ON APPLICATIONS OF APPROXIMATION THEORY TO IDENTIFICATION, CONTROL AND CLASSIFICATION

by

AJIT TRIMBAK DINGANKAR, B.Tech., M.S., M.S.

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

Acknowledgements

It is a great pleasure to express gratitude to my adviser, Professor Irwin W. Sandberg, for his invaluable guidance during the research period which culminated in this dissertation. This work took place in the Information and Systems Engineering Area within the Department of Electrical and Computer Engineering.

From June 1989 through August 1994, I was employed at the Advanced Workstations Division of IBM Corporation, Austin. I would like to thank my colleagues and supervisers for continually asking me the question "when are you going to finish that PhD?", thus reminding me constantly what I was in (school) for.

My parents provided a rationalistic development environment which shaped my life probably to a much greater extent than even they could have anticipated. My wife, Manjusha, patiently waited through my changes of departments, topics, employers and moods all these years. During the course of the doctorate we had the pleasure to welcome a new addition to our family, our son Chaaru, who reintroduced me to learning in a fresh, fun, noncomputational way.

A special note of thanks goes to my friends too numerous to enumerate here. Most of all, I appreciate the true spirit of brotherhood from Satish Dingankar.

ON APPLICATIONS OF APPROXIMATION THEORY TO IDENTIFICATION, CONTROL AND CLASSIFICATION

Publication No. _____

Ajit Trimbak Dingankar, Ph. D.
The University of Texas at Austin, 1995

Supervisor: Irwin W. Sandberg

Applications of approximation theory to some problems in identification of dynamic systems, their control, and to problems in signal classification are studied. First, an algorithm is given for constructing approximations in a wide variety of settings, and a corresponding error bound is derived. Then weak sufficient conditions for perfect classification of signals are studied. Next the problem of approximating linear functionals with certain sums of integrals is studied, alongwith its relation to the approximation of nonlinear functionals. Then an approximation theoretic characterization of continuity of nonlinear maps is given. As another application of function approximation, the problem of universally approximating controllers for discrete-time continuous plants is studied. Finally, error bounds for approximation of functions defined on finite dimensional Hilbert spaces are given.

Table of Contents

Acknowledgements ii

Abstract iii

List of Tables vii

List of Figures viii

Chapter 1. Introduction 1

1.1 Overview . . . 1

Chapter 2. A Constructive Algorithm 4

2.1 Introduction . . . 4

2.2 The algorithm . . . 5

Chapter 3. Classifiers on Relatively Compact Sets 7

3.1 Introduction . . . 7

3.2 Classification theorem . . . 7

3.2.1 Comments . . . 8

3.2.2 Example of C_0 . . . 9

3.2.3 Examples of Y . . . 10

Chapter 4. Approximation of Bounded Linear Functionals 12

4.1 Introduction . . . 12

4.2 Approximation Theorem . . . 13

4.2.1 Notation and Definitions . . . 13

4.2.2 The Theorem . . . 14

4.3 Comments . . . 15

Chapter 5. Approximation of Nonlinear Maps **16**
5.1 Introduction . . . 16
5.2 Preliminaries . . . 17
5.2.1 The inputs and outputs . . . 17
5.2.2 Localizable systems . . . 17
5.2.3 Motivation . . . 18
5.2.4 Approximating structures . . . 21
5.3 Characterization of Continuity . . . 21
5.3.1 The approximation theorem . . . 23
5.4 Applications to Time-invariant Systems . . . 25
5.4.1 Notation . . . 25
5.4.2 Approximation of discrete-time causal time-invariant systems . . . 26
5.4.3 A conjecture about nonlinear feedback systems . . . 27
5.4.4 Continuous-time systems on the half-line . . . 27
5.4.5 Approximation of stable nonlinear feedback systems . . . 29
5.5 An Algorithm and its Error Analysis . . . 30
5.5.1 Remark . . . 31
5.6 Applications . . . 31
5.6.1 Dynamic signal classification . . . 32
5.6.2 Image processing . . . 32
5.7 Conclusion . . . 32

Chapter 6. Robust Universal Neural Controllers **33**
6.1 Introduction . . . 33
6.2 Simultaneous Approximation Theorem . . . 34
6.3 Robust Retrainable Neural Controllers . . . 36
6.4 Comments . . . 39
6.5 Conclusion . . . 40

Chapter 7. Error Bounds for Function Approximation **41**
7.1 Introduction . . . 41
7.2 Notation . . . 41
7.3 Relation to previous work . . . 43
7.4 Error Bounds: Preliminaries . . . 44

7.4.1 The classes Π and Δ_c . . . 44
7.4.2 A consequence of c-integrability . . . 45
7.5 Approximation of complex-valued functionals . . . 45
7.6 Approximation of real-valued functionals . . . 49
7.7 Applications . . . 49
7.7.1 Bandlimited Functions . . . 49
7.7.2 Schwartz class . . . 50
7.8 Approximation of functionals in L_2 . . . 51
7.8.1 Classification . . . 51
7.9 Discussion of results . . . 52

Appendix A. Concerning Chapter 4 **53**
A.1 Uniform Riemann-Stieltjes Integrability . . . 53

Appendix B. Concerning Chapter 5 **54**
B.1 Proof of the Completeness of X_w . . . 54
B.1.1 Weighted continuous functions . . . 54
B.1.2 Weighted L_p space . . . 54
B.2 Verification of the Hypotheses of Theorem 6 . . . 54
B.2.1 Hypothesis (i) . . . 54
B.2.2 Hypothesis (ii) . . . 55

Appendix C. Concerning Chapter 6 **57**
C.1 Lemma on Equicontinuous Extensions . . . 57
C.2 Uniform Riemann Integrability . . . 59

Bibliography **61**

Vita **65**

List of Tables

7.1 An analogy between analogies. 43

List of Figures

3.1 Architecture of the classifier. 9
3.2 Neural network implementation of the map h. 9

5.1 A tensor product neural network M. 22
5.2 Nonlinear control system. 27

6.1 Unity feedback configuration for tracking. 38

7.1 Architecture of the complex Hilbert space network. 42
7.2 Architecture of the real Hilbert space network. 43

Chapter 1

Introduction

The problems of identification and control of dynamic systems, and of signal classification are of importance in many engineering settings. Several techniques of their solution have been proposed in the literature. Here we treat these problems from the viewpoint of applications of approximation theory.

1.1 Overview

The earliest proofs of approximation with neural networks of continuous functions on compact sets are nonconstructive and do not address the question of how one determines the number of nonlinearities needed for a given degree of approximation. In an interesting recent paper, Barron [3] shows that a lemma attributed to B. Maurey can be used in some cases to obtain a bound on the number of nonlinearities. In Chapter 2 we make an important simple observation concerning the proof of the lemma described above, and we give a corresponding simpler algorithm for achieving the error bound. The observation is that the proof of Lemma 1 given in [3] shows that each approximating function f_n there can in fact be taken to an *arithmetic average* of n points of the approximating set G. In particular, we describe an algorithm to compute such n-point averages. Chapter 2 is based on the paper

> A. T. Dingankar and I. W. Sandberg, "A Note on Error Bounds for Approximation in Inner Product Spaces", To appear in *Circuits, Systems and Signal Processing*, 1996.
>
> A conference version apppears in "On Error Bounds for Neural Network Approximation", *Proceedings of the International Symposium on Circuits and Systems (ISCAS-95)*, pp. 490–492, Seattle, Washington, April 30–May 3, 1995.

The problem of classifying signals is of interest in several engineering contexts, e.g., automatic target recognition and symbol detection in digital communication using analog channels. Typically we are given a finite number m of pairwise disjoint sets $C_1, \ldots, C_m$ of signals, and we would like to synthesize a system that maps the elements of each C_j into a real number a_j, such that the numbers $a_1, \ldots, a_m$ are distinct. In a recent paper [51] it is shown that this classification can be performed by certain simple structures involving linear functionals and memoryless nonlinear elements, assuming that the C_j are compact subsets of a real normed linear space. In Chapter 3 we give a similar solution to the problem under the considerably weaker assumption that the C_j are relatively compact (i.e., have compact closure) and are of positive distance

from each other. An example is given in which the C_j are subsets of $L_p(a,b)$, $1 \leq p < \infty$. This example shows that a very wide class of signals can be classified using the structures described. Chapter 3 is based on the paper

> I. W. Sandberg and A. T. Dingankar, "Classifiers on relatively compact sets", *IEEE Transactions on Circuits and Systems-I: Fundamental Theory and Applications*, 42(1):57, January 1995.

One of the earliest results in the area of neural networks is the proposition that any continuous real function defined on a compact subset of $\mathbb{R}^k$ (k an arbitrary positive integer) can be approximated arbitrarily well using a single-hidden-layer network with sigmoidal nonlinearities (see, for example, [18]). Among other results in the literature concerning arbitrarily good approximation that concern more general types of "target" functionals, different network structures, other nonlinearities, and various measures of approximation errors is the proposition in [49], [51] that any continuous real nonlinear functional on a compact subset of a real normed linear space can be approximated arbitrarily well using a single-hidden-layer neural network with a linear functional input layer and exponential (or polynomial or sigmoidal) nonlinearities. In an interesting paper [11] similar results concerning single-hidden-layer neural networks with sigmoidal nonlinearities are proved for continuous real functionals on a compact subset of an L_p space ($1 < p < \infty$). One of the main results in [11] is the proposition that the linear functionals in the input layer may have the special form of a finite sum of integrals of a certain type. In Chapter 4 we connect these results by showing that every continuous linear functional can be approximated arbitrarily well on compact sets by finite sums of the type used in [11]. Besides relating the two works [11] and [51], the theorem given here is felt to be interesting in itself. Chapter 4 is based on the paper

> I. W. Sandberg and A. T. Dingankar, "On Approximation of Linear Functionals on L_p Spaces", *IEEE Transactions on Circuits and Systems-I: Fundamental Theory and Applications*, 42(7):402, July 1995.

A modified version of Chapter 5 is under preparation as the paper

> Ajit T. Dingankar and Irwin W. Sandberg, "Tensor Product Neural Networks are Universal Approximators of Dynamical Systems".

Related results, in a general sense, on approximation of dynamic systems are reported in the work of Chen and Chen [12], where the approximating structures are certain forms of radial basis function neural networks. Approximation results using Volterra series for nonlinear time-varying systems are obtained in [47], and in Boyd's dissertation [6] for the time-invariant case. One of the principal results of the work reported on in Chapter 5, as an application of Theorem 5 there, states that the property of arbitrarily good approximation by finite sums of the form

$$\sum_{i=1}^{m} \sum_{j=1}^{k(i)} c_{ij} u_{ij}[y_{ij}(\cdot)] v_i$$

is equivalent to continuity of a certain nonlinear dynamic input-output map. Here the c_{ij} are real constants, the u_{ij} are certain continuous real-valued functions of the reals, the y_{ij} are continuous real functionals defined on the input space X, and the v_i belong to the output space E.

A modified version of Chapter 6 is under preparation as the paper

Ajit T. Dingankar and Irwin W. Sandberg, "Reconfigurable Universal Neural Controllers".

Here the problem is to approximate all members of a class of equicontinuous controllers arbitrarily well using a fixed dilation (called the "width") of a basis function in conjunction with a fixed finite set of translates called the "centers". This property is important in practical applications because of the need for robustness with respect to plant variations or modeling uncertainties [33]. In the case of discrete-time continuous controllers that are time-invariant and posess finite memory, our problem reduces to the problem of simultaneously approximating a class of continuous functions defined on a finite-dimensional Euclidean space. This is the setting adopted in this work. We show that the problem can be solved by radial basis function neural controllers (RBFNCs). Approximations with radial basis functions have been studied in the literature by many authors. See [36] for an early theoretical result, and [53] for a control application. However, in most cases only a single function (or a single controller) is approximated. Simultaneous approximation of a class of functions using radial basis functions was obtained by Chen and Chen [12]. In contrast to [12], we are not restricted to using an activation function that is radially symmetric with respect to the Euclidean norm. This has important practical implications for using a form of elliptic basis functions [37], [9]. Also, in our work the same width is used on all nodes, and hence the number of free parameters of the network is reduced. Hence the parameter estimation problem is simplified. Perhaps the most important consideration is the ease of retraining a network on a new target function, i.e., if a sufficiently accurate RBFNC has been obtained with our method, then it can be used with the same width and centers for any controller in the given equicontinuous class with the same error bound. The only parameters that need to be adapted to a new controller are the weights from the hidden nodes to the output, but this problem is linear. Together, these results indicate that the RBFNCs are a powerful and efficient architecture — they are linearly reconfigurable universal controllers.

Chapter 7 deals with the problem of function approximation using neural networks, and the associated error bounds. This area of study received a great impetus with the work of Barron [3]. Here we obtain similar error bounds using a different neural network architecture and a different class of "target" functions (defined via a different integrability condition than the condition in [3]). Most importantly, we give a much simplified construction of the approximating structures. We show that, if the target function f is integrable in the Lebesgue sense (which is the most commonly assumed condition to guarantee the existence of an integral representation of the type crucially used in [3]), then the integrability condition of [3] implies that of ours. We also give an example to show that there are functions f which satisfy our integrability condition but not the condition in [3]. An abridged version of Chapter 7 is under preparation for future publication.

Chapter 2

A Constructive Algorithm

2.1 Introduction

One of the earliest interesting results in the area of neural networks is the proposition that any continuous real function defined on a compact subset of $\mathbb{R}^k$ (k an arbitrary positive integer) can be approximated arbitrarily well using a single-hidden-layer network with sigmoidal nonlinearities [26]. There are several other results in the literature concerning arbitrarily good approximation that concern more general types of "target" functionals, different network structures, other nonlinearities, and various criteria by which the quality of approximation is gauged.[1] The earliest proofs of these results are nonconstructive and do not address the question of how one determines the number of nonlinearities needed for a given degree of approximation.

In an interesting recent paper, Barron [3] shows that a lemma attributed to B. Maurey can be used in some cases to obtain a bound on the number of nonlinearities. To discuss that lemma, let G be a subset of a real or complex inner product space H with norm $||\cdot||$, such that the elements of G are bounded in norm by some positive constant b, and let $\overline{co}(G)$ denote the closure of the convex hull of G. Maurey's lemma, which concerns the error in approximating an element of $\overline{co}(G)$ with a convex combination of n points in G, is the following:

Lemma 1 Let f be an element of $\overline{co}(G)$ and let $c > (b^2 - ||f||^2)$. Then for each positive integer n there is an f_n in the convex hull of n points of G such that

$$||f - f_n||^2 \leq \frac{c}{n}. \tag{2.1}$$

As mentioned in [3], this lemma is attributed in [39] to B. Maurey. In the main application given in [3] assumptions are made on an f so that $f(x)$ can be written as an integral with respect to a probability measure. This yields a setting in which f is an element of the closure of the convex hull of functions generated by the integrand. For a related later study concerning approximations using translates of a given function see [23]. An algorithm for achieving the bounds in Lemma 1 is also given in [3]. (The algorithm is an improved version of a result due to Jones [27].) In this algorithm, for each $n > 1$ one chooses an f_n of the form $\alpha_n f_{(n-1)} + (1 - \alpha_n) g_n$ such that

$$||f_n - f||^2 \leq \inf_{0 \leq \alpha \leq 1} \inf_{g \in G} ||\alpha f_{n-1} + (1 - \alpha) g - f||^2 + \epsilon_n$$

[1] An example of another result is the proposition [51] that any continuous real functional on a compact subset of a real normed linear space can be approximated arbitrarily well using a single-hidden-layer neural network with a linear functional input layer and exponential (or polynomial or sigmoidal) nonlinearities.

in which ϵ_n is a positive number that satisfies $\epsilon_n \leq O(n^{-2})$. In particular, for large n one chooses g as well as α so that $||\alpha f_{n-1} + (1-\alpha)g - f||$ is nearly minimal.

Here we direct attention to an observation concerning the proof of the lemma described above, and give a corresponding simpler algorithm for achieving bounds similar to (2.1). The observation is just that the proof of Lemma 1 given in [3] shows that each f_n in Lemma 1 can in fact be taken to be an *arithmetic average* of n points in G.[2] In the remainder of this chapter, we show how one can compute such n-point averages.

2.2 The algorithm

Let f be an element of $\overline{co}(G)$, and define

$$\gamma = \inf_{v \in H} \sup_{g \in G} \{||g - v||^2 - ||f - v||^2\}.$$

Choose $\delta > \gamma$ (it suffices to take $\delta > b^2 - ||f||^2$, because $\gamma \leq b^2 - ||f||^2$ since $0 \in H$).

Set $\epsilon_k = (\delta - \gamma)/k^2$ for $k = 1, 2, \cdots$.

Let $f_1, f_2, \cdots$ be any sequence generated as follows:

Choose $f_1 = g_1 \in G$ such that $||f - g_1||^2 \leq (\inf_{g \in G} ||f - g||^2) + \epsilon_1$.

For $k = 2, 3, \cdots$ choose $g_k \in G$ such that

$$||f - (1 - k^{-1})f_{k-1} - k^{-1}g_k||^2 \leq \inf_{g \in G} ||f - (1 - k^{-1})f_{k-1} - k^{-1}g||^2 + \epsilon_k \tag{2.2}$$

and set $f_k = (1 - k^{-1})f_{k-1} + k^{-1}g_k$.

Of course f_k is the arithmetic average $k^{-1}\sum_{j=1}^{k} g_j$ of k elements of G. We will prove the following:

Theorem 1 Under the conditions indicated,

$$||f - f_n||^2 \leq \frac{\delta}{n} \tag{2.3}$$

for all n.

Proof: We will use an induction argument based on the following lemma [14, Lemma 3 of Chapter 25] (which is related to a result in [3]).[3]

[2] As an aside, the proof is probabilistic and is reminiscent of one used by Shannon in his 1947 random coding argument, in that one observes that if an average is bounded from above by a certain number then at least one of the terms in the expression for the average must also be bounded in this way.

[3] The proof of Lemma 2 given in [14] and the proof of the related result in [3] assume that the inner product space is real, but obvious modifications show that both results hold also for complex spaces.

Lemma 2 Let G, f and γ be as above, and let h belong to the convex hull of G. Then

$$\inf_{g \in G} ||f - \lambda h - (1-\lambda)g||^2 \leq \lambda^2 ||f-h||^2 + (1-\lambda)^2 \gamma.$$

for $\lambda \in [0,1]$.

By Lemma 2 and the choice of f_1,

$$\begin{aligned} ||f - f_1||^2 &\leq \inf_{g \in G} ||f-g||^2 + \epsilon_1 \\ &\leq \inf_{g \in G} ||f-g||^2 + (\delta - \gamma) \\ &\leq \gamma + (\delta - \gamma) = \delta, \end{aligned}$$

showing that (2.3) holds for $n = 1$. Now suppose that $||f - f_{(n-1)}||^2 \leq \delta/(n-1)$ for some $n \geq 2$. By (2.2) and Lemma 2,

$$\begin{aligned} ||f - f_n||^2 &\leq \inf_{g \in G} ||f - (1-n^{-1})f_{n-1} - n^{-1}g||^2 + \epsilon_n \\ &\leq \left(\frac{n-1}{n}\right)^2 ||f - f_{n-1}||^2 + \frac{1}{n^2}\gamma + \epsilon_n \\ &\leq \frac{\delta(n-1)+\gamma}{n^2} + \epsilon_n \\ &\leq \frac{\delta}{n} - \frac{(\delta-\gamma)}{n^2} + \epsilon_n = \frac{\delta}{n} \end{aligned}$$

which completes the proof.

Besides showing that our algorithm may be used to generate simpler approximants with somewhat less computational cost, the theorem improves Lemma 1 in also the sense that (as simple examples show) δ can be less than $b^2 - ||f||^2$.

Chapter 3

Classifiers on Relatively Compact Sets

3.1 Introduction

The problem of classifying signals is of interest in several application areas. (E.g., it arises in connection with both automatic target recognition and symbol detection in digital communication using analog channels.) Typically we are given a finite number m of pairwise disjoint sets $C_1, \ldots, C_m$ of signals, and we would like to synthesize a system that maps the elements of each C_j into a real number a_j, such that the numbers $a_1, \ldots, a_m$ are distinct. In a recent paper [51] it is shown that this classification can be performed by certain simple structures involving linear functionals and memoryless nonlinear elements, assuming that the C_j are compact subsets of a real normed linear space[1]. In Section 3.2 we give a similar solution to the problem under the considerably weaker assumption that the C_j are relatively compact (i.e., have compact closure) and are of positive distance from each other. An example is given in which the C_j are subsets of $L_p(a,b)$, $1 \leq p < \infty$. This example shows that a very wide class of signals can be classified using the structures described. The signals need not be continuous, in contrast with those of the example given in [51].

3.2 Classification theorem

We need a few preliminaries:

Let $C_1, \ldots, C_m$ be subsets of a real normed linear space X, with $m > 1$. Let us say that the collection $\{C_1, \cdots, C_m\}$ is *strongly separated* if each C_i is nonempty and the distance between each pair of sets is positive in the usual sense that

$$0 < \inf\{\, ||a - b|| : a \in C_i \,, b \in C_j\}, \quad i \neq j.$$

We define a classifier on $\cup_i C_i$ to be a functional that for each i assigns a real number a_i to every $x \in C_i$ where $a_1, \cdots, a_m$ are distinct.

Let X^* be the set of bounded linear functionals on X (i.e., the set of bounded linear maps from X to the reals $\mathbb{R}$). Given a compact subset C of X, let Y be any set of continuous maps from X to $\mathbb{R}$ that is dense in X^* on C, in the sense that for each $\phi \in X^*$ and any $\epsilon > 0$ there is a $y \in Y$ such that $|\phi(x) - y(x)| < \epsilon$, $x \in C$. Examples of such sets Y are given in Section 3.2.3.

