The Capability of Feedforward Neural Networks with \( \Omega \)-Shaped Activation Functions

Yiwei Chen
Western Atlas International
P.O.Box 1407
Houston, Texas 77251

Farokh Bastani
Department of Computer Science
University of Houston
Houston, Texas 77204-3475

Abstract
The capability of the multilayer feedforward layered neural network with \( \Omega \)-shaped activation functions is studied in this paper. We prove that a three layer neural network with any continuous \( \Omega \)-shaped activation function can approximate any continuous function in the multi-dimensional real space. Further theoretical extensions to the generalized \( \Omega \)-shaped function are also explored. One example of such kind of neural networks is the Hermite network, for which there is one the efficient method to initialize the network in order to speed up convergence for the backpropagation learning algorithm.

1 Introduction
Artificial neural network models are composed of many simple computational elements operating in parallel and connected in structures reminiscent of biological neural networks. Computational elements or neural units (neurons) are connected via weighted links that are typically adapted during use to improve performance. There has been a resurgence in the field of artificial neural networks during the last decade, caused by new network topologies, adaptive and learning algorithms, hardware implementation techniques, and the belief that massive parallelism is essential for high performance pattern recognition.

The multilayer perceptron is a feedforward neural network. The term “feedforward” implies that the output of an element cannot be the input for another element on the same layer or a previous layer, i.e., all connections are unidirectional. The input layer normally performs no function other than the buffering of the input signals. The next layer is named the hidden layer because its outputs are internal to the neural network. Typically only one hidden layer exists, but for complex mappings or situations where the network size is restricted, additional hidden layers might exist between the input and the output layers. Special learning algorithms are used to adapt the connection weights between consecutive layers to achieve a satisfactory solution for a given problem.

Mathematically, a feedforward layered network is formed from the composition and superposition of simple nonlinear functions used as the neuron activation functions. Accordingly, the output at each output unit is the value of a function resulting from that particular composition and superposition of the nonlinearities.

We are mainly interested in the capability of the three layer feedforward neural network which typically consists of \( n+1 \) input units with identity activation function, \( N \) hidden units with a univariate activation function \( f \), and one output unit with identity activation function. The last input unit always has the constant value 1 as the input since the weights from it to hidden units are the variation of the thresholds for all the hidden units [1]. The range of input for each input unit is the unit interval \( I = [0,1] \), and the domain of the output is the real space \( \mathbb{R} \). The representation of the output is the linear combination of the form

\[
G(f, z) = \sum_{j=1}^{N} \alpha_j f(y_j^T x + \theta_j)
\]

where \( y_j \in \mathbb{R}^n \) is the weight vector from the input layer to the \( j \)th hidden unit, \( \alpha_j \in \mathbb{R} \) is the weight from \( j \)th hidden unit to the output unit, and \( z \in \mathbb{R}^m \) is the input vector for the first \( n \) input units, and \( \theta_j \) is the weight from the \( (n+1) \)th input unit to the \( j \)th hidden unit.

A diverse range of real functions have been used as the activation function for artificial neural network units. Sigmoidal functions are the most popularly used activation functions [2, 3, 4]. The general mathematical definition of the sigmoidal function is given below:

**Definition** A real function \( \sigma \) on \( \mathbb{R} \) is called a sigmoidal function, if

\[
\sigma(z) \rightarrow \begin{cases} 
1 & \text{as } z \to +\infty \\
0 & \text{as } z \to -\infty 
\end{cases}
\]

Figure 1 shows several examples of the sigmoidal function.

In the past few years, a number of results on the sigmoidal functions as the activation function have
been published. Cybenko shows that the three layer network with any continuous sigmoidal function can approximate all the continuous functions on $\mathbb{P}$ to an arbitrary precision [5]. Funahashi obtains the same result for the monotonic sigmoidal functions [6]. Cybenko [5] also proves that the network with any bounded measurable sigmoidal function can realise all the functions in $L^1(\mathbb{P})$. Cotter [7] recently proposed modified logistic networks with exponential functions, polynomials, partial fractions, and Boolean functions and proved that they are capable of approximating arbitrary bounded measurable functions. A wide variety of other functions can also be employed as activation functions to construct networks that approximate some class of functions. Geometrical analysis of the capabilities of three layer networks on forming disconnected decision regions can be also found in the references [8] and [9]. Cybenko [5] also indicated that the Wiener Tauberian theorem [10] implies that the above network with any function in $L^1(\mathbb{R})$, with non-zero integral, can approximate all the functions in $L^1(\mathbb{R})$.

