CHAPTER 6

BEYOND VOLTERRA AND WIENER: OPTIMAL MODELING OF NONLINEAR DYNAMICAL SYSTEMS IN A NEURAL SPACE FOR APPLICATIONS IN COMPUTATIONAL INTELLIGENCE

RUI J. P. DE FIGUEIREDO

6.1 INTRODUCTION

Nonlinear dynamical systems are playing a major role in a number of applications of computational intelligence. In order to maintain the current growth of the technologies supporting these applications into this new century, it is essential to develop rigorous, accurate, efficient, and insightful models for describing nonlinear dynamical systems' behavior, including *adaptation*, *learning*, and *evolution*, based on input–output observations or input–output specifications. If based on observations, the modeling process is called *system identification*, and if based on specifications, it is called *system realization* or *design*.

In this chapter, we present optimal solutions to both of the preceding problems in the setting of a *Neural Space* $\mathcal{N}$ introduced by the author in 1990 [1,2]. $\mathcal{N}$ is a separable Hilbert space of nonlinear[1] maps, f, that map a given vector x from a data space, X, which itself is a separable Hilbert or Euclidean space, to an m-vector y of m scalar outputs $y_j = f_j(x), j = 1, \ldots, m$, and f_j are bounded analytic functionals on X expressible as Volterra functional series on X [3]. The f_j belong to an appropriately constructed reproducing kernel Hilbert space, F, also introduced by de Figueiredo et al. in [4] in 1980, as a generalization of the symmetric Fock space. Details on this formulation as well as applications have been presented and discussed elsewhere [5–24].

[1]Throughout this chapter, by "nonlinear" we mean "not necessarily linear."

Computational Intelligence: The Experts Speak. Edited by D. B. Fogel and C. J. Robinson
ISBN 0-471-27454-2

Our objective here is to provide an overview and a further extension, oriented toward *computational intelligence*, of the underlying concepts and methodology for a mixed engineering/mathematical audience. The presentation will be from an approximation-theoretic rather than random-field viewpoint. The latter will be presented separately [24].

First and foremost, it is worthwhile pointing out the following two special features of our formulation.

First, our approach is *nonparametric* and leads to a simultaneous determination of an optimal structure as well as optimal values of parameters for the model of the system to be identified or realized. This is done by minimizing the maximum error (with respect to (w.r.t.) a metric in $\mathcal{N}$) between a desired but unknown nonlinear map, f, to be identified or realized, and its best estimate (model), $\hat{f}$, under prescribed prior uncertainty conditions on f and subject to the input–output observed or specified data constraints on f. This type of estimation embodies the notion of a *best robust approximation* of f by $\hat{f}$.

Second, even though no a priori structure is assumed for the model, the optimal solution appears in the form of a neural system. Thus an additional feature of our formulation is that it provides a mathematical justification for why biological systems, like the human brain, that perform tasks requiring computational intelligence have a neural system structure; and it points a way to model artificial and natural neural systems rigorously under a common framework.

Despite the power and richness in their description, prior works in the area of Volterra series [3,28–32] and its variants, such as Wiener–Bose series [25–32], had the following shortcomings, which we have attempted to overcome.

6.1.1 Shortcomings of the Previous Volterra and Wiener Formulations

The Volterra functional series (VS) representation of a nonlinear map, f, from a function space, X, to the complex plane, $\mathcal{C}$, is an abstract power series in the input $x \in X$ of the form

$$y(t) = f(t; x) = \sum_{n=0} \frac{1}{n!} f_n(t; x) \tag{6.1}$$

where t is an indexing time variable for the scalar output variable $y(t)$, and if $X = L^2(I)$, I being an interval of the real line,

$$f_n(t; x) = \int_I \cdots \int_I h_n(t; t_1, \ldots, t_n) x(t_1) \cdots x(t_n)\, dt_1 \cdots dt_n \tag{6.2}$$

where the kernels $h_n(t; \cdots)$ belong to an appropriate space like $L^2(I^n)$.

An expression for f_n when $x \in E^N = X$, where E^N is an N-dimensional complex Euclidean space, is given in a later section (6.31).

The VS representation (6.1)–(6.2), which has been used widely as a feedforward model for continuous time-parameter nonlinear dynamical systems [3,28–32], has

some serious limitations. The multiple integrals in (6.2) are difficult to implement computationally. If, in order to mitigate this difficulty the series is truncated except for a few terms, (1) the resulting truncation errors are significant except when the amplitude of the input x is small, and (2) any least-squares approximation of the model, to satisfy the input–output data constraints, takes place in a finite dimensional space spanned by the truncated series rather than in an infinite-dimensional space to which the series may belong. This difficulty does not occur in our formulation.

To mitigate these difficulties, Wiener and Bose [25,26] proposed a Gram–Schmidt orthogonalization of the VS in the space of the output random variable $y(t)$, with the input as white Gaussian noise (WGN).

