Circuits Systems Signal Processing
Vol. 15, No. 6, 1996, Pp. 711–725

Network Approximation of Input-Output Maps and Functionals*

Irwin W. Sandberg[1] *and Lilian Xu*[1]

Abstract. We give results concerning the problem of approximating the input-output maps of nonlinear discrete-time approximately finite-memory systems. Here the focus is on the linear dynamical parts of the approximating structures, and we give examples showing that these linear parts can be derived from a *single* prespecified function that meets certain conditions. This is done in the context of an approximation theorem in which attention is focused on what we call "basic sets." We also consider the related but very different problem of approximating nonlinear functionals using lattice operations or the usual linear ring operations. For this problem we give criteria, not just sufficient conditions, for approximation on compact subsets of reflexive Banach spaces (any Hilbert space is a reflexive Banach space).

1. Introduction

In the study of control systems it is almost always possible to view a system, or a part of a system (e.g., a controller), as a nonlinear map from one particular space to another. For this reason, and for a variety of problems concerning compensation, adaptivity, or identification, results concerning the representation and approximation of nonlinear maps can be of particular interest to control engineers. Especially pertinent is much of the progress that has been made in recent years in the neural networks area. For example, Cybenko [4] proved that any real-valued continuous map defined on a compact subset of $\mathbb{R}^n$ (n an arbitrary positive integer) can be uniformly approximated arbitrarily well using a neural network with a single hidden layer using any continuous sigmoidal activation function. And subsequently Mhaskar and Micchelli [6], and also Chen and Chen [3], showed that even any nonpolynomial continuous function would do. There are also general studies of approximation in $L_p(\mathbb{R}^n)$, $1 \leq p < \infty$ (and on compact subsets of $\mathbb{R}^n$), using

* Received April 7, 1995; accepted August 10, 1995.
[1]Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, Texas 78712.

radial basis functions as well as elliptical basis functions. It is known [7], for example, that arbitrarily good approximation of a general $f \in L_1(\mathbb{R}^n)$ is possible using uniform smoothing factors and radial basis functions generated in a certain natural way from a single g in $L_1(\mathbb{R}^n)$ if and only if g has a nonzero integral.

The material described above is concerned with the approximation of systems without dynamics. At about the same time there began [9] a corresponding study of the network (e.g., neural network) approximation of functionals and approximately-finite-memory maps. It was shown that large classes of approximately-finite-memory maps can be uniformly approximated arbitrarily well by the maps of certain simple nonlinear structures using, for example, sigmoidal nonlinearities or radial basis functions.[2] This is of interest in connection with, for example, the general problem of establishing a comprehensive analytical basis for the identification of dynamic systems. It was also observed that any continuous real functional on a compact subset of a real normed linear space can be uniformly approximated arbitrarily well using only a feedforward neural network with a linear-functional input layer and one hidden nonlinear (e.g., sigmoidal) layer.[3] This has applications concerning, for instance, the theory of classification of signals.

In a very real sense, the results in [9] are limited by their generality. Little is said, for example, about the details of the linear parts of the network structures. Here, for discrete-time approximately-finite-memory systems, we give results in which the focus is on these linear parts, and we give examples showing that the linear parts can be derived from a *single* prespecified function that meets certain conditions. This is done in the context of an approximation theorem in which attention is focused on what we call "basic sets." We also consider in this paper the related but very different problem of approximation of nonlinear functionals using lattice operations or the usual linear ring operations. For this problem we give criteria, not just sufficient conditions, for approximation in reflexive Banach spaces (any Hilbert space is a reflexive Banach space). All proofs are given in the Appendix.

2. Approximation of nonlinear discrete-time approximately finite-memory systems

2.1. *Preliminaries.*

Concerning notation, we use $\mathbb{R}$, $\mathcal{Z}$, $\mathcal{Z}_+$, and $\mathbb{N}$ to denote, respectively, the set of real numbers, the set of all integers, $\{0, 1, 2, \dots\}$, and $\{1, 2, 3, \dots\}$. As usual, $l_\infty(\mathcal{Z}_+)$ is the set of bounded $\mathbb{R}$-valued functions defined on $\mathcal{Z}_+$ (with the usual

[2] It was later found [12] that the approximately-finite-memory condition is met by the members of a certain familiar class of stable continuous-time systems.

[3] For related work, see [2].

sup metric). Let $r > 0$, and let U be the closed ball of radius r centered at the origin in $l_\infty(\mathcal{Z}_+)$.

Let X be the collection of all $\mathbb{R}$-valued functions on $\mathcal{Z}_+$. For each $\beta \in \mathcal{Z}_+$, let T_β, T^β, and P_β denote, respectively, the *delay, advance*, and *truncation* maps from X into itself defined by

$$(T_\beta x)(n) = \begin{cases} 0 & n < \beta \\ x(n-\beta) & n \geq \beta \end{cases}$$

$$\left(T^\beta x\right)(n) = x(n+\beta), \qquad n \in \mathcal{Z}_+$$

$$\left(P_\beta x\right)(n) = \begin{cases} x(n) & n \leq \beta \\ 0 & n > \beta \end{cases} .$$

With $\beta \in \mathcal{Z}_+$ and $\alpha \in \mathbb{N}$, the *window map* $W_{\beta,\alpha} : X \to X$ is defined by

$$\left(W_{\beta,\alpha} x\right)(n) = \begin{cases} x(n) & \beta - \alpha \leq n \leq \beta \\ 0 & \text{otherwise} \end{cases} .$$

By $G : U \to X$ *causal* and *time invariant*, respectively, we mean that, for each $\alpha \in \mathcal{Z}_+$ and each pair of elements x and y in U such that $P_\alpha x = P_\alpha y$, one has $(Gx)(n) = (Gy)(n)$ for $n = 0, 1, \ldots, \alpha$, and that for each $\beta \in \mathcal{Z}_+$ and $u \in U$

$$\left(GT_\beta u\right)(n) = \begin{cases} 0 & n < \beta \\ (Gu)(n-\beta) & n \geq \beta \end{cases} .$$

A map $F : U \to X^m (m \in \mathbb{N})$ is *causal* or *time invariant* if each of its m component maps is causal or time invariant, respectively.

