Neural Networks, Vol. 6, pp. 845–853, 1993
Printed in the USA. All rights reserved.

0893-6080/93 $6.00 + .00

ORIGINAL CONTRIBUTION

On the Derivatives of the Sigmoid

ALI A. MINAI AND RONALD D. WILLIAMS

University of Virginia, Charlottesville

(*Received* 6 *February* 1992; *revised and accepted* 13 *January* 1993)

Abstract—*The sigmoid function is very widely used as a neuron activation function in artificial neural networks, which makes its attributes a matter of some interest. This paper presents some general results on the derivatives of the sigmoid. These results relate the coefficients of various derivatives to standard number sequences from combinatorial theory, and thus provide a standard efficient way of calculating these derivatives. Such derivatives can be used in approximating the effect of input perturbations on the output of single neurons and could also be useful in statistical modeling applications.*

Keywords—Neuron activation function, Sigmoid function, Logistic function, Eulerian numbers, Stirling numbers.

1. INTRODUCTION

The sigmoid function has found extensive use as a nonlinear activation function for neurons in artificial neural networks. Thus, it is of some interest to explore its characteristics. In this paper, we study the derivatives of the 1-dimensional sigmoid function

$$y = \sigma(x; w) = \frac{1}{1 + e^{-wx}}, \tag{1}$$

where $x \in \mathcal{R}$ is the independent variable and $w \in \mathcal{R}$ is a weight parameter. In particular, we present recurrence relations for calculating derivatives of any order, and show that the coefficients generated by these recurrences are directly related to standard number theoretic sequences. Thus, our formulation can be used to determine higher derivatives of the sigmoid directly from standard tables.

2. MOTIVATION

The sigmoid in eqn (1) is also called the *logistic function*, and can be seen as representing a "neuron" with one input. However, it can also be given a more general interpretation. Let i be a neuron with n inputs x_j, $1 \leq j \leq n$, and let the output y of the neuron be given by:

$$y = \frac{1}{1 + e^{-z}} \tag{2.1}$$

$$z = \sum_j w_{ij} x_j + \theta = \mathbf{wx} + \theta, \tag{2.2}$$

where θ is a bias value, $\mathbf{x} = [x_1, x_2, \ldots, x_n]$ and $\mathbf{w} = [w_{i1}, w_{i2}, \ldots, w_{in}]$. This defines a sigmoidal step in $n + 1$ dimensions (henceforth called an *n-dimensional sigmoid*), oriented in the direction of the n-dimensional vector $\mathbf{w}$ (Figure 1).

The use of this sigmoid function in neural networks derives in part from its utility in Bayesian estimation of classification probabilities. It arises as the expression for the posterior probability when the weights are interpreted as logs of prior probability ratios (Stolorz et al., 1992). The sigmoid is also attractive as an activation function because of its monotonicity and its simple form. The form of its lower derivatives also makes it attractive for learning algorithms like back propagation (Werbos, 1974).

It is clear from eqns (2.1, 2.2) that the value of y remains unchanged along n-dimensional hyperplanes orthogonal to $\mathbf{w}$. If $\theta = 0$, the n-dimensional hyperplane $z = 0$ ($y = 0.5$), passes through the origin $\mathbf{x} = 0$. If θ is not 0, the sigmoid is shifted along $\mathbf{w}$ by a distance $-\theta / \|\mathbf{w}\|$, where $\| \quad \|$ is the Euclidean norm. The shape of the sigmoid (specifically, the sharpness of its nonlinearity) is determined by $\|\mathbf{w}\|$, with a larger value giving a sharper sigmoid. Figure 2 shows a 1-dimensional sigmoid with different weight values.

It can be shown quite easily (Minai, 1992) that trac-

Acknowledgements: This research was supported by the Center for Semicustom Integrated Systems at the University of Virginia and the Virginia Center for Innovative Technology. We would like to thank Prof. Worthy N. Martin and Prof. Bruce A. Chartres for their valuable suggestions. Dr. Minai would also like to thank the Department of Neurosurgery, University of Virginia, and especially Prof. William B. Levy for their support. The paper has also benefited considerably from the suggestions of two anonymous reviewers.

Requests for reprints should be sent to Dr. Ali A. Minai, Department of Neurosurgery, Box 420, Health Sciences Center, University of Virginia, Charlottesville, VA 22908.