Specifically, the Wiener–Bose model is expressed in the form [25–28]

$$y(t) = H(Lx(t)) \tag{6.3}$$

where L is a linear differential dynamical system that enables the expansion of the input signal into Laguerre functions $l_i(\cdot)$, and $H(\cdot)$ represents a zero-memory nonlinear system that expands the range of L (i.e., in terms of the scalar variable $z = Lx(t)$) into orthogonal Hermite functions[2] $\varphi_k(z)$ (see Fig. 6.1). This leads to an expansion in the form (6.1) and (6.2), where the functionals $f_n(t;x)$ are expressed in terms of the Laguerre and Hermite basis functions appearing in the representation (6.3). The coefficients associated with the kernels in such a parametric representation are obtained by using WGN as the test input x and performing a Gram–Schmidt

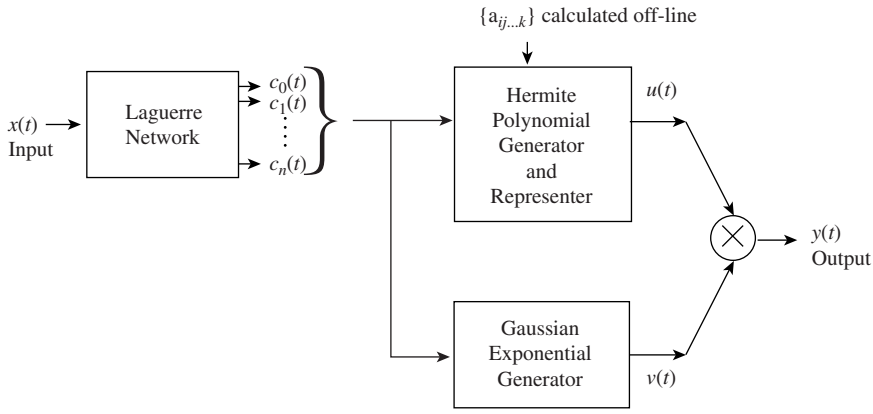


Figure 6.1 The Wiener–Bose nonlinear functional series model. In this model, the functional expansion is $y(t) = \sum_{k_0=0}^{\infty} \cdots \sum_{k_n=0}^{\infty} a_{k_0 \cdots k_n} \eta_{k_0}(c_0(t)) \eta_{k_1} \ (c_1(t)) \cdots \eta_{k_n}(c_n(t))$ $\exp\left[-\frac{1}{2}\sum_{i=0}^{n} c_i^2(t)\right]$, where $c_i(t) = \int_{-\infty}^{t} l_i(t-\tau)x(\tau)\, d\tau$.

[2] Orthogonal Hermite functions $\varphi_k(z)$ are of the form $\varphi_k(z) = \eta_k(z)\exp\left(-\frac{1}{2}z^2\right)$, where $\eta_k(z)$ is a Hermite polynomial in z of degree k.

orthogonalization of the output random variable $y(t)$ with respect to the random variable output of each kernel.

Thus the parameters in the model are determined so that, when $x(t)$ is WGN, the following (known as Wiener G-functional decomposition) holds:

$$E[y(t)] = f_0(t;x) \tag{6.4}$$

$$E[f_0(t;x)f_1(t;x)] = 0 \tag{6.5}$$

$$E[f_0(t;x)f_2(t;x)] = E[f_1(t;x)f_2(t;x)] = 0 \tag{6.6}$$

$$E[f_0(t;x)f_3(t;x)] = E[f_1(t;x)f_3(t;x)] = E[f_2(t;x)f_3(t;x)] = 0 \tag{6.7}$$

$$\text{etc} \ldots$$

This orthogonalization guarantees that, when the input is WGN, truncation gives the minimum mean-square estimate of the output using the untruncated terms.

Despite this property, the Wiener–Bose model also has some fundamental limitations.

First, the model presents a conceptual difficulty posed by its use of WGN as a test signal. While WGN is an ideal test signal for probing linear time-invariant (LTI) systems, because different frequency components pass through an LTI system without mutual interference, WGN appears the least desirable one for testing nonlinear dynamical systems because of the effects of this interference.

Second, a truncation of the series optimized w.r.t. WGN input need not be optimal with respect to any nonwhite or/and non-Gaussian input. Of course, the Wiener–Bose procedure could be repeated for a given nonwhite or/and non-Gaussian input signal, but then the model would not be optimal w.r.t any other type of input statistics.

These considerations and the computational effort in the implementation of the model mentioned previously point to the need of looking for other approaches.

Numerous papers and treatises have appeared on the analysis and control of nonlinear dynamical systems (see, e.g., [28–51]) that in one way or another relate to the approach presented here. Limitations in space do not permit us to review them here.

6.1.2 Summary of This Chapter

In Section 6.2 we briefiy describe the three basic types of nonlinear dynamical system models grouped according to their configuration and description. They are *feedforward*, *recurrent*, and *state-space models*. In these models we indicate the generic nonlinear maps that appear in their description. These are in general m-tuples f_j, $j = 1, \ldots, m$, of nonlinear maps from a complex separable Hilbert space X (which could be the N-dimensional Euclidean space E^N) to the complex plane $\mathcal{C}$.

In Section 6.3 we make the fundamental and very general assumption that the maps f_j are bounded analytic functionals on X expressible as an abstract power series (VS) in $x \in X$. We construct a reproducing kernel Hilbert space (RKHS), F, to which

the maps f_j can be made to belong, and study the properties of this space, F, needed in the modeling of the nonlinear dynamical systems described in Section 1.2.

In Section 6.4 we show how these properties provide a rationale for the derivation/design of sigmoid functions as elements in the nonlinear dynamical-system modeling process.