We say that the map $G : U \to X$ has *approximately finite memory* on U, denoted by $G \in \mathcal{A}(U)$, if given $\epsilon > 0$ there exists an $\alpha \in \mathbb{N}$ such that

$$\left|(Gu)(n) - \left(GW_{n,\alpha} u\right)(n)\right| < \epsilon , \quad n \in \mathcal{Z}_+$$

for all $u \in U$. For any $\alpha \in \mathbb{N}$, define

$$U_\alpha = \left\{ u|_{c_\alpha} : u \in U \right\}$$

where $c_\alpha = \{0, 1, \ldots, \alpha\}$, and view U_α as a metric space with metric ρ_α given by $\rho_\alpha(v_1, v_2) = \max_{n \in c_\alpha} |v_1(n) - v_2(n)|$. For any causal $G : U \to X$, G_α denotes the functional defined on U_α by

$$G_\alpha v = (Gy)(\alpha) , \quad v \in U_\alpha$$

where $y \in U$ is given by

$$y(n) = \begin{cases} v(n) & \text{if } n \leq \alpha \\ 0 & \text{if } n > \alpha \end{cases} , \quad n \in \mathcal{Z}_+ .$$

The *lattice operations* $a \vee b$ and $a \wedge b$ on pairs of real numbers a and b are defined by $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$. We say that a map $L : \mathbb{R}^m \to \mathbb{R}$, where $m \in \mathbb{N}$, is a *lattice map* if Lx for $x \in \mathbb{R}^m$ is generated from the components $x_1, \ldots, x_m$ of x by a finite number of lattice operations that do not depend on x.

Let K be a nonempty subset of X. We say that K is a *basic set* if given $\alpha \in \mathbb{N}$, $\epsilon > 0$, and $f \in X$, there is a g in the set

$$\left\{ \sum_{j=1}^{n} a_j h_j : a_j \in \mathbb{R},\quad h_j \in K,\quad n \in \mathbb{N} \right\}$$

of finite linear combinations of elements of K such that

$$|f(n) - g(n)| < \epsilon$$

for $n = 0, 1, \ldots, \alpha$.

2.2. Results concerning approximately finite-memory systems.

Basic sets play a central role in Theorem 1 below, which is our main result concerning discrete time systems.

Theorem 1. *Let $G : U \to X$ be causal and time invariant. Suppose that K is a basic set. Then statements* (i) *and* (ii) *below are equivalent, and statement* (iii) *holds if* (i) *holds.*

(i) *For any $\epsilon > 0$ there are $\alpha \in \mathbb{N}$, $m \in \mathbb{N}$, $p \in \mathbb{N}$, $c \in \mathbb{R}^m$, a real $m \times p$-matrix A, elements $h_1, \ldots, h_p$ of K, and a lattice map $L : \mathbb{R}^m \to \mathbb{R}$ such that*

$$\left|(Gu)(n) - L\left[c + A\left(HW_{n,\alpha}u\right)(n)\right]\right| < \epsilon\ , \quad n \in \mathcal{Z}_+ \qquad (1)$$

for all $u \in U$, where $H : U \to X^p$ is given by

$$(Hu)_j(n) = \sum_{l=0}^{n} h_j(n-l)u(l), \quad n \in \mathcal{Z}_+, \quad u \in U, \quad j = 1, \ldots, p\,.$$

(ii) *$G \in \mathcal{A}(U)$ and G_a is continuous on U_a for $a \in \mathbb{N}$.*

(iii) *$M : U \to X^p$ defined by*

$$(Mu)(n) = (HW_{n,\alpha}u)(n), \quad n \in \mathcal{Z}_+, \quad u \in U$$

is causal and time invariant, and (with $|\cdot|$ the Euclidean norm on $\mathbb{R}^p$)

$$\sup\{|(Mu)(n)| : u \in U,\quad n \in \mathcal{Z}_+\} < \infty\,.$$

For each $m \in \mathbb{N}$, let C_m denote the set of all continuous maps from $\mathbb{R}^m$ to $\mathbb{R}$, and let D_m stand for any subset of the set of maps from $\mathbb{R}^m$ to $\mathbb{R}$ that is dense in C_m on compact sets, in the (usual) sense that given $\epsilon > 0$, $f \in C_m$, and a compact $E \subset \mathbb{R}^m$, there is a $g \in D_m$ such that $|f(x) - g(x)| < \epsilon$ for $x \in E$. The set of polynomial maps from $\mathbb{R}^m$ to $\mathbb{R}$ is an example of a subset D_m. Because lattice maps are continuous, and because (ii) of Theorem 1 implies (i) and (iii) there, we have the following corollary to Theorem 1.

Corollary 1. *Let K be a basic set. Suppose that $G : U \to X$ is causal and time invariant, that $G \in \mathcal{A}(U)$, and that G_a is continuous on U_a for $a \in \mathbb{N}$. Then for any $\epsilon > 0$ there are $\alpha \in \mathbb{N}$, $n \in \mathbb{N}$, $p \in \mathbb{N}$, elements $h_1, \ldots, h_p$ of K, and a map $F \in D_p$ such that*

$$\left|(Gu)(n) - F\left[\left(HW_{n,\alpha}u\right)(n)\right]\right| < \epsilon\,, \quad n \in \mathcal{Z}_+$$

for all $u \in U$, where $H : U \to X^p$ is as described in Theorem 1.

Theorem 2 below is a partial converse to Corollary 1.

