A New Approximating Model for the Time Invariant Nonlinear Operators with Fading Memory

ADRIAN BUDURA(1), SILVIU CRISAN(2), GEORGETA BUDURA(3),
(1) ETA2U, Timisoara, (2) Identification Systems Division, Canadian Bank Note Company, Ottawa
(3) Communication Department “Politehnica” University of Timisoara
Bd. V. Parvan. No. 2
ROMANIA
abudura@eta2u.ro, georgeta.budura@etc.upt.ro, scrisan@cbnco.com

Abstract: - The paper presents a construction theorem for a class of operators dense over the set of causal, time invariant fading memory operators. In this sense, it extends the classical results of S. Boyd and L.O. Chua that the Volterra series operators are universal approximators for this set of nonlinear operators often encountered in the theory of dynamical systems. This new representation is based on the remarkable property of the neural network ΣΠ functions to be a dense algebra in the set of continuous functions over compacta in \( \mathbb{R}^n \). More, this class of functions is known to allow effective approximations of non-analytical type non-linearities and, as a consequence, to avoid higher order terms else way present in a polynomial decomposition. It is expected that with a proper choice of the ΣΠ base functions this property transfers to the non-linear operator representation. Following this reasoning, we are able to prove the inclusion of the Volterra series in this richer set of nonlinear operators.

Key-Words: - Non-linear operators, Fading memory, Volterra series, ΣΠ algebras, Neural networks

1 Introduction

The Volterra series approximation of time-invariant non-linear continuous operators comes as an intuitive extension to the convolution operators associated with the time-invariant linear dynamical systems (Eq.1).

\[
V\{u(t)\} = h_0 + \sum_{n=1}^{\infty} H_n\{u(t)\}
\]

where \( H_n\{u(t)\} \) represents the n order Volterra operator:

\[
H_n\{u(t)\} = \int h_n(\tau_1, \ldots, \tau_n) u(t-\tau_1) \cdots u(t-\tau_n) d\tau_1 \cdots d\tau_n
\]

To our knowledge, it is also the only explicit result in the approximation theory of the non-linear dynamical systems that are time invariant and physically realizable, i.e. causal. The scope of our article is to formulate an alternative series of operators that are generated by algebras of functions encountered in the neural network theory. These sets of functions commonly noted as ΣΠ in the literatures are known to supersede the polynomials. We will prove that this result encountered in the approximation theory of the continuous functions defined over real compacta will be transferred to the inclusion of the operators sub-algebra generated from a polynomial expansion i.e the Volterra series and the one constructed from the neural networks ΣΠ functions.

Long time present and tacitly accepted, the folk theorem stating the approximation of our class of operators by the Volterra series has been rigorously proved by Rugh in [1] for \( h_n \) kernels with a compact support and then by S. Boyd and L.O. Chua in [2] for \( h_n \) kernels having a fading memory. For the clarity of our exposure, we will highlight in the next section this last case as it will become the departing point in our development of a new approximation model. The proof given in [2] offers also the construction scheme for the Volterra approximating series as a linear dynamical system followed by a memoryless polynomial non-linearity. Consequently, any time-invariant non-linear continuous operators can be represented up to an arbitrary level of accuracy by such a structure: a linear dynamical system followed by a memoryless non-linearity.

It is the polynomial representation of this last read-out map that constitutes an argument in the favor of our search for a less regular, larger class of functions that are dense in the space of continuous functions over compact metric spaces. In this sense, it is known the notorious difficulty of polynomial series to accurately approximate non-linearities with a less than \( C^\infty \) regularity i.e. compact support, discontinuities in the higher derivatives etc. However, it is important to remark that we need to follow a path in our proof based only on the general properties of the class of operators we want to...
approximate without any prior knowledge of their physical realization. Our demonstration follows the Stone-Weierstrass and hence the ability to define an algebra on such a class of functions is strictly necessary.

