Chapter 8

SIGMOIDAL FUNCTION CLASSES FOR
FEEDFORWARD ARTIFICIAL NEURAL NETWORKS

8.1 Introduction

The popularity of the sigmoidal feedforward artificial neural networks (FFANN’s) may be traced to the universal approximation results for these networks [10, 38, 40, 43, 49, 66, 67, 104, 105, 196, 219]. These results have been used to justify sigmoidal FFANN usage for solving tasks (apparently) as diverse as - (a) Function approximation, (b) Regression, (c) Classification and (d) Density Estimation [41].

The universal approximation results (UAR) are primarily existential in nature. They assure that if a given function is to be approximated, there exists an (appropriate) FFANN that approximates it arbitrarily well; but leave the task of finding the “appropriate” network unsolved. This “problem” of finding the appropriate network is “solved” by the FFANN training procedures. The training algorithms involve arbitrary choices for the parameters of the network. Some of these choices are:

1. *The architecture (graph) of the network*. The number of input(s) and the output(s) define the number of the nodes in the input and the output layers, the UARs indicate that at least one hidden layer is required. Usually, the
number of nodes in the hidden layer is fixed (though, there are algorithms that allow pruning of nodes and/or growth of nodes [87]). The pruning and the constructive algorithms may be viewed as adaptive algorithms that adapt the architecture itself.

2. *The weights of the network.* The usual practice is to initialize the weights to (small) uniform random values in the range $[-r, r]$, where $r > 0$ and $r \in \mathbb{R}$. However, other weight initialization techniques have also been developed [183, 286].

3. *Optimization Algorithm.* The choice of the specific optimization algorithm to be used is another parameter for which heuristic guidelines are available [144].

4. *The activation function.* The specific activation function(s) used at the (active) nodes of the FFANN are fixed before the FFANN is trained. This is primarily due to the limited number of sigmoidal functions that are utilized in practice [58].

Data preprocessing and postprocessing techniques are not considered in this chapter or elsewhere in this thesis, as conceptually these may be considered as operations performed by additional layers present in the FFANN.

The quality of the approximation/network obtained by the training mechanisms depends on the initialization conditions and the choices (above mentioned) made [58, 131]. The activation functions, also called squashing functions, used in the learning algorithms for FFANN training, play an important role in determining the speed of training [144, 58]. The role of activation functions in FFANN training has not been investigated to the desired extent [58]. The usage of anti-symmetric activation functions have been favoured on symmetry grounds [144]. In chapter 7, a class of sigmoidal functions was defined as a generalization of the log-sigmoid function. In the present work, the results
are extended so that any sigmoidal function may be used as a generator of a sigmoidal class of functions.

In the next Section 8.2, the properties of the commonly used activation functions are summarized and a physically relevant formulation of the universal approximation theorem is given for sigmoidal activations. This allows the abstraction of the properties that any activation must satisfy to be both practically useful and also satisfy the requirements to act as activation in networks with universal approximation capability. In Section 8.3, two methods are presented to generate a corresponding class of sigmoidal functions from any given sigmoidal function or its derivative, and the properties of these classes are enumerated. In Section 8.4, examples are provided using the commonly used activations to generate the corresponding sigmoidal classes. Conclusions are presented in Section 8.5.

8.2 Sigmoidal Activation Functions

A sigmoidal (or sigmoid) function class may be defined as (reproduced from definition 2.1):

**Definition 8.1.** Let $S$ represent a class of functions whose elements are real functions such that, $\sigma(x) \in S$, $\sigma(x) : \mathbb{R} \rightarrow \mathbb{R}$, with the property that its limits for $x \rightarrow \pm \infty$ are:

\[
\lim_{x \to -\infty} \sigma(x) = \alpha \tag{8.1}
\]

\[
\lim_{x \to -\infty} \sigma(x) = \beta \tag{8.2}
\]

with $\alpha > \beta \tag{8.3}$
The usual values of are \( \alpha = 1 \) and \( \beta = 0 \) or \( -1 \), however, any value satisfying (8.3) are admissible.