Theorem 2. *Let $G : U \to X$ be causal and time invariant, and let K be a basic set. Suppose that for each $\epsilon > 0$ there are $\alpha \in \mathbb{N}$, $n \in \mathbb{N}$, $p \in \mathbb{N}$, elements $h_1, \ldots, h_p$ of K, and a map $F \in C_p$ such that*

$$\left|(Gu)(n) - F\left[\left(HW_{n,\alpha}u\right)(n)\right]\right| < \epsilon\,, \quad n \in \mathcal{Z}_+$$

for all $u \in U$, where $H : U \to X^p$ is as described in Theorem 1. Then $G \in \mathcal{A}(U)$, and G_a is continuous on U_a for $a \in \mathbb{N}$.

We next give examples of basic sets with the special property that each is generated from a single map. Proofs of the statements made below are given in the Appendix.

Example 1. Let h be any element of $l_1(\mathcal{Z}_+)$[4] such that $\hat{h}(\theta) \neq 0$ for all $\theta \in [-\pi, \pi)$, where

$$\hat{h}(\theta) = \sum_{n=0}^{\infty} h(n)e^{in\theta}\,,$$

and for $j \in \mathcal{Z}$ and $n \in \mathcal{Z}_+$ let

$$h_j(n) = \begin{cases} 0 & \text{if } n < j \\ h(n-j) & \text{if } n \geq j \end{cases}.$$

Then it follows from a version of Wiener's tauberian theorem that $\{h_j : j \in \mathcal{Z}\}$ is a basic set.

Example 2. Let $h \in X$ be such that $h(n) \neq 0$ for all n. Define $h_j(n) = n^j h(n)$ for $n \in \mathcal{Z}_+$ and $j \in \mathcal{Z}_+$. It follows from a simple property of Vandermonde matrices that $\{h_j : j \in \mathcal{Z}_+\}$ is a basic set.

Example 3. Let $h \in X$ satisfy the condition that $h(0) \neq 0$, and define $h_j = T_j h$ (see the definition of T_β) for $j \in \mathcal{Z}_+$. Then $\{h_j : j \in \mathcal{Z}_+\}$ is a basic set.

Example 4. Let μ be a real number not equal to zero or unity, and define $h \in X$ by

$$h(n) = \begin{cases} 0 & \text{if } n = 0 \\ \mu(1-\mu)^{n-1} & \text{if } n \geq 1 \end{cases}.$$

[4] Of course, $l_1(\mathcal{Z}_+) = \left\{f : \sum_{n=0}^{\infty} |f(n)| < \infty\right\}$.

Let $h_j \in X$ be defined for each $j \in \mathcal{Z}_+$ by

$$h_0(n) = \begin{cases} 1 & \text{if } n = 0 \\ 0 & \text{if } n \geq 1 \end{cases}$$

and $h_{j+1} = h_j * h$ for $j \in \mathcal{Z}_+$, where $*$ stands for the discrete convolution operator.[5] Here again, $\{h_j : j \in \mathcal{Z}_+\}$ is a basic set.

Each example above, together with Corollary 1, gives rise to an approximation structure. In all cases the structure consists of a linear section that accounts for dynamics, followed by a memoryless nonlinear section. And in each case the linear section is generated by simple operations on a single "impulse response" function. This simplification can be of considerable value in applications. The term $W_{n,\alpha}$ in Corollary 1 serves merely to truncate the tail of the vector-valued "impulse response" function associated with H. In other words, referring to

$$(Hu)_j(n) = \sum_{l=0}^{n} h_j(n-l)u(l), \quad n \in \mathcal{Z}_+, \quad u \in U, \quad j = 1, \ldots, p$$

of Theorem 1, we have

$$\left(HW_{n,\alpha}u\right)_j(n) = \sum_{l=0}^{n} h_{\alpha j}(n-l)u(l), \quad n \in \mathcal{Z}_+, \quad u \in U, \quad j = 1, \ldots, p$$

where $h_{\alpha j}(n) = h_j(n)$ for $n = 0, 1, \ldots, \alpha$ and $h_{\alpha j}(n) = 0$ for $n > \alpha$. The corresponding approximation structure is shown in Figure 1.

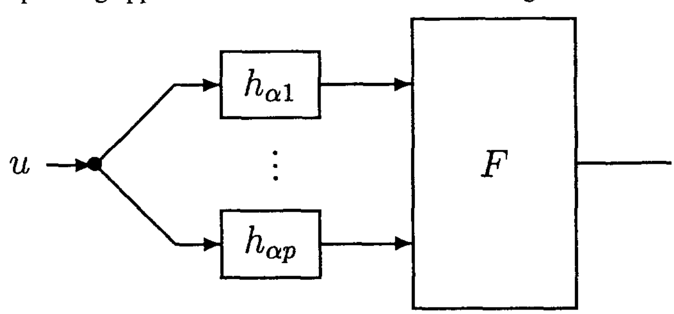

Figure 1. Approximation structure associated with Corollary 1.

In the case of Example 3, if we choose $h(0) = 1$ and $h(n) = 0$ for $n > 0$, then the $h_{\alpha j}$ represent just delays. This time-delay structure is general enough for arbitrarily good approximations,[6] but it can happen in applications that a given degree of approximation can be achieved with considerably fewer dynamical elements (i.e.,

[5] The elements of $\{h_j : j \in \mathbb{N}\}$ are called *gamma delay kernels* in [5].

[6] This was shown differently in [10].

with a much smaller p) if these elements area allowed to be of the type considered in Examples 1, 2, and 4.[7]

3. Approximation of continuous functionals on compact sets

Although any map from a set to the real or complex numbers is a functional, we are primarily interested here in cases in which the set is a subset of an infinite-dimensional space. In this context, the study of approximations of nonlinear functionals was initiated by Fréchet (see [16], p. 20), who considered the approximation of continuous functionals on compact subsets of continuous $\mathbb{R}$-valued functions defined on a finite interval $[a, b]$. Fréchet showed that these functionals can be approximated arbitrarily well by certain finite sums of integrals. The possibility of approximating general nonlinear functionals using linear functionals and a memoryless nonlinear network was noted in [9] and studied further in, for example, [13], [14].[8,9] The response of a control system at a given point in time is often of interest.[10] This is one reason for our interest in functionals. Another reason is that functionals play a central role in the problem of classifying signals [13].