In Section 6.5 we formally introduce the neural space $\mathcal{N}$ and present, as mentioned earlier, an explicit expression for the best robust approximation $\hat{f}$ of f in $\mathcal{N}$. We show that such an $\hat{f}$ appears as an abstract two-hidden-layer artificial neural network, called by us an *optimal interpolating* (*OI*) *neural network*, so obtained without prior assumption that $\hat{f}$ have a neural structure. This motivates our calling $\mathcal{N}$ a neural space. These theoretical developments also provide new rationales for representation of neural systems as linear combinations of shifted sigmoid functions and as linear combinations of radial basis functions (RBF). Also in that section, an optimal solution, $\hat{f}$, for the case in which the data are corrupted by WGN is given, with a two-layer *optimal smoothing* (*OS*) *neural network* as a special case.

In section 6.6, we port the developments of the preceding section to feedforward, recurrent and state-space models of nonlinear dynamical systems in the neural space $\mathcal{N}$. Extensions to complex models, such as OMNI (optimal multilayer neural interpolating) net and OSMAN (optimal smoothing multilayer artificial neural) net, which may include feedback, are also presented.

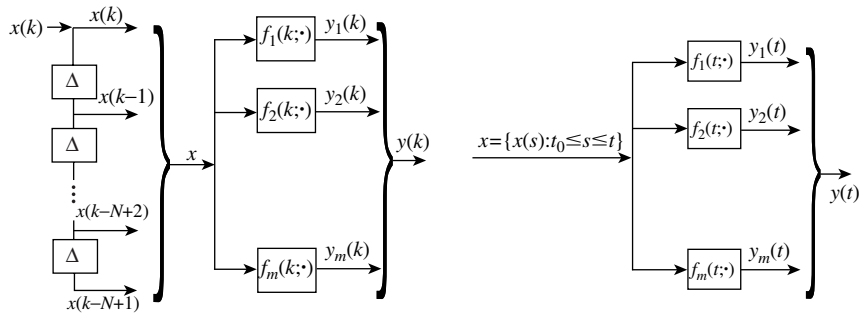
Section 6.7 provides a framework for porting the technology developed in the present chapter to computationally intelligent systems by modeling these systems as *mixed* (continuous/discrete) *systems*. The synthesis of these systems is achieved through appropriate application-specific combinations of MOI (mixed OI) and MOS (mixed OS) nets. Finally, in this section, a framework is presented for modeling what we call *intelligent learning* by CI systems as a combined adaptation and evolution process, and discovery as a consequence of augmentation of a higher-level neuronal layer in the system.

We end, in Section 6.8, with concluding remarks on some current and potential applications of this technology.

6.2 CLASSES OF NONLINEAR DYNAMICAL SYSTEM MODELS

There are in general three categories of basic models of nonlinear dynamical systems, namely *feedforward*, *recurrent*, and *state-space* models. Furthermore, by interconnection of such models (see, e.g., [51]), more complex models with any appropriate degree of complexity can be obtained.

In this section we discuss descriptions of the basic models just listed at the block-diagram level, both for the discrete-time-parameter (DTP) and continuous-time-parameter (CTP) cases. For simplicity in presentation, we restrict discussion to single-input/multiple-output systems in the feedforward case, to single-input/single-output systems in the recurrent case, and to multiple-input/multiple-output systems in the state-variable case. For the descriptions of the three basic models we will show in Section 6.6 how the nonlinear maps that appear in the various blocks



(a) DTP Case (b) CTP Case

Figure 6.2 Feedforward models for single-input/multiple-output dynamical systems.

in the diagrams of this section can be optimally approximated and realized using our approach. *A Note Regarding Notation*: Even though we will call X the input data space, it will stand for the domain of the functionals that appear in various blocks. However, the meaning of X will be clear from the context.

6.2.1 Feedforward Models

For the single-input/multiple-output feedforward model we have the following descriptions, as depicted in Figure 6.2.

DTP Case

$$\begin{aligned} y(k) &= f(k;x) \\ &= (f_1(k;x),\ldots,f_m(k;x))^T \end{aligned} \tag{6.8}$$

where the superscript T denotes the transpose, $x(k) \in E^1$ and $y(k) \in E^m$ are scalar input and vector output samples at the instant k, and we use the notation

$$x = (x(k), x(k-1),\ldots,x(k-N+1))^T \tag{6.9}$$

for the input data string of length N up to and including k, and $f_j(k;\cdot)$, $j = 1,\ldots,m$, are bounded analytic functionals (VS) on E^N. Thus the input data space, X, for this case is the space E^N of strings x. If the strings, x, are square summable and of infinite length, X is l^2.

CTP Case

$$\begin{aligned} y(t) &= f(t;x) \\ &= (f_1(t;x),\ldots,f_m(t;x))^T \end{aligned} \tag{6.10}$$

where the input signal (data) x on an interval $I = [t_0, t]$ belongs to a complex-valued function Hilbert space X such as $L^2(I)$ or[3] $W_N^2(I)$, and $f_j(t, \cdot)$ are bounded analytic functionals (VS) on X.

