sum_{k=1}^{n} F_k(x_k)$, where $F_k(x_k)$ is the residue or pole-coefficient of $f(x)\exp(xt)$. If the poles of $f(x)$ are of higher order, then a similar formula is available [3, Equation 2-25, pp. 22]

it follows that,

$$\begin{aligned}
O(u) &= \sum_{i=1}^{n} c_i \mathcal{F}_s(-h'(t); u + r_i) \\
&= \sum_{i=1}^{n} 2c_i \left\{\mathcal{F}_s(-h'(t)\cos(r_i t); u) + \mathcal{F}_c(-h'(t)\sin(r_i t); u)\right\} \\
&= \mathcal{F}_s\left\{2\sum_{i=1}^{n} c_i \left(-h'(t)\cos(r_i t); u\right)\right\} + \mathcal{F}_c\left\{2\sum_{i=1}^{n} c_i \left(-h'(t)\sin(r_i t); u\right)\right\} \\
\mathcal{F}^{-1}(O(u)) &= -h'(t)\sum_{i=1}^{n} c_i \sin(r_i + u)t \qquad (6.22)
\end{aligned}$$

But we may choose u arbitrarily, we set $u = 0$, implying $r_i = u_i = \sum_{j=1}^{k} w_{ij}x_j + \theta_i$, and Equation (6.22) becomes,

$$\mathcal{F}^{-1}(O(u)) = -4h'(t)\sum_{i=1}^{n} c_i \sin((\sum_j w_{ij}x_j + \theta_i)t) \qquad (6.23)$$

Equation (6.23) may be used as a starting point for an analysis identical to that adopted by Gallant and White in their study of 1-HL nets with "cosine squashing" functions [13]. It is then straightforward to show that the weights may be so chosen (hardwired) so that the 1-HL nets embeds as a special case a Fourier network, which yields a Fourier series approximation to a given function as its output. In this sense, the results of this section extend the study of Gallant and White.

More generally, one can draw similar conclusions by considering sigmoids that are the *Laplace transforms* of some function; for example $\tanh(x)/x$ is the Laplace transform of sgn $\left\{\sin(\frac{\pi t}{2})\right\}$, where $\mathrm{sgn}(x)$ is $+1$, 0 or -1 depending on whether x is greater, equal or lesser than zero [32, pp. 248]. An analysis similar to the one described above, would lead to a connection with real *exponential* approximation (rather than trigonometric approximation). Efficient algorithms, such as Prony's, exist for certain restricted forms of the exponential approximation problem [34, pp. 82-101].

Also related are the considerations of Marks and Arabshahi on the multidimensional Fourier transforms of the output of a 1-HL feedforward net; they showed that the transform of the output is the sum of certain scaled Dirac delta functions [24]. Here, we view the sigmoid *itself* as the Fourier transform of some function; the main advantage of our interpretation is the algorithms it suggests for training 1-HL nets of the type considered in this section. Extensions to multiple layer nets, while not trivial, should not present undue difficulties.

Another potential use of Equation (6.23) is its possible use in exploring the "goodness" of the approximation obtained by a 1-HL net with simple sigmoidal transfer functions. In the last 200 years, much has been learned about the errors associated with exponential and trigonometric approximation, and ways to deal with it; however, consideration of these issues is beyond the scope of this paper.

7 Conclusion

We have analyzed the behavior of important classes of sigmoid functions, called *simple* and *hyperbolic* sigmoids, instances of which are extensively used as node transfer functions in artificial neural network implementations. We have obtained a complete characterization for the inverses of hyperbolic sigmoids using Euler's incomplete beta functions, and have described composition rules that illustrate how such functions may be synthesized from others. We have obtained power series expansions of hyperbolic sigmoids, and suggested procedures for obtaining coefficients of the expansions. For a large class of node functions, we have shown that the continuous Hopfield net equations can be reduced to Legendre differential equations. Finally, we have shown that a large class of feedforward networks represent the output function as a Fourier series sine transform evaluated at the hidden layer node inputs, thus extending an earlier result due to Gallant and White.