Fortunately, an excellent example exists for such an approximating set of functions. It is the \( \Sigma II \) neural network functions as described in [3] or [4] to constitute an algebra capable to approximate as well as needed any continuous function defined on a compact metric space.

The structure of our paper is as follows: the first section introduces the necessary concepts and gives a brief summary of the results found in [2] relevant to our theory development. The second section states and proves the theorem allowing a new approximating class for the time invariant non-linear operators with fading memory. At the end of this part, we prove the inclusion of the Volterra series in this class of approximators for a particular choice of the \( \Sigma II \) functions. An example is given where the multidimensional linear system is based on the Laguerre polynomials while the \( \Sigma II \) non-linearity is formed upon the third order spline functions.

2 Notations, definitions, preliminary results

Let us identify with \( C(R) \) the Banach space of the continuous bounded functions defined over the negative part of the real axis with the uniform norm given by relation (2).

\[
\|u(t)\| = \sup |u(t)|, \quad t \in R
\]  

(2)

Over this space, \( C(R) \) we consider a real valued functional \( F : C(R) \rightarrow R \). We shall also suppose it to be continuous. From this functional we can always construct a causal, continuous, time invariant (TI) operator \( O : C(R) \rightarrow C(R) \) with values in the \( C(R) \) space of bounded continuous functions over \( R \). The property of time invariance can be expressed in a synthetic form as the commutativity between this operator \( O \) and the delay operator \( U \).

\[
U_z \{ u(t) \} = u(t - \tau)
\]

(3)

The causality of the \( O \) operator leads to the following implication: if \( u(t) = v(\tau) \) for \( \tau \leq t \) then \( O \{ u(t) \} = O \{ v(\tau) \} \).

The causal and TI operator \( O \) will be written with the help of the functional \( F \) as follows:

\[
O \{ u \} = F \{ T U_z \{ u \} \}
\]

(5)

where:

\[
T : C(R) \rightarrow C(R_+), \quad T \{ u(t) \} = u(t) \quad \text{for} \quad t \leq 0
\]

is the trivial truncation operator.

We may see this relation as a continuous mapping through the functional \( F \) of an advancing input \( u(\tau) \) to the operator output \( O \{ u(\tau) \} \).

The continuity of this operator appears as an immediate consequence of its construction and the continuity of the functional \( F \).

The relation (5) is important as it will allow us to work on the approximation theory for a set of continuous functionals rather than operators.

At this point of our exposure it is useful to write the Stone-Weierstrass theorem in a convenient form:

Let \( A \) be a set of continuous functionals on a compact subspace \( K \subset C(R_+) \) which fulfill the Stone-Weierstrass theorem hypothesis - see for example [5].

(i) \( A \) is an algebra together with the \((+,\times)\) operations

(ii) \( A \) separates points in \( K \):

\[ \forall u,v \in K, u \neq v \text{ there is an } F \in A \text{ so that } F(u) \neq F(v). \]

(iii) the \( A \) algebra vanishes at no point in \( K \), in other words, for any \( u \in K \) it exists an \( E \in A \) so that \( E(u) \neq 0 \)

Then, \( A \) is dense in the set of continuous functionals over \( K \).

As \( C(R) \) is not compact under the uniform norm given in (2) it is necessary to define a subset \( K \subset C(R_+) \) that is seen as compact under a different metric. The property of the compact support or fading memory of our causal, TI operators facilitates the definition of such a metric.

Let \( K \) be defined as the functions from \( C(R) \) that are square integrable and fulfill a Lipschitz type condition with a bound \( M \):

\[
|u(t) - u(\tau)| \leq M|t - \tau| \quad \forall u \in K
\]

(6)

We define a norm over this space with the help of a weighting function \( w \).

\[
\|u\|_w = \sup_{t \in K} |u(t)| \cdot w(-t)
\]

(7)

The weighting function is considered to be continuous over the \( C(R_+) \) and with values in the \((0,1]\) interval. We shall suppose it also to be monotonically decreasing toward + \( \infty \).

i. \( w : C(R_+) \rightarrow (0,1]\)

ii. \( \lim_{t \to -\infty} w(t) = 0 \) with \( w(s) < w(t), \forall (s > t) \)
Examples of such functions are multiple, we resume only to $e^{-\lambda t}$, $\lambda > 0$ exponential typical for the relaxation processes of linear dynamical systems.