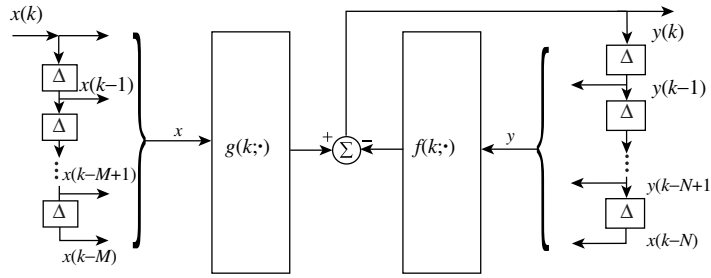
Note that (6.8) and (6.10) can be viewed as discrete- and continuous-time parameter nonlinear convolutions with the input x.

6.2.2 Recurrent Models

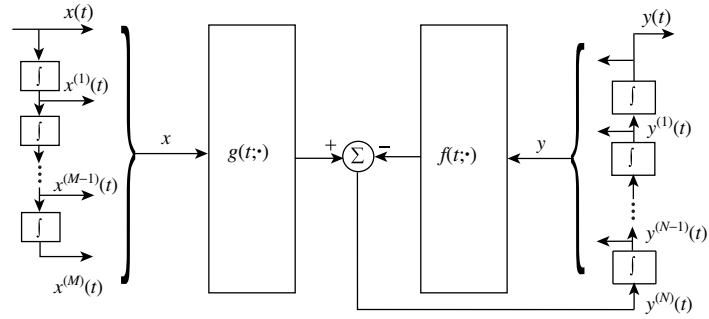
The general description for a recurrent single-input/single-output nonlinear dynamical system in Direct Form I, as depicted in Figure 6.3, is as follows.

DTP Case

$$y(k) + f(k; y) = g(k; x) \tag{6.11}$$



(a) DTP Case



(b) CTP Case

Figure 6.3 Recurrent models for single-input/single-output dynamical systems.

[3] W_N^2 is the Sobolev space of complex-valued functions, f, on I such that $f^{(i)}$, $i = 1, 2, \ldots, N-1$ are absolutely continuous and $f^{(N)} \in L^2(I)$.

where $x(k)$ and $y(k)$ denote the scalar input and output samples at instant k, and x and y are given by

$$x = (x(k), \ldots, x(k-M))^T \in E^{M+1} \tag{6.12}$$

$$y = (y(k-1), \ldots, y(k-N))^T \in E^N \tag{6.13}$$

Thus, for this case, the input data space, X, to which x belongs, is E^{M+1}, and $g(k;\cdot) : E^{M+1} \to \mathcal{C}$ is a bounded analytic functional on X. Similarly, y is the string of previous output samples from $k-1$ to $k-N$, and $f(k;\cdot) : E^N \to \mathcal{C}$ is a bounded analytic functional (VS) on E^N.

CTP Case

$$y^{(N)}(t) + f(t;y) = g(t;x), \qquad (\cdot)^{(j)} = \frac{d^j}{dt^j} \tag{6.14}$$

where $x(\cdot)$ and $y(\cdot)$ are appropriate scalar input and output functions of the continuous, t, variable, and, for convenience we use the abbreviated notation[4]

$$x = (x(t), x^{(1)}(t), \ldots, x^{(M)}(t))^T \in E^{M+1} \tag{6.15}$$

$$y = (y(t), y^{(1)}(t), \ldots, y^{(N-1)}(t))^T \in E^N \tag{6.16}$$

where $f(t;\cdot)$ and $g(t;\cdot)$ are defined as in the DTP case. Thus the space, X, to which x belongs is E^{M+1}, the set of values at time, t, of the input $x(\cdot)$ and its derivatives up to order M, together constituting the space E^{M+1}. Note also that in (6.14), $f(t;\cdot)$ is a nonlinear analytic function of the values $y^{(1)}(t), \ldots, y^{(N)}(t)$ constituting the space E^N, while in (6.10) $f_j(t;\cdot)$ represents a nonlinear convolution with the input function $x(\cdot)$.

6.2.3 State-Space Models

For a general M-input/N-output state-space model, we have, as indicated below.

DTP Case

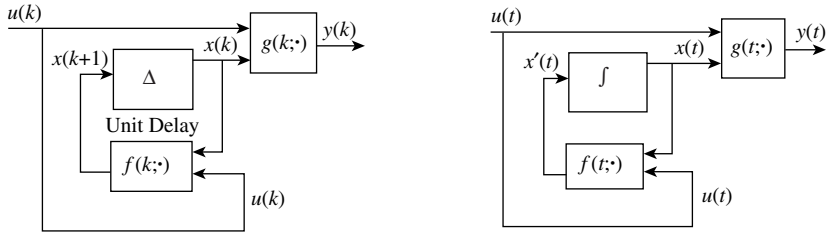
As depicted in Figure 6.4a,

$$x(k+1) = f(k;x(k),u(k)) \tag{6.17}$$

$$y(k) = g(k;x(k),u(k)) \tag{6.18}$$

where $x(k) \in E^s$, $u(k) \in E^M$, and $y(k) \in E^N$ denote the state, input, and output vectors at time k, and $f(k;\cdot,\cdot)$ and $g(k;\cdot,\cdot)$ are, respectively, s-tuples and N-tuples

[4] Note that $x(\cdot)$ needs to be sufficiently smooth by belonging to a function space such as $W_M^2(I)$ in order for the representation (6.15) to be possible.