Appendix I

Theorem 4.1: Let $y = \sigma(x)$ be a hyperbolic sigmoid, and let $\eta : (-1, 1) \rightarrow \Re$ be its inverse. Then, either

$$\eta(y) = yF(\alpha, \frac{1}{2}; \frac{3}{2}; y^2) = y \sum_{k=0}^{\infty} \frac{(\alpha)_k}{(2k+1)} \frac{y^{2k}}{k!} \qquad \alpha \geq 1 \tag{7.1}$$

or,

$$\eta(y) = yF(\alpha, -; -; y^2) = \frac{y}{(1 - y^2)^{\alpha}} \qquad \alpha > 0 \tag{7.2}$$

where, by $F(\alpha, -; -; y^2)$, we mean $F(\alpha, \beta; \beta; y^2)$ $(\beta \in \Re)$.

Proof: Since $\sigma(\cdot)$ is hyperbolic, by definition $\eta(\cdot)/x$ is described by a GH series with at most three parameters. There are then four major possibilities:

$$\begin{aligned}
\eta(x) &= x\ {}_3F_0(\alpha_1, \alpha_2, \alpha_3; ; x^2) && \leftarrow \text{Case 1} && (7.3)\\
\eta(x) &= x\ {}_2F_1(\alpha_1, \alpha_2; \gamma_1; x^2) && \leftarrow \text{Case 2} && (7.4)\\
\eta(x) &= x\ {}_1F_2(\alpha_1; \gamma_1, \gamma_2; x^2) && \leftarrow \text{Case 3} && (7.5)\\
\eta(x) &= x\ {}_0F_3(; \gamma_1, \gamma_2, \gamma_3; x^2) && \leftarrow \text{Case 4} && (7.6)\\
& && && (7.7)
\end{aligned}$$

The following proposition shows why there is no need to consider cases 1, 3 and 4, as possible forms for hyperbolic sigmoids.

Proposition A: [32, pp. 155] Let ${}_pF_q(\alpha_1, \ldots, \alpha_p; \gamma_1, \ldots, \gamma_q; z)$, be a GH series in z, with $p + q$ parameters. If none of the numeratorial parameters are non-positive integers, i.e. $\forall\, i : \alpha_i \neq 0, -1, -2, \cdots,$, then convergence behavior of ${}_pF_q$ is as follows:

$$\begin{array}{ll}
p < q + 1 & {}_pF_q \text{necessarily converges for all finite } z. \\
p = q + 1 & \text{convergence of } {}_pF_q \text{ is limited to } -1 < z < 1, \\
 & \text{and depends on the parameters } \alpha_i\text{'s and } \gamma_i\text{'s.} \\
p > q + 1 & {}_pF_q \text{necessarily diverges for all nonzero } z.
\end{array} \tag{7.8}$$

Since $\lim_{z \rightarrow \pm 1} \eta(z) \rightarrow \pm\infty$, but is finite in the interval $(-1, 1)$, it follows that if a GH series is to represent $\eta(\cdot)$, then it has to converge in the interval $(-1, 1)$, but diverge at $z = \pm 1$.

This rules out non-positive *integral* values for the numeratorial parameters; otherwise, the series would converge for *all* $z \in \Re$ (and not just in the interval $(-1, 1)$). Yet, even if the numeratorial parameters do not have non-positive integral values, in three of the above cases, the number of numeratorial parameters to denominatorial ones is such that either series again converges for all z (case 1), or diverges for all z (case 3, 4). That leaves just one case to consider, viz . the classical series, ${}_2F_1(\alpha_1, \alpha_2; \gamma_1, z) = F(\alpha, \beta; \gamma; z)$, i.e. we may take $\eta(x) = x\, F(\alpha, \beta; \gamma; x^2)$.