This weighting function will model the fading memory concept and it is easy to see that in the norm defined by (7), two functions are close if they are close for $t$ argument values close to $0$. The continuity of the functional $F$, and by extension (5) of the operator $O$, now restricted to $(K, \| \cdot \|_{w})$ can be formulated in very intuitive terms proper to dynamical systems: inputs that are almost identical in the recent past will lead to an almost identical output in the present. In a close form, we shall write that for any $\delta > 0$ we may find a positive $\varepsilon$ so that if $\|u - v\|_{w} < \varepsilon$ then implies $|F(u) - F(v)| < \delta$.

We shall give in the remaining part of this section the main steps that lead to the Volterra series rigorous formulation.

**Proposition 1:**
The $K$ space endowed with the $\| \cdot \|_{w}$ norm, has a compact structure. The demonstration of this proposition can be found in [2].

Over the space $K$ is defined a set $L$ of numerable functionals $L_{m} : K \to R$. The functionals $L_{m}$ are expressed with the help of the integral kernels $l_{m}$:

$$L_{m}(u) = \int_{0}^{\infty} l_{m}(\tau) u(-\tau) d\tau$$  (8)

The $l_{m}$ kernel functions need to have two important properties:

i. they are integrable with the weight $w^{-1}$ over: $(0, \infty)$; $B$ is their positive bound.

$$\int_{0}^{\infty} |l_{m}(\tau)| \cdot w^{-1}(\tau) d\tau < B$$  (9)

ii. the $l_{m}$ set of functions forms a complete basis over the $L^{2}(R_{+})$.

At this point, it is important to make the following observation. It is the fading function present in the integrability condition (9) that makes $K$ to appear compact to the $L_{m}$ functionals. We can restate this observation in a relation if we rewrite the expression (8) as follows:

$$L_{m}(u) = \int_{0}^{\infty} \left[l_{m}(\tau) \cdot w^{-1}(\tau)\right] w(\tau) u(-\tau) d\tau =$$

$$= \int_{0}^{\infty} [l_{m}(\tau) \cdot w^{-1}(\tau)] u(-\tau) d\tau$$  (10)

Now, the linear functional with the kernel $l_{m} = l_{m} w^{-1}$ will be defined over elements of $K$ having the form $u' = u \cdot w$.

The definition of the $L_{m}$ functionals together with the compact structure of the $(K, \| \cdot \|_{w})$ space allows the construction of a polynomial algebra fulfilling the Stone-Weierstrass theorem. This is the theoretical result that leads to the Volterra series. It is proved and exemplified in [2].

We shall rework the demonstration in the next section of the paper for an algebra based on the $\Sigma \Pi$ functions encountered in the neural network theory.

### 3 $\Sigma \Pi$ approximation model for the non-linear time invariant operators with fading memory

We shall begin by giving a few results found to be useful in the application of the Stone-Weierstrass theorem.

**Proposition 2:**
The $L_{m}$ functionals are continuous over the $(K, \| \cdot \|_{w})$ space.

**Proof:**

$$|L_{m}(u) - L(v)| \leq \int_{0}^{\infty} |l_{m}(\tau)| \cdot w(\tau) \cdot |u(-\tau) - v(-\tau)| d\tau =$$

$$= \int_{0}^{\infty} |l_{m}(\tau)| \cdot w(\tau) \cdot |u(-\tau) - v(-\tau)| \cdot w(\tau) d\tau \leq$$

$$\leq \sup_{\tau \in R_{+}} |u(-\tau) - v(-\tau)| \cdot w(\tau) \cdot \int_{0}^{\infty} |l_{m}(\tau)| \cdot w(\tau) d\tau \leq$$

$$B \cdot \|u - v\|_{w}$$

**Proposition 3:**
The $L$ set of functionals separates points in $K$.

**Proof:**

Let $u$, $v \in K$ and $\sup_{\tau \in R_{+}} |u(\tau) - v(\tau)| > 0$.