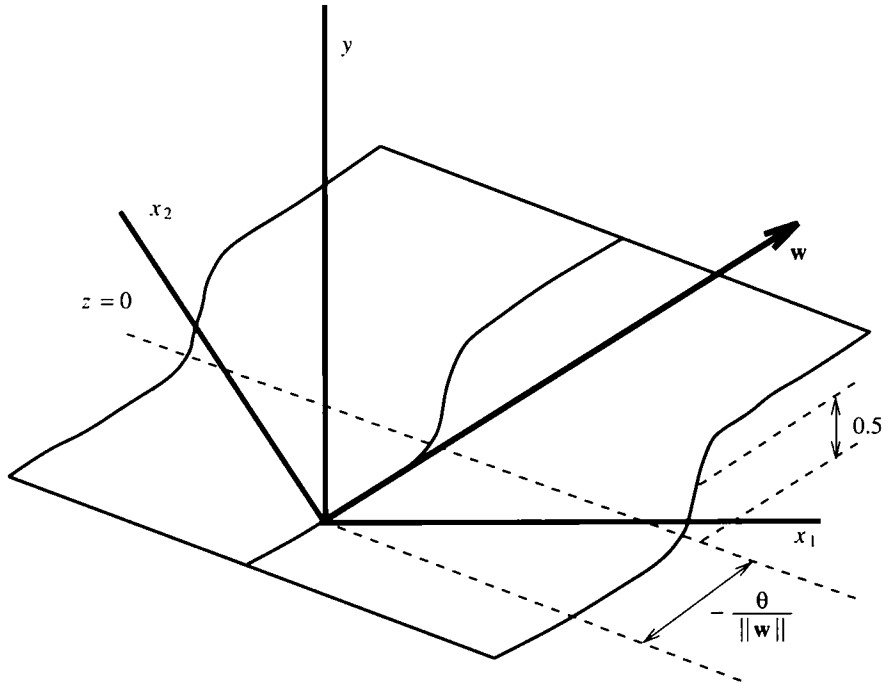

FIGURE 1. A 2-dimensional sigmoid function: $y(x_1, x_2) = 1/[1 + \exp(-wx + \theta)]$; θ = threshold, $z = w_1x_1 + w_2x_2 + \theta$, $w = [w_1w_2]$.

ing y along any line not orthogonal to $\mathbf{w}$ in the input space yields a 1-dimensional sigmoid with the effective weight $w = \|\mathbf{w}\|\cos\beta$, where β is the angle between $\mathbf{w}$ and the direction of the trace. Thus, in order to determine the change in y due to an arbitrary change in $\mathbf{x}$, it is sufficient to consider the corresponding change in a 1-dimensional sigmoid. A Taylor series expansion of the 1-dimensional sigmoid can be used to approximate this change, and this requires the calculation of higher derivatives, depending on the desired level of accuracy.

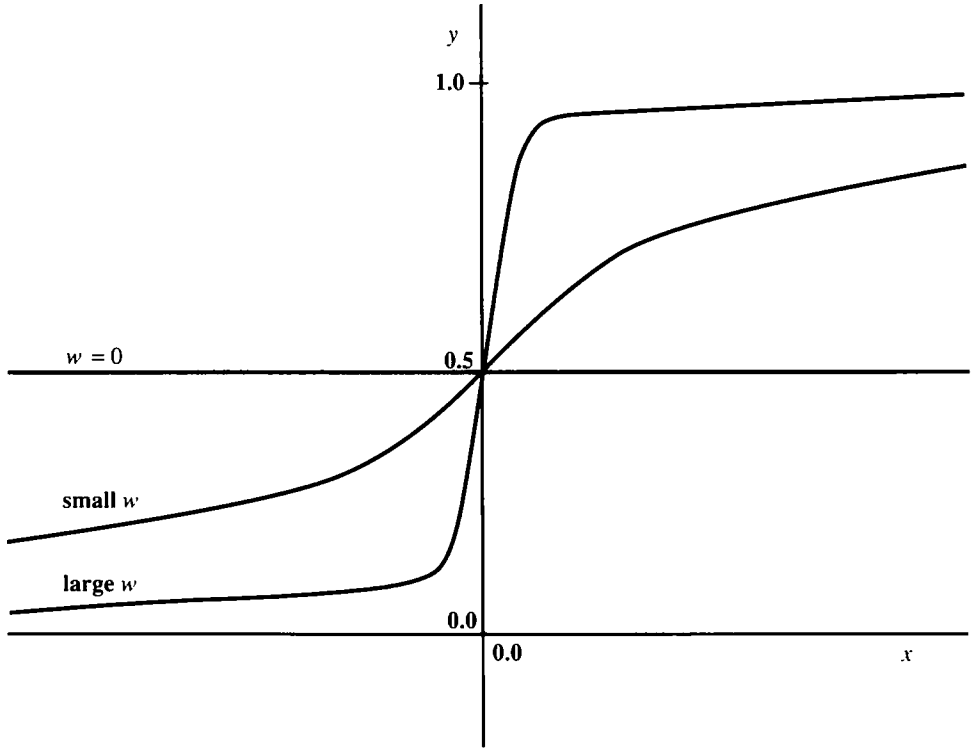

FIGURE 2. The shape of a 1-dimensional sigmoid function as it varies with the weight w. In the large limit of w, the sigmoid becomes a threshold function.

Expanding the function of eqn (1) around x gives:

$$\sigma(x + \delta x) = \sigma(x) + \sum_{n=1}^{\infty} \frac{1}{n!} \frac{d^n \sigma(x)}{dx^n} (\delta x)^n$$

$$= \sigma(x) + \sum_{n=1}^{\infty} \frac{1}{n!} \sigma^{(n)}(x)(\delta x)^n, \qquad (3)$$

where $\sigma^{(n)}(x) \equiv d^n\sigma(x)/dx^n$. The issue is to determine $\sigma^{(n)}$.

One interesting application of expansions such as these might be in the analysis of layers of sigmoid neurons. Using Taylor series expansions of a relatively high order, one could accurately write the transformation performed by a layer of sigmoid neurons locally as a polynomial. Such a description could then be used in analysing the effects of changes in the layer's input, thus characterizing the smoothness of the mapping and its robustness to perturbations. Such analysis could be very useful in studying the fault-tolerance of feed forward networks and, perhaps, even their ability to generalize (which is related to the smoothness of the induced mappings). Another application for such polynomial approximations might be in studying the dynamics of recurrent networks with sigmoid neurons, particularly in characterizing the sensitivity to small perturbations. The Appendix gives further details on estimating the effect of perturbations.