Since $\eta(\cdot)$ has to be a GH series with at *most* three parameters, some of the parameters are allowed to be "missing". In other words, Case 2 spawns in turn, the following possibilities:

$$\begin{aligned}
\eta(x) &= x\,F(\alpha,\beta;\gamma;x^2) && \leftarrow \text{ Case 2(a)} && (7.9)\\
\eta(x) &= x\,F(\alpha,\beta;-;x^2) && \leftarrow \text{ Case 2(b)} && (7.10)\\
\eta(x) &= x\,F(\alpha,-;\gamma;x^2) && \leftarrow \text{ Case 2(c)} && (7.11)\\
\eta(x) &= x\,F(\alpha,-;-;x^2) && \leftarrow \text{ Case 2(d)} && (7.12)\\
\eta(x) &= x\,F(-,-;\gamma;x^2) && \leftarrow \text{ Case 2(e)} && (7.13)\\
\eta(x) &= x\,F(-,-;-;x^2) && \leftarrow \text{ Case 2(f)} && (7.14)\\
& && && (7.15)
\end{aligned}$$

Proposition A can be used once again to weed out all but two of the above set, viz. Cases 2(a) and 2(d). The rest lead to inappropriate divergence or convergence behavior in the interval. The following property of GH functions will be needed.

Proposition B: [32, pp. 606] If $y = F(\alpha,\beta;\gamma;x)$, then $\dfrac{dy}{dx} = \dfrac{\alpha\beta}{\gamma}F(\alpha+1,\beta+1;\gamma+1;x)$. ∎

(i) 3-parameter GH series:

$$\begin{aligned}
\eta(x) &= x\,F(\alpha,\beta;\gamma;x^2)\\
&= x\sum_{k\,\geq\,0}\frac{(\alpha)_k(\beta)_k}{(\gamma)_k}\,\frac{x^{2k}}{k!}
\end{aligned} \qquad (7.16)$$

Let $\lambda : (-1,1) \rightarrow \Re^+$, with $\lambda(x) = \dfrac{d\eta(x)}{dx}$. Then,

$$\begin{aligned}
\lambda(x) &= \frac{\eta(x)}{dx} = \frac{d}{dx}\,\{x\,F(\alpha,\beta;\gamma;x^2)\}\\
&= F(\alpha,\beta;\gamma;x^2) + 2x\frac{dF(\alpha,\beta;\gamma;x^2)}{dx}\\
&= F(\alpha,\beta;\gamma;x^2) + 2x^2\frac{\alpha\beta}{\gamma}F(\alpha+1,\beta+1;\gamma+1;x^2) \quad \leftarrow \text{Prop. B}\\
&= \left\{\sum_{k\,\geq\,0}\frac{(\alpha)_k(\beta)_k}{(\gamma)_k}\frac{x^{2k}}{k!} + 2\sum_{k\,\geq\,1}\frac{(\alpha)_k(\beta)_k}{(\gamma)_k}\frac{x^{2k}}{(k-1)!}\right\} \qquad (7.17)\\
&= \left\{1 + \sum_{n\,\geq\,1}\frac{(\alpha)_k(\beta)_k}{(k-1)!\,(\gamma)_k}\left(\frac{1}{k}+2\right)x^{2k}\right\} = \left\{\sum_{k\,\geq\,0}\frac{(\alpha)_k(\beta)_k}{(\gamma)_k}(2k+1)\frac{x^{2k}}{k!}\right\}\\
&= \left\{\sum_{k\,\geq\,0}\frac{(\alpha)_k(\beta)_k}{(\gamma)_k}\,\frac{(3/2)_k}{(1/2)_k}\,\frac{x^{2k}}{k!}\right\}
\end{aligned}$$

From the definition of hyperbolic sigmoids, $\lambda(x)$, is to representable by a GH function with at most three parameters; we must make therefore make the identification, $\beta = 1/2$ and $\gamma = 3/2$. From the symmetry properties of the GH function, we need not consider the case when $\alpha = 1/2, \gamma = 3/2$. It follows that,

$$\begin{aligned} y &= x\, F(\alpha, 1/2;\, 3/2;\, x^2) \\ \lambda &= \frac{d\eta(x)}{dx} = F(\alpha, -; -; x^2) = \frac{1}{(1 - x^2)^{\alpha}} \end{aligned} \tag{7.18}$$

The parameter α cannot take any arbitrary real value. The behavior of $\eta(x)$ at the endpoints of its interval, requires that,

$$\lim_{x \to \pm 1} \eta(x) \to \pm\infty \quad \Rightarrow \lim_{x \to \pm 1} \lambda(x) \to \pm\infty \tag{7.19}$$

Equation (7.18) and Equation (7.19) taken together imply that $\alpha > 0$. This is a necessary but not sufficient condition. The following two propositions allow us to pin down α's value more precisely.