Note $f(\tau) = u(-\tau) - v(-\tau)$ an element in the $L^{2}(R_{+})$ Hilbert space. Due to the continuity property of the $u$ and $v$ functions, the norm in $L^{2}(R_{+})$ of the function $f$ cannot be zero.

The functional $L_{m}(f) = \int_{0}^{\infty} l_{m}(\tau) \cdot f(\tau) d\tau$ is nothing else than the scalar product on $L^{2}(R_{+})$ between the $l_{m}$ kernel and the function $f$: $\langle l_{m}, f \rangle$. As $\|f\|_{L^{2}} \neq 0$ and $L$ is a complete set over $L^{2}(R_{+})$ it leads to the conclusion that exist at least one kernel $l_{n}$ so that $\langle l_{n}, f \rangle \neq 0$.

Before enunciating our approximation theorem we should make a few brief remarks regarding the class of the $\Sigma \Pi$ functions encountered in the theory of the neural networks. The interested reader may consult the classical paper of Hornick, Stinchcombe
and White [3] or the alternative demonstration given by Cybenko [4]. The presentation we adopt follow loosely that on Hornick’s article.

For a continuous, non-constant function \( \varphi : R \rightarrow R \) we will define a set of functions:
\[
\Sigma \Pi(\varphi) : C \subset R' \rightarrow R
\]
with \( C \) a compact.

The functions belonging to this set have the general expression:
\[
g_{\varphi}(x) = \sum_{j=1}^{J} b_j \prod_{k=1}^{K} \varphi[a_{jk}(x)]
\]
(12)
where: \( a_{jk} : R' \rightarrow R \) are affine functions:
\[
a_{jk}(x) = \sum_{i=1}^{R} \alpha_{jki} x_i + \gamma_{jk}, \quad \alpha_{jki}, \gamma_{jk} \in R
\]
\[b_j \in R, \quad J \text{ and } K_j \in N^*\]
The set of functions \( \Sigma \Pi(\varphi) \) becomes an algebra with the \((+\times)\) operations and it is dense in the Banach space of continuous functions defined over compacta \( C \) in \( R' \). In other words, any continuous function \( f : C \subset R' \rightarrow R \) can be approximated with an arbitrarily accuracy by a function in the set \( \Sigma \Pi(\varphi) \). The sequence of functions converging towards the continuous function \( f \) is indexed by the integers \( J \) and \( K_j \) that model the size of our \( \Sigma \Pi(\varphi) \) network.

An important simplification in the structure of the \( \Sigma \Pi(\varphi) \) network is allowed when a supplementary condition is imposed to the function \( \varphi \). If \( \varphi \) is a “squashing function”, i.e. \( \varphi : R \rightarrow [0,1] \), non-decreasing and \( \lim_{x \to +\infty} \varphi(x) = 0 \), \( \lim_{x \to -\infty} \varphi(x) = 1 \). Then the \( \Sigma \Pi(\varphi) \) structure could be simplified to a \( \Sigma(\varphi) \) network \( (K_j = 1) \) while preserving the property of being dense over \( R' \) compacta. The classical structure of a three layers feed-forward neural network is found to be the natural model for this \( \Sigma(\varphi) \) class of functions. The archetype example for the squashing functions is the sigmoid. Remarkable \( \varphi \) functions that don’t have a squashing behavior and yet are used in \( \Sigma(\varphi) \) type networks are the radial basis functions. The \( n \) time convolution of the identity \( \{E_{[a,b]}\} \) over a compact \( [a,b] \) is a good example in this sense: \( \varphi(x) = (\ast)^n E_{[a,b]} \). The linear combination of the properly translated \( \varphi \) recover the space of the spline functions of order \( n \). We shall not insist further on this subject well treated in the literature.