Another function class of interest is the $\Omega$-shaped function. Examples of such kind of functions which could be used as the neuron activation functions include the piece-wise polynomial [11], Gaussian functions [12], and some other functions shown in Figure 2. Intuitively, it is the kind of function converging to zero at infinity, in both the positive and negative directions, and reaching its absolute supreme only in one closed interval.

**Definition** A real function $\omega_\alpha(z)$ on $\mathbb{R}$ is called an $\Omega$-shaped function, for any given $\alpha \geq 0$, if all of the following conditions are satisfied,

1. $\omega_\alpha(z) = 1$ as $|z| \leq \alpha$,
2. $|\omega_\alpha(z)| < 1$ as $|z| > \alpha$,
3. $\lim_{|z| \to \infty} \omega_\alpha(z) = 0$.

We will show that the three layer feedforward network with a continuous $\Omega$-shaped function as the activation function can approximate $C(\mathbb{P})$, the space of continuous functions on $\mathbb{P}$. This result is different in that the Wiener Tauberian theorem [10] requires the function to have non-zero integral while an $\Omega$-shaped function can have zero integral. The $\Omega$-shaped function is more intuitive in the suppression and excitation of the input signal, since it simply functions like a bandwidth filter.

The rest of this paper is organized as follows. In Section 2 we provide and prove the basic function approximation result for the continuous $\Omega$-shaped function. All the extensions and further discussions are presented in Section 3. Section 4 describes an applicable network example which uses an instance of the $\Omega$-shaped function. Section 5 summarizes the whole paper.

2 Neural Network Capabilities

The capability of an artificial neural network architecture is measured by the set of all problems it can solve. Theoretically, we are more interested in seeing what kinds of function spaces can be approximately realised. In this section, we will present the main result on the feedforward neural network architecture defined in the previous section and with an arbitrarily given continuous $\Omega$-shaped function as the activation function for the hidden layer. The basic mathematical tool used is Real Variables and Functional Analysis. All the fundamental terminologies used in deduction and proofs can be found in the Appendix. A few more basic notations and terminologies to be used include $C_0(\mathbb{R})$, the space of all continuous $\Omega$-shaped functions; $||f|| = \sup\{|f(z)| : z \in \mathbb{P}\}$, the supremum norm of a function $f$; and the dense concept of a set of functions $E$ which is defined as follows: $E$ is said to be dense in another set of functions $D$, iff for any $f \in D$ and $\alpha > 0$, there is a function $g \in E$, such that $||f - g|| < \alpha$; and $I^+$, the set of positive integers.
Theorem 1 Let \( \omega \in C_0(\mathbb{R}) \) be a continuous \( \Omega \)-function. Then the set of \( G(\omega, z) \) for all \( N \in \mathbb{N}^+ - \{+\infty\} \), i.e., a positive and finite integer \( N, y_j \in \mathbb{R}^n \) and \( c_j \in \mathbb{R}, 1 \leq j \leq N \) is dense in \( C(I^m) \) with respect to the supreme norm.

Before we prove this theorem, we need to mention three important results already available to us. The Cybenko Theorem listed below is from [5]. The second and third, the Dominated Convergence Theorem, can be found in [10], [13] or other real analysis bibliographies.

**Definition** We say a function \( f \) is discriminatory if for a measure \( \mu \in M(I^m) \) which is the space of finite, signed regular Borel measures on \( I^m \), such that

\[
\int_{I^m} f(y^T z + \theta) d\mu(z) = 0
\]

for all \( y \in \mathbb{R}^n \) and \( \theta \in \mathbb{R} \), implies that \( \mu = 0 \).

Cybenko theorem Let \( f \) be any continuous discriminatory function. Then the set of \( G(f, z) \) for all \( N, y_j, c_j, \) and \( \theta_j \) is dense in \( C(I^m) \) with respect to the supreme norm.