Proposition C : [9, pp. 57-61] If α and β are different from $0, -1, \cdots$ then $F(\alpha, \beta; \gamma;\, z)$ converges absolutely for $z < 1$. For $z = 1$:

$$F(\alpha, \beta; \gamma;\, z)\text{converges } \textit{absolutely} \qquad \text{if} \quad (\alpha + \beta - \gamma) < 0 \tag{7.20}$$

$$F(\alpha, \beta; \gamma;\, z)\text{converges } \textit{conditionally} \qquad \text{if} \quad 0 \le (\alpha + \beta - \gamma) < 1 \tag{7.21}$$

$$F(\alpha, \beta; \gamma;\, z)\text{diverges} \qquad \text{if} \quad 1 \le (\alpha + \beta - \gamma) \blacksquare \tag{7.22}$$

Proposition D: [9, pp. 57-61] If $(\gamma - \alpha - \beta) > 0$ then $F(\alpha, \beta; \gamma; 1) = \dfrac{\Gamma(\gamma)\Gamma(\gamma - \alpha - \beta)}{\Gamma(\gamma - \alpha)\Gamma(\gamma - \beta)}$, where $\Gamma(x) = \int_o^\infty exp(-t)t^{x-1}$ is Euler's Gamma function. $\blacksquare$

If $\alpha < 1$, from Proposition C we see that the series converges absolutely at $z = x^2 = 1$. From Proposition D, this in turn implies that, $\eta(x)/x$ will have a *finite* value at the endpoint of its domain interval. Therefore, $\alpha \ge 1$. The final form for the three parameter GH representation for $\eta(x)$ is therefore, $x\, F(\alpha, 1/2; 3/2; x^2)$ where $\alpha \ge 1$.

(ii) 1-parameter GH series:

In this case, $\eta(x) = x\, F(\alpha, -; -; x^2) = x \sum_{k \ge 0} (\alpha)_k \dfrac{x^{2k}}{k!} = \dfrac{x}{(1 - x^2)^{\alpha}}$. The situation is much simpler, since we have to place bounds on the value of one parameter alone. An argument almost identical to the one above, allow us to conclude that for $\eta(x)/x$ to satisfy the properties of a hyperbolic sigmoid, it is both necessary and sufficient that we take $\alpha > 0$. $\blacksquare$

Theorem 4.2 Let $\sigma : \Re \to (-1, 1)$ be a real analytic, odd, strictly increasing sigmoid, such that its inverse $\eta : (-1, 1) \to \Re$ has a GH series expansion in some injective, odd, increasing C^1 function $g(\cdot)$, with at most three parameters, convergent in $(-1, 1)$. Also let η' have a GH series expansion *in* $g(\cdot)$, with *at most one* parameter. Then, either

$$\eta(y) = g(y)F(\alpha, \frac{1}{2}; \frac{3}{2}; (g(y))^2) = g(y)\sum_{k=0}^{\infty} \frac{(\alpha)_k}{2k+1}\frac{(g(y))^{2k}}{k!}, \qquad \text{for } \alpha \geq 1 \tag{7.23}$$

$$\text{or,} \quad \eta(y) = g(y)F(\alpha, -; -; (g(y))^2) = \frac{g(y)}{(1-(g(y))^2)^\alpha}, \qquad \text{for } \alpha > 0 \tag{7.24}$$

provided $\lim_{y \to 1} \dfrac{g'(y)}{(1-y^2)^\alpha} \to \infty$, where $g'(\cdot)$ is the first derivative of $g(\cdot)$.