(a) DTP Case (b) CTP Case

Figure 6.4 State-space models for multiple-input/multiple-output dynamical systems.

of bounded analytic functionals on E^{s+M}. We can label this space X for the purpose of modeling f and g.

CTP Case

As shown in Figure 6.4b,

$$\frac{dx(t)}{dt} = f(t; x(t), u(t)) \tag{6.19}$$

$$y(t) = g(t; x(t), u(t)) \tag{6.20}$$

where $f(t;\cdot,\cdot)$ and $g(t;\cdot,\cdot)$ are the same mappings as in the DTP case.

In each of the preceding cases our detailed modeling depends on the best approximation of a generic bounded analytic functional on a separable Hilbert space, X, different notations and interpretations being given for such functionals in the descriptions of the three categories of models just presented.

On this basis we now proceed to construct and study the properties of an RKHS, F, to which such functionals are made to belong and which will aid in achieving our objectives.

6.3 THE de FIGUEIREDO–DWYER–ZYLA SPACE F

6.3.1 Definition of the Space

Let X denote an abstract separable Hilbert space over $\mathcal{C}$, with the scalar product and norm in X being denoted by $\langle x; z\rangle$ and $\|x\| = \langle x, x\rangle^{1/2}$ for any x and z in X. For example, depending on the model under consideration, X could be a finite dimensional Euclidean E^N, l^2, $L^2(I)$, or $W_N^2(I)$ for some positive integer N.

Let there be given a bounded set Ω in X defined by

$$\Omega = \{x \in X : \|x\|^2 \leq \mu^2\} \tag{6.21}$$

for some positive μ, as well as a sequence of positive weights, expressing prior uncertainty in the model

$$\lambda = \{\lambda_0, \lambda_1, \ldots\} \tag{6.22}$$

satisfying

$$\sum_{n=0}^{\infty} \lambda_n \frac{\mu^{2n}}{n!} < \infty \tag{6.23}$$

Also, let $\eta_i : i = 0, 1, \ldots$ denote an orthonormal basis in X. Then a homogenous Hilbert–Schmidt (H-S) polynomial f_n of degree n in elements of X is defined by the tensor product

$$\begin{aligned} f_n &= \sum_{i_1=0}^{\infty} \cdots \sum_{i_n=0}^{\infty} c_{i_1 \cdots i_n} \eta_{i_1} \otimes \eta_{i_2} \otimes \cdots \otimes \eta_{i_n} \\ &= \sum_{i_1=0}^{\infty} \cdots \sum_{i_n=0}^{\infty} c_{i_1 \cdots i_n} \langle \eta_{i_1}, \cdot \rangle \langle \eta_{i_2}, \cdot \rangle \cdots \langle \eta_{i_n}, \cdot \rangle \end{aligned} \tag{6.24}$$

where $c_{i_1 \cdots i_n}$ are complex constants, symmetric in the indices, satisfying

$$\|f_n\|_n \triangleq \left[\sum_{i_1=0}^{\infty} \cdots \sum_{i_n=0}^{\infty} |c_{i_1 \cdots i_n}|^2 \right]^{1/2} < \infty \tag{6.25}$$

The completion X^n, under (6.25), of all homogeneous H-S polynomials of degree n in elements of X is a Hilbert space under the inner product

$$\langle f_n, g_n \rangle_n \triangleq \sum_{i_1=0}^{\infty} \cdots \sum_{i_n=0}^{\infty} c^*_{i_1 \cdots i_n} d_{i_1 \cdots i_n} \tag{6.26}$$

where $*$ denotes complex conjugation, and $d_{i_1 \cdots i_n}$ are the coefficients associated with the H-S representation of g_n.

In terms of the element of X^n

we can define $$x^n = x, x \otimes x, \ldots, x \otimes \cdots \otimes x \tag{6.27}$$

$$f_n(x) \triangleq \langle f_n, x^n \rangle_n \tag{6.28}$$

This leads to the following [4].

Definition 1 Under (6.21)–(6.23), the de Figueiredo–Dwyer–Zyla (dFDZ) space,[5] denoted by $F_\lambda(\Omega)$ or simply F, is the completion, under the norm (6.29), of the space spanned by the sequence $f = \{f_n \in X^n : n = 0, 1, \ldots\}$, satisfying

$$\|f\|_F \triangleq \left(\sum_{n=0}^{\infty} \frac{1}{n!\lambda_n} \|f_n\|_n^2 \right)^{1/2} < \infty \tag{6.29}$$

Remark 1 Clearly, the following developments hold if F stands for a closed subspace of the space F defined earlier with some of the terms in the power series missing. This may occur in some applications.

Remark 2 Belonging to F are the bounded analytic functionals on Ω expressed as VS in the form

$$f(x) = \sum_{n=0}^{\infty} \frac{1}{n!} f_n(x) \tag{6.30}$$

For the case in which $X = L^2(I), f(\cdot)$, the form of $f(t, \cdot)$, was expressed previously by (6.1), with t as an indexing variable.