**Theorem** The finite valued simple functions are dense in \( L^\infty(I^m) \).

**Dominated convergence theorem** If \( f_1, f_2, ..., f \), \( g \) are Borel measurable functions, \( |f_n| \leq g \) for all \( n \), where \( g \) is \( p \)-integrable, and \( f_n \to f \), then \( f \) is \( p \)-integrable and

\[
\int_{I^m} f_n d\mu \to \int_{I^m} f d\mu.
\]

Hence, if we can prove that any continuous \( \Omega \)-shaped function is discriminatory, the main result is merely an application of the Cybenko theorem. In the following, we first show that any continuous \( \Omega \)-shaped function \( \omega_\alpha \) with \( \alpha > 0 \) is discriminatory. We denote that set of functions by \( C_{N,\alpha}(\mathbb{R}) \). Next we prove the main result in Theorem 1. In fact, a superset of \( C_{N,\alpha}(\mathbb{R}) \) is of discriminatory property which is stated in the following lemma.

**Lemma 1** Any Borel measurable \( \Omega \)-shaped function \( \omega_\alpha \) with \( \alpha > 0 \) is discriminatory.

**Proof.** First, we prove that \( \omega_\alpha \) for any positive integer \( i \) is also Borel measurable. According to the definition of Borel measurable, the set \( \{ z| c < \omega_\alpha(z) \} \) should be a Borel set for all \( c \in \mathbb{R} \) in order to prove the hypothesis. When \( i \) is odd, since we know that \( \{ z| c < \omega_\alpha(z) \} \) is a Borel set for all \( c \in \mathbb{R} \), it also holds for \( c^{1/i} \). Hence we can say that \( \{ z| c^{1/i} < \omega_\alpha(z) \} \) is a Borel set and so its identical set \( \{ z| c < \omega_\alpha(z)^i \} \). When \( i \) is an even number, the deduction is similar while \( c \geq 0 \). Otherwise, we have \( \{ z| c < \omega_\alpha(z)^i \} \) is its identical set.

Let \( \omega_\alpha \) be any continuous \( \Omega \)-shaped function, \( \omega_\alpha(z) \) is discriminatory, the main result is merely an application of the Cybenko theorem. In the following, we first show that any continuous \( \Omega \)-shaped function is discriminatory. We denote that set of functions by \( C_{N,\alpha}(\mathbb{R}) \). Next we prove the main result in Theorem 1. In fact, a superset of \( C_{N,\alpha}(\mathbb{R}) \) is of discriminatory property which is stated in the following lemma.

**Lemma 1** Any Borel measurable \( \Omega \)-shaped function \( \omega_\alpha \) with \( \alpha > 0 \) is discriminatory.

**Proof.** First, we prove that \( \omega_\alpha \) for any positive integer \( i \) is also Borel measurable. According to the definition of Borel measurable, the set \( \{ z| c < \omega_\alpha(z)^i \} \) should be a Borel set for all \( c \in \mathbb{R} \) in order to prove the hypothesis. When \( i \) is odd, since we know that \( \{ z| c < \omega_\alpha(z)^i \} \) is a Borel set for all \( c \in \mathbb{R} \), it also holds for \( c^{1/i} \). Hence we can say that \( \{ z| c^{1/i} < \omega_\alpha(z)^i \} \) is a Borel set and so its identical set \( \{ z| c < \omega_\alpha(z)^{i^2} \} \). When \( i \) is an even number, the deduction is similar while \( c \geq 0 \). Otherwise, we have \( \{ z| c < \omega_\alpha(z)^i \} \) is its identical set.
Proof of Theorem 1. The lemma shows that any measurable \( \Omega \)-shaped function \( \omega_a \) is discriminatory if \( \alpha > 0 \). Since any continuous \( \Omega \)-function is measurable, the lemma also holds for any continuous \( \Omega \)-function \( \omega_a \) with \( \alpha > 0 \). Here we just need to show that the same result is true for \( \omega_0 \) also. In fact we just need to prove that the set \( \{ \omega_a : \alpha > 0 \} \) is dense in the set \( \{ \omega_a : \alpha \geq 0 \} \).