Proof: The proof for Theorem 4.2 is very similar to that for Theorem 4.1. If we start with $\eta(x) = g(x)\,F(\alpha; 1/2; 3/2; (g(x))^2)$, then we can show that:

$$\eta'(x) = \frac{d\eta}{dx} = \frac{g'(x)}{(1-x^2)^\alpha} \tag{7.25}$$

where $g'(x)$ is the first derivative of $g(x)$. Since $g'(x) > 0$ for all $x \in Dom(g)$, and $\alpha > 0$, it follows that $\eta'(x) > 0$ for all $x \in Dom(\eta)$, i.e. $\eta(x)$ is a strictly increasing function. The analyticity, continuity and oddness of $\eta(\cdot)$ follow from the respective properties of the GH function. We assure that $\lim_{x \to 1} \eta(x) \to \infty$, by forcing its derivative $\eta'(x)$ to go to infinity at the endpoints of its interval. ∎

Theorem 5.1 If the inverse sigmoid is given by $y/(1-y^2)^\alpha$, $\alpha > 0$, then in some neighborhood of the origin, we have the valid expansion $\sigma(x) = x\sum_{k=0}^{\infty} \dfrac{b_{2k+1}}{(2k+1)!} x^{2k}$ where,

$$b_{2k+1} = (-1)^k (2k+1)! \binom{(2k+1)\alpha}{k} \tag{7.26}$$

Proof : We will need the Lagrange inversion formula, stated below [39, pp. 138-141].

Consider the functional equation: $u = t\phi(u)$. Suppose $f(u)$ and $\phi(u)$ are analytic in some neighborhood of the origin (u-plane), with $\phi(0) = 1$. Then there is a neighborhood of the origin (in the t-plane) in which the equation $u = t\phi(u)$ has exactly one root for u. Let $\sum_{k \geq 0} a_k t^k$ be the Maclaurin expansion of $f(u(t))$ in t, and $\sum_{k \geq 0} c_k t^k$ be the Maclaurin expansion of the function $f'(u)[\phi(u)]^n$. Then: $a_n = \dfrac{1}{n} c_{n-1}$

Here, $y \equiv u$, $x \equiv t$, and $\phi(u) = (1-y^2)^\alpha$. Take $f(u) = u \equiv y$, and the theorem follows from the Lagrange inversion formula.

Theorem 5.3: Let $\sigma(x) = \sum_{k=0}^{\infty} \dfrac{b_{2k+1}}{(2k+1)!} x^{2k}$ be an expansion for a hyperbolic sigmoid, with an inverse of the form $yF(\alpha, 1/2; 3/2; y^2)$, valid in some neighborhood of the origin. Then, $b_{2k} = 0$ and, $b_{2k+1} = C(2k+1, k)$. where we define the sequence $C(n,k)$ as follows:

$$\begin{aligned} &C(1,0) = 1 \\ &C(n,k) = 0 \qquad \forall\, k \geq n,\, k < 0 \\ &C(n+1,k) = (2k-n+1)C(n,k) - 2(n\alpha - k + 1)C(n,k-1) \quad n \geq 1 \end{aligned} \tag{7.27}$$

n and k are natural numbers, $D^n(\sigma(x))$, the nth derivatives of σ, are given by:

$$D^n(y) = D^n(\sigma(x)) = \sum_{k=0}^{n-1} C(n,k)y^{2k-n+1}(1-y^2)^{n\alpha-k} \tag{7.28}$$

Proof: This theorem was obtained by a process almost identical to that described in Minai and Williams' work on the derivatives of the logistic sigmoid [26]. We therefore restrict ourselves to an outline.

It is given that $y = \eta(x) = x\,F(\alpha, 1/2; 3/2; x^2)$, and $x = \sigma(y)$. It can be shown that, $D(x) = \frac{d}{dy}\sigma(y) = 1/\eta'(x) = (1-x^2)^\alpha$. Consider the derivatives of the polynomial $f_{k,l}(x) = x^k(1-x^2)^l$,

$$\begin{aligned} D(f_{k,l}(x)) = \frac{d}{dy}f_{k,l}(x) &= kx^{k-1}(1-x^2)^{\alpha+l} + -2lx^{k+1}(1-x^2)^{\alpha+l-1} \\ &= (k)f_{k-1,\alpha+1}(x) + (-2l)f_{k+1,\alpha+l-1}(x) \\ &= L(f_{k,l}(x)) + R(f_{k,l}(x)) \end{aligned} \tag{7.29}$$