The class of sigmoidal functions, as defined by definition 8.1, is the set of all bounded functions satisfying the condition of relation (8.3). It contains discontinuous functions like the Heaviside function, the sign or the signum function etc. as well as continuous functions like the tangent inverse, hyperbolic tangent, log-sigmoid function etc. Continuity, differentiability and monotonic increasing conditions are imposed on the class defined to obtain the class of analytic sigmoid functions defined as:

**Definition 8.2.** Let \( S_n \subset S \) represent the class of analytic sigmoid functions whose elements are real functions such that, \( \sigma_n(x) \in S_n \) is a continuous, differentiable (and with bounded derivatives) and (strict) monotonic increasing function.

Only the functions belonging to the set \( S_n \) are considered in this chapter; as it is the functions belonging to this class that have come to play an important role in the application areas of FFANN’s. This has been primarily due to the algorithmic requirements for finding the weight updates. This imposes the requirement of differentiability (to evaluate weight updates). The universal approximation capability of these functions is summarized in a physically relevant form as [87, 240]:

**Theorem 8.1.** Let \( \sigma_n \in S_n \) be a bounded, strict monotonically increasing function. Let \( X \) be a n-dimensional compact subset of \( \mathbb{R}^n \). Let the space of continuous real functions defined on \( X \) be represented by \( C(X) \). Then, given any function \( f \in C(X) \), \( f : X \to \mathbb{R} \) and a real \( \epsilon > 0 \), there exists an integer \( k \) and a set of real values \( \alpha_i, \beta_i, \gamma \), and \( w_{ij} \) where \( i = 1, 2, \ldots, k \) and \( j = 1, 2, \ldots, n \) such that one may define (8.4):

\[
F(x) = F(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{k} \alpha_i \sigma \left( \sum_{j=1}^{n} w_{ij} x_j + \beta_j \right) + \gamma \tag{8.4}
\]

220
where \( x \in \mathcal{X} \) and \( x \equiv (x_1, x_2, \ldots, x_n) \), as an approximate realization of the function \( f \); that is
\[
|F(x_1, x_2, \ldots, x_n) - f(x_1, x_2, \ldots, x_n)| \leq \epsilon 
\] \hspace{1cm} (8.5)

**Proof**: See [10, 38, 40, 43, 49, 66, 67, 104, 105, 196, 219] for proof(s) under various conditions. \( \square \)

The theorem 8.1 justifies the usage of the members of the set \( S_a \) as activation functions. This also enumerates the conditions that the activation functions should satisfy. The properties of a suitable activation function therefore may be summarized as (let \( s_a \in S_a \), and an implicit presence of the input domain \( \mathcal{X} \) is always assumed):

**Property 8.1.** The function \( s_a(x) \) is a real continuous function.

**Property 8.2.** The function \( s_a(x) \) is a real strictly increasing function.

**Property 8.3.** The function \( s_a(x) \) is a bounded function, the limits for \( x \to \pm \infty \) are given by (8.1) and (8.2).

**Property 8.4.** The function \( s_a(x) \) is a real differentiable function and satisfies the (simple) differential equation (8.6):
\[
\frac{ds_a(x)}{dx} = g(x) 
\] \hspace{1cm} (8.6)

**Property 8.5.** The function \( g(x) \) is a positive definite in the domain of the inputs.

**Proof**: Follows from the strict monotonic increasing requirement (8.2) on \( s_a(x) \) (see [115, p.36] theorem 3.11). \( \square \)

In the case of the generally used activation functions, one may impose the following conditions (together with the condition of property 8.5 which is specified as (8.7)):
\[ g(x) > 0; \ x \in \mathbb{R} \] \quad (Positivity) \quad (8.7)
\[ g(x) \leq M; \ x \in \mathbb{R}, \ M \in (0, \infty) \] \quad (Boundedness) \quad (8.8)
\[ \lim_{x \to \pm \infty} g(x) = 0; \ x \in \mathbb{R} \] \quad (Behaviour at infinity) \quad (8.9)
\[ \int_{-\infty}^{\infty} g(t) \ dt = \sigma(x) \in (\beta, \alpha); \ x \in \mathbb{R} \] \quad (Sigmoidal Integral) \quad (8.10)

### 8.3 Sigmoidal Function Classes

In this section, two procedures for the generation of sigmoidal classes are formulated.