Let B be a Banach space over $\mathbb{R}$, and let B^* be the set of bounded linear functionals on B. Denote by B^{**} the second conjugate space (i.e., the set of bounded linear functionals on B^*). B is said to be a *reflexive* Banach space if for each $\phi \in B^{**}$ there is an $x \in B$ such that $\phi(f) = f(x)$ for all $f \in B^*$. Examples of reflexive Banach spaces are the Hilbert spaces and all L_p and l_p spaces with $1 < p < \infty$. We give a theorem below in which attention is focused on certain operations that may be performed on a specified set of functionals. This theorem provides criteria (i.e., necessary and sufficient conditions) for achieving arbitrarily good approximation of all continuous functionals on arbitrary compact subsets of a reflexive Banach space.

With S a nonempty subset of B^*, we use $L(S)$ to denote

$$\left\{ \sum_{j=1}^{n} a_j s_j : a_j \in \mathbb{R}, \quad s_j \in S, \quad n \in \mathbb{N} \right\} .$$

[7] For experimental material related to Example 4, see [5]. Also, general approximation results are available for structures in which the input sections are nonlinear (see Theorem 5 of [11] and note that the F_λ there need not be linear). The use of nonlinear input sections can sometimes lead to a marked reduction in the complexity of the structure needed for a specified degree of approximation.

[8] There are some minor oversights (e.g., typos) in the proof of Theorem 1 of [13]. See the proof of the related results in [14].

[9] Additional related work can be found in [2] and [3].

[10] For example, often a system's control signal is a functional of the system's state, or is vector valued with each component a functional of the state. It is well known that there are important cases in which the state need not belong to a finite-dimensional space.

In the following, C denotes a nonempty compact subset of B, and $\mathcal{X}$ stands for the set of all $\mathbb{R}$-valued continuous functions on C. Clearly

$$\mathcal{X}_0 := \{h : C \to \mathbb{R} : h(x) = g(x) + \rho, \quad g \in L(S), \quad \rho \in \mathbb{R}\}$$

is a subset of $\mathcal{X}$. This set of affine (i.e., linear plus constant) functionals plays the role of our set of building blocks. Starting from these one may attempt to use certain operations to construct approximations to nonlinear functionals.

With regard to the operations we will consider, let the *lattice* operations $a \vee b$ and $a \wedge b$ on pairs of elements a and b of $\mathcal{X}$ be defined by $(a \vee b)(x) = \max[a(x), b(x)]$ for all x and $(a \wedge b)(x) = \min[a(x), b(x)]$ for all x. Finally, let $U_L(\mathcal{X}_0)$ be the uniform closure (i.e., the closure with respect to the usual sup norm) of the set of functionals generated from $\mathcal{X}_0$ by a finite number of lattice operations, and let $U_P(\mathcal{X}_0)$ be the uniform closure of the set of functionals generated from $\mathcal{X}_0$ by a finite number of the so-called linear-ring operations, i.e., by the usual rules of addition, pointwise multiplication, and multiplications by real scalars.

Our theorem below shows that certain statements are equivalent. It shows, for example, that using the linear-ring operations, arbitrarily good approximations of a general element of $\mathcal{X}$ defined on a general $C \subset B$ are achievable if and only if $L(S)$ is dense in B^*.

Theorem 3. *Assuming that B is reflexive, the following conditions are equivalent.*

(i) $U_L(\mathcal{X}_0) = \mathcal{X}$ *for each* $C \subset B$.

(ii) $U_P(\mathcal{X}_0) = \mathcal{X}$ *for each* $C \subset B$.

(iii) $L(S)$ *is dense in* B^*.

(iv) *For each nonzero* $x \in B$, *there is an* $s \in S$ *such that* $s(x) \neq 0$.

(v) *For each nonzero* $x \in B$, *there is a* $g \in L(S)$ *such that* $g(x) \neq 0$.

(vi) *Given distinct elements x and y of B and real scalars c_x and c_y, there are a real scalar ρ and a $g \in L(S)$ such that $\rho + g(x) = c_x$ and $\rho + g(y) = c_y$.*

This theorem, which concerns approximation structures of the form shown in Figure 2, where x is the input, $h_1, \ldots, h_q$ denote linear functionals, and M stands for a memoryless lattice or linear-ring network, has many applications.[11]

For example, it shows that any $f \in \mathcal{X}$ can be approximated arbitrarily well on any C using the lattice (or the linear-ring) operations if S is the set of all continuous linear functionals on B. And, we can do much better than this from the viewpoint of applications because all that is needed is that $L(S)$ is dense in B^*. To illustrate this, let $B = L_2(0, \infty)$ with norm $\| \cdot \|$, and let B_c and B_s, respectively, denote the intersection of B with the continuous functions and with the step functions (i.e.,

[11] In a limited but important context, it was Wiener who first introduced approximation structures for nonlinear systems consisting of a linear input section followed by a memoryless nonlinear section.