In Equation (7.29) we have split the effect of the operator $D \equiv \frac{d}{dy}$ into the sum of the actions of two operators L and R (Minai and Williams refer to them as Λ_0 and Λ_1). With respect to the polynomials $f_{k,l}$, these operators are defined by:

$$L(Af_{k,l}(x)) = Akf_{k-1,\alpha+l}(x) \tag{7.30}$$

$$R(Af_{k,l}(x)) = -2lAf_{k+1,\alpha+l-1}(x) \tag{7.31}$$

where A is a constant. The main advantage of introducing these operators is that they give a systematic way of visualizing the production of $D^{n+1}(x)$ from $D^n(x)$. L and R may be thought of as being applied to a binary tree of expressions, where each node is some polynomial $f_{k,l}(x)$, and the root is the polynomial $f_{0,\alpha} = (1-x^2)^\alpha$. The action of L on each node of this tree is to produce a left child, given by Equation (7.30), and that of R is to produce a right child, given by Equation (7.31). L acting upon $f_{k,l}(x)$ does three things: multiplies it by k (= the degree of x), *reduce* the degree of x by 1, and *increase* the degree of $(1-x^2)$ by α. On the other hand, R *increases* the degree of x by 1, that of $(1-x^2)$ by $(\alpha-1)$, and multiplies the operand by $-2l$, where l is the degree of $(1-x^2)$. Figure 7 depicts the process for the first four levels. By a detailed study of this "derivative" tree the following observations may be proved:

1. The nth level of the tree corresponds to the nth derivative of $\sigma(y)$, $D^n(x) = D^{n-1}(\sigma(y)) = L(D^{n-1}(x)) + R(D^{n-1}(x))$, (the root of the tree is designated $n = 1$, and $D^0(f_{k,l}(x)) = f_{k,l}(x)$).
2. At the nth level, the tree has n nodes, and the kth node (k runs from 0 through $n-1$), is a polynomial in x, given by $C(n,k)f_{2k-n+1,n\alpha-k} = C(n,k)x^{2k-n+1}(1-x^2)^{n\alpha-k}$, where $C(n,k)$ is a constant. It can be seen that the nth derivatives of σ satisfy: $D^n(k) = \sum_{k=0}^{n-1} C(n,k)f_{2k-n+1,n\alpha-k}$.
3. There are two sources contributing to the value of $C(n,k)$. One is the action of R on the $(k-1)$th term, and the other is that of L on the kth term on the $(n-1)$th level.

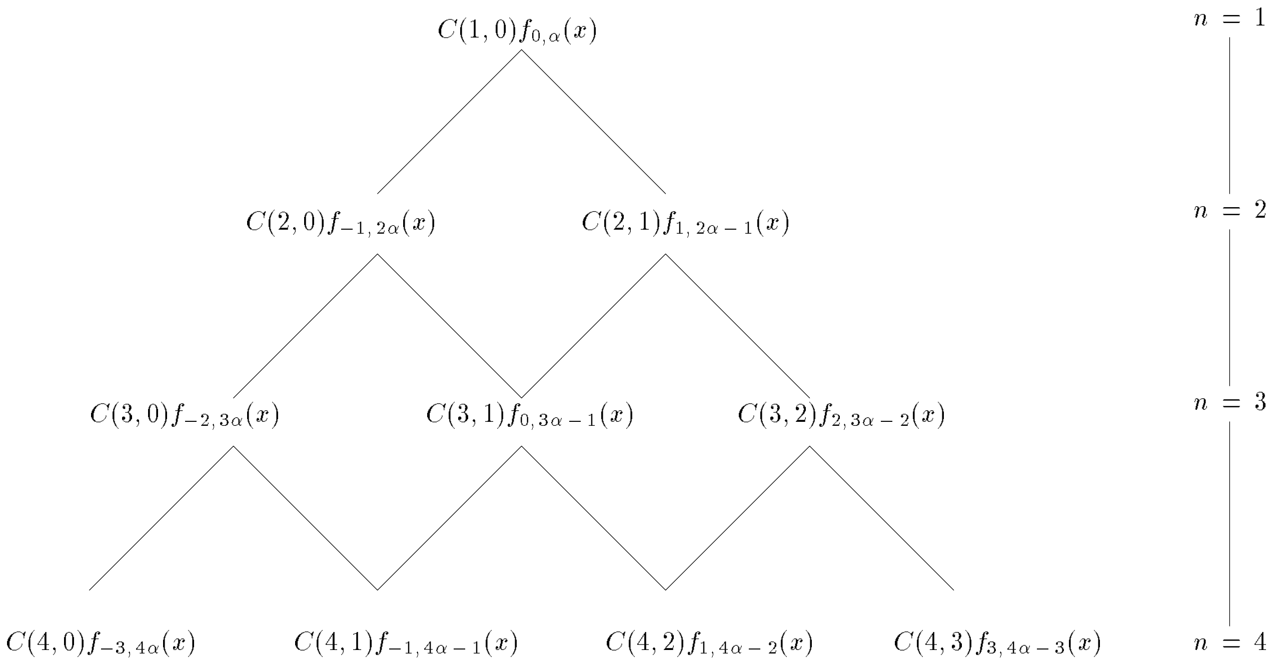
$C(1,0)f_{0,\alpha}(x)$
$n = 1$
$C(2,0)f_{-1,2\alpha}(x)$
$C(2,1)f_{1,2\alpha-1}(x)$
$n = 2$
$C(3,0)f_{-2,3\alpha}(x)$
$C(3,1)f_{0,3\alpha-1}(x)$
$C(3,2)f_{2,3\alpha-2}(x)$
$n = 3$
$C(4,0)f_{-3,4\alpha}(x)$
$C(4,1)f_{-1,4\alpha-1}(x)$
$C(4,2)f_{1,4\alpha-2}(x)$
$C(4,3)f_{3,4\alpha-3}(x)$
$n = 4$