#### 8.3.1 First method

A family of sigmoidal functions is generated from any given sigmoidal function \( s_a \in \mathcal{S}_a \) using a power equation. The generated family of sigmoids all satisfy the same bounds (eq. 8.1 and 8.2) as the function \( s_a \). The family of sigmoids is represented as \( \tilde{\mathcal{S}}_a \). Its members are parameterized by the parameter \( m \) and are given by the following relation:

\[ \tilde{s}_a^m = (\alpha - \beta) \left[ \frac{s_a - \beta}{\alpha - \beta} \right]^m + \beta; \ m \in (0, \infty) \] \quad (8.11)

The properties 8.1-8.5 have to be verified before the members of \( \tilde{\mathcal{S}}_a \) can be used as activation functions.

**Proposition 8.1.** The function \( \tilde{s}_a^m (x) \in \tilde{\mathcal{S}}_a \) is a real continuous function.

**Proof:** Follows from property 8.1 and the properties of the power function and relation (8.11) \( \square \)
Proposition 8.2. The function $\tilde{s}^m_a (x) \in \tilde{S}_a$ is a real strictly increasing function.

Proof: The function $s_a \in S_a$ is a real strictly increasing function (by assumption). Using (8.3), one can establish that the function (8.12) is strictly increasing and has a range $(0, 1)$.

$$s_{aux} (x) = \frac{s_a - \beta}{\alpha - \beta} \quad (8.12)$$

For any real $y_1, y_2 \in \mathbb{R}$, with $y_1 > y_2 > 0$, then for any real $m > 0$ one has $y_1^m > y_2^m$.

If $f$ is any strictly increasing function, then $af + b$ with $a > 0; a, b \in \mathbb{R}$ is a strictly increasing function. $\alpha - \beta > 0$, hence proved. \hfill $\square$

Proposition 8.3. The function $\tilde{s}^m_a (x) \in \tilde{S}_a$ is a bounded function, the limits for $x \rightarrow \pm \infty$ are given by (8.1) and (8.2).

Proof: It can easily be established from (8.1), (8.2) and (8.12) that for the function $s_{aux} (x) (8.12)$, the limits for $x \rightarrow \pm \infty$ are given by:

$$\lim_{x \rightarrow \infty} s_{aux} = 1 \quad (8.13)$$

$$\lim_{x \rightarrow -\infty} s_{aux} = 0 \quad (8.14)$$

The proof follows from the analytical properties of the power function. \hfill $\square$

Proposition 8.4. The function $\tilde{s}^m_a (x) \in \tilde{S}_a$ is a real differentiable function and satisfies the (simple) differential equation (8.15):

$$\frac{d\tilde{s}^m_a (x)}{dx} = \tilde{g} (x) \quad (8.15)$$
Proof: Differentiating (8.11) w.r.t. $x$ one obtains:

$$
\frac{d s^m_a (x)}{dx} = m \left[ \frac{s_a - \beta}{\alpha - \beta} \right]^{m-1} \frac{ds_a (x)}{dx} \quad \text{(Use (8.6)and(8.11))}
$$

$$
= \frac{m}{\alpha - \beta} s^{m-1}_a (x) \ g(x) - \frac{m \beta}{\alpha - \beta} g(x)
$$

(8.16)

Using (8.6) in (8.16), the proof is completed with $\hat{g} (x)$ given by:

$$
\hat{g} (x) = m \left[ \frac{s_a - \beta}{\alpha - \beta} \right]^{m-1} g(x)
$$

(8.17)

$$
\hat{g} (x) = \frac{m}{\alpha - \beta} s^{m-1}_a (x) \ g(x) - \frac{m \beta}{\alpha - \beta} g(x)
$$

(8.18)

where $g(x)$ is given by (8.6).

□

Proposition 8.5. The function $\hat{g} (x)$ is a positive definite in the domain of the inputs.

Proof: Follows from (8.12) and the positivity of the $s_{aux}$ function, property 8.5 (positivity of $g(x)$), positivity of $m, m \in (0, \infty)$ and relation (8.17).