[1] The main result in [51] expands on an earlier observation in [48] regarding approximation of continuous nonlinear functionals using feedforward neural networks with linear (or affine) functional input layer and one hidden (e.g., sigmoidal) layer.

Finally, let U be any set of continuous maps $u : \mathbb{R} \to \mathbb{R}$ such that given $\sigma > 0$ and any bounded interval $(\beta_1, \beta_2) \subset \mathbb{R}$ there exists a finite number of elements $u_1, \ldots, u_\ell$ of U for which $|\exp(\beta) - \sum_j u_j(\beta)| < \sigma$ for $\beta \in (\beta_1, \beta_2)$.[2]

Our classification result is as follows.

Theorem 2 Let $C_0 = C_1 \cup \ldots \cup C_m$ where $\{C_1, \ldots, C_m\}$ is a strongly-separated collection of relatively compact subsets of X, and let f be a classifier on C_0 that takes the value a_i on C_i, $\quad 1 \leq i \leq m$. Also, let $\epsilon > 0$ be given. Then there exists a positive integer k, real constants $c_1, \ldots, c_k$, elements $u_1, \ldots, u_k$ of U, and elements $y_1, \ldots, y_k$ of Y with C the closure of C_0 such that

$$a_i - \epsilon < \sum_{j=1}^{k} c_j u_j[y_j(x)] < a_i + \epsilon \quad (x \in C_i\,, \quad 1 \leq i \leq m).$$

Proof:

We will use the following.

Lemma [51]: Let g be a real-valued continuous map defined on a compact subset K of X. Then given $\epsilon > 0$ there are a positive integer k, real numbers $c_1, \ldots, c_k$, elements $u_1, \ldots, u_k$ of U, and elements $y_1, \ldots, y_k$ of Y with $C = K$ such that

$$|g(x) - \sum_j c_j u_j[y_j(x)]| < \epsilon$$

for $x \in K$.

Continuing with the proof and using the strong separation of the C_i, we see that f satisfies

$$x_1, x_2 \in C_0, \quad ||x_1 - x_2|| < \delta \Rightarrow |f(x_1) - f(x_2)| = 0$$

for any sufficiently small positive δ, showing that f is uniformly continuous on C_0.

Note that $K := \overline{C}_1 \cup \overline{C}_2 \ldots \cup \overline{C}_m$, (with $\overline{C}_i$ the closure of C_i), being the union of finitely many compact sets, is compact. Since C_0 is dense in K, there exists [42, pp. 99, Problem 13] a unique continuous extension of the functional f from C_0 to K. Let g denote this extension. Using the lemma we obtain the desired result.

3.2.1 Comments

Theorem 2 shows that the simple structure illustrated in Figure 3.1 consisting of elements y_i of a class of functionals,[3] a memoryless nonlinear map h, and a quantizer (here a map Q that for each j takes numbers in the interval $(a_j - 0.25\rho, a_j + 0.25\rho)$ into a_j, where $\rho = \min_{i \neq j} |a_i - a_j|$) is capable of implementing any classifier on strongly separated relatively compact sets. The theorem shows also that neural networks of a certain architecture with a single hidden layer may be used to implement the map h. This architecture is illustrated in Figure 3.2 where every u_j block represents the evaluation of a

[2] Of course we can take U to be the set whose only element is $\exp(\cdot)$, or the set $\{u : u(\beta) = (\beta)^n/n!,\ n \in \{0, 1, \ldots\}\}$.

[3] The functionals can be taken to be linear. As mentioned earlier, examples are given in Section 3.2.3.

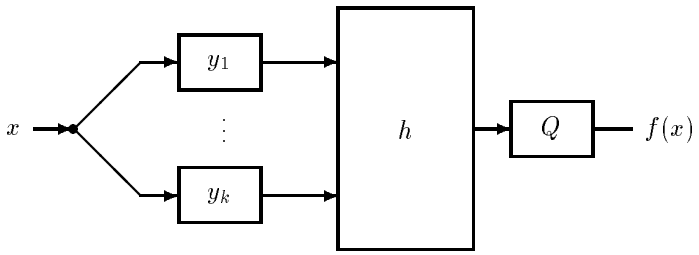


Figure 3.1: Architecture of the classifier.

real-valued function u_j defined on the reals. As described above, the u_j are drawn from a set U of continuous functions whose finite sums can approximate the exponential function $\exp(\cdot)$ on bounded intervals. In particular, u_j can be taken to be $\exp(\cdot)$ for all j.

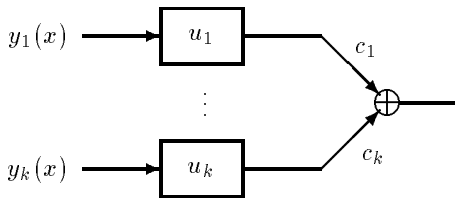


Figure 3.2: Neural network implementation of the map h.

3.2.2 Example of C_0

Here we prove a theorem which provides a class of examples of relatively compact sets of particular interest concerning the classification of signals. Given $\alpha, \beta, \lambda > 0$, let $A(\lambda, \alpha, \beta)$ denote any set of real-valued functions defined on a fixed finite interval $[a, b]$ such that for every $x \in A(\lambda, \alpha, \beta)$:

1) The number of discontinuities of x is at most λ.

2) For every interval of continuity $\mathcal{I}$ of x,

$$|x(t_1) - x(t_2)| \leq \alpha |t_1 - t_2|, \quad \forall\, t_1, t_2 \in \mathcal{I}.$$

3) $|x(t)| \leq \beta, \quad \forall\, t \in [a,b]$.

Let $L_p(a,b)$, with $1 \leq p < \infty$, denote the usual normed space of real-valued Lebesgue-measurable functions defined on $[a,b]$ that are p^{th} power integrable.

Theorem 3 $A(\lambda, \alpha, \beta)$ is a relatively compact subset of $L_p(a,b)$.

Proof:

We use the bound (see the proof of Lemma 3 of [50])

$$\int_a^b |x(t) - x(t+\delta)|^p dt \leq [(2\beta)^p (\lambda + 1) + \alpha(b-a)]|\delta|$$

for $x \in A(\lambda, \alpha, \beta)$ and $\delta \in \mathbb{R}$ (by an abuse of notation x in the integrand is an extension of x that vanishes outside $[a,b]$). This bound shows that given an $\epsilon > 0$ there exists a $\sigma > 0$ that is independent of x such that

$$\int_a^b |x(t) - x(t+\delta)|^p dt < \epsilon^p$$

for $|\delta| < \sigma$. Also, $A(\lambda, \alpha, \beta)$ is uniformly bounded because

$$\int_a^b |x(t)|^p dt \leq (b-a)\beta^p .$$

Hence, by a theorem due to Riesz [31, p. 44], $A(\lambda, \alpha, \beta)$ is relatively compact.

Theorem 3 shows that with $X = L_p(a,b)$, the C_i can be taken to be any strongly separated family in which each member is some $A(\lambda, \alpha, \beta)$. This example illustrates that a very wide class of signals can be classified using the structures described in the preceding section. The signals need not be continuous, in contrast with those of the example given in [51].

3.2.3 Examples of Y

Let $[a,b]$ and $L_p(a,b)$ be as described in the preceding section. For the case in which $X = L_p(a,b)$, each y_i in Figure 3.1 can be taken to be represented by integration over $[a,b]$ after pointwise multiplication by a kernel in the conjugate space, since all linear functionals can be represented in that manner.

The following example of an alternative simpler representation is taken from Chapter 4. For each integrable $x\colon [a,b] \to \mathbb{R}$ and each $\delta > 0$, let the so-called *Steklov function* be defined by

$$x_\delta(t) = \frac{1}{2\delta} \int_{t-\delta}^{t+\delta} x(\tau)\, d\tau, \quad t \in [a,b]$$

where (abusing the notation) x in the integrand is an extension of x that vanishes outside $[a,b]$. Note that x_δ is continuous on $[a,b]$ and that, in particular, it belongs to $L_p(a,b)$. For a given $n \in \mathbb{N}$ and $\delta > 0$, we use $\mathcal{S}_{n,\delta}$ to denote

$$\left\{ s\colon L_p(a,b) \to \mathbb{R} \mid s(x) = \sum_{j=1}^{n} c_j\, x_\delta(\alpha_j), \quad c_j \in \mathbb{R}, \quad 1 \leq j \leq n \right\}$$

in which the α_j are the elements of the uniformly spaced grid of n points $\alpha_j = a + j(b-a)/n, \quad 1 \leq j \leq n$. The set $\mathcal{S}_{n,\delta}$ is a family of what might be called Steklov functionals of order n and width δ. The class $\mathcal{S}$ of all Steklov functionals is defined by

$$\mathcal{S} = \bigcup_{\substack{n \in \mathbb{N} \\ \delta > 0}} \mathcal{S}_{n,\delta}.$$

The following theorem is proved in Chapter 4. It justifies taking $\mathcal{S}$ as an example of a set of continuous maps from $L_p(a,b)$ to $\mathbb{R}$ that is dense in $L_p(a,b)^*$ over compacta. (That is, it justifies the statement that given any compact $C \subset L_p(a,b)$ we can take $Y = \mathcal{S}$.)

Theorem 4 Let $f \in L_p(a,b)^*$ and let C be a compact subset of $L_p(a,b)$. Then for each $\epsilon > 0$ there exists an $s \in \mathcal{S}$ such that

$$|f(x) - s(x)| < \epsilon, \quad x \in C.$$

Note that the elements of $\mathcal{S}$ are simpler in form than the integrals involving a kernel mentioned at the beginning of this section, in that a member of $\mathcal{S}$ is specified by just a finite number of parameters.

Chapter 4

Approximation of Bounded Linear Functionals

4.1 Introduction

One of the earliest results in the area of neural networks is the proposition that any continuous real function defined on a compact subset of $\mathbb{R}^k$ (k an arbitrary positive integer) can be approximated arbitrarily well using a single-hidden-layer network with sigmoidal nonlinearities (see, for example, [18]). Among other results in the literature concerning arbitrarily good approximation that concern more general types of "target" functionals, different network structures, other nonlinearities, and various measures of approximation errors is the proposition in [49], [51] that any continuous real nonlinear functional on a compact subset of a real normed linear space can be approximated arbitrarily well using a single-hidden-layer neural network with a linear functional input layer and exponential (or polynomial or sigmoidal) nonlinearities.

In an interesting paper [11] by Chen and Chen similar results concerning single-hidden-layer neural networks with sigmoidal nonlinearities are proved for continuous real functionals on a compact subset of an L_p space ($1 < p < \infty$). One of the main results in [11] is the proposition that the linear functionals in the input layer may have the special form of a finite sum of integrals of a certain type. Here we connect these results by showing that every continuous linear functional can be approximated arbitrarily well on compact sets by finite sums of the type used in [11].

More specifically, the main result in [51] shows that any real continuous functional defined over a compact subset of a real normed linear space can be approximated arbitrarily well with a neural network employing an input layer of functionals that can be taken to be linear, and a single-hidden-layer implementing nonlinearities that, for example, can be taken to be sigmoidal. To describe the result proved there, let X be a real normed linear space, and let X^* be the set of bounded linear functionals on X (i.e., the set of bounded linear maps from X to the reals $\mathbb{R}$). Given a compact subset C of X, let Y be any set of continuous maps from X to $\mathbb{R}$ that is dense in X^* on C, in the sense that for each $\phi \in X^*$ and any $\rho > 0$ there is a $y \in Y$ such that $|\phi(x) - y(x)| < \rho$, $x \in C$. Also, let U be any set of continuous maps $u : \mathbb{R} \to \mathbb{R}$ such that given $\sigma > 0$ and any bounded interval $(\beta_1, \beta_2) \subset \mathbb{R}$ there exists a finite number of elements $u_1, \ldots, u_\ell$ of U for which $|\exp(\beta) - \sum_j u_j(\beta)| < \sigma$ for $\beta \in (\beta_1, \beta_2)$.[1] The result in [51] is this: Let g be a real-valued continuous map defined on C. Then given $\epsilon > 0$ there are a positive integer k, real numbers $c_1, \ldots, c_k$, elements $u_1, \ldots, u_k$ of U, and elements $y_1, \ldots, y_k$ of Y such that

$$|g(x) - \sum_j c_j u_j[y_j(x)]| < \epsilon$$

[1] Of course we can take U to be the set whose only element is $\exp(\cdot)$, or the set $\{u : u(\beta) = (\beta)^n/n!,\ n \in \{0, 1, \ldots\}\}$.

for $x \in C$.

The proof in [51] of the result described above shows that the result can be slightly sharpened in that "$y_1, \ldots, y_k$ of Y" can be replaced with "$y_1, \ldots, y_k$ of Y_ρ for some $\rho > 0$," where each Y_ρ is any set of continuous maps from X to $\mathbb{R}$ with the property that given $\phi \in X^*$ with $||\phi|| \leq 1$ there is a $y \in Y_\rho$ such that $|\phi(x) - y(x)| < \rho$, $x \in C$, provided that we also replace "such that given $\sigma > 0$ and any bounded interval $(\beta_1, \beta_2) \subset \mathbb{R}$ there exists a finite number of elements $u_1, \ldots, u_\ell$ of U for which $|\exp(\beta) - \sum_j u_j(\beta)| < \sigma$ for $\beta \in (\beta_1, \beta_2)$" with "such that given $\alpha \geq 1, \sigma > 0$ and any bounded interval $(\beta_1, \beta_2) \subset \mathbb{R}$ there exists a finite number of elements $u_1, \ldots, u_\ell$ of U for which $|\exp(\alpha\beta) - \sum_j u_j(\beta)| < \sigma$ for $\beta \in (\beta_1, \beta_2)$". It can be verified that for the modified result the elements of U need not be continuous, thus yielding an even stronger result.

Now suppose that $X = L_p$ with $1 < p < \infty$, and that $U = \{u : u(\beta) = cs(w\beta + \gamma),\ c, w, \gamma \in \mathbb{R}\}$, where s is a sigmoidal function. In this setting we obtain one of the main results in [11] as a special case of the above slightly-sharpened result, since our theorem below and a simple observation in Section 4.3 about its proof permit us to take the Y_ρ to be sets of finite linear combinations of integral functionals of the form considered in [11].[2] That is, in the L_p setting, for each ρ there is an $h > 0$ such that we can take the value at x of the elements y of Y_ρ to be given by certain finite linear combinations of integrals of the form

$$\frac{1}{2h} \int_{t-h}^{t+h} x(\tau)\, d\tau$$

with different values of t.

Besides relating the two works [11] and [51], the theorem given in the next section is felt to be interesting in itself.

4.2 Approximation Theorem

4.2.1 Notation and Definitions

Let L_p, $1 \leq p < \infty$ denote the space of real-valued Lebesgue-measurable functions defined on a finite interval $[a, b]$ that are p^{th} power integrable, and let $||\cdot||$ denote the usual norm. For each integrable function $x\colon [a, b] \to \mathbb{R}$ and each $h > 0$, let the so-called *Steklov function* be defined by

$$x_h(t) = \frac{1}{2h} \int_{t-h}^{t+h} x(\tau)\, d\tau, \quad t \in [a, b]$$

where (abusing the notation) x in the integral is an extension of x that vanishes outside $[a, b]$. Note that x_h is continuous on $[a, b]$ and that, in particular, it belongs to L_p. For a given $n \in \mathbb{N}$ and $h > 0$, we use $\mathcal{S}_{n,h}$ to denote

$$\left\{ s\colon L_p \to \mathbb{R} \,|\, s(x) = \sum_{j=1}^{n} c_j\, x_h(\alpha_j), \quad c_j \in \mathbb{R}, \quad 0 \leq j \leq n \right\}$$

in which the α_j are the elements of the uniformly spaced grid of n points $\alpha_j = a + j(b-a)/n$, $1 \leq j \leq n$. The set $\mathcal{S}_{n,h}$ is a family of what might be called Steklov functionals of order n and width h.

[2] We also extend the result in [11] to $p = 1$.

The class $\mathcal{S}$ of all Steklov functionals is defined by

$$\mathcal{S} = \bigcup_{\substack{n \in \mathbb{N} \\ h > 0}} \mathcal{S}_{n,h}.$$

4.2.2 The Theorem

Our result is the following.

Theorem: Let f be a real continuous linear functional defined on L_p and let C be a compact subset of L_p. Then for each $\epsilon > 0$ there exists an $s \in \mathcal{S}$ such that

$$|f(x) - s(x)| < \epsilon$$

for all $x \in C$.

Proof:

Let $\epsilon > 0$ be given. By the characterization of relatively compact sets[3] in L_p due to Kolmogorov as extended to $p = 1$ by A. N. Tulaikov (see [35], pp. 212-216), for any $\delta > 0$ there exists an $h > 0$ such that $||x - x_h|| < \delta$ for all $x \in C$. Using the uniform continuity of f, choose h so that

$$|f(x) - f(x_h)| < \epsilon/3, \ x \in C. \tag{4.1}$$

Since f is a continuous linear functional on L_p, we have

$$f(x_h) = \int_a^b k(\tau) x_h(\tau)\, d\tau \tag{4.2}$$

for some $k \in L_q$ where q is the conjugate index of p. Since the map from $x \in C$ to x_h in the normed space of continuous functions is continuous, we see that $\{x_h \mid x \in C\}$ is compact, and so $M = \sup_{x \in C} \sup_{t \in [a,b]} |x_h(t)|$ is finite. Using the integrability of k, let g be a step function on $[a, b]$ such that

$$M \int_a^b |k(\tau) - g(\tau)|\, d\tau < \epsilon/3. \tag{4.3}$$

By (4.1), (4.2), (4.3) and the triangle inequality in $\mathbb{R}$,

$$\left| f(x) - \int_a^b g(\tau) x_h(\tau)\, d\tau \right| < 2\epsilon/3, \ x \in C. \tag{4.4}$$

We will now show that there is an n and an $s \in \mathcal{S}_{n,h}$ such that

$$\left| \int_a^b g(\tau) x_h(\tau)\, d\tau - s(x) \right| < \epsilon/3, \ x \in C, \tag{4.5}$$

[3] In [35] the term "compact" means relatively compact.

and this will complete the proof. (Recall that h has been chosen.) We use the following lemma.

Lemma ([51]): Let A and B be normed linear spaces, and let K be a compact subset of A. Suppose that $G : K \to B$ is Lipschitz. Let $\mathcal{F}$ be any family of Lipschitz maps $H : K \to B$ such that given any $\delta > 0$ and any finite set $\{x_1, \ldots, x_r\}$ of points of K there is an $H \in \mathcal{F}$ for which

$$||Gx_i - Hx_i|| < \delta, \qquad 1 \le i \le r.$$

Then given $\epsilon > 0$ there is an $H \in \mathcal{F}$ such that $||Gx - Hx|| < \epsilon$ for all $x \in K$.

It is easy to check that the map from $x \in C$ to $\int_a^b g(\tau)x_h(\tau)\,d\tau$ as well as the elements of $\cup_{n\in\mathbb{N}}\mathcal{S}_{n,h}$ are Lipschitz. Referring to the lemma, take $A = L_p, K = C, B = \mathbb{R}$, $\mathcal{F} = \cup_{n\in\mathbb{N}}\mathcal{S}_{n,h}$ and let G be defined by

$$Gx = \int_a^b g(\tau)x_h(\tau)\,d\tau. \tag{4.6}$$

Now choose any $x \in C$ and any $\delta > 0$. Since gx_h is bounded and sectionally continuous on $[a,b]$, there is an $n_0(x) \in \mathbb{N}$ with the following property. For $n \in \mathbb{N}$ with $n \ge n_0(x)$ the Riemann partition $\alpha_j = a + j(b-a)/n, \quad 0 \le j \le n$, of $[a,b]$ is such that

$$\left| \int_a^b g(\tau)x_h(\tau)\,d\tau - \sum_{j=1}^{n} g(\alpha_j)x_h(\alpha_j)(b-a)/n \right| < \delta.$$

Thus, given a set $\{x_1, \ldots, x_r\}$ of points of C we have

$$\left| Gx_i - \sum_{j=1}^{n} c_j\, x_{ih}(\alpha_j) \right| < \delta, \qquad 1 \le i \le r.$$

for $n = \max_i\{n_0(x_i)\}$, in which $c_j = g(\alpha_j)(b-a)/n$. By the lemma there is an $s \in \cup_{n\in\mathbb{N}}\mathcal{S}_{n,h}$ such that (4.5) holds. This completes the proof of the theorem.

4.3 Comments

A check of the proof shows that h (which depends on C and ϵ) can be chosen to be independent of f for f of norm at most unity, and that the theorem holds also for the case in which f is complex valued, L_p is a complex space, and the coefficients c_j are complex numbers.

The last part of the proof can be generalized to yield an interesting result concerning what might be called the "uniform Riemann integrability" of certain integrands. This result, which might be useful in other approximation studies, is discussed in Appendix A.

Chapter 5

Approximation of Nonlinear Maps

Here we show that nonlinear dynamic systems defined on various spaces can be approximated arbitrarily well with a variety of neural network structures. We give two schemes for such approximation, and also provide error analysis for one scheme. We demonstrate the applicability of our results to a large class of stable nonlinear feedback systems in continuous-time, and causal time-invariant systems in the discrete-time case.