Another motivation for studying the logistic function and its derivatives comes from the area of logistic regression. The logistic function is widely used as the canonical link function in the statistical modeling of binary data with a Bernoulli distribution (McCullagh & Nelder, 1989). Specifically, when generalized linear models on binary data, $\mathbf{x} = [x_1 \ldots x_m]$, $x_i \in \{0, 1\}$, are used to estimate probabilities for classification, it is convenient to relate a probability, p, to the estimator, $\eta \equiv \mathbf{wx}$ by

$$\eta(\mathbf{x}; \mathbf{w}) \equiv \mathbf{w} \cdot \mathbf{x} = \log \frac{p}{1 - p} \equiv logit(p).$$

This, in turn, implies the linking function

$$p = \frac{e^{\eta}}{1 + e^{\eta}} = \frac{1}{1 + e^{-\eta}},$$

(i.e., the logistic function). There is considerable interest in the higher derivatives of the logistic function because they provide direct information about higher-order statistics of the underlying distribution.

3. DETERMINING $\sigma^{(n)}$

An interesting property of the sigmoid is that the derivatives of y can be written in terms of y and w only. The first derivative of y with respect to y_j is given by:

$$\sigma^{(1)} = \frac{dy}{dx} = w \cdot \frac{e^{-wx}}{(1 + e^{-wx})^2} = wy(1 - y). \qquad (4)$$

The second derivative can be obtained from this by differentiating again with respect to x. It is instructive to consider the following general case:

$$\frac{d}{dx}[y^k(1 - y)^l] = ky^{k-1}(1 - y)^l wy(1 - y) - l(1 - y)^{l-1}$$

$$\times[-(wy(1 - y)]y^k = wky^k(1 - y)^{l+1}$$

$$- wly^{k+1}(1 - y)^l, \qquad (5)$$

which immediately suggests the following recursive procedure for obtaining $\sigma^{(n+1)}$ from $\sigma^{(n)}$.
Let

$$\Lambda_0[Ay^k(1 - y)^l] \equiv Aky^k(1 - y)^{l+1} \qquad (6)$$

$$\Lambda_1[Ay^k(1 - y)^l] \equiv -Aly^{k+1}(1 - y)^l \qquad (7)$$

$$\Lambda_{ab}[G] \equiv \Lambda_b[\Lambda_a[G]], \qquad (8)$$

where a is a binary sequence (e.g., 0110101) and $b \in \{0, 1\}$. Thus,

$$\Lambda_{0110101}[G] \equiv \Lambda_1[\Lambda_0[\Lambda_1[\Lambda_1[\Lambda_0[G]]]]]$$

$$\Lambda[G] \equiv \Lambda_0[G] + \Lambda_1[G] \qquad (9)$$

$$\Lambda[G + H] \equiv \Lambda[G] + \Lambda[H] \qquad (10)$$

$$\Lambda^n[G] \equiv \Lambda[\Lambda^{n-1}[G]]. \qquad (11)$$

Then the following obtain:

LEMMA 1. $\sigma^{(n)} = w\Lambda[\sigma^{(n-1)}]$.

Proof. The conclusion follows from eqn (5) and the superposition of derivatives. □

LEMMA 2. $\sigma^{(n)} = w^{n-1}\Lambda^{n-1}[\sigma^{(1)}] = w^{n-1}\Lambda^{n-1}[wy(1 - y)] = w^n\Lambda^{n-1}[y(1 - y)]$

Proof. The conclusion follows by recursion on Lemma 1. □

REMARK 1. *Following Lemma 2, $\sigma^{(n)}$ can be written as $\sigma^{(n)} = w^n\zeta^{(n)}$, where $\zeta^{(n)} \equiv \Lambda^{n-1}[Y(1 - y)]$. Several observations can be made directly about the derivatives obtained using the above method. These are presented below in the form of lemmas and corollaries. Most of the proofs are straightforward, and are omitted to save space (see Minai, 1992, for all proofs). Some of the more significant ones are given in full.*

LEMMA 3. *$\sigma^{(n)}$, as generated by Λ, is the sum of 2^{n-1} terms. This form is called the* expanded *or* Λ-generated *form of $\sigma^{(n)}$. The process of obtaining $\sigma^{(n)}$ from $\sigma^{(1)}$ can be seen as the development of a* n-level binary tree (*Figure* 3), *with the* pth level nodes *representing the individual terms of $\sigma^{(p)}$.*

If, on each application of Λ to a term of $\sigma^{(p)}$, the left child node is generated by Λ_0 and the right child node by Λ_1, then this defines a unique ordering on all

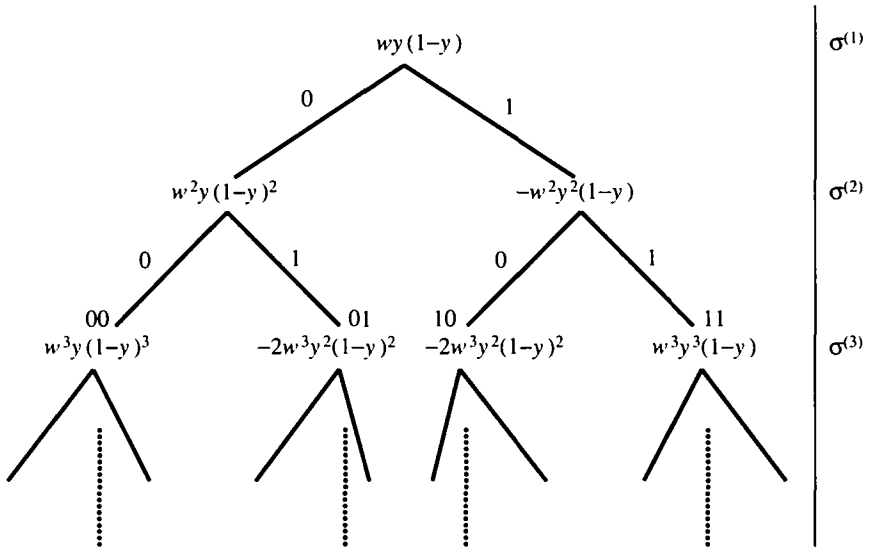

FIGURE 3. The binary tree formulation of the procedure for calculating successive derivatives of the sigmoid. The arcs are labeled with the subscript for the appropriate Λ operator. The derivative shown on the right is the sum of all the terms in the corresponding level of the tree.