One may establish the following result for the derivatives of the members of the class $\hat{S}_a$.

Proposition 8.6. The envelope of the derivatives of the members of the class $\hat{S}_a$ is given by:

$$
y(x) = \frac{-1}{e (\alpha - \beta)} \frac{1}{\ln (s_{aux} (x))} \ \frac{g(x)}{s_{aux} (x)}
$$

(8.19)

where $s_{aux} (x)$ is given by (8.12).

Proof: The proof is obtained by eliminating $m$ from the relation:

$$
U (m) = \frac{m}{\alpha - \beta} s^{m-1}_a (x) \ g(x) - \frac{m \beta}{\alpha - \beta} g(x) = 0
$$

(8.20)

$$
\frac{dU (m)}{dm} = \frac{s^{m-1}_a (x) - \beta}{\alpha - \beta} \left[ 1 + m \ln \left( \frac{s_a (x) - \beta}{\alpha - \beta} \right) \right] = 0
$$

(8.21)
From the relation (8.19), one can easily establish that the function $y(x)$ is continuous, differentiable and strictly positive. Thus, the boundedness and the increasing (monotonic) property of the function $y(x)$ should be investigated to conclude whether $y(x)$ is also sigmoidal in nature or not. The limits of the function $y(x)$ for $x \to \infty$ depends on the limit of the ratio:

$$\lim_{x \to \pm \infty} \frac{g(x)}{\ln \left( (s_{aux})^{-(s_{aux})} \right)} = (8.22)$$

If one assumes that the limit of (8.22) can be evaluated then the limits may be written as:

$$\lim_{x \to -\infty} \frac{g(x)}{\ln \left( (s_{aux})^{-(s_{aux})} \right)} = \alpha_1 \quad (8.23)$$

$$\lim_{x \to -\infty} \frac{g(x)}{\ln \left( (s_{aux})^{-(s_{aux})} \right)} = \beta_1 \quad (8.24)$$

In general, $\alpha_1$ and $\beta_1$ can take any real value (may be infinite or zero, separately or together). The function $y(x)$ is sigmoidal if $\alpha_1$ and $\beta_1$ are finite and $\alpha_1 > \beta_1$. The limits in (8.23) and (8.24) must be evaluated in particular cases. Moreover, if $\alpha_1$ and $\beta_1$ are finite and $y(x)$ is a general sigmoid, then one may establish that the condition for monotonic increasing property is:

$$\frac{dy(x)}{dx} > 0 \Rightarrow \frac{dg(x)}{dx} > 1 + \frac{1}{\alpha - \beta} \frac{g(x)^2}{\ln \left( (s_{aux})^{-(s_{aux})} \right)} \quad (8.25)$$

Thus, the functions defined by (8.11) satisfy the conditions to act as an activation function (or satisfies the requirements of theorem 8.1). Moreover, in
this subsection, it has established that under suitable conditions (8.25) the envelope (8.19) of the derivative functions of the class (8.11) is also sigmoidal in nature.

8.3.2 Second method

This method utilizes the derivative of a sigmoidal function to generate a family of sigmoidal functions. Let the derivative of a sigmoidal function \( s_a \in S_a \), be represented by \( g(x) \) (8.6). Let the function \( g(x) \) satisfy the conditions of equations (eq. 8.7 – 8.10). Then one may prove the following propositions:

**Proposition 8.7.** The function \( g(x) \) satisfies the relation (8.26) for \( m \in (0, \infty) \).

\[
(g(x))^m > 0, \ x \in \mathbb{R}, \ m \in (0, \infty)
\]  

(8.26)

**Proposition 8.8.** The function \( g(x) \) satisfies the relation (8.27) for \( m \in (0, \infty) \).

\[
(g(x))^m \leq M^m, \ x \in \mathbb{R}, \ m \in (0, \infty)
\]  

(8.27)

**Proposition 8.9.** The function \( g(x) \) satisfies the relation (8.28) for \( m \in (0, \infty) \).

\[
\lim_{x \to \pm \infty} (g(x))^m = 0, \ x \in \mathbb{R}, \ m \in (0, \infty)
\]  