We are ready now to state the theorem governing the \( \Sigma \Pi \) approximation of the continuous functionals defined over the compact metric space \( \{K_i, \| \| \} \).

**Theorem 1**
Let us consider the non-linear set of functionals
\[
G_{\varphi}(L_i(u)) : K \rightarrow R, G_{\varphi}(L_i) = \sum_{j=1}^{J} b_i \prod_{k=1}^{K} \varphi[A_{ik}(L_i)]
\]
where:
- \( L_i \) is an \( I \)-finite selection of functional \( L_i \in L \)
- \( \varphi \) is non-constant, bounded and continuous over \( R \)
- \( A_{ik}(L_i) = \sum_{i=1}^{I} \alpha_{ik} L_i + \gamma_{ik} I_d \), \( \alpha_{ik}, \gamma_{ik} \in R \)
- \( I_d(u) = 1 \) \( \forall u \in K \), is the unity functional \( 1_k \)
\[b_j \in R, \quad I, \ J \text{ and } K_j \in N^*\]
and let \( F \) be a continuous functional defined over the compact space \( (K_i, \| \|) \).

Then, for any \( \varepsilon > 0 \) and any \( u \in K \) it exists a functional \( S \in G_{\varphi}, \ G_{\varphi} \) constructed with sufficiently large \( I \), \( J \) and \( K_j \) numbers so that
\[
|F(u) - S(u)| < \varepsilon
\]

**Proof:**
a) The functionals belonging to \( G_{\varphi} \) are continuous over the \( K \) space.
The property is immediate as the \( L_i \in L \) functionals are continuous and the \( \varphi \) function is continuous and bounded on \( R \).

b) The closeness of the \( G_{\varphi} \) set with the \((+,\times)\) operations and scalars multiplication is obvious through definition.

c) The unity functional \( 1_k \) belongs to \( G_{\varphi} \).
We can always choose a functional \( E \in G_{\varphi} \) so that \( E(u) = \varphi(\gamma) = 1 \), \( \forall u \in K \).

d) The \( G_{\varphi} \) set separates points in \( K \).
Let \( u,v \in K \) and \( \sup_{r \in K} |u(r) - v(r)| > 0 \). The

**Proposition 3** ensures the existence of a functional \( L_n \in L \) so that \( L_n(u) \neq L_n(v) \). Consequently we may choose a functional \( P \in G_{\varphi} \) so that \( P \{1\} = \varphi(\alpha \cdot L_n \{1\}) \).
The parameter \( \alpha \) is chosen so that \( \varphi(\alpha \cdot L_n \{u\}) \neq \varphi(\alpha \cdot L_n \{v\}) \), a choice always possible when \( \varphi \) is not a constant.
e) The $K$ space seen through the $G_{\varphi}$ set of functionals is compact.

The construction of the $G_{\varphi}$ set preserves the fading memory property of the $L$ set of functionals. Consequently, the Proposition 1 and the arguments that followed lead to an effective compactness of the $K$ space with the norm $\|\cdot\|_w$.

In consequence of a) to e), the Stone-Weierstrass theorem applies to the set of functionals $L_i$ and the theorem is proved.

A simple corollary of the Theorem 1 provides the approximation result for the causal, TI operators generated by the functionals involved in this theorem. We may state the following:

**Corollary 1**

Any causal, TI, continuous operator with fading memory can be approximated with a level of an arbitrary accuracy by an element of the operators sub-algebra generated by the $G_{\varphi}$ set of functionals provided that sufficiently large $I$, $J$ and $K_j$ numbers are involved in the construction of $G_{\varphi}$.

**Corollary 2**

If $\varphi$ is as a polynomial function, the $G_{\varphi}$ set can be constructed as a $\Sigma(\varphi)$ structure:

$$G_{\varphi}(L_i) = \sum_{j=1}^{J} b_j \phi(\gamma_j(I_i))$$

with the affine mapping $A_j(L_i) = \sum_{j=1}^{J} \alpha_j(I_i) + \gamma_j(I_d)$.