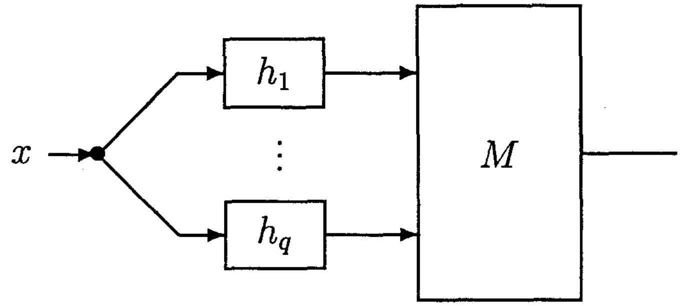

Figure 2. Approximation structure for nonlinear functionals.

the sectionally continuous functions having only a finite number of values). In this case each element y of B^* is given by

$$y(x) = \int_0^\infty k_y(\tau)x(\tau)\,d\tau \tag{2}$$

for some $k_y \in B$. And by the Schwarz inequality, $|y(x) - \tilde{y}(x)| \leq \|k_y - \tilde{k}_y\| \cdot \|x\|$, where $\tilde{y}(x)$ is the value of $y(x)$ in (2) when k_y is replaced with any $\tilde{k}_y \in B$. Therefore, as B_c and also B_s are known to be dense in B, we see that S can be chosen to be the family of y's of the form (2) with $k_y \in B_c$ (or with $k_y \in B_s$). These families are much simpler than the set of y's that have the representation (2) with k_y a general element of B. In fact, for the $k_y \in B_s$ case each element of S is specified by just a finite number of parameters. It is also clear that another acceptable S is the set of y's representable by (2) with the k_y drawn from any basis for B.

As an additional example of the way in which the theorem can be used, we note that it is not difficult to show directly that condition (iv) of the theorem is met when $B = L_2(0, \infty)$ and S is the set of linear functionals y given by (2) with $k_y \in B_c$ (or $k_y \in B_s$ or k_y drawn from a basis for B).

Condition (vi) is interesting; its presence on the list shows that the proposition that an arbitrary $f \in \mathcal{X}$ can be approximated arbitrarily well in the sense of (i) and (ii) is equivalent to the proposition that a certain two-point affine identification problem can be solved.

The problem of approximating a continuous functional on a given specific compact $C \subset B$ is, of course, also of interest. For this case we have the following.

Theorem 4. *Let compact C, $f \in \mathcal{X}$, and a nonempty subset S of $\mathcal{X}$ be given, and let $U_P(S)$ be the uniform closure of the set of functionals generated from S by a finite number of linear-ring operations. Define Z and E by $Z = \{x \in C : f(x) \neq 0\}$ and $E = \{(x, y) \in C^2 : f(x) \neq f(y)\}$. Then $f \in U_P(S)$ if and only if the following two conditions are satisfied.*

(i) *For each $x \in Z$, there is an $h \in S$ such that $h(x) \neq 0$.*

(ii) *For each $(x, y) \in E$, there is an $h \in S$ such that $h(x) \neq h(y)$.*

Theorem 4 is a version of the Stone–Weierstrass theorem that we have not seen before. Its advantage is that, unlike the usual versions found in books, it provides necessary as well as sufficient conditions for approximation. In particular, by choosing $\mathcal{S} = \mathcal{X}_0$ and observing that given $x, y \in C$ there is an $h \in \mathcal{X}_0$ such that $h(x) \neq h(y)$ if and only if there is an $s \in S$ for which $s(x - y) \neq 0$, we have the following.

Corollary 2. *Let compact C and $f \in \mathcal{X}$ be given. Then $f \in U_P(\mathcal{X}_0)$ if and only if for each $x, y \in C$ such that $f(x) \neq f(y)$, there is an $s \in S$ such that $s(x-y) \neq 0$.*

Because $f \in U_P(\mathcal{X}_0)$ when $U_P(\mathcal{X}_0) = \mathcal{X}$, this corollary and Theorem 3 are related in that the criterion of the corollary is met if any of the six equivalent conditions of the theorem hold.

Appendix

A.1. Proof of Theorem 1.

The proof is an application of Theorem 5 of [11]. The main ideas here are to establish (4) below and to observe that N below has the form AH.

Let $\alpha \in \mathbb{N}$, and notice first that the space U_α is compact with respect to the topology derived from the metric $\rho_\alpha(x, y) = \max\{|x(n) - y(n)| : n \in c_\alpha\}$. Let K be a basic set. Define $Q : U \to X$ by

$$(Qu)(n) = \sum_{m=0}^{n} q(n-m)u(m) , \quad n \in \mathcal{Z}_+ \tag{3}$$

with q an element of the set $\mathrm{span}(K)$ of all finite linear combinations of elements of K. It is clear that Q is causal and time invariant, and that the functional Q_α defined on U_α is uniformly continuous. Let u be a nonzero element of U_α, and choose $q_\lambda \in \mathrm{span}(K)$ so that $|q_\lambda(n) - u(\alpha - n)| < \epsilon$ for $n \in c_\alpha$, in which

$$\epsilon = \frac{\sum_{m=0}^{\alpha} u(m)^2}{2\sum_{m=0}^{\alpha} |u(m)|} .$$

Let $Q_\lambda : U \to X$ be the Q of (3) with $q = q_\lambda$. By the triangle inequality,

$$\begin{aligned} |(Q_\lambda u)(\alpha)| &\geq \sum_{m=0}^{\alpha} u(m)^2 - \sum_{m=0}^{\alpha} |q_\lambda(\alpha - m) - u(m)| \cdot |u(m)| \\ &> \sum_{m=0}^{\alpha} u(m)^2 - \epsilon \sum_{m=0}^{\alpha} |u(m)| \\ &> 0 \end{aligned}$$

showing that $Q_{\lambda\alpha} u \neq 0$. Thus, for any v and w in U_α and any real numbers r_v and r_w such that $r_v = r_w$ if $v = w$, there is an F of the form $\rho + Q$, with ρ a real constant and Q of the form (3), such that

$$F_\alpha v = r_v \text{ and } F_\alpha w = r_w \tag{4}$$

(because whenever $v \neq w$ there is a Q for which $Q_\alpha(v-w) = c_v - c_w$).