According to the definition, for any given \( \omega_0 \) and \( \epsilon > 0 \), we are going to construct an \( \omega_a \) such that

\[
||\omega_0 - \omega_a|| < \epsilon.
\]

Select a neighborhood \( O(0, 2a) \) such that

\[
||\omega_0 - 1|| < \epsilon
\]

inside \( (0, 2\alpha) \) since \( \omega_0 \) is continuous. It can be easily verified that an \( \omega_a \) constructed as in the following will be inside \( O(\omega_0, \epsilon) \).

\[
\omega_a(t) = \begin{cases} 
\omega_0(t) + \frac{1}{a} t & \text{as } |t| > 2a \\
\frac{1}{a} (\omega_0(-2a) + 2 - \omega_0(-2a)) & \text{as } -2a < t < -a \\
1 & \text{as } |t| \leq \alpha \\
\frac{1}{a} (\omega_0(2a) - 1) + 2 - \omega_0(2a) & \text{as } a < t < 2a \\
\omega_0(t) - \frac{1}{a} t & \text{as } |t| < a
\end{cases}
\]

Q.E.D.

The main result is thus completely proved. A straightforward conclusion is that the one-output-unit three layer feedforward neural network activated by an arbitrarily given continuous \( \Omega \)-shaped function for every hidden unit can approximate any continuous function to any arbitrary precision. Of course, the underlying assumption here is that there are no constraints on the size of the hidden layer.

3 Generalization and Further Discussion

A generalised \( \Omega \)-shaped function can be defined as any bounded function with \( \{ x \mid f(z) = \sup \{ f(z) \mid z \in R \} \} = [a, b] \) for two existing real numbers \( a < b \) and \( a, b \in R \). Then we have a generalization of Theorem 1 as follows.

Theorem 2. Let \( \omega \) be a continuous generalised \( \Omega \)-shaped function. Then the set of \( G(\omega, x) \) for all \( N \in I^+ - \{ +\infty \}, y_j \in R^n \) and \( \alpha_j, \theta_j \in R, 1 \leq j \leq N \) is dense in \( C(I^+) \) with respect to the supreme norm.

This theorem can be easily proven since a generalised \( \Omega \)-shaped function \( \omega \) is equivalent to an \( \Omega \)-shaped function \( \sup \{ (f(x)) \} \) with the scale factor \( \sup \{ |f(x)| \} \).

There is theoretical interest in the discontinuous \( \Omega \)-shaped function also. For an arbitrarily given generalised \( \Omega \)-shaped function, we have a theorem that is analogous to Theorem 2 but which holds in the space \( L^1(I^+) \) instead.

Theorem 3. Let \( \omega \) be a measurable generalised \( \Omega \)-shaped function. Then the set of \( G(\omega, x) \) for all \( N \in I^+ - \{ +\infty \}, y_j \in R^n \) and \( \alpha_j, \theta_j \in R, 1 \leq j \leq N \) is dense in \( L^1(I^+) \).

In order to prove Theorem 3 analogously to the approach for proving Theorem 1, we have to redefine the discriminatory function as in the following.

Definition. We say a function \( f \) is discriminatory if for a function \( h \in L^\infty(I^+) \),

\[
\int_{I^+} f(y^T z + \theta) h(z) dz = 0
\]

for all \( y \in R^n \) and \( \theta \in R \) implies that \( h(z) = 0 \) almost everywhere.

Measureable generalised \( \Omega \)-shaped functions with \( b > a \) are discriminatory as a result of Lemma 1 and its definition. Cybenko Theorem should be also changed to the following result:

Lemma 2. Let \( f \) be any discriminatory function. Then the set of \( G(f, x) \) for all \( N, y_j, \alpha_j, \theta_j \in R \) is dense in \( L^1(I^+) \) with respect to \( ||\cdot||_1 \).

Proof. This proof is analogous to that of the Cybenko Theorem. Let

\[
S = \{ G(f, x) \mid N \in I^+ - \{ +\infty \}, y_j \in R^n, \alpha_j, \theta_j \in R \}.
\]

Obviously, \( S \) is a linear subset of \( L^1(I^+) \) since \( |f| \) is \( \mu \)-integrable on \( I^+ \). If we take \( L^1(I^+) \) as the entire space, then we can claim that the closure of \( S, \overline{S}, \) is exactly \( L^1(I^+) \).