If $X = E^N$, the functional $f_n(x)$, where $x = (x_1, \ldots, x_N)^T$ takes the form

$$f_n(x) = \sum_{|k|=n} c_k \frac{|k|!}{k!} x^k \tag{6.31}$$

where

$$\begin{aligned} k &= (k_1 \cdots k_N), \\ |k| &= k_1 + k_2 + \cdots + k_N, \\ k! &= k_1! \cdots k_N! \\ c_k &= c_{k_1 \cdots k_N}, \\ x^k &= x_1^{k_1} \cdots x_N^{k_N}. \end{aligned}$$

Remark 3 F constitutes a generalization of the symmetric Fock space used in the representation of non-self-interacting Boson fields in quantum field theory [52–54]. Also, Hilbert spaces of analytic functions on $\mathcal{C}^n$ have been investigated extensively (see, e.g., [55–57]). Our approach considers the more general case of functionals

[5] Previously, we called this space *generalized Fock space*.

(rather than functions) on a Hilbert space. It is based on the work of Dwyer on differential operators of infinite order [58].

Finally, F has a unique reproducing kernel that often is available in closed form, as stated in the following theorem the proof of which is given elsewhere [4,5,7].

THEOREM 1 Under the scalar product

$$\langle f, g\rangle_F = \sum_{n=0}^{\infty} \frac{1}{n!}\frac{1}{\lambda_n}\langle f_n, g_n\rangle_n \tag{6.32}$$

where F is an RKHS with the reproducing kernel

$$K(u, v) = \varphi(\langle u, v\rangle) = \sum_{n=0}^{\infty} \frac{\lambda_n}{n!}\langle u, v\rangle^n \tag{6.33}$$

that is,

$$\varphi(s) = \sum_{n=0}^{\infty} \frac{\lambda_n s^n}{n!} \tag{6.34}$$

In the special case in which

$$\lambda_n = \lambda_0^n \tag{6.35}$$

the reproducing kernel takes the form

$$K(u, v) = \exp(\lambda_0\langle u, v\rangle) \tag{6.36}$$

and so

$$\varphi(s) = \exp(\lambda_0 s) \tag{6.37}$$

6.3.2 Properties of *F*

The following three propositions follow from the theory of reproducing kernels [61]. We will use them in the solution of the modeling problem.

PROPOSITION 1 As a function of x, $\varphi(\langle v, x\rangle)$, and in particular $\exp(\lambda_0\langle v, x\rangle)$, are members of F. We express this as

$$\varphi(\langle v, \cdot\rangle) \in F \tag{6.38}$$

and in particular,

$$\exp(\lambda_0\langle v, \cdot\rangle) \in F \tag{6.39}$$

PROPOSITION 2 $\varphi(\langle v, \cdot\rangle)$ is the representer (in the sense of the Riesz representation theorem) in F of the point evaluation functional on F, i.e.,

$$\langle \varphi(\langle v, \cdot\rangle), f(\cdot)\rangle_F = f(v) \tag{6.40}$$

$\forall f \in F$.

PROPOSITION 3 Let ξ be a continuous linear functional on F. Then a representer $\xi(\langle v, \cdot\rangle) \in F$ of ξ is obtained by the action of ξ on φ with φ as a function of its adjoint argument, this being denoted by a respective subscript on ξ, that is,

$$\xi(\langle v, \cdot\rangle) = \xi_v(\varphi(\langle v, \cdot\rangle)) \tag{6.41}$$

Remark 4 For the purpose of this chapter it will be sufficient to consider the class of linear functionals on F defined in terms of bounded sequences of constants $(\alpha_0, \alpha_1, \ldots)$ by

$$\xi_v(f) = \sum_{n=0}^{\infty} \frac{\alpha_n}{n!} f_n(v) \tag{6.42}$$

where f_n is as in (6.28) and (6.30).

By considering $\varphi(\langle v, x\rangle)$ as an element of F in terms of v, with x a fixed parameter, we have, according to (6.42), and (6.33),

$$\xi_v(\varphi(\langle v, x\rangle)) = \sum_{n=0}^{\infty} \frac{\alpha_n}{n!} \lambda_n \langle v, x\rangle^n \tag{6.43}$$

and hence, according to (6.33) and (6.42), (6.43) gives (6.41) explicitly, i.e.,

$$\langle \xi_v(\varphi(\langle v, \cdot\rangle)), f(\cdot)\rangle_F = \sum_{n=0}^{\infty} \frac{\alpha_n}{n!} f_n(v) \tag{6.44}$$

6.4 DERIVATION OF SIGMOID FUNCTIONALS

Sigmoid functions[6] [37] play an important role in computationally intelligent systems. In our formulation, they are *representers* of linear observation or specification

[6] A sigmoid functional is a composition of two maps: a nonlinear function and a scalar product in X. The first map is called a *sigmoid function*.

functionals on F. Hence their modeling is application specific. We derive expressions for such representers ((6.47), (6.50), (6.51), (6.54), (6.55)) for five important cases. We denote by superscripts the labels of the corresponding sigmoids and illustrate their well-known characteristics for some of them in Figure 6.5. In these expressions, the parameter λ_0 determines the reproducing kernel (RK) of F and hence the metric of F. So in a given application, by adjusting λ_0 one may make this metric match the prior mode uncertainty expressed by the RK.