5.1 Introduction

One of the principal results of this work is that finite sums of the form

$$\sum_{i=1}^{m} \sum_{j=1}^{k(i)} c_{ij} u_{ij}[y_{ij}(\cdot)] v_i$$

can approximate (arbitarily well in an important sense) the elements of a large class of input-output maps of nonlinear dynamical systems taking a certain subset of the linear space of $\mathbb{R}^n$-valued continuous functions defined over $\mathbb{R}$, into the linear space of $\mathbb{R}$-valued functions defined over $\mathbb{R}$. Here the c_{ij} are real constants, the u_{ij} are certain continuous real-valued functions of the reals, the y_{ij} are continuous real functionals, and the v_i are elements of the output space. A similar result follows also for the case of functions defined over $\mathbb{R}^n$ or over discrete domains. Corresponding results for complex-valued inputs and/or outputs can also be derived, but that is not considered here.

To the best of our knowledge, this is the first approximation result (using structures like neural networks) concerning input-output maps of systems that need not be time (or shift) invariant. Another fact that makes this result interesting is that the inputs are not restricted to be defined on finite intervals. Also, the set of inputs need not be compact in the usual topologies of the input spaces.

Results related in a general sense to ours can be found in the work of Chen and Chen [12], where the approximating structures are certain forms of radial basis function neural networks. In contrast to [12], our results do not restrict the form of the basis function to be radial with respect to the Euclidean norm, thus permitting very flexible choices of approximating structures. Approximation results using Volterra series for nonlinear systems were obtained in [47] (in this connection see also Boyd's dissertation [6]). In contrast to [6], we treat input and output functions with multidimensional domains and ranges, as well as provide approximations in the L_p norm. Unlike [47], we do not require Fréchet differentiability of the input-output map.

5.2 Preliminaries

5.2.1 The inputs and outputs

We consider nonlinear dynamical systems as input-output maps between appropriate input and output spaces as defined below. Throughout this work V denotes $\mathbb{R}^d$, for an arbitrary positive integer d. We denote the norm in V by $||\cdot||_V$. Let n be a positive integer, and let D denote the domain of the input functions, where $D = \mathbb{R}^n_+$ or $\mathbb{Z}^n_+$, or $\mathbb{R}^n$ or $\mathbb{Z}^n$. With D we associate the "natural" measure μ (the Lebesgue measure for $\mathbb{R}^n_+$ or $\mathbb{R}^n$, and the counting measure for $\mathbb{Z}^n_+$ or $\mathbb{Z}^n$). The norm in D is denoted by $|\cdot|$, taken to be the Euclidean norm for $\mathbb{R}^n_+$ or $\mathbb{R}^n$, and the inherited Euclidean norm for $\mathbb{Z}^n_+$ or $\mathbb{Z}^n$.

Let $\mathcal{X}$ denote the set of all V-valued μ-measurable functions defined on D. By a slight abuse of notation, let both the zero element of $\mathcal{X}$ and that of V be denoted by θ. We assume that the inputs are drawn from $\mathcal{X}$. We let X denote a subset of $\mathcal{X}$ satisfying one of the following assumptions:

A1. $X = \mathcal{C}_b(D)$, the set of all bounded continuous functions in $\mathcal{X}$, with the norm given by

$$||f|| = \sup_{t \in D} ||f(t)||_V .$$

A2. $X = L_p(D, \mu)$, $1 \le p < \infty$, with the norm defined by

$$||f|| = \left\{ \int_D ||f(t)||_V{}^p \mu(dt) \right\}^{1/p} .$$

Note that in each case X is a Banach space.

Let E be a real Banach space with norm $||\cdot||_E$, e.g., E can be taken to be the linear space of $\mathbb{R}$-valued bounded continuous functions defined over D with $||y||_E = \sup_{t \in D} |y(t)|$. We assume that our outputs belong to E.

5.2.2 Localizable systems

We first define the weighted norm spaces used in the development of our results.[1]

- If $X = \mathcal{C}_b(D)$, let W be the set of continuous maps $w\colon D \to (0, 1]$ with $\lim_{|t| \to \infty} w(t) = 0$. For $w \in W$, let X_w be the set of all continuous functions f in $\mathcal{X}$ for which

$$||f||_w = \sup_{t \in D} ||f(t) w(t)||_V < \infty .$$

[1] Related ideas (in a general sense) involving weighted norms on function spaces can be found in, for example, proofs of the boundedness of solutions of integral equations [45], [46, pp. 880–1], and in a large number of other publications (see, for instance, [6], [17], [15], [16]).

- If $X = L_p(D, \mu)$, let W be the set of μ-measurable maps $w: D \to (0, 1]$ with $\lim_{|t|\to\infty} w(t) = 0$. For $w \in W$, let X_w be the set of all functions f in $\mathcal{X}$ for which

$$||f||_w = \left\{ \int_D ||f(t)||_V{}^p w(t)\mu(dt) \right\}^{1/p} < \infty.$$

Later we will use the fact that in every case X_w is a Banach space (see Appendix B).

For every $a \in \mathbb{R}_+$, let c_a denote $[-a, a]^n \cap D$. Let λ be a function from $\mathbb{R}_+$ to $\mathbb{R}_+$, and assuming **A1** holds, let $R \in X_w$ be a function that satisfies

$$\lim_{|t|\to\infty} ||R(t)w(t)||_V = 0. \tag{5.1}$$

Let B_0 be defined as follows:

1. If $D = \mathbb{Z}_+^n$ or $\mathbb{Z}^n$, let B_0 be any nonempty subset of $\mathcal{X}$ satisfying the following condition: for every $f \in B_0$ and $t \in c_a$,

$$||f(t)||_V \le ||R(a)||_V.$$

2. If $D = \mathbb{R}_+^n$ or $\mathbb{R}^n$, let B_0 be any nonempty subset of $\mathcal{X}$ such that for every $f \in B_0$ and $t \in c_a$, $||f(t)||_V \le ||R(a)||_V$. We assume additionally that:

   - Under Assumption **A1**, $||f(t) - f(\tau)||_V \le \lambda(a)|t - \tau|$ for all $t, \tau \in c_a$, and every $f \in B_0$.
   - Under Assumption **A2**, let η be a function from $\mathbb{R}_+$ to $\mathbb{R}_+$, such that the discontinuities in c_a of any $f \in B_0$ lie on at most $\eta(a)$ hyperplanes perpendicular to the coordinate axes, and for every n-interval of continuity $\mathcal{I}$ of f, $||f(t) - f(\tau)||_V \le \lambda(a)|t - \tau|$ for all $t, \tau \in c_a$, and every $f \in B_0$.

Let $B = \overline{B_0}$ denote the closure of B_0 in the norm $||\cdot||_w$. We say that a map $M: B \to E$ is *w-localizable* on B if M is continuous in the norm $||\cdot||_w$.

5.2.3 Motivation

A characterization for time-varying kernels

Here we give a characterization of w-localizability in the case of maps defined by convolution with a time-varying kernel. Consider input-output maps M under Assumption **A1**, for the case $D = \mathbb{R}_+$, $V = \mathbb{R}$. Here E is the linear space of $\mathbb{R}$-valued bounded continuous functions defined over D with $||y||_E = \sup_{t\in D} |y(t)|$. Let the kernel $k(\cdot, \cdot)$ be a jointly measurable bounded real valued function on $D \times D$.

Proposition 1 Let M be a linear causal time-varying map M on B defined by

$$(Mx)(t) = \int_0^t k(t, \tau)x(\tau)d\tau, \tag{5.2}$$

for all $t \in \mathbb{R}_+$ and $x \in B$. M is w-localizable if and only if

$$\sup_{t \in \mathbb{R}_+} \int_0^t \left| \frac{k(t,\tau)}{w(\tau)} \right| d\tau \; < \; \infty. \tag{5.3}$$

Proof: Note that M may be reexpressed in terms of the new kernel $k_w(t,\tau) = k(t,\tau)/w(\tau)$ since w is continuous and does not vanish. Hence,

$$(Mx)(t) = \int_0^t k_w(t,\tau) x(\tau) w(\tau) d\tau.$$

We will use the following well-known result [4], [61] (see also [20] for a succinct proof). Let L_∞ be the linear space of all essentially bounded measurable real valued functions on D, and let $g(\cdot,\cdot)$ be a jointly measurable real valued function on $D \times D$. Then, G defined by

$$(Gx)(t) = \int_0^t g(t,\tau) x(\tau) d\tau$$

is a bounded map from L_∞ into L_∞ if and only if

$$\sup_{t \in \mathbb{R}_+} \int_0^t |g(t,\tau)| \, d\tau \; < \; \infty.$$

Sufficiency: Assume that k_w satisfies the condition (5.3), then by the above mentioned result there exists a constant $A > 0$ such that

$$y \in L_\infty \;\; \Rightarrow \;\; ||My||_\infty \leq A||y||_\infty.$$

Let $\epsilon > 0$ be given and choose $\delta = \epsilon/A$. Consider $x \in X_w$ such that $||x||_w < \delta$. Since $xw \in L_\infty$ and $||xw||_\infty < \delta$, $||Mx||_\infty \leq A\delta = \epsilon$. Hence M is w-localizable.

Necessity: Assume that M is w-localizable, i.e., it is continuous in the norm $||\cdot||_w$, hence bounded. Therefore, there exists a constant $A' > 0$ such that

$$x \in X_w \;\; \Rightarrow \;\; ||Mx||_\infty \leq A'||x||_w.$$

We will show that M as defined in (5.2) over L_∞ is bounded hence (5.3) is satisfied. Let $y \in L_\infty$ be arbitrary. We will use the following expressions for the output My at any arbitrary but fixed time $t \geq 0$:

$$\begin{aligned} (My)(t) &= \int_0^t k(t,\tau) y(\tau) d\tau \\ &= \int_0^t k_w(t,\tau) y(\tau) w(\tau) d\tau. \end{aligned}$$

Using Luzin's Theorem [29, p. 293], we can find a continuous real valued function y_c defined on $\mathbb{R}_+$ such that $\mu(L_t) < 1/K_t$ and $y_c(\tau) = y_c(t)$ for all $\tau > t$, where $L_t = \{\tau \in [0,t] \colon y_c(\tau) \neq y(\tau)\}$ and $K_t = \max_{\tau \in [0,t]} k(t,\tau)$. Hence

$$|(My)(t)| \;\; \leq \int_{L_t} |k(t,\tau) y(\tau)| \, d\tau + \int_{[0,t] \setminus L_t} |k_w(t,\tau) y_c(\tau) w(\tau)| \, d\tau$$

$$\begin{aligned} &\leq K_t||y||_\infty \mu(L_t) + A'||y_c||_w \\ &\leq ||y||_\infty + A'||y_c||_\infty \\ &\leq (1+A')||y||_\infty. \end{aligned}$$

The last inequality follows from the construction of y_c given in [21, Theorem 7.5.2, p. 190], since y_c may be chosen to satisfy $y_c(\tau) \leq ||y||_\infty$ for all $\tau \in [0,t]$. ■

Relationship of w-localizability to usual continuity

Here we consider linear input-output maps M under Assumption **A1**, for the case $D = \mathbb{R}_+$, $V = \mathbb{R}$, with E being the linear space of $\mathbb{R}$-valued bounded continuous functions defined over D with the usual sup norm. We show that w-localizability is stronger than continuity in the usual norm. We also give sufficient conditions on the set of inputs B to guarantee w-localizability given continuity in the usual norm.

Proposition 2 Let M be a w-localizable linear map $M: B \to E$, where the inputs are bounded in the usual sense, i.e., $||x||_\infty < \infty$ for any input $x \in B$. Then M is continuous in the usual norm $||\cdot||_\infty$.

Proof: Since M is w-localizable, it is continuous in the norm $||\cdot||_w$ by definition. Hence, given any positive ϵ there exists a positive δ such that $||x||_w < \delta \;\Rightarrow\; ||Mx||_\infty < \epsilon$. Since $w: D \to (0,1]$, we have $||x||_\infty < \delta \;\Rightarrow\; ||x||_w < \delta$. Hence M is continuous in the norm $||\cdot||_\infty$. ■

Now we give sufficient conditions on B so that w-localizability follows from continuity in the usual norm. Let B satisfy the following conditions:[2]

1. For any input $x \in B$, $||x||_\infty < \infty$.
2. There exists a positive T with the following property: for every $x \in B$ there exists a $\rho_x \in (0,1)$ such that $\sup_{\tau > T} |x(\tau)| < \rho_x ||x||_\infty$.

Proposition 3 Let M be a linear map $M: B \to E$, where B satisfies the above conditions. Assume that M is continuous in the usual norm $||\cdot||_\infty$. Then M is w-localizable for any nonincreasing $w \in W$.

Proof: Let $T > 0$ be chosen according to the above conditions. Let $w \in W$ be any nonincreasing weight. Note that

$$||x||_w = \max\{\max_{t\in[0,T]} |x(t)w(t)|, \sup_{t\in(T,\infty)} |x(t)w(t)|\}.$$

Since w is nonincreasing,

$$\max_{t\in[0,T]} |x(t)w(t)| \;\geq\; w(T) \max_{t\in[0,T]} |x(t)|,$$

[2] These conditions are met by, for example, a set B consisting of functions with support contained in a fixed finite interval.

and

$$\sup_{t\in(T,\infty)}|x(t)w(t)| \;\le\; w(T)\sup_{t\in(T,\infty)}|x(t)|.$$

Since $\sup_{t\in(T,\infty)}|x(t)|<\rho||x||_\infty<||x||_\infty$, it follows that $\max_{t\in[0,T]}|x(t)|=||x||_\infty$. Hence

$$\begin{aligned}||x||_w &= \max_{t\in[0,T]}|x(t)w(t)|\\ &\ge w(T)\max_{t\in[0,T]}|x(t)|\\ &= w(T)||x||_\infty.\end{aligned}$$

It follows that

$$||x||_\infty \le w(T)^{-1}||x||_w,\quad x\in B.$$

Since M is continuous in the usual norm $||\cdot||_\infty$, given any positive ϵ there exists a positive δ such that $||x||_\infty<\delta \;\Rightarrow\; ||Mx||_\infty<\epsilon$. Let $\eta=\delta w(T)$. Hence, $||x||_w<\eta \;\Rightarrow\; ||x||_\infty<\delta$. Combining the above implications, we have the desired result. ∎

5.2.4 Approximating structures

We denote by X_w^* the set of bounded linear functionals on X_w (i.e., the set of bounded linear maps from X_w to the reals $\mathbb{R}$). Given a compact subset C of X_w, let Y be any set of continuous maps from X_w to $\mathbb{R}$ that is dense in X_w^* on C, in the sense that for each $\phi\in X_w^*$ and any $\rho>0$ there is a $y\in Y$ such that $|\phi(x)-y(x)|<\rho$, $x\in C$. Unless otherwise mentioned, $w\in W$ will henceforth be fixed.

Also, let U be any set of continuous maps $u:\mathbb{R}\to\mathbb{R}$ such that given $\sigma>0$ and any bounded interval $(\beta_1,\beta_2)\subset\mathbb{R}$ there exists a finite number of elements $u_1,\ldots,u_\ell$ of U for which $|\exp(\beta)-\sum_j u_j(\beta)|<\sigma$ for $\beta\in(\beta_1,\beta_2)$.[3]

Let $\mathcal{T}$ denote the set of maps from B to E of the form

$$\sum_{i=1}^{m}\sum_{j=1}^{k(i)} c_{ij}u_{ij}[y_{ij}(\cdot)]v_i \tag{5.4}$$

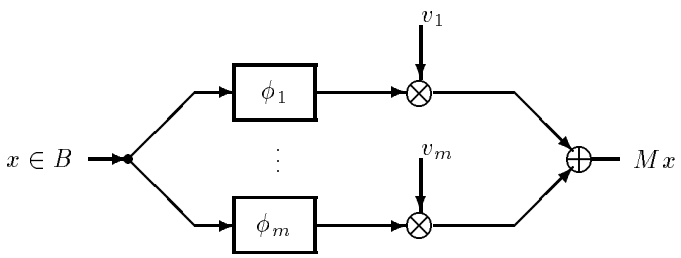
where $m\in\mathbb{N}$, $k(i)\in\mathbb{N}$ for $1\le i\le m$, and $c_{ij}\in\mathbb{R}$, $u_{ij}\in U$, $y_{ij}\in Y$, $v_i\in E$ for $1\le i\le m$, $1\le j\le k(i)$. A typical element M of $\mathcal{T}$ can be realized by what may be called a *tensor product neural network*, since, by defining the continuous functionals $\phi_i(\cdot)=\sum_{j=1}^{k(i)}c_{ij}u_{ij}[y_{ij}(\cdot)]$, we see immediately that $M(\cdot)=\sum_{i=1}^{m}\phi_i(\cdot)v_i$ is an element of the so-called tensor product [13], and that M can be implemented as a neural network as shown in Figure 5.1.

5.3 Characterization of Continuity

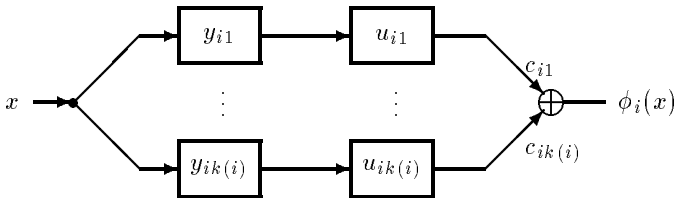
Let H be a compact metric space. The linear space $\mathcal{G}$ is the set of all maps $G\colon H\to E$ such that

$$||G||_{\mathcal{G}}=\sup_{x\in H}||Gx||_E<\infty.$$

[3] Of course we can take U to be the set whose only element is $\exp(\cdot)$, or the set $\{u:u(\beta)=(\beta)^n/n!,\ n\in\{0,1,\ldots\}\}$.



(a) Network representation of the tensor product $M(x) = \sum_{i=1}^{m} \phi_i(x) v_i$.



(b) Neural network implementation of one of the input functionals $\phi_i(x) = \sum_{j=1}^{k(i)} c_{ij} u_{ij}[y_{ij}(x)]$.

Figure 5.1: A tensor product neural network M.

Note that the set of continuous maps, denoted by $\mathcal{C}(H, E)$, is a subspace of $\mathcal{G}$. To give an example of a discontinuous map $G \in \mathcal{G}$ consider disjoint nonempty compact subsets $C_1, \cdots, C_m$ of H with metric ρ and the distance between $A, B \subset H$ given by $d(A, B) = \inf\{\rho(a, b) : a \in A,\ b \in B\}$. With $C_0 = H \setminus \cup_{i=1}^{m} C_i$, let G be defined by

$$Gx = i, \text{ if } x \in C_i,\ i \in \{0, 1, \cdots, m\},$$

where it is assumed that

$$d(C_i, C_0) = 0,\ i = 1, \cdots, m.$$

5.3.1 The approximation theorem

Theorem 5 Let $G : B \to E$ be given, and let $\mathcal{T}$ be defined as in (5.4) with $C = B$. Then the following are equivalent.

1. G is w-localizable on B.
2. For any $\epsilon > 0$, there exists $M \in \mathcal{T}$ such that $||G - M||_{\mathcal{G}} < \epsilon$.

Proof: (1) $\Rightarrow$ (2). We first prove that B_0 is relatively compact in the norm $||\cdot||_w$. Define a sequence of operators $T_1, T_2, \ldots$ from X_w into itself for the continuous and discrete input domains separately:

1. ($D = \mathbb{Z}_+^n$ or $D = \mathbb{Z}^n$)

$$(T_k s)(t) = \begin{cases} s(t), & \text{if } t \in c_k \\ \theta, & \text{otherwise.} \end{cases} \tag{5.5}$$

2. ($D = \mathbb{R}_+^n$ or $D = \mathbb{R}^n$)
   Under assumption **A1**, we define T_k by the following extension. Note that the norm $|\cdot|$ in D is uniformly convex. Since c_k is closed and convex, for any $x \in D$ there exists a unique $\hat{x} \in c_k$ such that $||x - \hat{x}|| = \inf_{y \in c_k} ||x - y||$. Let

$$(T_k s)(x) = s(\hat{x}) \tag{5.6}$$

   Under assumption **A2**, we let T_k be defined as in (5.5).

We use the following tool theorem.

Theorem 6 ([52]) Let S be a subset of a complete metric space A with metric ρ, and let $T_1, T_2, \ldots$ be maps of A into itself such that

(i) $T_k(S)$ is a relatively compact subset of A for each k, and

(ii) $\rho(s, T_k s) \to 0$ as $k \to \infty$ uniformly for $s \in S$.

Then S is a relatively compact subset of A.

With $A = X_w$, $S = B_0$, and ρ the metric induced by the norm $||\cdot||_w$, the theorem implies (by the verification of its hypotheses in Appendix B) that B_0 is relatively compact in the norm $||\cdot||_w$.

Since G is w-localizable, G is continuous in the norm $||\cdot||_w$ by definition, i.e., $G \in \mathcal{C}(B, E)$.

The space of all continuous functionals on B, $\mathcal{C}(B, \mathbb{R})$, is denoted by $\mathcal{C}(B)$ as usual. Let A_1 be a subspace of $\mathcal{C}(B)$. The so-called *tensor product* [13] of A_1 and E (denoted by $A_1 \otimes E$) is the linear manifold of $\mathcal{C}(B, E)$ consisting of finite sums of the form $\Phi(\cdot) = \sum_{j=1}^{k} a_j(\cdot) v_j$, where $a_j \in A_1$ and $v_j \in E$. We need the following technical proposition.

Proposition 4 $\mathcal{C}(B) \otimes E$ is dense in $\mathcal{C}(B, E)$.

Proof of the Proposition: Let $\mathcal{M} = \mathcal{C}(B) \otimes E$, and let $\mathcal{M}(x) = \{f(x) : f \in \mathcal{M}\}$. Since $\mathcal{C}(B)$ contains the constant function taking the value 1, $\mathcal{M}(x) = E$ for every $x \in B$. The proposition follows from [7, Corollary 1].