Assume that \( \overline{S} \neq L^1(I^+) \). Then \( \overline{S} \) is a closed proper subspace of \( L^1(I^+) \). By the Hahn-Banach Theorem, there is a bounded linear functional \( F \) on \( L^1(I^+) \), with the property that \( F(L^1) \neq 0 \), but \( F(\overline{S}) = F(S) = 0 \).

By the Riesz Representation Theorem, there is a unique \( h \in L^\infty(I^+) \) and \( h \neq 0 \), such that

\[
F(g) = \int_{I^+} g(z) h(z) dz,
\]

for all \( g \in L^1(I^+) \). In particular, since \( f(y^T z + \theta) \in \overline{S} \subset L^1(I^+) \) for all \( y \) and \( \theta \), and \( F(\overline{S}) = 0 \), we have

\[
F(f(y^T z + \theta) = \int_{I^+} f(y^T z + \theta) h(z) dz = 0
\]

for all \( y \) and \( \theta \).

Since \( f \) is discriminatory, we conclude that \( h = 0 \) almost everywhere. This contradicts the assumption that \( \overline{S} \neq L^1(I^+) \).

Q.E.D.

Theorem 3 is merely the result from both Lemma 1 and Lemma 2.

All the discussion above is about the three layer feedforward network with one output unit. For a general feedforward network with \( l \) output units, Theorems 2 and 3 should be restated as...
Theorem 4. Let \( \omega \in C^0(\mathbb{R}^n) \) be a continuous \( \Omega \)-shaped function. Then any vector of functions in \( (C^0(\mathbb{R}^n))^m \) can be approximated by a three layer feed-forward network of \( l \) output units, with \( \omega \) as the activation function of the hidden layer, at any given precision with respect to the supreme norm.

Theorem 5. Let \( \omega \) be a measurable generalized \( \Omega \)-shaped function on \( \mathbb{R}^n \). Then any vector of functions in \( (L^1(\mathbb{R}^n))^m \) can be approximated by a three layer feed-forward network of \( l \) output units, with \( \omega \) as the activation function of the hidden layer, at any given accuracy with respect to \( L^1 \) norm.

Since we are only studying the capability of the feedforward network class, the existence of such networks as in Theorem 4 and 5 is obvious. They can be the integration of \( l \) independent networks each with one output unit.

4 Network Example

An example of such a class of artificial neural networks is the Hermite network. A Hermite network is a three layer feedforward neural network. The activation function of its hidden layer unit is the base function for the segmented piece-wise Hermite spline of degree three,

\[
H(z) = \begin{cases} 
\phi_0\left(\frac{z}{h}\right) & 0 \leq z \leq h \\
\phi_0\left(-\frac{z}{h}\right) & -h \leq z \leq 0 \\
0 & \text{otherwise}
\end{cases}
\]

an \( \Omega \)-shaped function, where

\[
\phi_0 = 2z^3 - 3z^2 + 1.
\]

Figure 3 illustrates the shape of the activation function. The advantages of such an activation function are that in addition to its continuity, differentiability and nonlinearity, the computational complexity is almost negligible comparing with an exponential function. Further, we find that this kind of neural network can be initialized with optimized weights other than starting the learning procedure with randomly selected initial weights.

![Figure 3: The Hermite base function.](image-url)

Without loss of generality, we will just describe how the initialization algorithm works for the network with \( m \) input units, one extra input unit \((m+1)\)th with constant input of 1, \( n \) hidden units, and just one output unit.

We first group the hidden layer into \( l \) groups, with \( k \) units in each group, assuming \( n = lk \) for simplicity. Within each hidden layer group, we enforce the initial input to each unit to be identical for the same input to the network. This can be achieved simply by assigning the same weight vector to every hidden unit of the same group.

Now we can start to initialize the weights from the first \( m \) input units to the hidden layer. The idea behind this is the feature selection. By picking up an optimized feature selection method, such as the Walsh transform or the Fisher classifier, we can optimally (under certain criteria) project the \( m \)-dimensional input space onto an feature space with \( l \) bases. Note that there should not be restrictions such as \( l \leq m \). When \( l < m \), we normally require that all \( l \) bases in the feature space are orthogonal vectors. If we denote \( w_1, w_2, ..., w_l \) to be the bases for the feature space, we simply scale them and assign each to every respective hidden unit group, i.e., the weight vector from the \( m \) input units to every unit within \( j \)th group is assigned with the scaled \( w_j \). The purpose of scaling is to guarantee that every input to the hidden unit will be inside \([0, 1] \).