6.4.1 Exponential Activation Functional

The exponential activation functional is the point evaluation functional ξ_v^p on F, i.e.,

$$\xi_v^p(f) = f(v) \tag{6.45}$$

and according to Proposition 1, its representer is

$$\xi^p(\langle v, \cdot \rangle) = \varphi(\langle v, \cdot \rangle) \tag{6.46}$$

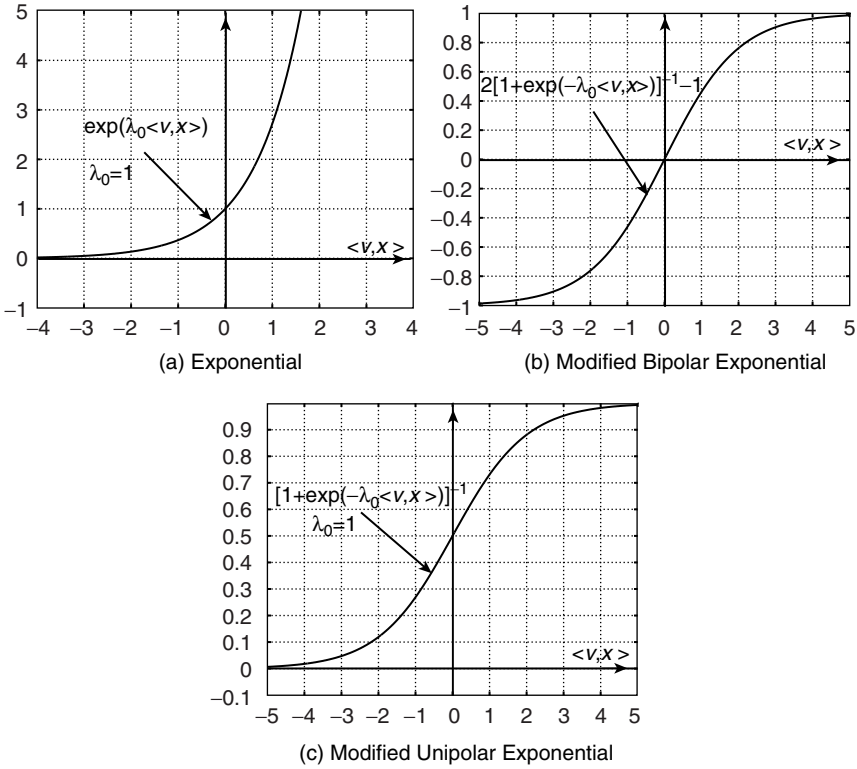


Figure 6.5 Examples of sigmoid functions.

and in the case of (6.37)

$$\xi^{p}(v,\cdot) = \exp(\lambda_0\langle v,\cdot\rangle) \tag{6.47}$$

Note that the constants α_n in (6.42) are equal to 1 for this functional.

6.4.2 Modified Exponential Sigmoid Functional

This functional corresponds to the unipolar activation functional

$$g_{eu}(x) = \frac{1}{1+\exp(-\rho\langle v,x\rangle)} \tag{6.48}$$

where ρ is a scaling parameter and v is an appropriate element of X.

This functional corresponds to the one defined by (6.42) with the constants α_n:

$$\alpha_n = \frac{1}{2}(-1)^n E_n(0) \tag{6.49}$$

where $E_n(0)$ denotes the coefficient of the zeroth-order term of the nth-degree Euler polynomial, i.e., α_n is the zeroth-order nth-degree Euler number.

With this agreement, the representer in F for this functional is as in (6.42) with the α_n as in (6.49), and in the special case of the exponential reproducing kernel (6.37),

$$\xi^{eu}(\langle v,\cdot\rangle) = \frac{1}{1+\exp(-\lambda_0\langle v,\cdot\rangle)} \tag{6.50}$$

where the a priori uncertainty weight λ_0 in F corresponds to the scaling parameter ρ in (6.48).

In a similar way, one can derive the expression for the representer of the bipolar-modified exponential sigmoid functional

$$\xi^{eb}(\langle v,\cdot\rangle) = \frac{2}{1+\exp(-\lambda_0\langle v,\cdot\rangle)} - 1 \tag{6.51}$$

6.4.3 Hyperbolic Tangent Sigmoid Functionals

To obtain the representer in F corresponding to the bipolar hyperbolic tangent activation functional

$$g_{tb}(v) = \tanh(\rho\langle v,x\rangle) \tag{6.52}$$

where ρ and v are the same as defined in connection with (6.48), assume that in (6.42) the even power coefficients are zero. Then the desired representer is obtained by letting the odd power coefficients in (6.42) be

$$\alpha_{2n-1} = \frac{2^{2n}(2n-1)}{n} B_{2n}, \qquad n = 1, 2, \ldots \tag{6.53}$$

where B_{2n} denotes the Bernoulli [62] number of degree $2n$. The corresponding representer, in the special case of the exponential reproducing kernel (6.37), is

$$\begin{aligned} \xi^{tb}(\langle v, \cdot \rangle) &= \sum_{n=1}^{\infty} \frac{1}{(2n-1)!} \alpha_{2n-1} (\lambda_0 \langle v, \cdot \rangle^{2n-1}) \\ &= \tanh(\lambda_0 \langle v, \cdot \rangle) \end{aligned} \tag{6.54}$$

where the scaling weight λ_0 is equal to ρ in (6.52).