This is nothing else than another formulation of the polynomial Volterra algebra.

An interesting direction of study in this sense may investigate the possibility to adapt the demonstration of the theorem 2.3 from the Hornik’s article [3] to our case when the arguments are in the $L_i$, $I$ - finite selection of functional $L_i \in L$ rather than in $R^\beta$. A larger class of $\varphi$ functions less regular than the polynomials such as the squashing functions may be found to be acceptable.

It is worth mentioning a renewed interest in this subject. Recently, A.M. Schäfer and H.G Zimmermann are approaching the approximation of the non-linear dynamical systems using the recurrent neural networks (RNN). Their work [6] uses also the main results from [3] but follow the ability of RNN state-space model to generate an arbitrary time serie. We will end this section with an example where the functionals $L_i \in L, L \subset L$ are constructed with the help of the Laguerre set of polynomials:

$$p_i(t) = \frac{e^t}{\gamma_i} \frac{d^{\gamma_i}}{dt^{\gamma_i}}(t^\gamma_i e^{-t})$$

The $p_i$ polynomials are orthonormal over the $[0, \infty)$ interval with the weight $e^{-t}$:

$$\int_0^\infty p_i(t) \cdot p_j(t) \cdot e^{-t} = \delta_{ij}$$

It is known that the functions $l_i(t) = p_i(t) \cdot e^{-t/2}$ span a dense subspace in $L_2_{\infty, \infty}$, a result given by V.A. Steklov or M. Riesz. See for example [7].

We write the $L_i$ functionals as follows:

$$L_i(u) = \int_0^\infty l_i(t) e^{-t/2} \cdot u(-t) e^{-t/2} dt$$

This relation leads to the definition of the weighting function: $w(t) = e^{-t/2}$ and consequently to a norm $\|u\|_{\infty} = \sup [u(t) \cdot w(-t)] t \in R_{-}$ with the properties requested by the concept of the fading memory.

The set of approximating operators will be generated by a $\Sigma(\varphi)$ non-linear mapping constructed on the base of a third order splines generator $\phi(x) = (\phi)^{3}(10,1)$. The Fig.1 gives a graphical representation of such a structure for the particular case when $J=2$ and $I=3$. A non-linear, time-invariant dynamical system with the input-output relation modeled by a causal operator will have a closest representation in the $G_{\varphi}(L_i)$ set for a particular choice of the coefficients $\alpha_j$, $\gamma_j$ and $\beta_j$.

As a final remark we should note the theory we developed does not provide any relation on how large the $G_{\varphi}(L_i)$ class should be in terms of $\alpha_j$, $\gamma_j$, $p_j$ and $\beta_j$ parameters or on the choice of the $L_i$ set for an approximation bounded by a preset level of accuracy. The Volterra series share the same existence-only default. An interesting and vast subject lies here.

### 4 Conclusions

We have demonstrated a theorem allowing the approximation of the causal, time invariant, non-linear operators with fading memory by elements of a set of operators constructed as a multidimensional linear convolution followed by a memoryless non-linearity.

The linear convolution part requests to be made with kernels that have a fading memory and span a
complete set of functions in $L^2(0, \infty)$. The memoryless non-linearity is modeled by a $\Sigma \Pi$ structure of a classical feed-forward neural network. This two stage model of these approximating operators is a direct consequence of the Stone-Weierstrass theorem used in the demonstration.

Prior to our article, a similar result has been shown regarding the structure and the completeness of the Volterra series.

It is proved that the model we are proposing includes, as a particular case, the Volterra approximators. In this sense, the algebra of operators generated by the $\Sigma \Pi$ class of functions is richer than the polynomial Volterra counterpart.

References:


Fig. 1 Model structure for the $\Sigma(\varphi)$ in the particular case when $J=2$ and $I=3$