The weights from the \((m+1)\)th input unit to the hidden layer is related to the segmented intervals for the Hermite spline. We divide the range \([0, 1] \) into \( k-1 \) isometric segments of length

\[
h = \frac{1}{(k-1)},
\]

where \( k > 1 \) is the size of the hidden group. Let \( z_j = jh \), for \( j = 1, 2, \ldots, k \), be knots of the spline. Then the weight from the extra input unit to the \( j \)th unit within each hidden group is \(-z_j\), in order to have the activation function of that unit be the Hermite base function centered at \( z_j \).

For each hidden group, the weights from that group to the output unit should make the network output the result of the best fit Hermite spline to the input, i.e.,

\[
o(z) = \sum_{j=1}^{k} y_j H(z - z_j),
\]

where \( z \) is the input to that hidden group, and \( y_j \) is naturally the weight for the connection from the \( j \)th unit of the group to the output unit.

The rest of this section is devoted to a least square error solution to the coefficients \( y_j \) in \( o(z) \). For the given training set of size \( r \), let

\[
(u_1, u_2, \ldots, u_r)
\]

be all transformed input data to group \( g \) in the hidden layer and

\[
(x_1, x_2, \ldots, x_r)
\]

be the desired corresponding output set. If every \( u_i \) is equal to an interpolation knot, say, \( z_i \), for some \( i \),
of the optimal spline approximation theory, there is a unique optimal solution in our inner product space, according to the optimal approximation theory, there is a unique optimal solution. Let us assume that the function to be approximated is given by $f(z)$, where $f(z)$ is the Hermite spline function, and $o(z)$ is the Hermite spline function, and $P(z)$ is the linear interpolating function for $H(z)$.

The norm of a function $f$ is defined as $\|f\|_2 = \left(\int_{-h}^{h} f(x)^2 dx\right)^{1/2}$. The inner product of two functions $f_1$ and $f_2$ is defined as the integration of their product, that is $\langle f_1, f_2 \rangle = \int_{-h}^{h} f_1(x)f_2(x)dx$.

The norm of a function $f$ is given by $\|f\|_2 = \left(\sum_{i=1}^{n} |f_i|^2 \right)^{1/2}$, where $n$ is the number of training patterns.

The optimal solution to this scaling vector is estimated with the $LSE$ as $c = A^\dagger z$, where $A^\dagger = [A^T A]^{-1}A^T$, assuming $r \gg l$, i.e., the number of training patterns are far more than the number of hidden layer groups. In the actual computation, it is not necessary to know the inverse matrix. We can get $c$ by solving the linear system $[A^T A]c = A^T z$.

Now we can assign initial values to $w(r)$, i.e., $w(r) = c(r)y(r)$.

The weight initialisation is accomplished and, therefore, the network can then be trained with the generalised delta rule.

If the Walsh transform is used, the whole initialization cost is comparable to that of one sweep of the backpropagation algorithm [11]. Figure 4 shows a scenario of how a function is approximated by a Hermite Network.

Although the neural network with the $\Omega$-shaped function as the activation function has been proved to have at least the same capability as that using the sigmoidal function, it gives us more opportunities to search for more applicable network models. This particular Hermite network is an example of a network that can be easily initialized for real applications, without losing the capability of mapping from one set to another.
5 Conclusion

We have studied a class of functions called $\Omega$-shaped functions. We proved that a three layer neural network with any continuous $\Omega$-shaped function as its activation function can approximate any continuous function in the multi-dimensional real space. One conclusion of this on the application of decision problem is that such a network can solve an arbitrarily partitioned multi-dimensional space to any desired accuracy, provided that there is no limit on the number of hidden units. In other words, any decision functions on a multi-dimensional space can be arbitrarily well approximated by such a neural network.

Some extensions to the main result are also presented. Theorems 4 and 5 provide the capabilities for the general type of three layer feedforward neural networks.

One such kind of neural network is the Hermite network, which has an efficient method to initialize works. Theorems 4 and 5 provide the capabilities for the network in order to have a faster convergence versus other types such as sigmoidal functions as the activation function, under the assumption that the size of the hidden layer is fixed.