Continuing with the proof of the theorem, and using the above Proposition, there exist continuous functionals a_i and elements v_i of E such that

$$||Gu - \textstyle\sum_{i=1}^{m} a_i(u) v_i||_E < \epsilon/2, \quad u \in B.$$

By Theorem 1 of [51], there exist $k(i) \in \mathbb{N}$, $c_{ij} \in \mathbb{R}$, $u_{ij} \in U$, $y_{ij} \in Y$, $1 \le j \le k(i)$ such that

$$\left| a_i(u) - \sum_{j=1}^{k(i)} c_{ij} u_{ij}[y_{ij}(u)] \right| < \epsilon/(2mv), \quad u \in B, \ 1 \le i \le m,$$

where $v = \max_{1 \le i \le m} ||v_i||_E$. Combining the above estimates using the triangle inequality, we get the desired result.

(2) $\Rightarrow$ (1). Given $\epsilon > 0$, let $\hat{G} \in \mathcal{T}$ be chosen such that

$$||Gu - \hat{G}u||_E < \epsilon/3, \quad u \in B. \tag{5.7}$$

Note that $\hat{G}$ is continuous on B which is compact, hence $\hat{G}$ is uniformly continuous, i.e., there exists $\delta > 0$ such that

$$u, v \in B, \ ||u - v||_w < \delta \ \Rightarrow \ ||\hat{G}u - \hat{G}u||_E < \epsilon/3. \tag{5.8}$$

By (5.7), (5.8) and the triangle inequality,

$$||Gu - Gv||_E \le ||Gu - \hat{G}u||_E + ||\hat{G}u - \hat{G}v||_E + ||\hat{G}v - Gv||_E < \epsilon. \tag{5.9}$$

From the above, G is continuous in the norm $||\cdot||_w$, hence G is w-localizable.

■

More generally, the following result is true. Again, with H a compact metric space, let $\mathcal{P}$ be the set of all maps of the form $x \mapsto \sum_{i=1}^{m} \phi_i(x) v_i$ where $\phi_i \in \mathcal{C}(H)$, and $v_i \in E$.

Scholium 1 A map $G: H \to E$ is continuous if and only if for any $\epsilon > 0$, there exists $M \in \mathcal{P}$ such that $||Gx - Mx||_E < \epsilon,\ x \in H$.

Proof: Note that the proof of sufficiency can be given exactly analogous to that of Theorem 5, since any $M \in \mathcal{P}$ is (uniformly) continuous on H.

Similarly for the necessity, note that we can stop after using Proposition 4 in the proof of Theorem 5. ∎

5.4 Applications to Time-invariant Systems

Here we exhibit a large class of applications of our last result in the area of approximation of causal time-invariant systems.

5.4.1 Notation

Let $D = \mathbb{Z}_+$, and let X be the linear space of real-valued sequences with the metric

$$\rho(x, y) = \sum_{k \in D} 2^{-k} \frac{|x(k) - y(k)|}{1 + |x(k) - y(k)|}.$$

With an arbitrary positive number c, let $S = \{x \in X: |x(t)| \leq c,\ t \in \mathbb{Z}_+\}$. From the criteria of compactness in X [35, Theorem 1, p. 202, vol. II], S is compact.

For every $\beta \in D$, let the delay and advance operators on X be defined by

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

$$(T^\beta x)(t) = x(t + \beta),\ t \geq \beta.$$

Note that S is closed under T_β and T^β. Let $U = S$. We say that $M: U \to X$ is *time-invariant* if $MT_\beta = T_\beta M$ for every β.

Let $s|_{c_a}$ stand for the restriction of $s \in X$ to c_a. As usual, we say that a map $M: U \to X$ is *causal* if for all $a \in D$ and $x, y \in S$, $x|_{c_a} = y|_{c_a}$ implies $(Mx)(a) = (My)(a)$.

Let $W_{t,a}: X \to X$ for $t, a \in D,\ a > 0$ be a windowing operator defined by

$$(W_{t,a} x)(\alpha) = \begin{cases} x(\alpha), & t - a \leq \alpha \leq t \\ 0, & \text{otherwise.} \end{cases}$$

We say that $M: U \to X$ has *approximately finite memory* on S if given any $\epsilon > 0$ there is an $a \in D,\ a > 0$ such that

$$|(Gs)(t) - (GW_{t,a} s)(t)| < \epsilon,\ \ t \in D,\ s \in S.$$

Let $S_a = \{s|_{c_a} : s \in S\}$ be topologized by a metric ρ_a. For any $a \in D$, the *associated functional* M_a for a causal operator $M: U \to X$ is defined on S_a by

$$M_a x = (My)(a),\ \ x \in S_a$$

where $y \in S$ is chosen such that $y|_{c_a} = x$.

5.4.2 Approximation of discrete-time causal time-invariant systems

In this section, by "G meets hypothesis H" we mean that $G: U \to X$ is a causal, time-invariant map with approximately finite memory on S, and G_a is uniformly continuous on S_a for every a. Here,

$$\rho_a(x, y) = \sum_{k=1}^{a} 2^{-k} \frac{|x(k) - y(k)|}{1 + |x(k) - y(k)|}.$$

Lemma 3 If $G: U \to X$ meets hypothesis H, then for any $\eta > 0$, there exist $m \in \mathbb{N}$, $\phi_i \in \mathcal{C}(S)$, and $v_i \in X,\ i = 1, \ldots, m$, such that

$$\left| (Gs)(t) - \sum_{i=1}^{m} \phi_i(s) v_i(t) \right| < \eta, \quad t \in D,\ s \in S.$$

Proof: We only need to show that $G \in \mathcal{C}(S, X)$, since the result then follows from Scholium 1. Let ϵ be an arbitrary positive number. Using approximately finite memory of G, choose $a \in D$ such that

$$|(Gs)(t) - (GW_{t,a}s)(t)| < \epsilon/3, \quad t \in D,\ s \in S. \tag{5.10}$$

Hence, for $s_1, s_2 \in S$,

$$\begin{aligned} |(Gs_1)(t) - (Gs_2)(t)| &\leq |(Gs_1)(t) - (GW_{t,a}s_1)(t)| \\ &\quad + |(GW_{t,a}s_1)(t) - (GW_{t,a}s_2)(t)| \\ &\quad + |(Gs_2)(t) - (GW_{t,a}s_2)(t)| \\ &< 2\epsilon/3 + |(GW_{t,a}s_1)(t) - (GW_{t,a}s_2)(t)| \end{aligned}$$

Now we consider two cases:

1. If $t \geq a$, let $v \in S_a$ be defined by

   $$v(\alpha) = (T^{t-a} W_{t,a} s)(\alpha), \quad \alpha \in c_a.$$

   Using the causality and time-invariance of G,

   $$(GW_{t,a}s)(t) = (GT_{t-a}T^{t-a}W_{t,a}s)(t) = (GT^{t-a}W_{t,a}s)(a) = G_a(v).$$

2. If $t < a$, let $v \in S_a$ be defined by

   $$v(\alpha) = (T_{a-t} s)(\alpha), \quad \alpha \in c_a,$$

   since S is closed under T_β. Now use the causality and time-invariance of G again,

   $$(GW_{t,a}s)(t) = (Gs)(t) = (GT_{a-t}s)(t) = G_a(v).$$

Since G_a is uniformly continuous in the ρ_a metric on S_a, we can choose a $\delta > 0$ such that $|G_a(w) - G_a(w')| < \epsilon/3$ for $w, w' \in S_a$ satisfying $\rho_a(w, w') < \delta$. It is clear that

$$\rho(s_1, s_2) < \delta \;\Rightarrow\; \rho_a(s_1|_{c_a}, s_2|_{c_a}) < \delta$$

hence

$$|(Gs_1)(t) - (Gs_2)(t)| \;<\; 2\epsilon/3 + |G_a(v_1) - G_a(v_2)| \;<\; \epsilon.$$

Thus, $G: S \to X$ is continuous. ∎

5.4.3 A conjecture about nonlinear feedback systems

Here the writer would like to conjecture that our result above can be used to approximate members of a large class of discrete-time stable nonlinear feedback systems using networks implementing the so-called tensor products. Consider the closed-loop input-output maps of feedback control systems (see Figure 5.2) containing a time-invariant sector nonlinearity represented by the operator N and a linear time-invariant system L for which the circle condition [45] is met. The conjecture is that, under these conditions, the map G taking the input r to the output y meets hypothesis H (similar to [50, Theorem 4], but in discrete-time setting) of Lemma 3, hence it can be approximated uniformly with tensor product networks.

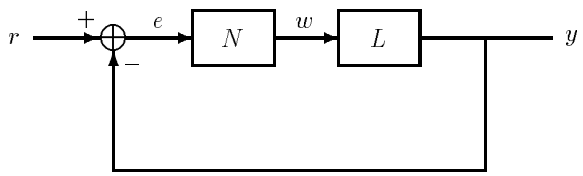


Figure 5.2: Nonlinear control system.

Results along similar lines can be proved for the case of continuous-time systems, as well as for more general domains of input functions, e.g., $D = \mathbb{R}_+^n$ or $\mathbb{Z}_+^n$, or $\mathbb{R}^n$ or $\mathbb{Z}^n$, but only $D = \mathbb{R}_+$ is considered next.

5.4.4 Continuous-time systems on the half-line

Here we treat the case $D = \mathbb{R}_+$. We use the same formal definitions for the delay, advance and windowing operators and the concepts of time-invariance, causality, approximately finite memory and associated functionals as in Section 5.4.1, keeping in mind that now $X = L_2(0, \infty)$. We also abbreviate the latter to just L_2. Let U be the collection of equivalence

classes of Lebesgue measurable real-valued functions on D. Note that U is closed under the T_t and T^t. We assume that S, a nonempty subset of U, satisfies the following conditions:

1. S is closed under the T^t.

2. S is a compact subset of L_2. (It follows that S_a is a compact subset of U_a in the metric

$$\rho_a(x,y) = \left(\int_0^a |x(t)-y(t)|^2 dt\right)^{1/2}$$

for every $a > 0$, since the map $x \mapsto x|_{c_a}$ is continuous.)

To get a class of examples of S, we introduce a concept reminiscent of the Hilbert cube [29, Example 5, p. 98] in one respect as follows: Let λ be any nondecreasing nonnegative function defined on $\mathbb{R}_+$. Assume that $\beta \in L_2$ is a bounded function, and let $\alpha > 0$ be given. $D(\lambda,\alpha,\beta) \subset U$ is called a *β-fundamental set* if every $x \in D(\lambda,\alpha,\beta)$ satisfies the following:

1) For any positive T, the number of discontinuities of x in $[0,T]$ is at most $\lambda(T)$.

2) On every interval of continuity $\mathcal{I}$ of x,

$$|x(t_1) - x(t_2)| \le \alpha |t_1 - t_2|, \ \forall\, t_1, t_2 \in \mathcal{I}.$$

3) $|x(t)| \le |\beta(t)|, \ t \in \mathbb{R}_+$.

It is clear that a β-fundamental set is a subset of L_2. The proof of the following proposition can be given using some of the ideas from Section 3.2.2 and from [29, Theorem 3, p. 101].

Proposition 5 $D(\lambda,\alpha,\beta)$ is relatively compact in $L_2(0,\infty)$.

It is clear that S can be taken to be the closure of any $D(\lambda,\alpha,\beta)$ that is closed under the T^t (which holds trivially if β is nonincreasing).

Lemma 4 If $G\colon U \to X$ meets hypothesis H and S satisfies the two conditions stated at the beginning of this subsection, then for any $\eta > 0$, there exist $m \in \mathbb{N}$, $\phi_i \in \mathcal{C}(S)$, $v_i \in X$, $i = 1,\ldots,m$, and $a > 0$ such that

$$\left| (Gs)(t) - \sum_{i=1}^{m} \phi_i(s) v_i(t) \right| < \eta, \quad t \ge a, \ s \in S.$$

The proof follows along the same lines as that of Lemma 3, except that the case $t < a$ there is not needed, and the uniform continuity of G_a on S_a is now in the ρ_a metric as defined above.

5.4.5 Approximation of stable nonlinear feedback systems

Here we show that our result above is applicable to the problem of approximating members of a large class of stable nonlinear feedback systems using tensor product networks.

Consider a closed-loop input-output map G of a feedback control system (see Figure 5.2) containing a time-invariant sector nonlinearity for which the circle condition [44] is met.[4] In particular, assume that the signals r, e, w and y belong to $L_{\infty e}$, the space of real-valued Lebesgue measurable functions on $\mathbb{R}_+$ that are bounded on bounded subintervals. Also suppose that the initial conditions are zero, and that w and y are related by

$$y(t) = (Lw)(t) = \int_0^t k(t-\tau)w(\tau)d\tau, \quad t \in D.$$

Here the kernel $k: \mathbb{R}_+ \to \mathbb{R}$ is such that its Laplace transform K is a strictly proper rational function without any pole-zero cancellation. The nonlinear part of the system is assumed to be given by the operator

$$(Nx)(t) = \eta[x(t)], \quad t \in D,$$

where $\eta: \mathbb{R} \to \mathbb{R}$ is a time-invariant nonlinearity such that $\eta(0) = 0$, and there are two real constants α, β with the following property:

$$\alpha \leq \frac{\eta(a) - \eta(b)}{a - b} \leq \beta, \quad a \neq b.$$

We assume that one of the following "circle conditions" holds:

$0 < \alpha < \beta$. The locus of $K(i\omega)$ for $-\infty < \omega < \infty$ is bounded away from the circle in the complex plane C_1 of radius $(\alpha^{-1} - \beta^{-1})/2$ centered at $((-\alpha^{-1} - \beta^{-1})/2, 0)$, and encircles C_1 in the counterclockwise direction n_p times, where n_p is the number of poles of K with positive real parts.

$0 = \alpha < \beta$. K has no poles in the open right-half plane, and $Re[K(i\omega)] > -\beta^{-1}$ for all real ω for which $Re[K(i\omega)]$ is finite.

$\alpha < 0 < \beta$. K has no poles in the closed right-half plane, and the locus of $K(i\omega)$ for $-\infty < \omega < \infty$ is contained within the circle in the complex plane C_2 of radius $(\beta^{-1} - \alpha^{-1})/2$ centered at $((-\alpha^{-1} - \beta^{-1})/2, 0)$.

Let V be the solution map (taking u into v) of the integral equation

$$u(t) = v(t) + \int_0^t k(t-\tau)\eta[v(\tau)]d\tau, \quad t \in D,$$

and let the input-output map G of the closed-loop feedback system in Figure 5.2 be given by $y = Gr = LNVr$.

Theorem 7 Assume that the inputs r to feedback system in Figure 5.2 belong to S that satisfies the conditions stated at the beginning of Section 5.4.4. If a circle condition is met, then the input-output map G can be approximated uniformly with tensor product networks.

[4] For more information about approximation of causal time-invariant systems using a different network architecture, and about the relationship of the circle condition with approximately finite memory, see [50], on which the subsequent discussion is based.

Proof: Note that under these conditions G meets hypothesis H [50, Theorem 4 and comments in that section], hence the result follows by Lemma 4. ∎

5.5 An Algorithm and its Error Analysis

In this section we give another scheme for approximation of input-output maps. The result given here has the advantage over the one proved in the previous section that an error analysis is provided. Let F be a Banach space with a basis denoted by $\{y_k\}_{k=1}^{\infty}$, i.e., every $v \in F$ has a unique representation $v = \sum_{k=1}^{\infty} a_k(v) y_k$, in the sense of [31]. The requirement of the existence of a basis is not very restrictive, since bases have been constructed for most of the useful separable Banach spaces; however, it rules out nonseparable spaces such as L_∞.

Without loss of generality, we assume that the set $\{y_k\}_{k=1}^{\infty}$ is linearly independent. We give error bounds for an algorithm to construct an approximation to a continuous operator $G: B \to F$. Let the remainder $R_m(v)$ after m terms of the expansion of a $v \in F$ with respect to the basis be defined by

$$R_m(v) = \sum_{k=m+1}^{\infty} a_k(v) y_k .$$

Since B is compact, $G(B)$ is a compact subset of F which has a basis, hence given $\epsilon > 0$, there exists $m \in \mathbb{N}$ such that (see [31, p. 136])

$$||R_m(Gu)||_F < \epsilon, \quad u \in B.$$

Let Y_m denote the span of $y_1, \ldots, y_m$. Define

$$S_m(v) = \sum_{k=1}^{m} a_k(v) y_k , \quad v \in F.$$

Note that S_m is a projection of F onto Y_m which is continuous in $||\cdot||_F$, since the a_k are continuous functionals [31]. Hence $S_m(G(B))$ is compact, thus bounded.

Note that the coordinate isomorphism σ given by $\sigma(v) = (a_1(v), \ldots, a_m(v))$ for $v = \sum_{k=1}^{m} a_k(v) y_k$ from Y_m onto $\mathbb{R}^m$ is linear and one-to-one. Hence the function

$$\langle v, w \rangle_{Y_m} = \langle \sigma v, \sigma w \rangle , \quad v, w \in Y_m \tag{5.11}$$

defines an inner product on Y_m. The inner product induces a norm $||\cdot||_{Y_m}$ which, due to finite-dimensionality of Y_m, is equivalent to $||\cdot||_F$ in Y_m, i.e., there exist positive constants α, β such that

$$\alpha ||v||_{Y_m} \le ||v||_F \le \beta ||v||_{Y_m}, \quad v \in Y_m.$$

Let $r > 0$ be chosen such that $B_r = \{f \in Y_m \mid ||f||_{Y_m} < r\}$ contains $S_m(G(B))$. Let $\partial B_r = \{f \in Y_m \mid ||f||_{Y_m} = r\}$. For a subset A of a real topological vector space, let $co(A)$ denote the convex hull of A, and let $\overline{co}(A)$ denote the closure of the convex hull of A. Note that $B_r = \overline{co}(\partial B_r)$, hence by the algorithm given in Chapter 2 there exists $c > 0$ such that given $u \in B$, $S_m(Gu)$ can be approximated by $f_n(u) \in co(\partial B_r)$ such that

$$||S_m(Gu) - f_n(u)||_{Y_m} \le c/\sqrt{n}.$$

Also note that for every $u \in B$, $f_n(u)$ has the form

$$f_n(u) = \sum_{k=1}^{m} \sum_{i=1}^{n} \lambda_i(u) a_k(z_i(Gu)) y_k, \quad \lambda_i(u) \geq 0, \ \sum_{i=1}^{n} \lambda_i(u) = 1,$$

where $z_i \in \partial B_r$.

Now it is easy to derive the error bound for such an approximation scheme. For any $u \in B$,

$$\begin{aligned} ||Gu - f_n(u)||_F &= ||S_m Gu + R_m Gu - f_n(u)||_F \\ &\leq ||S_m Gu - f_n(u)||_F + ||R_m Gu||_F \\ &\leq \beta ||S_m Gu - f_n(u)||_{Y_m} + \epsilon \\ &\leq \frac{\beta c}{\sqrt{n}} + \epsilon. \end{aligned} \tag{5.12}$$

We have shown the existence for every $\epsilon > 0$ of a number k_ϵ such that $||Gu - f_n(u)||_F \leq \epsilon + k_\epsilon n^{-1/2}$ uniformly for $u \in B$.

Note that the approximating structures f_n have the form

$$f_n(u) = \sum_{k=1}^{m} \phi_k(u) y_k,$$

where

$$\phi_k(u) = \sum_{i=1}^{n} \lambda_i(u) a_k(z_i(Gu))$$

and that f_n may be implemented as a neural network.

This approximation is similar to that considered in the previous section, but is different in the detailed form of the weighting functionals and the requirement of the existence of a basis for the output space.

5.5.1 Remark

If F has an inner-product structure, then from boundedness of $G(B)$ and the algorithm given in Chapter 2, we can find a constant $c > 0$ such that $||Gu - f_n(u)||_F \leq cn^{-1/2}$ uniformly for $u \in B$.

5.6 Applications

Theorem 5 is a result to the effect that under mild conditions finite sums of the form

$$\sum_{i=1}^{m} \sum_{j=1}^{k(i)} c_{ij} u_{ij}[y_{ij}(\cdot)] v_i$$

can approximate the input-output maps of a certain kind of nonlinear dynamical systems taking a subset B of X, the linear space of $\mathbb{R}^n$-valued functions defined over $\mathbb{R}$, into E, the linear space of $\mathbb{R}$-valued functions defined over $\mathbb{R}$. Here the c_{ij} are real constants, the u_{ij} are certain continuous real-valued functions of the reals, the y_{ij} are continuous real functionals defined on X, and the v_i are elements of E. A similar result for the case of discrete domain systems can also be deduced.

5.6.1 Dynamic signal classification

Many authors have considered dynamic neural network classifiers of signals in the literature, e.g. [22]. In a typical classifier of this kind, the inputs are bounded $\mathbb{R}^d$-valued sequences of feature vectors and the outputs are real valued sequences which indicate how closely a segment of the past input matches a class, where classes are some predetermined nonoverlapping sets of input sequences with finite support. Usually such classifiers have exactly finite memory, and are time-invariant, but they have never been studied from the viewpoint of localizability or continuity of the associated functionals. Our result shows that a neural network of the type described above can approximate such a dynamic classifier arbitrarily well if it is w-localizable, or has continuous associated functionals.

5.6.2 Image processing

Shift-invariant adaptive filters [2] are often used in image processing applications. In this setting, D is usually taken to be $\mathbb{Z}^2$ (however, it could be $\mathbb{R}^2$), and assumption **A1** holds. The most commonly used such filters have exactly finite memory, but their localizability (or continuity of the associated functionals) has never been studied. It follows that they can be approximated arbitratily well by neural network structures if they are w-localizable or have continuous associated functionals.