With $m \in \mathbb{N}$, let $N : U \to X^m$ be defined by

$$(Nu)_i = Q_i u, \quad u \in U, \quad i = 1, \ldots, m$$

where each $Q_i : U \to \mathcal{X}$ is given by (3) with $q = q_i \in \operatorname{span}(K)$. Let $p \in \mathbb{N}$ be such that each of the q_i is a linear combination of elements $h_1, \ldots, h_p$ of K. Then $N = AH$ with $H = (h_1, \ldots, h_p)^{\mathrm{tr}}$ and A some real $(m \times p)$-matrix. The rest of the proof follows directly from the statement of Theorem 5 of [11].[12] □

A.2. Proof of Theorem 2.

Let $\epsilon > 0$, and let $\alpha \in \mathbb{N}$, $m \in \mathbb{N}$, $p \in \mathbb{N}$, elements $h_1, \ldots, h_p$ of K, and a map $F \in C_p$ be such that

$$|(Gu)(n) - (\tilde{G}u)(n)| < \epsilon/2, \quad n \in \mathcal{Z}_+ \tag{5}$$

for $u \in U$, where $\tilde{G} : U \to X$ is defined by

$$(\tilde{G}u)(n) = F\left[\left(HW_{n,\alpha}u\right)(n)\right], \quad n \in \mathcal{Z}_+ \tag{6}$$

with H as in Theorem 1. Choose $u \in U$ and $n \in \mathcal{Z}_+$. By the causality of G,

$$\left|(Gu)(n) - \left(GW_{n,\alpha}u\right)(n)\right| = 0, \quad n \leq \alpha .$$

Let $n > \alpha$. Note that $(\tilde{G}u)(n) = (\tilde{G}W_{n,\alpha}u)(n)$, and so

$$\begin{aligned}\left|(Gu)(n) - \left(GW_{n,\alpha}u\right)(n)\right| &\leq \left|(Gu)(n) - (\tilde{G}u)(n)\right| \\ &\quad + \left|(\tilde{G}u)(n) - \left(GW_{n,\alpha}u\right)(n)\right| \\ &< \epsilon/2 + \left|\left(\tilde{G}W_{n,\alpha}u\right)(n) - \left(GW_{n,\alpha}u\right)(n)\right| .\end{aligned}$$

By the causality of G and H,

$$\begin{aligned}\left(GW_{n,\alpha}u\right)(n) &= \left(Gu_{n,\alpha}\right)(n) \\ \left(\tilde{G}W_{n,\alpha}u\right)(n) &= \left(\tilde{G}u_{n,\alpha}\right)(n),\end{aligned}$$

where $u_{n,\alpha} = T_{n-\alpha}T^{n-\alpha}u$. Because $u_{n,\alpha} \in U$, it follows from (5) that

$$\left|\left(\tilde{G}W_{n,\alpha}u\right)(n) - \left(GW_{n,\alpha}u\right)(n)\right| = \left|\left(\tilde{G}u_{n,\alpha}\right)(n) - \left(Gu_{n,\alpha}\right)(n)\right| < \epsilon/2 .$$

Thus $\left|(Gu)(n) - \left(GW_{n,\alpha}u\right)(n)\right| < \epsilon$ for all u and n, showing that $G \in \mathcal{A}(U)$.

[12] We take this opportunity to improve the wording of Lemma 2 of [11]: "has the property" should be replaced with "and". The same change should be made in Lemma 1 of [9].

Now choose any $a \in \mathbb{N}$. By the form of H and the boundedness of U, $\tilde{G}$ is causal and

$$b := \sup\left\{\left|\left(HW_{a,\alpha}u\right)(a)\right| : u \in U\right\} < \infty ,$$

where $|\cdot|$ is the maximum-of-components norm in $\mathbb{R}^p$. By the continuity (and thus the uniform continuity) of F on $[-b, b]^p$, there is a $\delta > 0$ such that $|F(r_1) - F(r_2)| < \epsilon/2$ for r_1 and r_2 in $[-b, b]^p$ with $|r_1 - r_2| < \delta$. It is not difficult to check that the functionals $H_{1a}, \ldots, H_{pa}$, with $H_j u = (Hu)_j$ for $j = 1, \ldots, p$, are uniformly continuous. Choose δ_1 so that

$$\left|H_{ja}v_1 - H_{ja}v_2\right| < \delta, \quad j = 1, \ldots, p$$

for v_1 and v_2 in U_a with $\rho_a(v_1, v_2) < \delta_1$. Observe that, with $v_1, v_2 \in U_a$ and $u_1, u_2 \in U$ such that $v_1(n) = u_1(n)$ and $v_2(n) = u_2(n)$ for $n = 0, 1, \ldots, a$,

$$\begin{aligned}\left|\tilde{G}_a v_1 - \tilde{G}_a v_2\right| &= \left|F\left[\left(HW_{a,\alpha}u_1\right)(a)\right] - F\left[\left(HW_{a,\alpha}u_2\right)(a)\right]\right| \\ &= \left|F\left[(Hw_1)(a)\right] - F\left[(Hw_2)(a)\right]\right| ,\end{aligned}$$

where $w_1 = u_1$ and $w_2 = u_2$ if $a \leq \alpha$, and

$$w_1 = T_{a-\alpha}T^{a-\alpha}u_1 \text{ and } w_2 = T_{a-\alpha}T^{a-\alpha}u_2$$

if $a > \alpha$. In either case, $\rho_a(w_1|c_\alpha, w_2|c_\alpha) \leq \rho_a(v_1, v_2)$, showing that

$$\left|\tilde{G}_a v_1 - \tilde{G}_a v_2\right| < \epsilon/2 \tag{7}$$

for $\rho_a(v_1, v_2) < \delta_1$. Therefore, for elements v_1 and v_2 of U_a with $\rho_a(v_1, v_2) < \delta_1$,

$$|G_a v_1 - G_a v_2| \leq \left|G_a v_1 - \tilde{G}_a v_1\right| + \left|\tilde{G}_a v_1 - \tilde{G}_a v_2\right| + \left|G_a v_2 - \tilde{G}_a v_2\right| < 3\epsilon/2$$

by (5) and (7). This shows that G_a is uniformly continuous on U_a, and it completes the proof. □

A.3. Comments concerning Examples 1–4.