(8.28)

**Proposition 8.10.** The function defined by the relation (8.29) exists for \( m \in [1, \infty) \), and is monotonically increasing.

\[
\int_{-\infty}^{x} (g(t))^m \ dt = \sigma_g(x), \ x \in \mathbb{R}, \ m \in [1, \infty)
\]  

(8.29)
**Proof**: Since, \( g(x) \) satisfies the relations (8.7) and (8.8), for any \( m \in [1, \infty) \) one may write:

\[
\sigma_g(x) = \int_{-\infty}^{x} (g(t))^m \, dt \\
\leq M^{m-1} \int_{-\infty}^{x} g(t) \, dt \\
(\text{Using the analytical properties of } g(x)) \\
\leq M^{m-1} s_n(x) \\
\text{where } s_n(x) \text{ is the sigmoid integral of } g(x) \text{ (8.10).}
\]

The above result establishes that the integral (8.29) exists. The monotonic increasing property follows from the fact that the derivative of \( \sigma_g(x) \) is \((g(x))^m > 0.1\)

\[
\square
\]

### 8.4 Examples

In this section, the two methodologies proposed in Section 8.3 are used to construct sigmoidal classes which are generated from the commonly used sigmoidal functions [58]. The sigmoidal functions that are used are the log - sigmoidal function (8.34), the error function (8.35), and another function defined by (8.36).

\[
s_1 = \frac{1}{1 + e^{-x}} \\
\]

\[
s_2 = \text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_{0}^{x} e^{-t^2} \, dt \\
\]

\[
\text{In some cases it is possible to extend the result to hold for the values of } m \in (0, \infty), \text{ but this requires a case by case analysis.}
\]
### Learning in Sigmoidal Feedforward Artificial Neural Networks

<table>
<thead>
<tr>
<th>Function</th>
<th>( \alpha )</th>
<th>( \beta )</th>
<th>( s_{a}^{m} )</th>
<th>( y^* )</th>
<th>( \alpha_1 )</th>
<th>( \beta_1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( s_1 )</td>
<td>1</td>
<td>0</td>
<td>((s_{1})^{m})</td>
<td>( e^{-x^{-1}} )</td>
<td>(1/e )</td>
<td>0</td>
</tr>
<tr>
<td>( s_2 )</td>
<td>1</td>
<td>-1</td>
<td>(2\left(\frac{s_{2}+1}{2}\right)^{m}-1)</td>
<td>(2 e^{-x^{2}-1} / \left(\text{erf}(x) + 1\right))</td>
<td>(\infty)</td>
<td>0</td>
</tr>
<tr>
<td>( s_3 )</td>
<td>(\kappa)</td>
<td>(-\kappa)</td>
<td>(2\kappa\left(\frac{s_{3}+\kappa}{2\kappa}\right)^{m}-\kappa)</td>
<td>(\sqrt{\pi} \ln \left(\frac{1}{2} \text{erf}(x) + \frac{1}{2}\right)) (x - \sqrt{1 + x^2}) (\sqrt{1 + x^2})</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 8.1: The properties of the sigmoidal functions of eq. (8.34), (8.35) and (8.36).

\( y^* \) represents the envelope of the derivatives of the function class(es) generated (eq. 8.11).

\[
s_3 = \frac{\kappa x}{\sqrt{1 + x^2}} \quad \text{(where } \kappa \text{ is a positive and finite real.)} \tag{8.36}
\]

Table 1 summarizes the property of these sigmoidal functions. It also represents the sigmoidal classes generated from them by the first method.

Fig. 8.1 - 8.6 show the functions and their derivatives obtained by the first method. From the graphs it is clear that three types of functional classes are obtained, namely – (a) The function set whose derivative’s envelope is sigmoidal, (b) The function set whose derivative’s envelope unbounded, and (c) The function set whose derivative’s envelope is the x-axis, that is at infinity the derivatives of the activation functions vanishes.