5.7 Conclusion

In the settings of both discrete and continuous domains for inputs and outputs, we have proved results concerning the approximation of certain nonlinear dynamic systems using neural network architectures introduced here. We have characterized the time-varying systems for which our results hold, and demonstrated a large class of time-invariant systems which are of importance in engineering applications. The inputs are not restricted to be defined on finite intervals. Input and output functions with multidimensional domains and ranges, as well as L_p norms are treated. We have given two schemes for approximation of such maps, and have provided error analysis for one of the schemes using the algorithm described in Chapter 2.

Chapter 6

Robust Universal Neural Controllers

6.1 Introduction

Here we give a solution to the problem of approximating all members of a class of equicontinuous controllers arbitrarily well using a fixed basis function. Traditionally a common choice of the basis function has been the Gaussian, hence the translates of a basis function are usually called "centers", and the dilation is called the "width". We use a fixed finite set of centers and a fixed width to approximate any member of the given class. This property is important in practical applications because of the need for robustness with respect to plant variations or modeling uncertainties [33]. In the case of discrete-time continuous controllers that are time-invariant and posess finite memory, our problem reduces to the problem of simultaneously approximating a class of continuous functions defined on a finite-dimensional Euclidean space. This is the setting adopted in this work. The continuous functions to be approximated are called "target functions" or "target controllers", with a slight abuse of notation. We show that the problem can be solved by the so-called radial basis function neural controllers (RBFNCs).

Approximations with radial basis functions have been studied in the literature by many authors. See [36] for an early theoretical result, and [53] for a control application. However, in most cases only a single function (or a single controller) is approximated. Simultaneous approximation of a class of functions using radial basis functions was obtained by Chen and Chen [12]. In contrast to [12], we are not restricted to using an activation function that is radially symmetric with respect to the Euclidean norm. This feature allows the use of elliptic basis functions which are more suitable for certain applications in practice [37], [9]. Also, in our work the same width is used on all nodes, and hence the number of free parameters of the network is reduced. This results in a simplification of the parameter estimation problem. A minor distiction is that in our case the class of target functions need not be compact, relative compactness is sufficient. A particularly important difference is our emphasis on the application of our simultaneous approximation result to the engineering problem of discrete-time control.

In [53] adaptive control applications of radial basis function networks are considered. The principal difference between the present work and [53] is that we give a robust design of the width and centers (and a prescription for the weights) for a class of plants, as compared to an adaptive approach for a single plant. Also, the basis function in our case is not restricted to be Gaussian, nor is it required to be radial with respect to the Euclidean norm. An important difference is that our choice of centers for the radial basis function network is not restricted to a uniform grid. Hence our designs are not subject to an exponential growth in the number of nodes used in the network to achieve increasingly accurate approximations, a

phenomenon usually described as "the curse of dimensionality". In fact, our work in conjunction with the error bounds for radial basis function network approximation in [23] constitutes an efficient approximation of controllers in terms of the network size.

Our result also addresses the problem of retraining a network on a new target function (other than the original one used in initial training). In particular, if a sufficiently accurate RBFNC has been obtained with our method for one controller, then it can be used with the same width and centers to implement any other controller in the given equicontinuous class with the same error bound. The only parameters that need to be changed for the new controller are the weights from the hidden nodes to the output, but the problem of determining these weights is linear. Together, these results indicate that the RBFNCs are a powerful and efficient architecture — they are linearly retrainable universal controllers.

6.2 Simultaneous Approximation Theorem

For $A \subset \mathbb{R}^n$, let $\mathcal{C}(A)$ denote the linear space of all real-valued bounded continuous functions defined on A. Let $\phi \in \mathcal{C}(\mathbb{R}^n)$ be an integrable function such that $\int \phi(x)dx = 1$. For any $y > 0$, let C_y denote the cube $[-y, y]^n$.

For any $T > 0$, and a multiindex $\gamma = (m_1, \ldots, m_n) \in \mathbb{N}^n$, let $P = \{(I_\beta, z_\beta)\}_{\beta \leq \gamma}$ be a partition of C_T of size γ, with the n-intervals I_β. P is called a Perron partition [38] if the centers z_β are any points in I_β. The norm of the partition P, denoted by $|P|$, is given by

$$|P| = \max_{\beta \leq \gamma} \text{diam } I_\beta,$$

where the diameter of the n-interval I_β is given by

$$\text{diam } I_\beta = \sup_{x,y \in I_\beta} ||x - y||$$

(here $||\cdot||$ denotes the Euclidean norm on $\mathbb{R}^n$). The volume V_β of I_β is given by $V_\beta = \Pi_{i=1}^n V_{\beta_i}$, where V_{β_i} is the length of I_β along i^{th} axis.

Theorem 8 Let $K \subset \mathcal{C}(C_1)$ be relatively compact, and let $\epsilon > 0$ be given. For any $T > 1$ there exist $\sigma, \delta > 0$ and a certain continuous extension $\tilde{f}$ of f to $\mathbb{R}^n$ such that

$$\left| f(x) - \sum_{\beta \leq \gamma} w_\beta(f) \phi\left(\frac{x - z_\beta}{\sigma}\right) \right| < \epsilon, \quad x \in C_1, \ f \in K,$$

for any Perron partition $P = \{(I_\beta, z_\beta)\}_{\beta \leq \gamma}$ of C_T with $|P| < \delta$, where the weights w_β are given by $w_\beta(f) = \frac{1}{\sigma^n} V_\beta \tilde{f}(z_\beta)$.

Proof: Note that by the Ascoli-Arzelà theorem K is uniformly bounded, i.e., there exists an $M > 0$ such that

$$\sup_{x \in C_1} |f(x)| < M, \quad f \in K.$$

Extend every $f \in K$ by the construction given in (C.1) of Appendix C to $\tilde{f} \in \mathcal{C}(\mathbb{R}^n)$. Let $\tilde{K} \triangleq \{\tilde{f}\colon f \in K\}$. By the construction it is clear that

$$\sup_{x \in \mathbb{R}^n} |\tilde{f}(x)| < M, \quad \tilde{f} \in \tilde{K}. \tag{6.1}$$

Using the integrability of ϕ, pick $T_0 > 0$ such that

$$\int\limits_{\mathbb{R}^n \setminus C_{T_0}} |\phi(x)|dx < \frac{\epsilon}{8M}. \tag{6.2}$$

By its construction, $\tilde{K}$ is equicontinuous. Hence there exists $\delta > 0$ such that for any $x, y \in \mathbb{R}^n$, and arbitrary $\tilde{f} \in \tilde{K}$,

$$||x - y|| < \delta \;\Rightarrow\; |\tilde{f}(x) - \tilde{f}(y)| < \frac{\epsilon}{4||\phi||_1}, \tag{6.3}$$

where $||\phi||_1 = \int |\phi(x)|dx$. Choose $\sigma > 0$ such that

$$||\sigma x|| < \delta \quad \text{for all} \quad x \in C_{T_0}. \tag{6.4}$$

Define $\phi_\sigma \in \mathcal{C}(\mathbb{R}^n)$ by

$$\phi_\sigma(x) = \frac{1}{\sigma^n}\phi\left(\frac{x}{\sigma}\right), \quad x \in \mathbb{R}^n.$$

Let $\alpha \in C_1$. From a change of variables,

$$(\phi_\sigma * \tilde{f})(\alpha) = \int_{\mathbb{R}^n} \tilde{f}(\alpha - x)\phi_\sigma(x)dx = \int_{\mathbb{R}^n} \tilde{f}(\alpha - \sigma x)\phi(x)dx.$$

Hence, for arbitrary $\tilde{f} \in \tilde{K}$,

$$\begin{aligned}
&|(\phi_\sigma * \tilde{f})(\alpha) - \tilde{f}(\alpha)| \\
&\quad\leq \int\limits_{\mathbb{R}^n} |\tilde{f}(\alpha - \sigma x) - \tilde{f}(\alpha)| \cdot |\phi(x)|dx \\
&\quad\leq \int\limits_{C_{T_0}} |\tilde{f}(\alpha - \sigma x) - \tilde{f}(\alpha)| \cdot |\phi(x)|dx \\
&\qquad + \int\limits_{\mathbb{R}^n \setminus C_{T_0}} 2||\tilde{f}||_\infty \cdot |\phi(x)|dx \\
&\quad< \frac{\epsilon}{4||\phi||_1} \int\limits_{C_{T_0}} |\phi(x)|dx + 2M \int\limits_{\mathbb{R}^n \setminus C_{T_0}} |\phi(x)|dx \\
&\quad< \frac{\epsilon}{4} + 2M\,\frac{\epsilon}{8M} = \frac{\epsilon}{2}
\end{aligned} \tag{6.5}$$

where in the last two inequalities we have used (6.1), (6.3), (6.4), and (6.2).

Since $\tilde{f}(x) = 0$ for all $x \in \mathbb{R}^n \setminus C_T$,

$$(\phi_\sigma * \tilde{f})(\alpha) \quad = \quad \int_{C_T} \phi_\sigma(\alpha - x)\tilde{f}(x)dx.$$

Define $\psi_{f,\alpha}: C_T \to \mathbb{R}$ by

$$\psi_{f,\alpha}(x) = \phi_\sigma(\alpha - x)\tilde{f}(x).$$

so that

$$(\phi_\sigma * \tilde{f})(\alpha) \quad = \quad \int_{C_T} \psi_{f,\alpha}(x)dx \tag{6.6}$$

Let $\Psi \stackrel{\triangle}{=} \{\psi_{f,\alpha}: f \in K, \alpha \in C_1\}$. Since the set of pointwise products of functions belonging to relatively compact sets is itself relatively compact[1], and $\tilde{K}$ as well as $\{\phi_\sigma(\alpha - \cdot): \alpha \in C_1\}$ are relatively compact, it follows that Ψ is relatively compact.

By the lemma on uniform Riemann integrability in Appendix C.2, there exists a $\delta > 0$ such that for any Perron partition $P = \{(I_\beta, z_\beta)\}_{\beta \leq \gamma}$ of C_T with $|P| < \delta$,

$$\left| \int_{C_T} \psi_{f,\alpha}(x)dx - \sum_{\beta \leq \gamma} \psi_{f,\alpha}(z_\beta)V_\beta \right| < \frac{\epsilon}{2} \tag{6.7}$$

for any $\psi_{f,\alpha} \in \Psi$, and arbitrary centers $z_\beta \in I_\beta$.

From (6.7), (6.5) and (6.6), we have for arbitrary $\alpha \in C_1$,

$$\left| \tilde{f}(\alpha) - \sum_{\beta \leq \gamma} \phi_\sigma(\alpha - x)\tilde{f}(z_\beta)V_\beta \right| < \epsilon, \quad \tilde{f} \in \tilde{K}.$$

But $\tilde{f}(\alpha) = f(\alpha)$ for all $\alpha \in C_1$. Hence, by defining

$$w_\beta(f) = \frac{1}{\sigma^n} V_\beta \tilde{f}(z_\beta), \quad \beta \leq \gamma,$$

we have the desired result

$$\left| f(\alpha) - \sum_{\beta \leq \gamma} w_\beta(f)\phi\left(\frac{x - z_\beta}{\sigma}\right) \right| < \epsilon, \quad x \in C_1, \ f \in K.$$

■

6.3 Robust Retrainable Neural Controllers

The relevance of Theorem 8 to Robust Retrainable Neural Control is the following: if a sufficiently accurate RBFNC has been obtained with our method for one equicontinuous class of discrete-time continuous plants, then its width and centers can be used to construct a controller for any plant belonging to that class. The only parameters that need to be adapted during retraining for a new controller are the weights from the hidden nodes to the output. Note that this latter problem is

[1] Consider the set H of functions of the form $h(x) = a(x)b(x), x \in C_T$, with $a \in A$ and $b \in B$ where A and B are relatively compact. If M_A and M_B are uniform bounds for A and B, then it is obvious that $M_A M_B$ is a uniform bound for H. Also, if $\eta > 0$ is arbitrary, by the equicontinuity of B, there exists $\delta_1 > 0$ such that for any $x, y \in C_T$, and for arbitrary $b \in B$, $||x - y|| < \delta_1 \ \Rightarrow \ |b(x) - b(y)| < \frac{\eta}{2M_A}$. Similarly, by the equicontinuity of A, there exists $\delta_2 > 0$ such that for any $x, y \in C_T$, and $a \in A$, $||x - y|| < \delta_2 \ \Rightarrow \ |a(x) - a(y)| < \frac{\eta}{2M_B}$. Hence for any $a \in A, b \in B$, and $x, y \in C_T$ satisfying $||x - y|| < \delta_0 \stackrel{\triangle}{=} \min(\delta_1, \delta_2)$ we have $|h(x) - h(y)| = |a(x)b(x) - a(y)b(y)| = |\,[a(x) - a(y)]b(x) - a(y)[b(x) - b(y)]\,| < \frac{\eta}{2M_B}|b(x)| + |a(y)|\frac{\eta}{2M_A} \leq \eta$.

linear, and hence frequently involves tractable computations. Also note that the error bound for the new controller remains the same.

To state this result precisely, we need some preliminary material. Let $X = \ell_\infty$, the linear space of bounded real-valued sequences defined over $\mathbb{Z}_+$, with the norm $||x||_\infty = \sup_{t\in\mathbb{Z}_+} |x(t)|$. For a positive number γ, let X_γ denote a compact set uniformly bounded by γ. With $\alpha, \beta > 0$, we assume that the plants to be controlled can be viewed as input-output maps with inputs drawn from X_α, and outputs belonging to X_β. Let $\mathcal{P}$ denote the set of such maps that are continuous, i.e.,

$$\mathcal{P} = \{P\colon X_\alpha \to X_\beta \mid P \text{ is continuous}\}.$$

Let $\mathcal{C}$ denote the set of time-invariant continuous maps from X_β to X_α having exactly finite memory $n-1$ for a fixed $n \in \mathbb{N}$, i.e.,

$$x'(t_0 - t) = y'(t_0 - t), \text{ for all } t = 0, \ldots, n-1 \quad \Rightarrow \quad (Cx)(t_0) = (Cy)(t_0)$$

for all $t_0 \in \mathbb{Z}_+$, and all $C \in \mathcal{C}$, where x' is an extension of x to $\mathbb{Z}$ given by

$$x'(t) = \begin{cases} x(t), & t \in \mathbb{Z}_+ \\ 0 & \text{otherwise.} \end{cases}$$

Let the nominal plant $P' \in \mathcal{P}$ and the desired output y_d be fixed henceforth. Let the perturbed family of plants $\mathcal{P}_\Delta$ be any set of plants in $\mathcal{P}$ containing P' such that $\mathcal{P}_\Delta$ is equicontinuous, i.e., for any positive ϵ there exists a positive δ with the property that

$$x, y \in X_\alpha, \ ||x - y||_\infty < \delta \quad \Rightarrow \quad ||Px - Py||_\infty < \epsilon, \quad P \in \mathcal{P}_\Delta.$$

Such $\mathcal{P}_\Delta$ always exists; a trivial example is $\{P'\}$, since P' is continuous. More interesting examples are finite subsets of $\mathcal{P}$ containing P', since these are usually of importance in model matching problems [33].

A "design procedure" $\mathcal{D}$ is an algorithm that accepts a plant $P \in \mathcal{P}_\Delta$ and produces a controller[2] $C \in \mathcal{C}$ for the desired output $y_d \in X_\beta$. We will always use such a controller in the unity feedback configuration to track y_d (see Figure 6.1 where C_e denotes the error controller used in conventional design), so that the output of the controlled plant at time t, $(P(\mathcal{D}P)y)(t)$, is intended to be as close to $y_d(t)$ as possible for large t.

Define a norm on $\mathcal{P}$ by $||P|| = \sup_{x\in X_\alpha} ||Px||_\infty$ and similarly on $\mathcal{C}$. We say that $\mathcal{D}$ is a "stable design procedure" if it is a continuous map on $\mathcal{P}_\Delta$. If $\mathcal{D}$ has the additional property that the controller designed by it steers the plant to the desired output uniformly aymptotically, i.e., for any positive ϵ there exists an $m \in \mathbb{Z}_+$ (that may depend on ϵ but not on the plant P in the class $\mathcal{P}_\Delta$) such that

$$|(P(\mathcal{D}P)y)(t) - y_d(t)| < \epsilon, \quad t > m,$$

we say that $\mathcal{D}$ is a "stable tracking design procedure".

Define the maps $V_t\colon X \to \mathbb{R}^n$ by

$$V_t x = (x'(t), x'(t-1), \cdots, x'(t-n+1)), \quad t \in \mathbb{Z}_+.$$

[2] See Section 6.4 for a comment on bounded controllability of nonlinear systems.

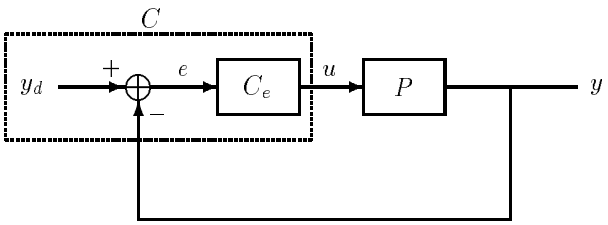


Figure 6.1: Unity feedback configuration for tracking.

Let $P_t : X \to X$ be defined by

$$(P_t x)(\tau) = \begin{cases} x(\tau), & \tau \leq t \\ 0 & \text{otherwise.} \end{cases}$$

We hereafter assume that

$$\text{(i)} \quad V_{n-1} X_\beta = C_\beta, \qquad \text{(ii)} \quad X_\beta \text{ is closed under } P_{n-1}. \tag{6.8}$$

Let $\mathcal{N}$ be an RBFNC generator that accepts a plant P in $\mathcal{P}_\Delta$, and returns a map of the form

$$y \mapsto \sum_{i=1}^{N} w_i(P) \phi \left(\frac{V_t y - z_i}{\sigma} \right), \quad y \in X_\beta,$$

where $N \in \mathbb{N}$, the z_i are real n-vectors, the $w_i(P)$ are real numbers, and σ is a positive constant. We say that the outputs of $\mathcal{N}$ are radial basis function neural controllers (RBFNCs), because they may be implemented by networks of processors computing weighted combinations of translates of the so-called radial basis function ϕ.

We are now ready to state and prove our result.

Theorem 9 Let $\mathcal{P}_\Delta$ be a given equicontinuous perturbed family of plants containing the nominal plant $P' \in \mathcal{P}$. Let a stable tracking design procedure $\mathcal{D}$ be given for $\mathcal{P}_\Delta$. Then there exists an RBFNC generator $\mathcal{N}$ with the following property: for any positive ϵ there exists an $m \in \mathbb{Z}_+$ for which

$$|(P(\mathcal{N}P)y)(t) - y_d(t)| < \epsilon, \quad t > m, \ P \in \mathcal{P}_\Delta.$$

Proof: Since $\mathcal{P}_\Delta$ is equicontinuous it is relatively compact by the generalization of the Ascoli-Arzelà theorem to continuous maps between compact metric spaces [29, Problem 5, p. 107]. Since the design procedure $\mathcal{D}$ is a continuous map, the set of controllers produced by $\mathcal{D}$ for $\mathcal{P}_\Delta$,

$$\mathcal{C}_\Delta = \{\mathcal{D}P \mid P \in \mathcal{P}_\Delta\},$$

is also relatively compact. By the generalization of the Ascoli-Arzelà theorem cited above, $\mathcal{C}_\Delta$ is equicontinuous. Since any controller $C \in \mathcal{C}_\Delta$ is time-invariant and has exactly finite memory $n-1$, it determines the function $f_C : C_\beta \to \mathbb{R}$ given by

$$f_C(z) = (Cx)(t), \quad z \in C_\beta,$$

where $x \in X_\beta$ and $t \in \mathbb{Z}_+$ are chosen such that $z = V_t x$. Let $\mathcal{F}_\Delta$ be the set of such functions f_C as C varies over the designed controllers for the perturbed plants, i.e.,

$$\mathcal{F}_\Delta = \{f_C \mid C \in \mathcal{C}_\Delta\}.$$

Since $\mathcal{C}_\Delta$ is equicontinuous, for any positive η there exists a positive ζ with the property that

$$x, y \in X_\beta, \ ||x-y||_\infty < \zeta \quad \Rightarrow \quad ||Cx - Cy||_\infty < \eta, \quad C \in \mathcal{C}_\Delta.$$

Let $z_1, z_2 \in C_\beta$ be any points such that $||z_1 - z_2|| < \zeta$. Using (6.8) choose $x_1, x_2 \in X_\beta$ such that $V_{n-1}x_i = z_i$ and $x_i(t) = 0$, $t \geq n$ for $i = 1, 2$. It is obvious that $||x_1 - x_2||_\infty < \zeta$, hence $|(Cx_1)(n-1) - (Cx_2)(n-1)| < \eta$, from which it follows that $|f_C(z_1) - f_C(z_2)| < \eta$ for any $C \in \mathcal{C}_\Delta$. Hence $\mathcal{F}_\Delta$ is an equicontinuous subset of $\mathcal{C}(C_\beta)$. Since the outputs of any $C \in \mathcal{C}_\Delta$ belong to X_α, $\mathcal{F}_\Delta$ is obviously uniformly bounded.