A.3.1. Example 1. Let $\alpha \in \mathbb{N}$, $\epsilon > 0$, and $f \in X$ be given. Define $\tilde{f}$ and $\tilde{h}$ on $\mathcal{Z}$ by

$$\tilde{f}(n) = \begin{cases} f(n) & \text{if } 0 \leq n \leq \alpha \\ 0 & \text{if } n < 0 \text{ or } n > \alpha \end{cases}$$

and

$$\tilde{h}(n) = \begin{cases} 0 & \text{if } n < 0 \\ h(n) & \text{if } n \geq 0 \end{cases}$$

Clearly, if $\tilde{f} \in l_1(\mathcal{Z})$ and

$$\hat{\tilde{h}}(\theta) := \sum_{n \in \mathcal{Z}} \tilde{h}(n)e^{in\theta} \neq 0$$

for all $\theta \in [-\pi, \pi)$. By a general form of Wiener's tauberian theorem (see p. 162 of [8]), there are $p \in \mathbb{N}$, real numbers $a_1, \dots, a_p$, and integers $\tau_1, \dots, \tau_p$ such that

$$\sum_{n\in\mathcal{Z}} \left| \tilde{f}(n) - \sum_{j=1}^{p} a_j \tilde{h}(n-\tau_j) \right| < \epsilon .$$

For $j \in \mathcal{Z}$, let $h_j : \mathcal{Z}_+ \to \mathbb{R}$ be defined by

$$h_j(n) = \begin{cases} 0 & \text{if } n < j \\ h(n-j) & \text{if } n \geq j \end{cases} .$$

Note that $h_{\tau_j}(\cdot) = \tilde{h}(\cdot - \tau_j)$ on $\mathcal{Z}_+$ for $j = 1, \dots, p$. So, for $n = 0, 1, \dots, \alpha$,

$$\left| f(n) - \sum_{j=1}^{p} a_j h_{\tau_j}(n) \right| \leq \sum_{n=0}^{\infty} \left| f(n) - \sum_{j=1}^{p} a_j \tilde{h}(n-\tau_j) \right| < \epsilon$$

showing that $K = \{h_j : j \in \mathcal{Z}\}$ is a basic set.

A.3.2 Example 2. Let $\alpha \in \mathbb{N}$, $\epsilon > 0$, and $f \in X$ be given. By the invertibility of the Vandermonde matrix under the usual distinctness condition, there are real numbers $a_0, a_1, \dots, a_\alpha$ such that

$$\sum_{j=0}^{\alpha} a_j n^j - \frac{f(n)}{h(n)} = 0 , \quad n \in c_\alpha .$$

Thus, for $n = 0, 1, \dots, \alpha$,

$$\left| \sum_{j=0}^{\alpha} a_j n^j h(n) - f(n) \right| = 0 < \epsilon$$

showing that $K = \{h_j : j \in \mathcal{Z}_+\}$ is a basic set.

A.3.3. Example 3. The validity of Example 3 is clear because in this case for each $j \in \mathcal{Z}_+$ we have $h_j(j) = h(0)$ and $h_{j+1}(j) = 0$. In particular, here given α and f there are $a_0, \dots, a_\alpha$ for which

$$\left| f(n) - \sum_{j=0}^{\alpha} a_j h_j(n) \right| = 0$$

for $n = 0, 1, \dots, \alpha$.

A.3.4. Example 4. It is not difficult to verify that for each $j \in \mathbb{N}$ we have[13]

$$h_j(n) = \begin{cases} 0 & \text{if } n < j \\ \frac{(n-1)!}{(j-1)!(n-j)!} \mu^j (1-\mu)^{n-j} & \text{if } n \geq j \end{cases} .$$

Thus, for all $j \in \mathcal{Z}_+$ we have $h_j(j) = \mu^j \neq 0$ and $h_j(n) = 0$ for $0 \leq n < j$. By a slight modification of the observation above concerning Example 3, $K := \{h_j : j \in \mathcal{Z}_+\}$ is a basic set.[14]

[13] In the expression for $h_j(n)$ we use the usual convention that $0! = 1$.

[14] And note that Examples 3 and 4 are special cases of the proposition that $\{h_j \in X : j \in \mathcal{Z}_+\}$ is a basic set if for all j, $h_j(j) \neq 0$ and $h_j(n) = 0$ for $0 \leq n < j$.

A.4. Proof of Theorem 3.

(iii) $\Rightarrow$ (iv) $\Rightarrow$ (v) $\Rightarrow$ (iii): Suppose that $L(S)$ is dense in B^*. Let x be a nonzero element of B. By a corollary of the Hahn–Banach theorem there is an $f \in B^*$ such that $f(x) = \|x\|$. Choose any such f and select a $g \in L(S)$ for which $\|f - g\|_* < 1/2$ (where $\| \cdot \|_*$ is the norm in B^*). This gives

$$|f(x) - g(x)| \leq \|f - g\|_* \|x\| < \|x\|/2\,,$$

and also

$$|g(x)| \geq |f(x)| - |f(x) - g(x)| > \|x\| - \|x\|/2 > 0\,.$$

Thus, some finite linear combination g of elements of S evaluated at x is nonzero, and hence $s(x) \neq 0$ for some $s \in S$. This shows that (iii) $\Rightarrow$ (iv). That (iv) implies (v) follows from $S \subseteq L(S)$. Now suppose that (iii) is not met. Then the closure $cl(L(S))$ of $L(S)$ with respect to the norm in B^* is a closed proper linear manifold of B^*. By another corollary of the Hahn–Banach theorem (see, for example, [1], p. 199), there exists an $x^{**} \in B^{**}$ such that $x^{**}(g) = 0$ for all $g \in cl(L(S))$ but $x^{**}(f) \neq 0$ for some $f \in (B^* - cl(L(S))$. So, by the representation of the elements of B^{**}, there is a nonzero $x \in B$ such that $g(x) = 0$ for all $g \in L(S)$, showing that (v) is not satisfied. Thus, (v) $\Rightarrow$ (iii).