From the table it’s clear that the derivatives envelope may be a sigmoidal function \((s_{1})\), may be a unbounded increasing function \((s_{2})\) and may have the limiting value equal to zero \((s_{3})\). Also the derivatives of the functions of the classes are skewed except for the function(s) corresponding to the parameter value \(m = 1\).

For the functions \(s_{i} \in S_{a}, i = 1, 2, 3\), the corresponding derivatives are:
Figure 8.1: The functions obtained from $s_1$. 
Figure 8.2: The derivatives of the functions obtained from $s_1$. 
Figure 8.3: The functions obtained from $\varepsilon_2$. 
Figure 8.4: The derivatives of the functions obtained from $s_2$. 
Figure 8.5: The functions obtained from $s_3$. 
Figure 8.6: The derivatives of the functions obtained from $s_3$. 
\[ g_1 = \frac{e^{-x}}{1 + e^{-x}} = s_1 (1 - s_1) \quad (8.37) \]

\[ g_2 = \frac{2}{\sqrt{\pi}} e^{-x^2} \quad (8.38) \]

\[ g_3 = \kappa (1 + x^2)^{-\frac{3}{2}} \quad (8.39) \]

The second method is illustrated by using the above three (derivative) functions or the function \( g_i, i = \{1, 2, 3\} \) (8.37)-(8.39). The class generated is represented as \( S_{g_i} \), and the members of \( S_{g_i} \)'s are represented as:

\[ \sigma_{g_i}^m (x) = \int_{-\infty}^{x} \left( \frac{e^{-t}}{(1 + e^{-t})^2} \right)^m dt; \quad m \in [1, \infty) \quad (8.40) \]

and it was not possible to integrate this integral explicitly, and may have to be numerically evaluated.

\[ \sigma_{g_2}^m (x) = \frac{2}{\sqrt{\pi}} \int_{-\infty}^{x} e^{-mt^2} dt; \quad m \in [1, \infty) \quad (8.41) \]

\[ = \frac{1}{\sqrt{m}} \text{erf} (\sqrt{m} x); \quad m \in [1, \infty) \quad (8.42) \]

\[ \sigma_{g_3}^m (x) = (\kappa)^m \int_{-\infty}^{x} \left( \frac{1}{1 + x^2} \right)^{\frac{3m}{2}} dt; \quad m \in [1, \infty) \quad (8.43) \]

\[ = x_2 \mathcal{F}_1 \left( \frac{1}{2}, \frac{3m}{2}, -x^2 \right); \quad m \in [1, \infty) \quad (8.44) \]

where \( _2\mathcal{F}_1(a, b, c, x) \) represents the three parameter hyper-geometric series for \( x \).
[1, 253].

In the illustrations for the second method, in two cases the method did give an expression for the sigmoidal family generated from the associated $g$’s ($g_2$ and $g_3$). In the second case ($g_2$) it is apparent that the range of $m$ can be analytically continued to the range $(0, \infty)$.

Moreover, it is apparent from the form of the expressions obtained for the sigmoidal families by the two procedures, that the first method varies the shape of the functions while leaving the range invariant, the second method may change both the shape as well as the range (amplitude).

8.5 Conclusions

In the current chapter, two methodologies for the creation of parameterized sigmoidal function classes were demonstrated. The first method is guaranteed to give a parameterized family which is not too computationally expensive. While the second method also provides a methodology for the creation of sigmoidal families of functions, the method may entail numerical evaluation of an integral which can be substantially costly (in terms of efficiency). The first method provides a mechanism where the generating and the generated sigmoids have the same range, whereas in the second method the amplitude may also change.

The universal approximation results only assume a sigmoidal activation to be present in the network. No specific activation is favoured. Thus, it appears that self-adaptation of the sigmoidal activation may lead to data-dependent activations being used. The adaptation of activation functions have been used and shown to be more efficient in training than the static (activation) networks [265, 287, 37, 107]. The existing works adapt either the amplitude [265], or the parameters of the standard log-sigmoid function (the $a$ and $b$ parameters of the the function $a/(1 + e^{-bx})$ [107] or use auxiliary tuning parameters [287, 37]. An activation function auto-tuning approach may be constructed for
the self-adaptation of the parameter $m$.

The results of this chapter are reported in [249].