terms of $\sigma^{(p+1)}$. If this scheme is used at each level in the generation of $\sigma^{(n)}$, the terms of $\sigma^{(n)}$ are uniquely ordered as follows:

LEMMA 4. *The* m*th term of* $\sigma^{(n)}$ *is given by* $w^n\Lambda_S[y(1-y)]$, *where* m $= 0, 1, 2, \ldots, 2^{n-1} - 1$ *and* S *is the* $n-1$ *digit binary representation of* m, *and is called the* generating sequence *for the term.*

REMARK 2. $\sigma^{(n)}$ *can also be generated by* $w^n\Lambda^n[y]$, *because* $\Lambda[y^1(1-y)^0] = y(1-y)$. *However, the* Λ_1 *branch operating on* y *generates a* 0*-term, producing a single-term* $\sigma^{(1)}$. *Because this introduces a needless asymmetry into the procedure, the formulation of Lemma* 2 *is used throughout.*

LEMMA 5. *The term generated by* $\Lambda_S[y(1-y)]$ *has the form* $Cy^k(1-y)^l$, *where*

$$k = 1 + \text{the number of 1's in } S$$

$$l = 1 + \text{the number of 0's in } S$$

Proof. The powers of both y and $(1-y)$ are initially 1. Thereafter, each operation by Λ_1 adds 1 to the power of y [eqn (7)], and each operation by Λ_0 does the same to the power of $(1-y)$ [eqn (6)]. The conclusion follows immediately. □

COROLLARY 1. *In the* Λ*-generated expression for* $\sigma^{(n)}$, *two terms* $\sigma_p^{(n)}$ *and* $\sigma_q^{(n)}$ *can be combined if their generating sequences* S_p *and* S_q *have the same number of* 1*'s and* 0*'s.*

COROLLARY 2. *Each term in* $\sigma^{(n)}$ *has the form* $Cy^k(1-y)^l$, *where* $l + k = n + 1$ *and* C *is a coefficient.* $\{k, l\}$ *is called the* power pair *of the term.*

COROLLARY 3. *In the* Λ*-generated expression for* $\sigma^{(n)}$, *there are* $\binom{n-1}{k-1} = \binom{n-1}{n-k}$ *terms with the power pair* $\{k, n+1-k\}$.

Because $\sigma^{(n)} = w^n\zeta^{(n)}$, $\zeta^{(n)}$ completely determines $\sigma^{(n)}$. Thus, only $\zeta^{(n)}$ will be studied henceforth. Note that, like $\sigma^{(n)}$, it has Λ-generated and simplified forms, obtained by dividing $\sigma^{(n)}$ by w^n.

Another fact about $\sigma^{(n)}$ and $\zeta^{(n)}$ is that they possess *odd symmetry* about the origin for even n, and *even symmetry* for odd n. This follows trivially from the fact that the sigmoid has odd symmetry. Also, $\sigma^{(n)}$ has $n+1$ roots (zero-crossings), since it is a polynomial of degree $n+1$ in y. The first six derivatives are shown in Figure 4, plotted over the range $0 \le y \le 1$, which corresponds to $-\infty \le x \le \infty$.

4. CALCULATING THE COEFFICIENTS

LEMMA 6. *If the* m*th term of the* Λ*-generated* $\zeta^{(n)}$ *has the generating sequence* $S_m^{(n)} = \{d_i\}$ (d_i *is the* i*th digit of the binary sequence*), *the coefficient* $c_m^{(n)}$ *of the term can be obtained by the following equations:*

$$c_m^{(n)} = \prod_{p=0}^{n-2} [d_p(-(n+1-k_p)) + |d_p - 1|k_p]$$

$$k_p = 1 + \sum_{q=0}^{p-1} d_q,$$

or, equivalently, by the equations:

$$c_m^{(n)} = \prod_{p=0}^{n-2} [d_p(-l_p) + |d_p - 1|(n+1-l_p)]$$

$$l_p = 1 + \sum_{q=0}^{p-1} |d_q - 1|.$$

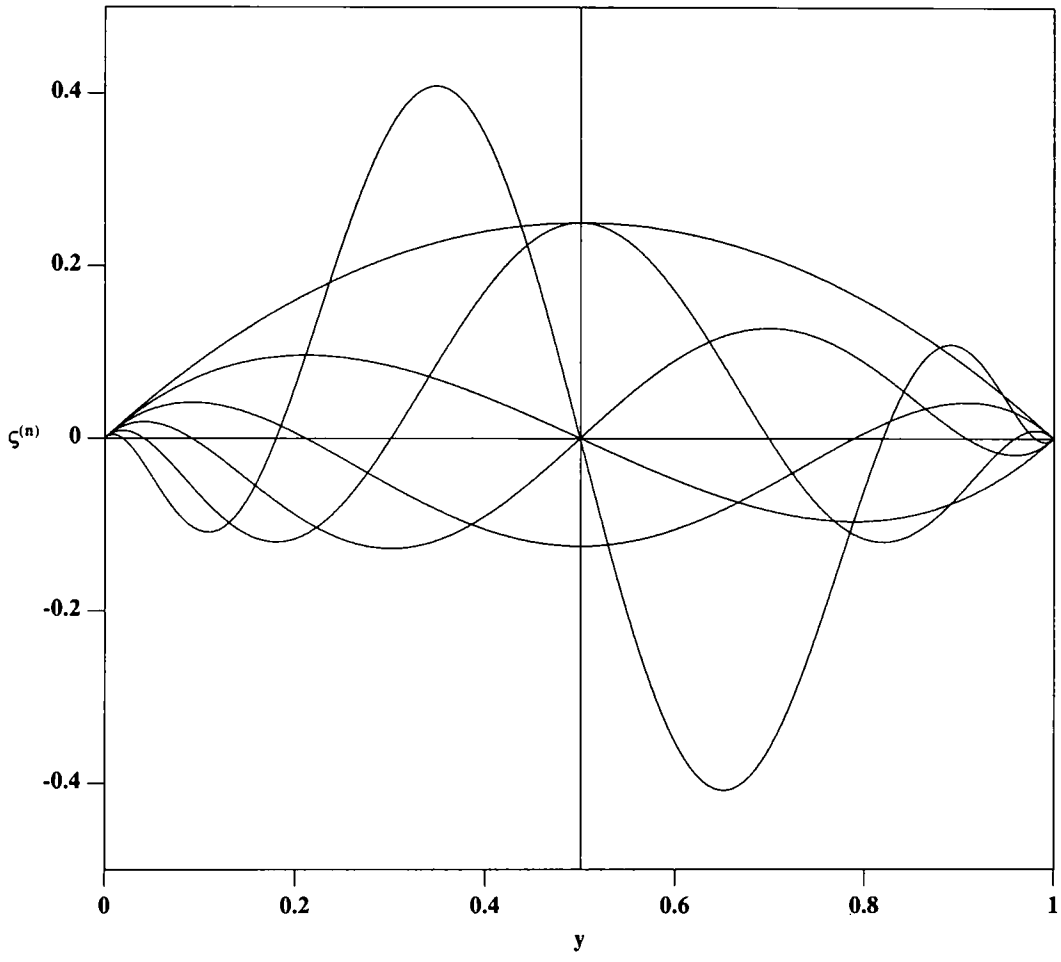

FIGURE 4. The first six derivatives of the sigmoid for $w = 1$. *The domain* $0 \le y \le 1$ corresponds to $-\infty \le x \le \infty$. The total number of maxima and minima equals the order of the derivative in each case.

Proof. The coefficient in question can be generated as follows:
Initialize $k = 1$, $l = 1$, $c = 1$ (since $c_1^{(1)} = 1$).
For each digit d_i of $S_m^{(n)}$

if d_i is 0

$c = c \times k$.

$l = l + 1$.

else if d_i is 1

$c = c \times -l$.

$k = k + 1$.

$c_m^{(n)} = c$.
The equations given in the lemma implement this algorithm, using the fact that $l + k = n + 1$ (Corollary 2). □

COROLLARY 4. *All terms of the* Λ*-generated form of* $\varsigma^{(n)}$ *which have the same power pair* $\{k, l\}$ *have coefficients with the same sign.*

COROLLARY 5. *In the* Λ*-generated form of* $\varsigma^{(n)}$, *all terms with an odd number of* 1*'s in their generating sequence have negative coefficients, while all terms with an even number of* 1*'s have positive coefficients.*

COROLLARY 6. *All terms in* $\varsigma^{(n)}$ *which have odd k in their power pair will have positive coefficients, while all terms with even k will have negative coefficients.*

LEMMA 7. *In the* Λ*-generated form of* $\varsigma^{(n)}$, *the mth and the* $(2^{n-1} - m)$*th terms have coefficients with the same absolute value* $(0 \le m \le 2^{n-1} - 1)$. *They are called* duals *of each other.*

COROLLARY 7. *If n is even, the coefficient* c_m *of the mth term and coefficient* c_M *of the Mth term,* $M = 2^{n-1} - m$, *are related by* $c_m = -c_M$. *If n is odd, they are related by* $c_m = c_M$.

LEMMA 8. *Of all the terms in the* Λ*-generated form of* $\varsigma^{(n)}$, *those with generating sequences of the form* 010101 . . . *or* 101010 . . . *have the coefficients with the highest absolute value.*

Proof. This is easily proved by induction. Suppose, without loss of generality, that the first (leftmost) ele-

ment of the sequence is 0, giving a coefficient of 1. For the next sequence digit, choosing 0 again gives the coefficient 1×1, while chosing 1 gives the coefficient 1×-2 (because the power of $(1 - y)$ is now 2 due to the initial operation by Λ_0). The coefficient with the largest absolute value that can be produced with a 2-digit sequence is, therefore, -2, obtained by choosing 1, giving the sequence 01. At the next step, choosing 0 gives coefficient $1 \times -2 \times 2$, while choosing 1 gives the coefficient $1 \times -2 \times -2$, both of which have the same absolute value. Thus, the largest absolute value coefficient for a 3-digit sequence belongs to the term generated by 010 (and by 011). Let the greatest absolute value coefficient for a p-digit sequence be associated with the term 010101 . . . 010. The coefficient in question is given by $1 \times -2 \times 2 \times -3 \times 3 \times \ldots \times -m \times m$, where $m = \lceil 0.5p \rceil$. The power of y is m, while that of $(1 - y)$ is $m + 1$ (because there are m 0's and $m - 1$ 1's in the sequence). Thus, a choice of 0 as the next digit would multiply the coefficient by m, while a choice of 1 will multiply it by $-(m + 1)$. Clearly, choosing 1 maximizes the absolute value, and the highest absolute value coefficient for a $p + 1$ digit sequence belongs to the term generated by 010101 . . . 0101. At the next step, choosing 0 will multiply the coefficient by $m + 1$ (because the power of y is now also $m + 1$), while choosing 1 will multiply it by $-(m + 1)$. Clearly, choosing 0 is sufficient to ensure the maximization of absolute value, and the $p + 2$ digit sequence producing this value is 010101 . . . 01010 (and also 010101 . . . 01011). It is also obvious that no future gain in coefficient multipliers can be realized by choosing the smaller multiplier at any step. For example, choosing 0 as the $p + 1$st digit multiplies the coefficient by m (instead of $-(m + 1)$, and the choices at the $p + 2$nd step become m if 0 is chosen and $-(m + 2)$ if 1 is chosen. Because $|m \times -(m + 2)| < |-(m + 1) \times (m + 1)|$, this produces a loss rather than a gain. An identical argument can be made if the first digit chosen is 1. Thus, terms with generating sequences of the form 010101 . . . 101010 . . . have the largest absolute value coefficients among all terms with generating sequences of the same length. □