Figure 2: Binary "Derivation" tree for Hyperbolic Sigmoids

Induction arguments in conjunction with the above arguments then give:

$$\begin{aligned} &C(1,0) = 1 \\ &C(n,k) = 0 \qquad \forall\, k \geq n,\, k < 0 \\ &C(n+1,k) = (2k-n+1)C(n,k) - 2(n\alpha - k + 1)C(n,k-1) \quad n \geq 1 \end{aligned} \tag{7.32}$$

Now, all terms in $D^n(x)$, with a x term having positive degree will vanish, when evaluated at $x = 0$. For even n, all the nodes have an x term with an odd degree, and hence $D^n(x)$ vanishes identically at $x = 0$. For odd n, all terms, excepting the term corresponding to $k = (n+1)/2$, vanish at $x = 0$. Since $b_n = D^n(x)\,|_{x=0}$, it follows that $b_{2k} = 0$ and $b_{2k+1} = C(2k+1,k)$.

References

[1] F. Albertini, E. Sontag, and V. Maillot. Uniqueness of weights for neural networks. In R. Mammone, editor, *Artificial Neural Networks with Applications in Speech and Vision*. Chapman and Hall, 1993.

[2] A. W. Babister. *Transcendental Functions Satisfying Nonhomogeneous Linear Differential Equations*. Macmillan Co., New York, 1967.

[3] E. V. Bohn. *The Transform Analysis of Linear Systems*. Addison-Wesley, U.S.A, 1963.

[4] L. Carlitz. The inverse of the error function. *Pacific J. of Math.*, 13(2):459–470, 1963.

[5] G. Cybenko. Approximation by superposition of a sigmoidal function. *Math. Control, Signals and Systems*, 2:303–314, 1989.

[6] B. Davies. *Integral Transforms & their Applications.* Springer-Verlag, New York, 1978.

[7] P. Diaconis and M. Shahshahani. On nonlinear functions of linear combinations. *SIAM J. Sci. Stat. Comput.*, 5:175–191, 1984.

[8] L. D. Elliot. A better activation function for artificial neural networks. Technical Report TR 93-8, Inst. for Systems Research, Univ. of Maryland, College Park, MD, Jan. 1993.

[9] A. Erdélyi, W. Magnus, F. Oberhettinger, and F. G. *et. al.* Tricomi. *Higher Transcendental Functions*, volume 1. McGraw-Hill, New York, 1953.

[10] W. Feller. *An Introduction to Probability Theory and Its Applications*, volume II. John Wiley & Sons, New York, 1965.