Let any positive ϵ be given. Since $\mathcal{P}_\Delta$ is equicontinuous, there exists a positive δ such that

$$x, y \in X_\alpha, \ ||x-y||_\infty < \delta \quad \Rightarrow \quad ||Px - Py||_\infty < \epsilon/2, \quad P \in \mathcal{P}_\Delta. \tag{6.9}$$

By Theorem 8, there exists a radial basis function network such that

$$\left| f_C(z) - \sum_{i=1}^{N} w_i(f_C)\phi\left(\frac{z - z_i}{\sigma}\right) \right| < \delta, \quad z \in C_\beta, \ C \in \mathcal{C}_\Delta.$$

Setting $z = V_t y$ we get $\mathcal{N}P$ such that

$$|((\mathcal{N}P)y)(t) - ((\mathcal{D}P)y)(t)| < \delta, \quad t \in \mathbb{Z}_+, \ y \in X_\beta, \ P \in \mathcal{P}_\Delta. \tag{6.10}$$

Since $\mathcal{D}$ is a stable tracking design procedure, there exists an $m \in \mathbb{Z}_+$ such that

$$|(P(\mathcal{D}P)y)(t) - y_d(t)| < \epsilon/2, \quad t > m, \ P \in \mathcal{P}_\Delta. \tag{6.11}$$

Combining (6.11), (6.9) and (6.10) we have the desired result. ∎

6.4 Comments

First we provide motivation for our definition of a stable tracking design procedure. In the case of linear time-invariant plants having finite-dimensional state space descriptions, there is extensive literature on discrete-time control design. The most commonly used design procedure in this case is stabilization with PID controllers, which is stable in our sense because it relies only on the technique of pole placement which is inherently robust [58]. Also, since we employ only unity gain

output feedback, the traditional methods of solving tracking problems with full or reduced-order observers can be used as stable tracking design procedures.

Extension to the case of continuous-time systems is straightforward, with the corresponding change of the design procedure to generate fixed order controllers. Again, in the case of linear time-invariant systems there is a plethora of results (in particular see [8, Chapter 7] for design of robust aymptotic tracking for a *class* of inputs; see also [32]) in the literature providing design procedures which are stable in our sense because of the robustness of pole placement designs. In some cases approximate pole placement can even be achieved with only first order controllers using output feedback [30].

If the plants are nonlinear, then the theory of control design is not so simple. However, under certain mild assumptions a design procedure may be formulated. In this regard, see [54] where sufficient conditions are given for the existence of piecewise linear controllers which regulate the nonlinear plant in a fixed number of time steps. Also, in [55] it is shown that bounded control inputs may be used to control the plant under certain conditions on linear or nonlinear plants in continuous or discrete-time. In [59] stabilizability of a certain nonlinear time-varying plant with output feedback is considered, and in [60] a similar problem is solved for the case in which the nonlinear system has a right coprime factorization. A model reference adaptive control scheme was suggested in [34].

For the case of the perturbations to the nominal system arising out of parametric uncertainties, there are many recent results in the literature about the design and analysis of robust controllers. For example, a nice parametrization of robust stabilizing controllers is given in [40]. Continuous (indeed, analytic or rational or polynomial) dependence of controllers on plant parameters is shown to be possible in [19], [56] under the assumption of constant McMillan degree of the plants. In [10] compact sets of linear time-invariant single-input single-output systems are studied where the numerator and the denominator polynomials of the system transfer function have uncertain coefficients belonging to a box around the nominal plant in the coefficient space. In this connection see also the pioneering work of Kharitonov [28] where Hurwitz stability of such interval polynomials was shown to be equivalent to that of just four extreme polynomials. This work gave a strong impetus to research in robust stability of interval systems (see [5] and references therein).

In summary, results in the literature indicate that our assumption of the existence of a design procedure is justified in a variety of settings. The stability and tracking properties of such design procedures are topics of future research.

In our result stated above, we assume that there exists an $m \in \mathbb{Z}_+$ independent of the plant $P \in \mathcal{P}_\Delta$ for asymptotic tracking. This requirement may be relaxed if it poses a strong restriction on the design procedure. For such cases we can obtain another result in which asymptotic tracking is still guaranteed with the RBFNCs, but the length of the transient period may vary with the perturbation.

6.5 Conclusion

Our results provide an analytical basis for separating the problem of parameter determination of the RBFNCs into a nonlinear part involving the determination of the width and the centers, and a linear part concerning the estimation of the output weights. In particular, our results show that the computationally expensive nonlinear part needs to be solved only once for the given class of equicontinuous controllers. The remaining part (which occurs frequently due to retraining) usually involves relatively easy computations, because the problem reduces to linear estimation of parameters.

Chapter 7

Error Bounds for Function Approximation

Here we prove results concerning the error bound in the mean-square sense of functionals defined on finite-dimensional Hilbert spaces by neural network structures. The functionals belong to a certain general class introduced here involving a condition on the integrability of the Fourier transform. The bound on the error of approximation depends on the size of the network but not on the dimension of the underlying space. It is valid for all functionals in the class mentioned above. This class for which the theorems are applicable is shown to include some important interesting sets of functionals, e.g., the bandlimited signals on $\mathbb{R}^n$ and the Schwartz class.

7.1 Introduction

Most of the current literature on neural networks deals with approximation of functions defined on finite-dimensional spaces. Even though the existence of neural network structures for arbitrarily good approximations of functionals on compact subsets of these spaces has been known for some time [36, 26, 18], not much is known about the computational efficiency for given classes of approximation structures. Hence it is important to consider approximation structures (of the neural network type or otherwise) for which error bounds for significant classes of functionals are known.

7.2 Notation

This work is concerned with error bounds for approximation of a certain kind of functional defined on a finite-dimensional Euclidean space. Let $\mathbb{K}$ be $\mathbb{R}$ or $\mathbb{C}$, $d \in \mathbb{N}$, and let $\mathcal{H}$ be the finite-dimensional Euclidean space $\mathbb{K}^d$ over $\mathbb{K}$. The Lebesgue measure on $\mathcal{H}$ is denoted by m_d. The approximation error is specified in the $L_2(\mathcal{H}, \mu)$ norm, where μ is a finite positive Borel measure on $\mathcal{H}$ such that $\mu(\mathcal{H}) = 1$ (e.g., μ defined by

$$\mu(S) = m_d(S \cap B_r)/m_d(B_r), \quad S \in \mathcal{B}(\mathcal{H}) \tag{7.1}$$

on the σ-algebra of Borel sets $\mathcal{B}(\mathcal{H})$, where B_r is a ball of radius r centered at the origin). We also assume that μ satisfies the following conditions.

Condition 1 μ is absolutely continuous with respect to m_d.

Condition 2 The Radon-Nikodym derivative $[d\mu/dm_d]$ for $\mathcal{H}$ is bounded.[1]

Note that important practical choices of μ, e.g., the example given above in (7.1) satisfy these conditions. However, we rule out atomic measures by Condition 1.

As usual for a positive measure ν on $\mathcal{H}$,

$$L_p(\mathcal{H},\nu) \triangleq \{f:\mathcal{H}\to\mathbb{K} \mid \int |f(x)|^p \nu(dx) < \infty\}, \quad 1 \le p < \infty$$

and then

$$||f||_{p,\nu} \triangleq \left(\int |f(x)|^p \nu(dx)\right)^{1/p}.$$

If it is clear from the context, we leave out the measure with respect to which the norm is defined and write simply $||f||_p$. $\mathcal{F}f$ denotes the Fourier transform of a function $f \in L_1(\mathcal{H}, m_d)$.

For approximation of complex-valued functionals we use the same architecture as in [49], where nodes belonging to the single hidden layer implement exponentiation and the input nodes compute the inner product of $\mathcal{H}$ as shown in Figure 7.1. This architecture is remarkable in that it can be used for approximation both in the uniform norm for continuous functions on compact subsets [49] and in the L_2 norm for functions defined on the whole space. The class of such networks (with m hidden nodes) is denoted by $\mathsf{C}(m)$ defined below.

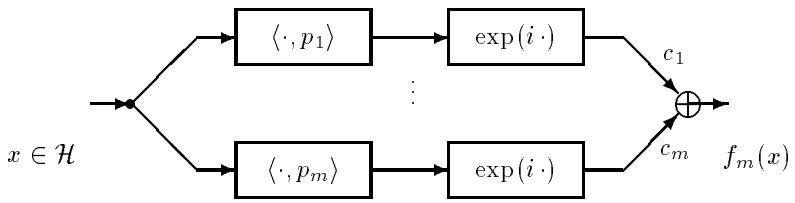


Figure 7.1: Architecture of the complex Hilbert space network.

Definition: A network $\mathcal{N}$ in the class $\mathsf{C}(m)$ implements a weighted finite sum of exponentials composed with inner products with fixed elements of $\mathcal{H}$, i.e., the output of $\mathcal{N}$ is a functional f_m given by

$$f_m(x) = \sum_{k=1}^{m} c_k \exp(i\,\langle x, p_k\rangle) \tag{7.2}$$

where the network parameters c_k belong to $\mathbb{C}$ and the network parameters p_k are drawn from $\mathcal{H}$. □

For approximation of real-valued functionals we adopt the architecture defined below and shown in Figure 7.2.

[1] $[d\mu/dm_d]$ exists because Condition 1 is satisfied. Moreover, it is real-valued and integrable [21, p. 134, Theorem 5.5.4].

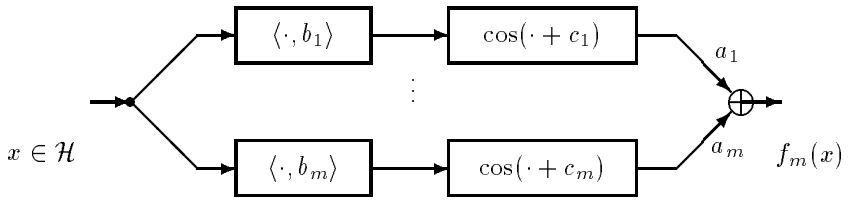


Figure 7.2: Architecture of the real Hilbert space network.

Definition: A network $\mathcal{N}$ in the class $\mathsf{R}(m)$ implements a weighted finite sum of cosines composed with inner products with fixed elements of $\mathcal{H}$, i.e., the output of $\mathcal{N}$ is a functional f_m given by

$$f_m(x) = \sum_{k=1}^{m} a_k \cos(\langle x, b_k \rangle + c_k) \tag{7.3}$$

where the network parameters a_k, c_k belong to $\mathbb{R}$ and the network parameters b_k are drawn from $\mathcal{H}$. □

7.3 Relation to previous work

It may be noted that previous results regarding approximation (in the uniform or the L_2 norm) of functionals on finite-dimensional spaces [36, 26, 18] are existence theorems. In practice it is useful to have some idea of how many terms are sufficient for a given accuracy of approximation. In particular it is important to ensure that the error of approximation decreases with increasing network size with some acceptable upper bound.

Sandberg [49] proved existence theorems regarding approximation of nonlinear functionals with the particular type of neural network architecture introduced in that paper. The relation between the applications of the Stone-Weirstrass Theorem [42, p. 122] and the Generalized Wiener Theorem [42, p. 228] in terms of the norms used in evaluating the error of approximation and the domains of definition of the functionals is analogous to that between this work and [49]. These analogies are tabulated for easy reference in Table 7.1.

	Stone–Weirstrass	Generalized Wiener
Norm	L_∞	L_1
Domain	Compacta in $\mathbb{R}^n$	All of $\mathbb{R}^n$
	Sandberg [49]	**Present Work**
Norm	L_∞	L_2
Domain	Compacta in $\mathcal{H}$	All of $\mathcal{H}$

Table 7.1: An analogy between analogies.

Barron [3] considered error bounds for the class of neural networks that employ the usual "sigmoidal" functions in their hidden nodes and established error bounds of $O(1/\sqrt{m})$ for L_2 approximation, where m is the number of hidden nodes. He used a lemma due to Jones [27] in an essential way along with a certain integrability condition on the Fourier transform of the functional.

Girosi and Anzellotti [23] mainly studied rates of convergence of approximation with translates of a function in L_2 and the Sobolev space, with emphasis on the Gaussian and the Bessel-MacDonalds kernels.

The present work was inspired by Barron's approach; here also an integrability condition on functionals on a Hilbert space $\mathcal{H}$ is used. But we do not use Jones' lemma. Instead we develop and use a variant of Maurey's lemma [39]. Our Lemma 5 is not just a special case of Maurey's lemma because we prove more than the existence of a convex combination; the combination is a special (equally weighted, i.e., arithmetic) average. Hence, the computational load is reduced in applying our lemma because only one combination of the chosen m functionals needs to be tested for the desired error bound (instead of finding a suitable combination from the set of all convex combinations). Also, our proof is more straightforward and does not use any probabilistic terminology. It may be noted that the architecture considered in [3] is different from that used in the present work (pictured in Figure 7.1 for the case of complex-valued functionals and Figure 7.2 for the case of real-valued functionals).

7.4 Error Bounds: Preliminaries

7.4.1 The classes Π and Δ_c

We consider functionals which belong to the class Δ_c defined below.

Definition: The class of permissible functionals is denoted by

$$\Pi \stackrel{\triangle}{=} \{f: \mathcal{H} \to \mathbb{K} | \quad f \neq 0; \quad f, \mathcal{F}f \in L^1(\mathcal{H}, m_d) \bigcap L^2(\mathcal{H}, m_d)\}$$

where the condition $f \neq 0$ is taken in the m_d-almost everywhere sense. □

Definition: The class of all c-integrable functionals is denoted by

$$\Delta_c \stackrel{\triangle}{=} \{g \in \Pi | \quad c_g \stackrel{\triangle}{=} \frac{1}{(\sqrt{2\pi})^d} \int |(\mathcal{F}g)(\omega)| m_d(d\omega) \leq c\}$$

for any $c > 0$. □

The most commonly assumed condition to guarantee the existence of an integral representation of the type crucially used in [3] is that f is integrable (in the Lebesgue sense). In this case it is easy to see that the integrability condition of [3] implies that of ours:

$$\begin{aligned} \int |\omega| \ |(\mathcal{F}f)(\omega)| m_d(d\omega) \quad &\leq \quad c \\ &\Rightarrow \quad \int_{\omega \notin [-1,1]^d} |\omega| \ |(\mathcal{F}f)(\omega)| m_d(d\omega) \leq c \end{aligned}$$

$$\Rightarrow \quad \int_{\omega \notin [-1,1]^d} |(\mathcal{F}f)(\omega)| m_d(d\omega) \leq c$$
$$\Rightarrow \quad \int |(\mathcal{F}f)(\omega)| m_d(d\omega) \leq 2^d F_{\max} + c.$$

Here the Fourier transform of the integrable function f is bounded by some constant $F_{\max}$.

However, the following example shows that there are functions f which satisfy our integrability condition but not the condition in [3]. Consider $f: \mathbb{R} \to \mathbb{R}$ given by

$$f(t) = \exp(-|t|).$$

In this case

$$(\mathcal{F}f)(\omega) = 2/(1+\omega^2),$$

and it is easy to verify that $c_f = \sqrt{2\pi}$ but the function $\omega \mapsto |\omega \cdot (\mathcal{F}f)(\omega)|$ is not integrable.

7.4.2 A consequence of c-integrability

The c-integrability condition on the Fourier transform implies the boundedness of the equivalent continuous functional $\tilde{f}$, since

$$\begin{aligned} |\tilde{f}(x)| &= \frac{1}{(\sqrt{2\pi})^d} \left| \int \exp(i\langle \omega, x \rangle)(\mathcal{F}\tilde{f})(\omega) m_d(d\omega) \right| \\ &\leq \frac{1}{(\sqrt{2\pi})^d} \int |(\mathcal{F}\tilde{f})(\omega)| m_d(d\omega) \qquad \forall x \in \mathcal{H}, \end{aligned}$$

Hence

$$\begin{aligned} \sup_x |\tilde{f}(x)| &\leq \frac{1}{(\sqrt{2\pi})^d} \int |(\mathcal{F}\tilde{f})(\omega)| m_d(d\omega) \\ &\triangleq \frac{1}{(\sqrt{2\pi})^d} ||\mathcal{F}\tilde{f}||_1 \triangleq c_{\tilde{f}} \leq c. \end{aligned} \tag{7.4}$$

7.5 Approximation of complex-valued functionals

In this section let $\mathbb{K} = \mathbb{C}$ in which case $\mathcal{H} = \mathbb{C}^d$. The main result of this work is the following theorem which says that with $2m$ network parameters we are guaranteed an L_2 (of measure μ) error bound of $O(1/\sqrt{m})$.

Theorem 10 Let $f: \mathcal{H} \to \mathbb{C}$ be a functional such that $f \in \Delta_c$ for some $c > ||f||_{2,\mu}$. Then, given $m \in \mathbb{N}$ there exists a network $\mathcal{N} \in \mathsf{C}(m)$ implementing f_m such that

$$\int_{\mathcal{H}} |f(x) - f_m(x)|^2 \mu(dx) \leq \frac{c^2 - ||f||_{2,\mu}^2}{m}.$$

First we prove a variant of Maurey's lemma [39] which guarantees the existence of a convex combination (in our case an equally weighted, i.e., arithmetic average) of functionals belonging to a certain subset of $L_2(\mathcal{H}, \mu)$ which satisfies the

error bound claimed in the theorem. We note that this lemma may be proved for any inner-product space (even in the infinite-dimensional case) because its proof does not depend on the dimensionality of $\mathcal{H}$ or even its completeness.

Lemma 5 Let $\mathcal{H}$ be an inner product space and $f: \mathcal{H} \to \mathbb{C}$ be a functional such that

$$f(x) = \int_{\mathcal{H}} d(x, \omega) \Lambda(d\omega) \tag{7.5}$$

where $\Lambda(d\omega)$ is a positive measure such that $\Lambda(\mathcal{H}) = 1$ and $d(\cdot, \cdot)$ belongs to a subset $\mathcal{D}$ of bounded functionals indexed by $\omega \in \mathcal{H}$, i.e.,

$$\mathcal{D} \subset \{d: \mathcal{H} \times \mathcal{H} \to \mathbb{C} | \quad |d(x, \omega)| \leq c \quad \forall x, \omega \in \mathcal{H}\} \tag{7.6}$$

such that $d(x, \cdot)$ is measurable with respect to the σ-algebra on which Λ is defined and $d(\cdot, \omega)$ is measurable with respect to the σ-algebra on which μ is defined. Then, given $m \in \mathbb{N}$ there exists an arithmetic average f_m of m functionals in $\mathcal{D}$ satisfying the inequality

$$\int_{\mathcal{H}} |f(x) - f_m(x)|^2 \mu(dx) \leq \frac{c^2 - ||f||_{2,\mu}^2}{m}.$$

Proof: Consider the arithmetic average

$$\frac{1}{m} \sum_{i=1}^{m} d(x, \omega_i)$$

of m functionals from $\mathcal{D}$ for any $m \in \mathbb{N}$. Let e_m denote

$$\int_{\mathcal{H}^m} \int_{\mathcal{H}} \left| f(x) - \frac{1}{m} \sum_{i=1}^{m} d(x, \omega_i) \right|^2 \mu(dx) \Lambda^m(d\omega_1 \cdots d\omega_m) \tag{7.7}$$

where $\Lambda^m(d\omega_1 \cdots d\omega_m)$ is the product measure. Observe that e_m is an *averaged* L_2 (of measure μ) error of arithmetic averages of m functionals from $\mathcal{D}$ as an approximant for f.

Since f and $d(x, \omega_i)$ are Borel measurable and the integrand is non-negative we may apply Fubini's theorem for the interchange of the order of integration, obtaining

$$e_m = \int_{\mathcal{H}} \int_{\mathcal{H}^m} \left| f(x) - \frac{1}{m} \sum_{i=1}^{m} d(x, \omega_i) \right|^2 \Lambda^m(d\omega_1 \cdots d\omega_m) \mu(dx).$$

By simple manipulations of the integrand,

$$\begin{aligned} e_m &= \int_{\mathcal{H}} \frac{1}{m^2} \int_{\mathcal{H}^m} \left| \sum_{i=1}^{m} d(x, \omega_i) - m f(x) \right|^2 \Lambda^m(d\omega_1 \cdots d\omega_m) \mu(dx) \\ &= \int_{\mathcal{H}} \frac{1}{m^2} \int_{\mathcal{H}^m} \left| \sum_{i=1}^{m} [d(x, \omega_i) - f(x)] \right|^2 \Lambda^m(d\omega_1 \cdots d\omega_m) \mu(dx) \end{aligned}$$

$$
\begin{aligned}
&= \int_{\mathcal{H}} \frac{1}{m^2} \int_{\mathcal{H}^m} \left\{ \left(\Re \sum_{i=1}^{m} [d(x,\omega_i) - f(x)] \right)^2 \right. \\
&\qquad \left. + \left(\Im \sum_{i=1}^{m} [d(x,\omega_i) - f(x)] \right)^2 \right\} \Lambda^m(d\omega_1 \cdots d\omega_m)\mu(dx) \\
&= e_m^{\Re} + e_m^{\Im}. \qquad (7.8)
\end{aligned}
$$

Considering the real part first,

$$
\begin{aligned}
e_m^{\Re} &= \int_{\mathcal{H}} \frac{1}{m^2} \int_{\mathcal{H}^m} \left(\sum_{i=1}^{m} [\Re d(x,\omega_i) - \Re f(x)] \right)^2 \Lambda^m(d\omega_1 \cdots d\omega_m)\mu(dx) \\
&= \int_{\mathcal{H}} \frac{1}{m^2} \int_{\mathcal{H}^m} \left(\sum_{i=1}^{m} (\Re d(x,\omega_i) - \Re f(x))^2 \right. \\
&\qquad \left. + \sum_{i=1}^{m} \sum_{j \neq i}^{m} (\Re d(x,\omega_i) - \Re f(x))(\Re d(x,\omega_j) - \Re f(x)) \right) \\
&\qquad \Lambda^m(d\omega_1 \cdots d\omega_m)\mu(dx).
\end{aligned}
$$

But

$$
\int_{\mathcal{H}} (\Re d(x,\omega_i) - \Re f(x))\Lambda(d\omega_i) = 0
$$

by (7.5). Hence the cross-terms in $e_m^{\Re}$ vanish after application of Fubini's Theorem, and we are left with

$$
\begin{aligned}
e_m^{\Re} &= \int_{\mathcal{H}} \frac{1}{m^2} \sum_{i=1}^{m} \int_{\mathcal{H}} (\Re d(x,\omega_i) - \Re f(x))^2 \Lambda(d\omega_i)\mu(dx) \\
&= \int_{\mathcal{H}} \frac{1}{m^2} \sum_{i=1}^{m} \int_{\mathcal{H}} (\Re^2 d(x,\omega_i) - 2\Re d(x,\omega_i)\Re f(x) + \Re^2 f(x))\Lambda(d\omega_i)\mu(dx)
\end{aligned}
$$

But

$$
\int_{\mathcal{H}} \int_{\mathcal{H}} \Re d(x,\omega_i)\Re f(x)\Lambda(d\omega_i)\mu(dx) = \int_{\mathcal{H}} \Re^2 f(x)\mu(dx)
$$

by (7.5). Hence

$$
e_m^{\Re} = \int_{\mathcal{H}} \frac{1}{m^2} \sum_{i=1}^{m} \int_{\mathcal{H}} (\Re^2 d(x,\omega_i) - \Re^2 f(x))\Lambda(d\omega_i)\mu(dx).
$$

Similarly

$$
e_m^{\Im} = \int_{\mathcal{H}} \frac{1}{m^2} \sum_{i=1}^{m} \int_{\mathcal{H}} (\Im^2 d(x,\omega_i) - \Im^2 f(x))\Lambda(d\omega_i)\mu(dx).
$$

Hence

$$\begin{aligned} e_m &= \int_{\mathcal{H}} \frac{1}{m^2} \sum_{i=1}^{m} \int_{\mathcal{H}} (|d(x,\omega_i)|^2 - |f(x)|^2) \Lambda(d\omega_i) \mu(dx) \\ &= \frac{1}{m} \int_{\mathcal{H}} \int_{\mathcal{H}} |d(x,\omega)|^2 \Lambda(d\omega) \mu(dx) - \frac{1}{m} \int_{\mathcal{H}} \int_{\mathcal{H}} |f(x)|^2 \Lambda(d\omega) \mu(dx) \\ &\leq \frac{1}{m} \left\{ c^2 - ||f||_{2,\mu}{}^2 \right\}. \end{aligned} \tag{7.9}$$

The last inequality results from the defining property of $\mathcal{D}$ and the membership of d in it.