REMARK 3. *It might be noted that, for a given n, the absolute values of the coefficients for the* 010101 . . . *term and the* 101010 . . . *term are the same. If n is even, they have opposite signs, and if odd, identical signs. This is just a special case of the duality results.*

REMARK 4. *The proof for Lemma* 8 *shows that the terms with generating sequences* 01010 . . . *and* 10101 . . . *are not the only ones with the highest coefficients. A necessary condition that a term have the highest absolute value coefficient is that its generating sequence begin with* 10 *or* 01, *and contain no subsequence of more than two consecutive* 0*'s or* 1*'s. This is a consequence of* Λ *being composed of two operators that increment one power while multiplying the coefficient by the other.*

5. SIMPLER FORMS OF THE DERIVATIVES

While the Λ-generated version of $\zeta^{(n)}$ is quite interesting algorithmically, it can be simplified to much more compact and manageable forms. This is discussed next.

LEMMA 9. *After simplification,* $\zeta^{(n)}$ *has n terms, and can be written in the form:*

$$\zeta^{(n)} = \sum_{k=1}^{n} (-1)^{k-1} C_k^{(n)} y^k (1 - y)^{n+1-k}$$

$$= \sum_{l=1}^{n} (-1)^{n-l} C_{n+1-l}^{(n)} y^{n+1-l} (1 - y)^l \quad (12)$$

This is called the simplified *form of* $\sigma^{(n)}$.

Proof. Because all terms of $\zeta^{(n)}$ have generating sequences of $n - 1$ digits, there are n different 0/1 distributions, ranging from terms with $n - 1$ 1's and no 0's to those with no 1's and $n - 1$ 0's. Thus, there will be n different power pair combinations in $\zeta^{(n)}$, and n irreducible terms. The form of eqn (12) follows from Corollary 2 and the alternating signs from Corollary 6. □

The coefficients $C_k^{(n)}$ for the first few values of n are given in Table 1 (also see Theorem 1).

LEMMA 10. *In the simplified version of* $\zeta^{(n)}$, *the coefficient* $C_k^{(n)}$ *of the* $y^k(1 - y)^{n+1-k}$ *term can be obtained by the recursion:*

$$C_k^{(n)} = 0 \ \forall n \quad \text{if} \quad k < 1 \quad \text{or} \quad n < 0$$

$$C_1^{(1)} = 1$$

$$C_k^{(n)} = kC_k^{(n-1)} + (n + 1 - k)C_{k-1}^{(n-1)}. \quad (13)$$

Proof. The only two ways to obtain power pair $\{k, n + 1 - k\}$ in the expression for $\zeta^{(n)}$ are: (1) Λ_0 operating on a $\{k, n - k\}$ term of $\zeta^{(n-1)}$; and (2) Λ_1 operating on a $\{k - 1, n + 1 - k\}$ term of $\zeta^{(n-1)}$. The former produces coefficient $k(-1)^k C_k^{(n-1)}$, and the latter $-(n + 1 - k)(-1)^{k-1} C_{k-1}^{(n-1)} = (n + 1 - k)(-1)^k C_{k-1}^{(n-1)}$. Thus, the two add in magnitude to give the coefficient $(-1)^k[kC_k^{(n-1)} + (n + 1 - k)C_{k-1}^{(n-1)}]$, which gives $C_k^{(n)}$ as in (13). □

THEOREM 1. *The coefficient* $C_k^{(n)}$ *in the simplified representation of* $\zeta^{(n)}$ *is equal to the* Eulerian number $A_{n,k-1}$ (*see Graham et al.*, 1989, *for a discussion*).

Proof. The recurrence for generating Eulerian numbers is:

$$A_{r,q} = (q + 1)A_{r-1,q} + (r - q)A_{r-1,q-1};$$

integer $q, r > 0$. (14)

TABLE 1
The Coefficients, $C_k^{(n)}$, for the Eulerian Form of the First Seven Derivatives

Derivative Number (n)	k						
	1	2	3	4	5	6	7
1	1						
2	1	1					
3	1	4	1				
4	1	11	11	1			
5	1	26	66	26	1		
6	1	57	302	302	57	1	
7	1	120	1191	2146	1191	120	1

Each $C_k^{(n)}$ is the coefficient for a term of the form $y^k(1-y)^{n+1-k}$; $1 \le k \le n$, and equals $A_{n,k-1}$, where $A_{n,k}$ are the Eulerian Numbers. The sign is given by $(-1)^{k-1}$.

with the conditions that $A_{1,0} = 0$ and $A_{r,q} = 0$ if either r or q is less than 0. Substituting $k-1$ for q, n for r, and $C_k^{(n)}$ for $A_{n,k-1}$ in eqn (14), eqn (13) is obtained and the result follows. □

REMARK 5. *Following Lemma* 10 *and Theorem* 1, $\zeta^{(n)}$ *can be written as:*

$$\zeta^{(n)} = \sum_{k=1}^{n} (-1)^{k-1} A_{n,k-1} y^k (1-y)^{n+1-k} \qquad (15)$$

which may be called the Eulerian form *of* $\zeta^{(n)}$. *Because the Eulerian numbers can be easily calculated (or looked up), any* $\sigma^{(n)}$ *can be calculated using eqn* (15).