So far, only the mapping capability of the neural network has been discussed. Further work needs to be done on the classification power of such a network versus other types such as with sigmoidal functions as the activation function, under the assumption that the size of the hidden layer is fixed.

References


Definition If \( S \) is a class of sets, \( \sigma(S) \) is the smallest \( \sigma \)-field containing the sets of \( S \), and sometimes called the minimal \( \sigma \)-field over \( S \).

Definition \( B(\mathbb{R}) \) is the collection of Borel sets of \( \mathbb{R} \) and defined as the smallest \( \sigma \)-fields of \( \mathbb{R} \) containing all intervals \((a, b], a, b \in \mathbb{R}\).

Definition A measure on a \( \sigma \)-field \( \mathcal{F} \) is a nonnegative, extended real-valued function \( \mu \) on \( \mathcal{F} \) such that whenever \( A_1, A_2, \ldots \in \mathcal{F} \) form a finite or countably infinite collection of disjoint sets in \( \mathcal{F} \), we have

\[
\mu(\bigcup_n A_n) = \sum_n \mu(A_n).
\]

Definition A topological space is a set \( \Omega \) with a collection \( \mathcal{T} \) (called topology of subsets of \( \Omega \)) such that \( \emptyset \) and \( \Omega \) belong to \( \mathcal{T} \) and \( \mathcal{T} \) is closed under finite intersection and arbitrary union. The members of \( \mathcal{T} \) are called open sets. A set is closed iff its complement is open.

Definition A topological space \( \Omega \) is said to a Hausdorff space iff given any two distinct points \( x \) and \( y \), both points have disjoint neighborhoods.

Definition \( \Omega \) is a normal topological space iff Omega is Hausdorff, and if \( A \) and \( B \) are disjoint closed subsets of \( \Omega \), there are disjoined open sets \( U \) and \( V \) with \( A < U \) and \( B < V \).

Definition If \( \mu \) is a measure on \( B(\Omega) \), where \( \Omega \) is a normal topological space, \( \mu \) is said to be regular iff for each \( A \in B(\Omega) \),

\[
\mu(A) = \inf\{\mu(V) : V > A, V \text{ open}\}
\]

and

\[
\mu(A) = \sup\{\mu(V) : V < A, V \text{ closed}\}.
\]

Definition A signed measure \( \mu \) is the difference between two measures \( \mu_1 \) and \( \mu_2 \).

Definition \( M(I^n) \) is the space of finite, signed regular Borel measures on \( I^n \).

Definition A real value function \( f : E \to \mathbb{R} \), \( E \in \mathbb{R} \), is said a Borel Measurable Function if for and \( c \in \mathbb{R} \), there must have

\[
ex|f(x)| \geq c, x \in E \in B(E).
\]

Definition A set of functions \( E \) is dense in another set of functions \( D \), if for any \( f \in E \) and \( \alpha > 0 \), there is a function \( g \in D \), such that

\[
||f - g|| < \alpha.
\]

Definition \( L^p(I^n) \) is the space of all Borel Measurable functions \( f \) on \( I^n \) such that \( |f|^p \) is \( \mu \)-integrable, i.e., both

\[
\int_{I^n} |f|^p d\mu
\]

and

\[
\int_{I^n} |f|^p d\mu
\]

are finite.

Definition \( L^{\infty}(\mathbb{R}) \) is the space of all functions \( f \) such that \( ||f|| < \infty \).

Definition Let \( (\Omega, \mathcal{F}, \mu) \) be a measurable space. If function \( h : \Omega \to \mathbb{R} \cup \{-\infty\} \cup \{\infty\} \) is said simple iff \( h \) is Borel measurable and takes on only finitely many distinct values. Equivalently, \( h \) is simple iff it can be written as a finite sum \( \sum_{i=1}^n a_i I_{A_i} \) where \( A_i \) are disjoint sets in \( \mathcal{F} \); \( I_{A_i} \) is the indicator of \( A_i \) and \( x_i \in \mathbb{R} \).

Definition A set of functions \( E \) is dense in another set of functions \( D \), if for any \( f \in E \) and \( \alpha > 0 \), there is a function \( g \in D \), such that

\[
||f - g|| < \alpha.
\]