Since the *averaged* L_2 (of measure μ) error of arithmetic averages of m functionals in $\mathcal{D}$ satisfies the inequality (7.9), there must be an arithmetic average of m functionals in $\mathcal{D}$ that satisfies it. ■

Now we prove our theorem.

Proof: Since $f \in \Pi$, we have $\mathcal{F}f \in L_1$, and by the Fourier inversion theorem

$$f(x) = \frac{1}{(\sqrt{2\pi})^d} \int_{\mathcal{H}} \exp(i\langle \omega, x \rangle) (\mathcal{F}f)(\omega) m_d(d\omega). \tag{7.10}$$

Using the polar representation of complex numbers

$$(\mathcal{F}f)(\omega) = |(\mathcal{F}f)(\omega)| \exp(i\theta(\omega))$$

we get

$$f(x) = \int_{\mathcal{H}} d(x,\omega) \Lambda(d\omega) \tag{7.11}$$

where

$$d(x,\omega) = c_f \exp(i[\langle \omega, x \rangle + \theta(\omega)]), \tag{7.12}$$

and

$$\Lambda(d\omega) = \frac{|(\mathcal{F}f)(\omega)| m_d(d\omega)}{(\sqrt{2\pi})^d c_f} \tag{7.13}$$

is a positive finite measure such that $\Lambda(\mathcal{H}) = 1$.

Note that $d(x,\omega)$ is bounded by c in absolute value, and let

$$E_c \stackrel{\triangle}{=} \{\gamma \exp(i\langle \omega, x \rangle) |\ \gamma \in \mathbb{C},\ |\gamma| \leq c\}. \tag{7.14}$$

Hence by Lemma 5 we get the desired result that given $m \in \mathbb{N}$ there exists $f_m(x)$, an arithmetic average of m points of E_c such that

$$\int_{\mathcal{H}} |f(x) - f_m(x)|^2 \mu(dx) \leq \frac{c^2 - ||f||_{2,\mu}^2}{m}.$$

But such a convex combination is implemented by $\mathcal{N} \in \mathsf{C}(m)$. ■

7.6 Approximation of real-valued functionals

In this section let $\mathbb{K} = \mathbb{R}$ in which case $\mathcal{H} = \mathbb{R}^d$. The architecture shown in Figure 7.2 is considered.

Theorem 11 Let $f: \mathcal{H} \to \mathbb{R}$ be a functional such that $f \in \Delta_c$ for some $c > ||f||_{2,\mu}$. Then, given $m \in \mathbb{N}$ there exists a network $\mathcal{N} \in \mathsf{R}(m)$ implementing f_m such that

$$\int_{\mathcal{H}} |f(x) - f_m(x)|^2 \mu(dx) \leq \frac{c^2 - ||f||_{2,\mu}^2}{m}.$$

Proof: Using the Fourier inversion theorem as in the proof of the previous theorem, and noting that f is real-valued,

$$f(x) = \Re \left(\frac{1}{(\sqrt{2\pi})^d} \int_{\mathcal{H}} \exp(i\langle \omega, x \rangle)(\mathcal{F}f)(\omega) m_d(d\omega) \right).$$

By an argument similar to that in the proof of the Theorem 10

$$f(x) = \int_{\mathcal{H}} d(x, \omega) \Lambda(d\omega)$$

where

$$d(x, \omega) = c_f \cos(\langle \omega, x \rangle + \theta(\omega)),$$

and Λ is as given in (7.13) with the property that it is a positive measure satisfying $\Lambda(\mathcal{H}) = 1$.

Note that $d(x, \omega)$ is bounded by c in absolute value, and let

$$C_c \triangleq \{\gamma \cos(\langle \omega, x \rangle + \alpha) | \ \gamma \in \mathbb{R}, \ |\gamma| \leq c, \ \alpha \in \mathbb{R}\}.$$

Applying Lemma 5 we get the desired result, since in this case

$$\Im[f] = 0 = \Im[d(x, \omega)].$$

■

7.7 Applications

7.7.1 Bandlimited Functions

Definition: With $\Omega > 0$, a continuous function $f \in L_2(\mathbb{R}^d, m_d)$ is said to be bandlimited to $[-\Omega, \Omega]^d$ if its L_2 Fourier transform vanishes outside of $[-\Omega, \Omega]^d$. □

Theorem 12 Let $f \neq 0$ be bandlimited to $[-\Omega, \Omega]^d$. Then $f \in \Delta_r$ for any

$$r \geq \left(\sqrt{\frac{\Omega}{\pi}} \right)^d ||f||_2.$$

Proof: Note that the L_2 Fourier transform of a bandlimited function is measurable and has compact support, hence is integrable. Therefore $0 < c_f \stackrel{\triangle}{=} \frac{1}{(\sqrt{2\pi})^d} \int |(\mathcal{F}f)(\omega)| m_d(d\omega) < \infty$. So $f \in \Delta_c$ for any $c \geq c_f$. The upper bound on c_f stated in the theorem can be obtained as follows:

$$\begin{aligned} c_f &\stackrel{\triangle}{=} \frac{1}{(\sqrt{2\pi})^d} \int\limits_{[-\Omega,\Omega]^d} |(\mathcal{F}f)(\omega)| m_d(d\omega) \\ &= \frac{1}{(\sqrt{2\pi})^d} \int\limits_{[-\Omega,\Omega]^d} |1 \cdot (\mathcal{F}f)(\omega)| m_d(d\omega) \\ &\leq \frac{1}{(\sqrt{2\pi})^d} \left[\int\limits_{[-\Omega,\Omega]^d} 1^2 m_d(d\omega) \right]^{1/2} \left[\int\limits_{[-\Omega,\Omega]^d} |(\mathcal{F}f)(\omega)|^2 m_d(d\omega) \right]^{1/2} \\ &= \frac{1}{(\sqrt{2\pi})^d} (2\Omega)^{d/2} \, ||f||_2 \end{aligned} \tag{7.15}$$

where the inequality is an application of the Schwarz inequality, and the last equality follows from the Parseval identity. ■

7.7.2 Schwartz class

Definition: A function $f: \mathbb{R}^d \to \mathbb{R}$ is said to be rapidly decreasing [43, 7.3] if it has partial derivatives of all orders and

$$\sup_{x \in \mathbb{R}^d} (1 + |x|^2)^N |(D^\alpha f)(x)| < \infty$$

for $|\alpha| \leq N$; $N = 0, 1, 2, \cdots$. Here $\alpha \in \mathbb{Z}_+^d$ is a multi-index, $|x|$ is the Euclidean norm on $\mathbb{R}^d$, and

$$D^\alpha = \left(\frac{\partial}{\partial x_1} \right)^{\alpha_1} \cdots \left(\frac{\partial}{\partial x_d} \right)^{\alpha_d}.$$

The class of such functions is called the Schwartz class and is denoted by $\mathcal{S}_d$. □

The Schwartz class has some important properties:

1. $\mathcal{S}_d$ is closed under the operations of addition, multiplication, convolution, multiplication by any power of an independent variable, and differentiation.
2. $\mathcal{S}_d \subset L_p$, $1 \leq p \leq \infty$.
3. $\mathcal{S}_d$ is dense in L_1 and L_2.
4. The elements of $\mathcal{S}_d$ are uniformly and absolutely continuous.
5. Fourier transformation is a homeomorphism on $\mathcal{S}_d$ [43, Theorems 7.7, 7.4].

Theorem 13 Let $f \in \mathcal{S}_d$ and $f \neq 0$. Then $f \in \Delta_c$ for any $c \geq \frac{1}{(\sqrt{2\pi})^d} ||\mathcal{F}f||_1$.

Proof: Since the L_2 Fourier transform of a function in the Schwartz class is another function in the Schwartz class, it is integrable; hence $0 < c_f \triangleq \frac{1}{(\sqrt{2\pi})^d}||\mathcal{F}f||_1 < \infty$. So $f \in \Delta_c$ for any $c \geq c_f$. That the particular "best" c is as stated in the theorem can be seen from the definition of Δ_c. ■

7.8 Approximation of functionals in L_2

Theorem 14 Let $f \in L_2(\mathbb{R}^d, m_d)$ and let $\epsilon > 0$ be given. Then there exists a $c > ||f||_{2,\mu}$ such that f can be approximated by a network $\mathcal{N} \in \mathsf{R}(m)$ with an error

$$||f - f_m||_{2,\mu} \leq \epsilon + \sqrt{\frac{c^2 - ||f||_{2,\mu}^2}{m}}$$

for all $m \in \mathbb{N}$.

Proof: Given $f \in L_2$ and $\epsilon > 0$ choose $g \in \mathcal{S}_d$ such that $||f - g||_{2,m_d} < \epsilon/\sqrt{M}$ (using the third property of $\mathcal{S}_d$ in §7.7.2) where M is an upper bound on the Radon-Nikodym derivative $[d\mu/dm_d]$ for the set $\mathcal{H}$. As a consequence of the Radon-Nikodym theorem applied to a change of measures with respect to which integration is performed [25, p. 134, Theorem B] we get

$$\begin{aligned} ||f - g||_{2,\mu} &\triangleq \left(\int_{\mathcal{H}} |f(x) - g(x)|^2 \mu(dx)\right)^{1/2} \\ &= \left(\int_{\mathcal{H}} |f(x) - g(x)|^2 [d\mu/dm_d] m_d(dx)\right)^{1/2} \\ &\leq \sqrt{M}||f - g||_{2,m_d} \\ &< \epsilon. \end{aligned}$$

By Theorem 13 there exists a $c > ||f||_{2,\mu}$ such that $g \in \Delta_c$. Hence there exists a $\mathcal{N} \in \mathsf{R}(m)$ with output f_m such that

$$||g - f_m||_{2,\mu}^2 \leq \frac{c^2 - ||f||_{2,\mu}^2}{m}.$$

Hence by the triangle inequality

$$||f - f_m||_{2,\mu} \leq ||f - g||_{2,\mu} + ||g - f_m||_{2,\mu} \leq \epsilon + \sqrt{\frac{c^2 - ||f||_{2,\mu}^2}{m}}.$$

■

7.8.1 Classification

Definition: A functional $f: \mathcal{H} \to \mathbb{R}$ that vanishes outside of a compact $A \subset \mathcal{H}$ is called a classifier on A or a decision function on A if, with $r \in \mathbb{N}$, $P_1, \ldots, P_r$ is a measurable partition of A such that

$$x \in P_j \;\Rightarrow\; f(x) = j$$

for all $j = 1, \cdots, r$. □

Corollary 1 Let f be a decision function on a compact set $A \subset \mathbb{R}^d$. Then, given $\epsilon > 0$ there exists a $c > ||f||_{2,\mu}$ such that f can be approximated by a network $\mathcal{N} \in \mathsf{R}(m)$ with an error

$$||f - f_m||_{2,\mu} \leq \epsilon + \sqrt{\frac{c^2 - ||f||_{2,\mu}^2}{m}}$$

for all $m \in \mathbb{N}$.

Proof: Note that

$$f(x) = \sum_{j=1}^{r} j \, \chi_{P_j}$$

is a bounded measurable function whose support is A which, being a compact subset of $\mathbb{R}^d$, has finite measure. Hence $f \in L_2$ and Theorem 14 applies. ■

7.9 Discussion of results

We have presented error bounds for approximating functionals belonging to a class introduced here with a certain kind of neural network with exponentiation (in the complex case) or cosine (in the real case) as the nonlinear function in the nodes belonging to a single hidden layer. This error decreases as $O(1/\sqrt{m})$ where m is the number of nodes in the hidden layer. This is the first error bound for this particular network architecture. We have shown that there are large, important classes of functionals (e.g., bandlimited functions and the Schwartz class) for which the results are applicable.

Based on these results, an error bound is obtained for arbitrarily good approximations of functionals in L_2. In particular, an application to classification problems is presented.

Appendix A

Concerning Chapter 4

A.1 Uniform Riemann-Stieltjes Integrability

Here let K be a compact subset of a normed linear space A, and consider Riemann-Stieltjes integrals of the form

$$\int_a^b u(x)(\tau)\, d\beta(\tau)$$

in which $x \in K$ and $u(x)(\cdot)$ and β are real valued. Suppose that the integral exists for each x, that it is Lipschitz with respect to x, and that $u(\cdot)(\tau)$ is Lipschitz for each $\tau \in [a,b]$. Then by an argument similar to that at the end of the proof in Section 2.2 it follows that $u(x)(\cdot)$ is uniformly integrable in (for example) the sense that for each $\epsilon > 0$ there is a partition $\alpha_j = a + j(b-a)/n, \quad 0 \le j \le n$, of $[a,b]$ such that

$$|\int_a^b u(x)(\tau)\, d\beta(\tau) - \sum_{j=1}^{n} u(x)(\alpha_j)[\beta(\alpha_j) - \beta(\alpha_{(j-1)})]| < \epsilon$$

for all $x \in K$. This result can be extended in a few directions. For example, A need not be a normed linear space; it can be a metric space. And $u(x)(\alpha_j)$ in the sum can be replaced with $u(x)(\lambda\alpha_j + (1-\lambda)\alpha_{(j-1)})$ for any given $\lambda \in [0,1]$.

Appendix B

Concerning Chapter 5

B.1 Proof of the Completeness of X_w

B.1.1 Weighted continuous functions

To see that X_w is complete in the norm $||f||_w = \sup_{t\in D} ||f(t)w(t)||_V$, consider a Cauchy sequence $\{x_n\}$ in X_w. Given an arbitrary positive ϵ there is a positive N such that

$$||x_m - x_n||_w < \epsilon, \quad m, n > N.$$

But this is clearly equivalent to

$$||wx_m - wx_n|| < \epsilon, \quad m, n > N.$$

By the completeness of X, there exists a $y \in X$ such that $\{wx_n\}$ converges to y in X, i.e., for an arbitrary positive ϵ there exists a positive N' such that

$$||y - wx_n|| < \epsilon, \quad n > N'.$$

Since the weighting function w never vanishes, the function $x = y/w$ can be seen to be the limit of $\{x_n\}$ in X_w.

B.1.2 Weighted L_p space

If $\{x_n\}$ is a Cauchy sequence in X_w, then given an arbitrary positive ϵ there is an $N > 0$ such that

$$||\sqrt[p]{w}x_m - \sqrt[p]{w}x_n|| < \epsilon, \quad m, n > N.$$

By the completeness of X, there exists a $y \in X$ such that $\{\sqrt[p]{w}x_n\}$ converges to y in X, i.e., for every positive ϵ there exists an $N' > 0$ such that

$$||y - \sqrt[p]{w}x_n|| < \epsilon, \quad n > N'.$$

Since the weighting function w never vanishes, the function $x = y/\sqrt[p]{w}$ can be seen to be the limit of $\{x_n\}$ in X.

B.2 Verification of the Hypotheses of Theorem 6

B.2.1 Hypothesis (i)

Here we show that $T_k(B^o)$ is relatively compact for every k. First we prove that it is relatively compact in X. Note that X is a subspace of X_w and that the identity imbedding of X in X_w is continuous. We treat the continuous and discrete

input domains separately:

1. ($D = \mathbb{Z}_+^n$ or $D = \mathbb{Z}^n$)

   (a) Assumption **A1**: Note that the proof of [35, Theorem 1, p. 202, vol. II] may be easily modified to cover the case of $D = \mathbb{Z}^n$ (the proof in [35] treats only $D = \mathbb{Z}_+^n$), and V-valued functions. Hence we only need to show that $||(T_k s)(t)||_V$ is bounded uniformly in s for every t. But this is obvious from $||(T_k s)(t)||_V \leq \sup_{\tau \in c_k} ||R(\tau)||_V$, since c_k is a finite set.

   (b) Assumption **A2**: Proceeding along the same lines as above, we need to show additionally that $\sum_{t \in D} ||(T_k s)(t)||_V{}^p$ converges uniformly in s. Since $\sum_{t \in D} ||(T_k s)(t)||_V{}^p \leq \sum_{t \in c_k} ||R(t)||_V{}^p$, the result follows.

2. ($D = \mathbb{R}_+^n$ or $D = \mathbb{R}^n$)
   Similarly as above, we observe that the proof of the classical theorem of Ascoli-Arzelà [35, Theorem 3, p. 207, vol. II] remains valid for functions with n-dimensional domains of definition and taking values in a topological vector space having the Heine-Borel property.

   (a) Assumption **A1**: We will first show that $T_k(B^o)|_{c_k}$ is relatively compact for every k in $\mathcal{C}(c_k)$. Let the map ι from $\mathcal{C}(c_k)$ onto X be given by $\iota(f) = g$, where $g \in T_k(B^o)$ satisfies $f = g|_{c_k}$. Note that ι is continuous, hence $T_k(B^o)$ is relatively compact for every k in X.

   By the observation regarding Ascoli-Arzelà theorem made above, we only need to show that $T_k(B^o)|_{c_k}$ is uniformly bounded and equicontinuous. For any $s \in T_k(B^o)$ and $t \in c_k$, we have $||(T_k s)(t)||_V \leq \sup_{\tau \in c_k} ||R(\tau)||_V$, hence uniform boundedness follows. Similarly, equicontinuity follows from the defining property of B^o, namely, $||s(t) - s(\tau)||_V \leq \lambda(k)|t - \tau|$, for any $s \in T_k(B^o)$ and $t, \tau \in c_k$.

   (b) Assumption **A2**: Note that the proof of Kolmogorov-Tulaikov theorem [35, Theorem 6, p. 215, vol. II] uses the Ascoli-Arzelà theorem and the fact that a relatively compact set in $\mathcal{C}(c_k)$ is also relatively compact in $L_p(c_k)$ when $D = \mathbb{R}$. Both of these assertions remain true in our setting, hence all we need to show is that $T_k(B^o)|_{c_k}$ is relatively compact in $L_p(c_k)$. Since the map $s|_{c_k} \mapsto s$ for $s \in T_k(B^o)$ is continuous, the relatively compactness in X follows.

   From Theorem 3 of Chapter 3, and the observation that the components of $s|_{c_k}$ satisfy the conditions of that theorem, we have that for every component $(T_k s)^i$ there exists h_i such that $\lim_{h_i \to 0} ||(T_k s)^i - (T_k s)^i_{h_i}||_{L_p(-k,k)} = 0$ uniformly in $T_k(B^o)|_{c_k}$. Hence, $\lim_{h \to 0} ||s - s_h|| = 0$. By a result of Sudakov [57], this also implies uniform L_p-boundedness, hence by the Kolmogorov-Tulaikov theorem $T_k(B^o)|_{c_k}$ is relatively compact in $L_p(c_k)$.

B.2.2 Hypothesis (ii)

We must show that $\rho(s, T_k s) \to 0$ as $k \to \infty$ uniformly for $s \in B^o$.

1. Assumption **A1**: Note that

$$\rho(s, T_k s) \quad = \quad \sup_{t \in D \setminus c_k} ||s(t) w(t)||_V$$

$$\leq \sup_{t \in D \setminus c_k} ||R(t)w(t)||_V .$$

By the condition (5.1) on the growth at infinity, it follows that $\rho(s, T_k s) \to 0$ uniformly in $s \in B^o$.

2. Assumption **A2**: Observe that

$$\begin{aligned} \rho(s, T_k s) &= \int_{t \in D \setminus c_k} ||s(t)||_V^p w(t) \mu(dt) \\ &\leq \int_{t \in D \setminus c_k} ||R(t)||_V^p w(t) \mu(dt). \end{aligned}$$

From the fact that $R \in X_w$, the desired result follows.