Other forms can also be found for $\zeta^{(n)}$, and these may provide further insight. One of these is briefly discussed next.
Using

$$\zeta^{(1)} = y(1-y) = y - y^2, \qquad (16)$$

$\zeta^{(n)}$ may be written in powers of y only (rather than powers of y and $1-y$). This is done as before by defining a recurrence that generates $\zeta^{(n)}$ from $\zeta^{(n-1)}$. Clearly, from (12), $\zeta^{(n)}$ has the form:

$$\zeta^{(n)} = \sum_{k=1}^{n+1} H_k^{(n)} y^k.$$

The following results provide specific information about $H_k^{(n)}$.

LEMMA 11. *If* $\zeta^{(n)}$ *is written in the form of eqn* (16), *all terms with odd powers have positive coefficients, while all terms with even powers have negative coefficients.*

Proof. In generating the terms of $\zeta^{(n)}$ from those of $\zeta^{(n-1)}$, the y^q term of the latter produces two terms in $\zeta^{(n-1)}$. Denoting this by an operator Γ,

$$\Gamma[y^q] = qy^{q-1}y(1-y) = qy^q - qy^{q+1}. \qquad (17)$$

Thus, the y^q term in $\zeta^{(n-1)}$ produces a y^q term in $\zeta^{(n)}$ with the same sign, and a y^{q+1} term with the opposite sign. Because $\zeta^{(1)} = y - y^2$, it is clear that, for all $\zeta^{(n)}$, odd powers of y will have positive coefficients and even powers will have negative ones. □

Thus, $\zeta^{(n)}$ can be written as:

$$\zeta^{(n)} = \sum_{k=1}^{n+1} (-1)^{k-1} K_k^{(n)} y^k, \qquad (18)$$

with $K_k^{(n)} = 0 \ \forall n < 0, k < 1, k > n+1$.

The coefficients $K_k^{(n)}$ for the first few values of n are given in Table 2.

TABLE 2
The Coefficients, $K_k^{(n)}$, for the Stirling Form of the First Six Derivatives

Derivative Number (n)	k						
	1	2	3	4	5	6	7
0	1						
1	1	1					
2	1	3	2				
3	1	7	12	6			
4	1	15	50	60	24		
5	1	31	180	390	360	120	
6	1	63	602	2100	3360	2520	720

Each $K_k^{(n)}$ is the coefficient for a term of the form y^k; $1 \le k \le n$, and equals $(k-1)!S_{n+1,k}$, where $S_{n,k}$ are the Stirling Numbers of the Second Kind. The sign is given by $(-1)^{k-1}$.

THEOREM 2. *The coefficient* $K_k^{(n)}$ *of* y^k *in* $\zeta^{(n)}$ *is equal to* $(k-1)!S_{n+1,k}$, *where* $S_{n,k}$ *are the* Stirling Numbers of the Second Kind (*see Graham et al.*, 1989 *for a description*).

Proof. Using eqn (17), the recurrence for generating $K_k^{(n)}$ is:

$$K_k^{(n)} = (k-1)K_{k-1}^{(n-1)} + kK_k^{(n-1)}. \tag{19}$$

The proof can be facilitated by considering the triangle of coefficients K generated by eqn (19), with the rows indexed by n and the columns by k (see Table 2). The assertion of the theorem can be proved by double induction on k and n, using the standard recursion for Stirling numbers of the Second Kind: $S_{n+1,k} = S_{n,k-1} + kS_{n,k}$.

$k = 1$: eqn (16) gives $K_1^{(1)} = 1$, and from eqn (19),

$$K_1^{(n)} = (1-1)K_0^{(n-1)} + 1K_1^{(n-1)} = 1 \ \forall n \geq 0$$

Because $S_{n+1,1} = 1 \ \forall n \geq 0$, the assertion is satisfied for $k = 1$.

Suppose the assertion holds for column $k-1$. Then, proceeding with the induction on n, consider the first non-zero element of column k. Because this corresponds to $n = k-1$, it is given by

$$\begin{aligned} K_k^{(k-1)} &= (k-1)K_{k-1}^{(k-2)} + kK_k^{(k-2)} \\ &= (k-1)K_{k-1}^{(k-2)} \quad \text{since} \quad K_k^{(k-2)} = 0 \\ &= (k-1)(k-2)!S_{k-1,k-1} \\ &\quad \text{(by the hypothesis of the induction)} \\ &= (k-1)!S_{k-1,k-1}, \end{aligned}$$

but $S_{k-1,k-1} = S_{k,k} = 1$. Which means that

$$K_k^{(k-1)} = (k-1)!S_{k,k}.$$

Thus, the assertion holds for the first non-zero element of column k (corresponding to $n = k-1$). Suppose the assertion holds for the $n-1$ term in column k, i.e., for $K_k^{(n-1)}$. Then, for the n term

$$\begin{aligned} K_k^{(n)} &= (k-1)K_{k-1}^{(n-1)} + kK_k^{(n-1)} \\ &= (k-1)(k-2)!S_{n,k-1} + k(k-1)!S_{n,k} \\ &= (k-1)![S_{n,k-1} + kS_{n,k}] \\ &= (k-1)!S_{n+1,k}, \end{aligned}$$

which means that the assertion holds for $K_k^{(n)}$. Thus, the double induction is complete, and the assertion is proved. □

REMARK 6. *Following the above results,* $\zeta^{(n)}$ *may be written as:*

$$\zeta^{(n)} = \sum_{k=1}^{n+1} (-1)^{k-1}(k-1)!S_{n+1,k}y^k \tag{20}$$

which may be called the Stirling form *of* $\zeta^{(n)}$.