(v) $\Rightarrow$ (vi) $\Rightarrow$ (i) $\Rightarrow$ (v): Suppose that (v) holds and let x and y be distinct elements of B. Given scalars c_x and c_y, there is a $g \in L(S)$ for which $g(x - y) = c_x - c_y$. Thus, with $\rho := c_x - g(x)$, we have $\rho + g(y) = c_y$, showing that (vi) is met. If (vi) holds, then for distinct elements x and y of C and scalars c_x and c_y, there exists a $k \in \mathcal{X}_0$ for which $k(x) = c_x$ and $k(y) = c_y$. Therefore, by the first corollary of Theorem 1 of [15], (i) follows from (vi). Suppose now that (v) is not satisfied. In other words, suppose that there exists a nonzero $x_0 \in B$ such that $g(x_0) = 0$ for all $g \in L(S)$. Select

$$C = \{\alpha x_0 : \alpha \in [a, b]\}$$

in which a, b are real numbers with $b > a$. In this case $\mathcal{X}_0$ contains only constant functions and hence $U_L(\mathcal{X}_0)$ too consists of only constant functions. By a corollary of the Hahn–Banach theorem, choose $f \in B^*$ so that $f(x_0) = \|x_0\| \neq 0$. By the linearity of f, it is not constant over C; so $f \in \mathcal{X}$ but $f \notin U_L(\mathcal{X}_0)$. Thus (i) implies (v).

(v) $\Rightarrow$ (ii) $\Rightarrow$ (v): The proof of (ii) $\Rightarrow$ (v) is essentially the same as the proof of (i) $\Rightarrow$ (v), and (v) $\Rightarrow$ (ii) follows from the proposition that $U_P(\mathcal{X}_0) = \mathcal{X}$ if $\mathcal{X}_0$ separates the points of C and does not vanish anywhere on C (see, for example, Theorem 5 of [15]). This completes the proof of Theorem 3. □

A.5. Proof of Theorem 4.

Note that (i) is true if and only if f vanishes whenever S does, and that (ii) is true if and only if $f(x) = f(y)$ whenever $h(x) = h(y)$ for all $h \in S$. Thus, the proof follows directly from Theorem 5 of [15]. □

Theorem 4 holds also if C is replaced with a compact subset C of a topological space, and $\mathcal{X}$ is replaced with the set of all $\mathbb{R}$-valued continuous maps defined on C. (And Corollary 2 is valid under the weaker hypothesis that B is just a real normed linear space.)

References

[1] G. Bachman and L. Narici, *Functional Analysis*, New York: Academic Press, 1966.

[2] T. Chen and H. Chen, Approximations of continuous functionals by neural networks with application to dynamical systems, *IEEE Transactions on Neural Networks*, vol. 4, no. 6, pp. 910–918, November 1993.

[3] ———, Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its applications to dynamic systems, *IEEE Transactions on Neural Networks*, vol. 6, no. 4, pp. 911–917, July 1995.

[4] G. Cybenko, Approximation by superposition of a single function, *Mathematics of Control, Signals and Systems*, vol. 2, pp. 303–314, 1989.

[5] B. De Vries, J. C. Principe, and P. Guedes de Oliveira, Adaline with adaptive recursive memory, *Proceedings of IEEE-SP Workshop on Neural Networks for Signal Processing*, pp. 101–110, 1991.

[6] H. N. Mhaskar and C. A. Micchelli, Approximation by superposition of sigmoidal and radial basis functions, *Advances in Applied Mathematics*, vol. 3, pp. 350–373, 1992.

[7] J. Park and I. W. Sandberg, Approximation and radial-basis function networks, *Neural Computation*, vol. 5, no. 2, pp. 305–316, March 1993.

[8] W. Rudin, *Fourier Analysis on Groups*, New York: Interscience, 1962.

[9] I. W. Sandberg, Structure theorems for nonlinear systems, *Multidimensional Systems and Signal Processing*, vol. 2, no. 3, pp. 267–286, 1991. (See also the Errata in vol. 3, no. 1, p. 101, 1992). A conference version of the paper appears in *Integral Methods in Science and Engineering-90*, Proceedings of the International Conference on Integral Methods in Science and Engineering, Arlington, Texas, May 15–18, 1990, ed. A. H. Haji-Sheikh, New York: Hemisphere Publishing, pp. 92–110, 1991.

[10] ———, Approximation theorems for discrete-time systems, *IEEE Transactions on Circuits and Systems-I: Fundamental Theory and Applications*, vol. 38, no. 5, pp. 564–566, May 1991.

[11] ———, Approximately-finite memory and the theory of representations, *International Journal of Electronics and Communication*, vol. 46, pp. 191–199, April 1992.

[12] ———, Uniform approximation and the circle criterion, *IEEE Transactions on Automatic Control*, vol. 38, no. 10, pp. 1450–1458, October 1993.

[13] ———, General structures for classification, *IEEE Transactions on Circuits and Systems-I: Fundamental Theory and Applications*, vol. 41, no. 5, pp. 372–376, May 1994.

[14] ———, Notes on weighted norms and network approximation of functionals, *IEEE Transactions on Circuits and Systems I*, vol. 43, no. 7, pp. 600–601, July 1996.

[15] M. H. Stone, A generalized Weierstrass approximation theorem, In *Studies in Modern Analysis*, ed. R. C. Buck, vol. 1 of *MAA Studies in Mathematics*, pp. 30–87, Englewood Cliffs, N.J.: Prentice-Hall, March 1962.

[16] V. Volterra, *Theory of Functionals and of Integral and Integro-Differential Equations*, New York: Dover, 1959.