Appendix C

Concerning Chapter 6

C.1 Lemma on Equicontinuous Extensions

Here we present a constructive method of extending members of a relatively compact set of functions defined on a compact convex subset of $\mathbb{R}^n$ which preserves the properties of uniform boundedness and equicontinuity.[1]

Lemma 6 Let $F \subset \mathcal{C}(C_1)$ be relatively compact and let $T > 1$. Any member f of F may be extended to $\tilde{f} \in \mathcal{C}(\mathbb{R}^n)$ such that:

1. $\tilde{f}(x) = f(x), \quad x \in C_1$.
2. $\tilde{f}(x) = 0, \quad x \in \mathbb{R}^n \setminus C_T$.
3. $\sup_{x \in C_T \setminus C_1} |\tilde{f}(x)| \leq \sup_{x \in C_1} |f(x)|$.
4. $\{\tilde{f} \mid f \in F\}$ is equicontinuous.

Proof: Set $\lambda = T - 1$. Note that the usual norm in $\mathbb{R}^n$ is uniformly convex. Since C_1 is closed and convex, for any $x \in \mathbb{R}^n$ there exists a unique $\hat{x} \in C_1$ such that $||x - \hat{x}|| = \inf_{y \in C_1} ||x - y||$.

Let $\ell: \mathbb{R}_+ \to \mathbb{R}_+$ be defined as

$$\ell(t) = \begin{cases} 1 - \frac{t}{\lambda}, & \text{if } t \leq \lambda \\ 0, & \text{otherwise.} \end{cases}$$

Extend $f \in F$ to $\tilde{f}: \mathbb{R}^n \to \mathbb{R}$ as follows:

$$\tilde{f}(x) = f(\hat{x})\ell(||x - \hat{x}||), \quad x \in \mathbb{R}^n. \tag{C.1}$$

Note that the first three properties required of the extension follow immediately from the definition.

Using the equicontinuity of F (which follows from relative compactness and the Ascoli-Arzelà theorem), we can define a nondecreasing function $\delta_F: \mathbb{R}_+ \to \mathbb{R}_+$ such that for any $x, y \in C_1$ and arbitrary $f \in F$,

$$||x - y|| < \delta_F(\zeta) \;\Rightarrow\; |f(x) - f(y)| < \zeta.$$

[1] A related result (which does not address the preservation of equicontinuity) is Rudin's version [41, p. 422] of the Tietze extension theorem.

Note that F is uniformly bounded by the Ascoli-Arzelà theorem, i.e., there exists an $M > 0$ such that

$$\sup_{x \in C_T} |f(x)| < M, \quad f \in F.$$

Define $\delta_{\tilde{F}} : \mathbb{R}_+ \to \mathbb{R}_+$ by

$$\delta_{\tilde{F}}(\zeta) = \min\left(\frac{\delta_F(\zeta/4)}{2}, \frac{\zeta\lambda}{4M}, \lambda\right).$$

To verify the claim of equicontinuity, let an arbitrary $\eta > 0$ be given, and consider the six cases where each of $x, y \in \mathbb{R}^n$ belong to $C_1, C_T \setminus C_1$, or $\mathbb{R}^n \setminus C_T$, satisfying the constraint

$$||x - y|| < \delta_{\tilde{F}}(\eta).$$

1. $x, y \in C_1$.
   Since $\delta_{\tilde{F}}(\eta) < \delta_F(\eta)$, it follows trivially that $|\tilde{f}(x) - \tilde{f}(y)| = |f(x) - f(y)| < \eta$.

2. $x \in C_1,\ y \in C_T \setminus C_1$.
   By the best approximation property of $\hat{y}$, $||y - \hat{y}|| \leq ||y - x||$. By the triangle inequality, $||x - \hat{y}|| \leq ||x - y|| + ||y - \hat{y}|| \leq 2||x - y||$. Now,

$$\begin{aligned} & |\tilde{f}(x) - \tilde{f}(y)| \\ = & \left| f(x) - f(\hat{y}) + f(\hat{y}) \frac{||y - \hat{y}||}{\lambda} \right| \\ \leq & |f(x) - f(\hat{y})| + \frac{|f(\hat{y})|}{\lambda} ||y - \hat{y}||. \end{aligned}$$

   Since $\delta_{\tilde{F}}(\eta) < \min\left(\frac{\delta_F(\eta/2)}{2}, \frac{\eta\lambda}{2M}\right)$, $||x - \hat{y}|| < \delta_F(\eta/2)$, and $||y - \hat{y}|| < \frac{\eta\lambda}{2M}$, hence

$$|\tilde{f}(x) - \tilde{f}(y)| < \frac{\eta}{2} + \frac{M}{\lambda}||y - \hat{y}|| \leq \eta.$$

3. $x \in C_1,\ y \in \mathbb{R}^n \setminus C_T$.
   Impossible, since $||x - y|| < \delta_{\tilde{F}}(\eta) < \lambda$.

4. $x, y \in C_T \setminus C_1$.
   By the best approximation property,

$$||x - \hat{x}|| \leq ||x - \hat{y}||, \quad \text{and} \quad ||y - \hat{y}|| \leq ||y - \hat{x}||.$$

   By the triangle inequality,

$$||x - \hat{y}|| \leq ||x - y|| + ||y - \hat{y}||, \quad \text{and} \quad ||y - \hat{x}|| \leq ||y - x|| + ||x - \hat{x}||.$$

   Hence

$$||x - \hat{x}|| \leq ||x - y|| + ||y - \hat{y}||, \quad \text{and} \quad ||y - \hat{y}|| \leq ||x - y|| + ||x - \hat{x}||,$$

implying

$$| \, ||y-\hat{y}|| - ||x-\hat{x}|| \, | \leq ||x-y||.$$

Let $\beta \stackrel{\triangle}{=} |\tilde{f}(x) - \tilde{f}(y)|$. We have

$$\beta = |f(\hat{x}) - f(\hat{y}) + [f(\hat{y})||y-\hat{y}|| - f(\hat{x}) \, ||x-\hat{x}|| \,]/\lambda \, |.$$

Adding and subtracting $f(\hat{x})||y-\hat{y}||/\lambda$,

$$\begin{aligned} \beta &\leq |f(\hat{x}) - f(\hat{y})| \\ &+ \quad \frac{1}{\lambda} \, | \, f(\hat{x}) \, (||y-\hat{y}|| - ||x-\hat{x}||) - [f(\hat{x}) - f(\hat{y})] \, ||y-\hat{y}|| \, | \\ &\leq \quad 2|f(\hat{x}) - f(\hat{y})| \\ &\quad + \frac{|f(\hat{x})|}{\lambda} | \, ||y-\hat{y}|| - ||x-\hat{x}|| \, |. \end{aligned}$$

By a well-known theorem on the Lipschitz property of a projection on a closed convex subset of a Hilbert space [24, p. 100], we have $||\hat{x}-\hat{y}|| < ||x-y||$. Now, $\delta_{\tilde{F}}(\eta) < \min\left(\delta_F(\eta/4), \frac{\eta\lambda}{2M}\right)$; implies

$$\beta < 2\,\frac{\eta}{4} + \frac{M}{\lambda} \, ||x-y|| \leq \eta.$$

5. $x \in C_T \setminus C_1, \; y \in \mathbb{R}^n \setminus C_T$.
Let $y^{\#}$ be the unique point of C_T which is closest to y. By the best approximation property, $||y-y^{\#}|| \leq ||y-x||$. By the triangle inequality, $||x-y^{\#}|| \leq ||x-y|| + ||y-y^{\#}|| \leq 2||x-y||$. Consider

$$\begin{aligned} \beta &= |\tilde{f}(x) - \tilde{f}(y)| \\ &= |\tilde{f}(x) - \tilde{f}(y^{\#})|. \end{aligned}$$

Since $\delta_{\tilde{F}}(\eta) < \min\left(\frac{\delta_F(\eta/4)}{2}, \frac{\eta\lambda}{4M}\right)$, $||x-y^{\#}|| < \min\left(\delta_F(\eta/4), \frac{\eta\lambda}{2M}\right)$, and from the previous case we have $\beta \leq \eta$.

6. $x, y \in \mathbb{R}^n \setminus C_T$.
Trivial, since $\tilde{f}(x) = \tilde{f}(y) = 0$.

C.2 Uniform Riemann Integrability

Here we present a lemma that may be of independent interest for other approximation studies. We assume that an arbitrary $T > 0$ is given. The Riemann integral of a function $f: C_T \to \mathbb{R}$ is denoted by $\Re(f)$ whenever f is Riemann integrable.

Lemma 7 Let $F \subset \mathcal{C}(C_T)$ be an equicontinuous family of functions. Given $\epsilon > 0$, there exists a $\delta > 0$ such that for any Perron partition $P = \{(I_\beta, z_\beta)\}_{\beta \leq \gamma}$ of C_T with $|P| < \delta$,

$$\left| \Re(f) - \sum_{\beta \leq \gamma} f(z_\beta) V_\beta \right| < \epsilon, \quad f \in F.$$

Proof: Let $P = \{(I_\beta, z_\beta)\}_{\beta \le \gamma}$ be a Perron partition of C_T of size γ, with the partition norm $|P|$. Let the corresponding Riemann sum for $\Re(f)$ be

$$v_P(f) = \sum_{\beta \le \gamma} f(z_\beta) V_\beta .$$

For $A \subset B \subset \mathbb{R}^n$ and a bounded function $g : B \to \mathbb{R}$, let the oscillation of g over A be defined as $\Delta(g; A) = \sup_{x \in A} g(x) - \inf_{x \in A} g(x)$. Let $\eta = \epsilon/(2T)^n$. Since F is equicontinuous, there exists $\delta > 0$ such that for any $x, y \in C_T$,

$$||x - y|| < \delta \;\Rightarrow\; |f(x) - f(y)| < \eta, \quad f \in F.$$

If $|P| < \delta$, then for any $x, y \in I_\beta$, and $f \in F$,

$$\begin{aligned} ||x - y|| \le |I|_\beta \quad &\le \quad |P| < \delta \\ &\Rightarrow \quad |f(x) - f(y)| < \eta \\ &\Rightarrow \quad |\Delta(f; I_\beta)| < \eta \end{aligned}$$

Using the convergence property of the Riemann sums [1, Theorem 14.4], for all $f \in F$,

$$\begin{aligned} |\Re(f) - v_P(f)| \quad &\le \quad \sum_{\beta \le \gamma} |\Delta(f; I_\beta)| V_\beta \\ &< \quad \eta \sum_{\beta \le \gamma} V_\beta \;=\; \epsilon. \end{aligned}$$

■

Bibliography

[1] T. M. Apostol. *Mathematical Analysis*. Addison-Wesley Series in Mathematics. Addison-Wesley, Reading, Massachusetts, Second edition, 1974.

[2] Mahmood R. Azimi-Sadjadi and Hongye Pan. Two-Dimensional Block Diagonal LMS Adaptive Filtering. *IEEE Transactions on Signal Processing*, 42(9):2420–2429, September 1994.

[3] A. R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. Department of Statistics Technical Report 58, University of Illinois at Urbana-Champaign, Champaign, Illinois, 1991.

[4] R. Bellman. On an application of a Banach-Steinhaus theorem to the study of the boundedness of solutions of nonlinear differential and difference equations. *Annals of Mathematics*, 49(3):515–522, July 1948.

[5] S. P. Bhattacharyya, H. Chapellat, and L. H. Keel. *Robust Control: The Parametric Approach*. Information and Systems Scinces Series. Prentice Hall, 1995.

[6] S. Boyd. *Volterra Series: Engineering Fundamentals*. PhD thesis, The University of California at Berkeley, Berkeley, California, 1985.

[7] R. C. Buck. Approximation properties of vector valued functions. *Pacific Journal of Mathematics*, 53(1):85–94, 1974.

[8] F. M. Callier and C. A. Deoser. *Multivariable Feedback Systems*. Springer-Verlag, New York, 1982.

[9] S. V. Chakravarthy. Capabilities of adaptive networks with Gaussian Basis Functions. Master's thesis, The University of Texas at Austin, Austin, Texas, December 1991.

[10] H. Chapellat and S. P. Bhattacharyya. Robust Stability and Stabilization of Interval Plants. In M. Milanese, R. Tempo, and A. Vicino, editors, *Robustness in Identification and Control*, pages 207–230. Plenum Press, New York, 1988.

[11] T. Chen and H. Chen. Approximations of continuous functionals by neural networks with application to dynamical systems. *IEEE Transactions on Neural Networks*, 4(6):910–918, November 1993.

[12] T. Chen and H. Chen. Approximation capability to functions of several variables, nonlinear functionals, and operators by radial basis function neural networks. *IEEE Transactions on Neural Networks*, 6(4):904–910, July 1995.

[13] E. W. Cheney. *Multivariate Approximation Theory: Selected Topics*, volume 51 of *The Regional Conference Series in Applied Mathematics*. SIAM, Philadelphia, Pennsylvania, 1986.

[14] E. W. Cheney. Topics in approximation theory. Course Notes for `M 393 D`, Department of Mathematics, The University of Texas at Austin, Spring 1993.

[15] B. D. Coleman and V. J. Mizel. Norms and semi-groups in the theory of fading memory. *Arch. Rational Mech. Anal.*, 23:87–123, 1966.

[16] B. D. Coleman and V. J. Mizel. On the general theory of fading memory. *Arch. Rational Mech. Anal.*, 29:18–31, 1968.

[17] B. D. Coleman and W. Noll. An approximation theorem for functionals, with applications in continuum mechanics. *Arch. Rational Mech. Anal.*, 6:355–370, 1960.

[18] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, 2:303–314, 1989.

[19] D. F. Delchamps. Analytic stabilization and the algebraic Riccati equation. In *Proceedings of the IEEE Conference on Decision and Control*, pages 1396–1401, 1983.

[20] C. A. Desoer and A. J. Thomasian. A Note on Zero-State Stability of Linear Systems. In *Proceedings of the First Annual Allerton Conference on Circuits and System Theory*, pages 50–52, November 1963.

[21] R. M. Dudley. *Real Analysis and Probability.* Wadsworth, Inc., 1989.

[22] J. Ghosh and L. Deuser. Classification of spatio-temporal patterns with applications to recognition of sonar sequences. In E. Covey, H. Hawkins, T. McMullen, and R. Port, editors, *Neural Representation of Temporal Patterns.* 1995.

[23] F. Girosi and G. Anzellotti. Rates of convergence for radial basis functions and neural networks. In *Artificial Neural Networks with Applications in Speech and Vision.* Chapman & Hall, 1993.

[24] A. A. Goldstein. *Constructive Real Analysis.* Harper's Series in Modern Mathematics. Harper & Row, New York, 1967.

[25] P. R. Halmos. *Measure Theory.* Van Nostrand, Princeton, New Jersey, 1950.

[26] K. Hornik, M. Stinchcombe, and H. White. Multi-layer feedforward networks are universal approximators. *Neural Networks*, 2:359–366, 1989.

[27] L. K. Jones. A simple lemma on iterative sequences in Hilbert space and convergence rates for projection pursuit regression. Technical Report 16, University of Lowell, Lowell, Massachusetts, 1989.

[28] V. L. Kharitonov. Asymptotic stability of an equilibrium position of a family of systems of linear differential equations. *Differentsial'nye Uravneniya*, 14:2086–2088, 1978.

[29] A. N. Kolmogorov and S. V. Fomin. *Introductory Real Analysis.* Dover, New York, 1975.

[30] B. K. Lande and S. D. Agashé. Pole Assignment using Output Feedback. In Mohan C. Joshi and A. V. Balakrishnan, editors, *Mathematical Theory of Control*, pages 225–232. Marcel Dekker, 1993.

[31] L. A. Liusternik and V. J. Sobolev. *Elements of Functional Analysis.* Frederick Ungar Publishing Co., New York, 1961.

[32] Jan Lunze. *Robust Multivariable Feedback Control.* Prentice Hall, New York, 1988.

[33] D. E. Miller and E. J. Davison. Adaptive Control of a Family of Plants. In D. Hinrichsen and B. Mårtensson, editors, *Adaptive Control of Uncertain Systems,* Progress in Systems and Control Theory, Volume 6, pages 197–219. Birkhäuser, Boston, 1990.

[34] K. Nam and A. Arapostathis. A model reference adaptive control scheme for pure-feedback nonlinear systems. *IEEE Transactions on Automatic Control*, 33(9):803–811, 1988.

[35] I. Natanson. *Theory of Functions of a Real Variable.* Frederick Ungar Publishing Co., New York, 1955.

[36] J. Park and I. W. Sandberg. Universal approximation using radial-basis-function networks. *Neural Computation*, 3:246–257, 1991.

[37] J. Park and I. W. Sandberg. Nonlinear approximations using elliptic basis function networks. *Circuits, Systems and Signal Processing*, 13(1):99–113, 1993.

[38] Washek F. Pfeffer. *The Riemann Approach to Integration: Local Geometric Theory*, volume 109 of *Cambridge Tracts in Mathematics.* Cambridge University Press, Cambridge (England), 1993.

[39] G. Pisier. Remarques sur un resultat non publié de B. Maurey. *Sem. d'Analyse Fonctionelle*, 1(12), 1980–81. Ecole Poly. Centre de Math., Palaiseau.

[40] A. Rantzer and A. Megretsky. A Convex Parameterization of Robustly Stabilizing Controllers. In M. Mansour, S. Balemi, and W. Truöl, editors, *Robustness of Dynamical Systems with Parameter Uncertainties*, pages 271–279. Birkhäuser, 1992.

[41] W. Rudin. *Real and Complex Analysis.* McGraw-Hill, Inc., New York, 1974.

[42] W. Rudin. *Principles of Mathematical Analysis.* International Series in Pure and Applied Mathematics. McGraw-Hill, Inc., New York, third edition, 1976.

[43] W. Rudin. *Functional Analysis.* McGraw-Hill, Inc., New York, 1991.

[44] I. W. Sandberg. A Frequency-Domain Condition for the Stability of Feedback Systems Containing a Single Time-Varying Nonlinear Element. *Bell System Technical Journal*, 43(4):1601–1608, July 1964.

[45] I. W. Sandberg. On the boundedness of solutions of nonlinear integral equations. *Bell System Technical Journal*, 44(3):439–453, March 1965.

[46] I. W. Sandberg. Some results on the theory of physical systems governed by nonlinear functional equations. *Bell System Technical Journal*, 44(5):871–898, May–June 1965.

[47] I. W. Sandberg. On Volterra expansions for time-varying nonlinear systems. *IEEE Transactions on Circuits and Systems*, 30:61–67, February 1983.

[48] I. W. Sandberg. Structure Theorems for Nonlinear Systems. *Multidimensional Systems and Signal Processing*, 2:267–286, 1991.

[49] I. W. Sandberg. Approximations for nonlinear functionals. *IEEE Transactions on Circuits and Systems-I. Fundamental Theory and Applications*, 39(1):65–67, January 1992.

[50] I. W. Sandberg. Uniform approximation and the circle criterion. *IEEE Transactions on Automatic Control*, 38(10):1450–1458, October 1993.

[51] I. W. Sandberg. General structures for classification. *IEEE Transactions on Circuits and Systems-I: Fundamental Theory and Applications*, 41(5):372–376, May 1994.

[52] Irwin W. Sandberg and Lilian Xu. Nonlinear functionals on spaces with weighted norms. *In preparation*, 1995.

[53] R. Sanner and J.-J. E. Slotine. Gaussian Networks for Direct Adaptive Control. *IEEE Transactions on Neural Networks*, 3(6):837–863, November 1992.

[54] E. D. Sontag. Nonlinear regulation: The piecewise linear approach. *IEEE Transactions on Automatic Control*, 26:346–358, 1981.

[55] E. D. Sontag. An algebraic approach to bounded controllability of nonlinear systems. *International Journal of Control*, 39:181–188, 1984.

[56] E. D. Sontag. Parametric stabilization is easy. *Systems and Control Letters*, 4:181–188, 1984.

[57] V. N. Sudakov. Concerning criteria for compactness in function spaces. *Uspekhi Mat. Nauk*, 12(3), 1957.

[58] Richard J. Vaccaro. *Digital Control*. McGraw-Hill, Inc., New York, 1995.

[59] M. Vidyasagar. On the stabilization of nonlinear systems using state detection. *IEEE Transactions on Automatic Control*, 25:504–509, 1980.

[60] N. Viswanadham and M. Vidyasagar. Stabilization of linear and nonlinear dynamical systems using an observer-controller configuration. *System Control Letters*, 1:87–91, 1981.

[61] D. C. Youla. On the stability of linear systems. *IEEE Transactions on Circuit Theory*, CT-10:276–279, June 1963.

Vita

Ajit, son of Mrs. Vijaya and Mr. Trimbak Dingankar, was born on May 25, 1960 in Bombay, India. He received the Secondary School Certificate from Balmohan Vidyamandir, Bombay, and the Higher Secondary Certificate from Jai Hind College, Bombay. After obtaining the Diploma in Computer Management from the Indian Institute of Computer Management, Ahmedabad, he earned the B. Tech. degree in Electrical Engineering from the Indian Institute of Technology, Bombay.

He went to Clemson University, South Carolina, in the Fall of 1986 where he was awarded an M. S. in Bioengineering. Afterwards he entered the University of Texas at Austin to participate in the Ph. D. program in Biomedical Engineering. After testing two departments for compatibility of interests he settled down in the Information and Systems Engineering Area in the Department of Electrical & Computer Engineering. He worked for International Business Machines Corporation, Austin, from June 1989 through August 1994.

Permanent address: 53 "Savitri"
Abhinav Nagar
Borivli (East), Bombay
INDIA 400 066

This dissertation was typeset with LaTeX[‡] by the author.

[‡]LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.