TABLE 3
The Coefficients, $t_k^{(n)}$, for the Hyperbolic Tangent Form of the First Six Derivatives

Derivative Number (n)	k			
	2	4	6	8
0	1			
1	2			
2	4	2		
4	8	16		
5	16	88	16	
6	32	416	272	
7	64	1824	2880	272

Defining $p = 0.5\,wx$, each $t_k^{(n)}$ is the coefficient for a term of the form $2^{-(n+1)}\text{sech}^{2k}(p)\tanh^{n+1-2k}(p)$; $1 \leq k \leq \lfloor (n+1)/2 \rfloor$, and equals $T_{n,k-1}$, where $T_{n,k}$ is the number of permutations of the first n integers with exactly k peaks. The sign is given by $(-1)^{n-k}$.

Many other equivalent forms can, of course, be found for the derivatives using standard combinatorial relationships. For example, by writing the sigmoid as a hyperbolic tangent function:

$$\frac{1}{1+e^{-wx}} = \frac{e^{wx}}{1+e^{wx}} = \frac{1}{2}[\tanh(wx/2) + 1], \tag{21}$$

and defining $p = wx/2$, it is possible to write $\zeta^{(n)}$ as

$$\zeta^{(n)} = \frac{1}{2^{n+1}} \sum_{k=1}^{\lfloor (n+1)/2 \rfloor} t_k^{(n)} \text{sech}^{2k}(p)\tanh^{n+1-2k}(p). \tag{22}$$

It appears that the coefficients $t_k^{(n)}$ are expressible in terms of a well-known group of numbers: permutations of the first n integers by the number of peaks:

$$t_k^{(n)} = (-1)^{n-k}T_{n,k-1}, \tag{23}$$

where $T_{n,k}$ are the number of permutations of the first n integers with exactly k peaks (see David et al., 1966, pp. 53, 260). The unsigned coefficients $t_k^{(n)}$ for the first few values of n are shown in Table 3. The sign is determined by $(-1)^{n-k}$. The study of this and other forms is left for the future.

6. CONCLUSION

In this paper, we have derived two very simple and elegant formulations for the derivatives of the standard sigmoid function used in neural network research. Not only do these formulations lead to very efficient methods for calculating higher derivatives, they also relate the sigmoid to well-known combinatorial number families, which can help in the analysis of systems with sigmoidal units.

REFERENCES

David, F. N., Kendall, M. G., & Barton, D. E. (1966). *Symmetric function and allied tables.* Cambridge, UK: Cambridge University Press.

Graham, R. L., Knuth, D. E., & Patashnik, O. (1989). *Concrete mathematics.* Reading, MA: Addison-Wesley.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models.* Cambridge, UK: Chapman & Hall.

Minai, A. A. (1992). *The robustness of feed forward neural networks: A preliminary investigation.* Ph.D. Dissertation. University of Virginia, Charlottesville, VA.

Stolorz, P., Lapedes, A., & Xia, Y. (1992). Predicting protein secondary structure using neural net and statistical methods. *Journal of Molecular Biology,* **225,** 363–377.

Werbos, P. (1974). *Beyond regression: New tools for prediction and analysis in the behavioral sciences.* Ph.D. Thesis. Harvard University, Cambridge, MA.

APPENDIX

In this appendix, we briefly describe how our formulation could be used to estimate the effect of input perturbations on the output of layers of sigmoid neurons.

Consider a layer of two neurons connected to the same two inputs, x_1 and x_2, through weight vectors $\mathbf{w}_1$ and $\mathbf{w}_2$, respectively. Let the neuron outputs by y_1 and y_2, and define $\mathbf{x} = [x_1 x_2]$. We can thus write

$$[y_1 y_2] = [\sigma(\mathbf{x}; \mathbf{w}_1)\sigma(\mathbf{x}; \mathbf{w}_2)].$$

Suppose there is a perturbation, $\delta\mathbf{x} \equiv [\delta x_1 \delta x_2]$ in the input. The change in output is then approximated by

$$\delta y_1 \approx \sum_{n=1}^{N} \frac{1}{n!} (\|\mathbf{w}_1\| \|\delta\mathbf{x}\| \cos \beta_1)^n \zeta^{(n)}(y_1)$$

$$\delta y_2 \approx \sum_{n=1}^{N} \frac{1}{n!} (\|\mathbf{w}_2\| \|\delta\mathbf{x}\| \cos \beta_2)^n \zeta^{(n)}(y_2),$$

where N is the order of the approximation and β_i is the angle between $\delta\mathbf{x}$ and $\mathbf{w}_i$. The approximation holds provided the perturbation lies within the radius of convergence for the sigmoid. This is $O(\|\mathbf{w}_i\|^{-1})$ (see Minai, 1992 for details). Thus, the magnitude of the effect at the layer's output is $\|\delta\mathbf{y}\|$, where $\delta\mathbf{y} = [\delta y_1 \delta y_2]$.