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

[12] Polya G. On the zeroes of the derivatives of a function and its analytic character. In R. P. Boas, editor, *George Polya: Collected Works*, pages 178–189. MIT Press, 1974.

[13] A. R. Gallant and H. White. There exists a neural network that does not make avoidable mistakes. In *IEEE International Conf. on Neural Networks*, volume 1, pages 657–664, San Diego, CA, 1988.

[14] W. A. Goodman. *Univalent Functions*, volume I. Mariner Publishing Co., New York, 1983.

[15] R. L. Graham, D. E. Knuth, and O. Patashnik. *Concrete Mathematics.* Addison-Wesley, Reading, MA, 1989.

[16] E. R. Hansen. *A Table of Series and Products.* Prentice-Hall, N.J., 1975.

[17] P. Harrington. Sigmoid transfer functions in backpropagation neural networks. *Analytical Chemistry*, 65(15):2167–2168, 1993.

[18] J. Hertz, A. Krogh, and G. R. Palmer. *An Introduction to the Theory of Neural Computation*, volume 1. Addison-Wesley, 1991.

[19] J. J. Hopfield and D. W. Tank. Computing with neural circuits: A model. *Science*, 233:625–633, 1986.

[20] K. Hornik, M. Stinchcombe, and H. White. Multi-layer feedforward networks are universal approximators. *Neural Networks*, 2:359–366, 1989.

[21] S. G. Krantz and H. R. Parks. *A Primer of Real Analytic Functions.* Birkhäuser Verlag, Berlin, 1992.

[22] R. P. Lippmann. An introduction to computing with neural nets. *IEEE ASSP Magazine*, 4:4–22, 1987.

[23] A. Macintyre and E. Sontag. Finiteness results for sigmoidal "neural" networks. In *Proc. 25th Annual Symp. Theory Computing*, San Diego, May 1993.

[24] R. J. Marks and P. Arabshahi. Fourier analysis and filtering of a single hidden layer perceptron. In *International Conference on Artificial Neural Networks (IEEE/ENNS)*, Sorrento, Italy, May 1994.

[25] R. E. McBride. *Obtaining Generating Functions.* Springer Verlag, Germany, 1970. Vol. 21.

[26] A. Minai and R. Williams. On the derivatives of the sigmoid. *Neural Networks*, 6:845–853, 1993.

[27] M. Minsky and Papert S. A. *Perceptrons, an Introduction to Computational Geometry.* The MIT Press, 1988.

[28] F. Oberhettinger. *Fourier Transforms of Distributions and their Inverses.* Academic Press, New York, 1973.

[29] Y. H. Pao and D. J. Sobajic. Metric synthesis and concept discovery with connectionist networks. In *Proc. of the IEEE Systems, Man and Cybernetics Conf.*, Alexandria, VA, October 1987.

[30] G. Polya. Remarks on the characteristic function. In *Proc. 4th Berkeley Symp. Math. Statist. & Probab.*, pages 115–123, 1949.

[31] E. D. Rainville. *Intermediate Differential Equations.* Macmillan Company, New York, 1964.

[32] J. Spanier and Oldham K. *An Atlas of Functions.* Hemisphere Pub. Corp., Washington, 1987.

[33] W. S. Stornetta and B. A. Huberman. An improved three layer back-propagation algorithm. In *Proc. of the IEEE First Intl. Conf. on Neural Networks*, 1987.

[34] K. L. Su. *Time-Domain Synthesis of Linear Networks.* Prentice-Hall, New Jersey, 1971.

[35] H. J. Sussmann. Uniqueness of weights for minimal feedforward nets with a given input-output map. *Neural Networks*, 5:589–593, 1992.

[36] H. White. Learning in artificial neural networks: A statistical perspective. *Neural Computation*, 1:425–464, 1989.

[37] E. T. Whittaker and G. N. Watson. *Modern Analysis.* Cambridge Univ. Press, Cambridge, fourth edition, 1927.

[38] D. V. Widder. *The Laplace Transform.* Princeton Univ. Press, Princeton, 1946.

[39] H. S. Wilf. *generatingfunctionology.* Academic Press, Inc., New York, 1989.