In a similar way, we can obtain the representer of the unipolar hyperbolic tangent functional in F:

$$\xi^{tu}(\langle v, \cdot \rangle) = \frac{1}{2}[1 + \tanh(\lambda_0 \langle v, \cdot \rangle)] \tag{6.55}$$

6.5 BEST ROBUST APPROXIMATION OF *f* IN THE NEURAL SPACE $\mathcal{N}$

We now introduce the neural space $\mathcal{N}$ by way of the following.

Definition 2 For a given positive integer, m, and space, F, the neural space $\mathcal{N}$ is the Hilbert space of m-tuples $f = (f_1, \ldots, f_m)$, with $f_j \in F$, $j = 1, \ldots, m$, with scalar product and norm in $\mathcal{N}$ for any f and g in $\mathcal{N}$ defined by

$$\langle f, g \rangle_{\mathcal{N}} = \sum_{j=1}^{m} \langle f_j, g_j \rangle_F \tag{6.56}$$

and

$$\|f\|_{\mathcal{N}} = [\langle f, f \rangle_{\mathcal{N}}]^{1/2} \tag{6.57}$$

Remark 5 The space, $\mathcal{N}$, is the direct product of the spaces F, i.e.,

$$\mathcal{N} = \overbrace{F \times F \times \cdots \times F}^{m}$$

so the members of $\mathcal{N}$ having a common domain X. On this basis, (6.56) and (6.57) make sense. We will show that members of $\mathcal{N}$ are optimally implemented as neural

networks. Therefore the scalar product (6.56) measures the similarity between the two neural networks that the maps f and g represent, and the norm (6.57) when used as $\|f-g\|_{\mathcal{N}}$ expresses a metric distance between these two networks.

We now state the following best robust approximation problem of a nonlinear map $f \in \mathcal{N}$ (with the notation $f(x)=y$) based on an ellipsoidal prior uncertainty model in $\mathcal{N}$ and a set of q observations or specifications' constraints on f. Even though f may be a component of a larger dynamical system, such as those described in Section 6.2, we call, for convenience, its domain and range spaces, input and output spaces.

PROBLEM 1 Let there be given the input–output data pairs

$$(x^i \in X, y^i \in E^m), \qquad i=1,\ldots,q \tag{6.58}$$

where $x^i, i=1,\ldots,q$ are linearly independent, and a set of q functionals of the type (6.42) with the representers in F of the form (6.43) expressed by

$$\xi_{x^i}(\varphi(\langle x^i,\cdot\rangle)), \qquad i=1,\ldots,q \tag{6.59}$$

with regard to which f satisfy input–output data constraints that confine f to the set

$$\begin{aligned}\Phi = \{f \in \mathcal{N} : \langle \xi_{x^i}(\varphi(\langle x^i,\cdot\rangle)), f_j\rangle_F = y^i_j,\\ i=1,\ldots,q,\ j=1,\ldots,m\}\end{aligned} \tag{6.60}$$

and assume that f lies on a prior uncertainty ellipsoidal set in Ω

$$\Gamma = \{f \in \mathcal{N} : \|f\|_{\mathcal{N}} \le \gamma\} \tag{6.61}$$

for some $\gamma > 0$ sufficiently large so that the set

$$\chi = \Phi \cap \Gamma \tag{6.62}$$

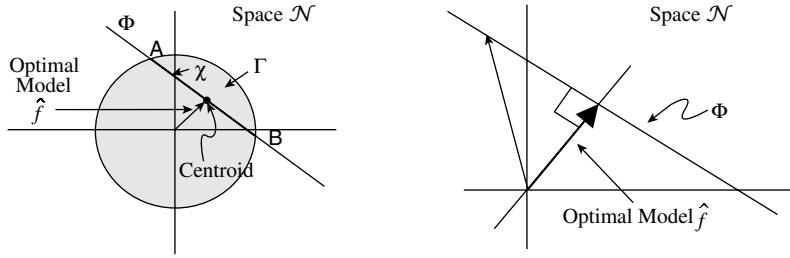
is nonempty.

Find the *best robust approximation* $\hat{f}$ of f as the solution of the min-max optimization problem

$$\sup_{\tilde{f}\in\chi} \|\hat{f}-\tilde{f}\|_{\mathcal{N}} \le \sup_{\tilde{f}\in\chi} \|f-\tilde{f}\|_{\mathcal{N}} \qquad \forall\, f \in \chi \tag{6.63}$$

Remark 6 Figure 6.6 provides a geometrical illustration of the sets ϕ, Γ, and χ in $\mathcal{N}$, where ϕ is a hyperplane, Γ an ellipsoid, and χ a subset of ϕ. It is also clear from this figure that the point in χ for which the maximum distance from all other points in χ is minimum is the centroid of χ, which thus corresponds to the solution, $\hat{f}$, of the minimum norm problem

$$\min_{f\in\chi} \|f\|_{\mathcal{N}} \tag{6.64}$$



(a) Problem Formulation and Solution (b) Optimization Problem

Figure 6.6

Remark 7 Since, according to (6.56) and (6.57),

$$\|f\|_{\mathcal{N}} = \left[\sum_{j=1}^{m} \|f_j\|_F^2\right]^{1/2} \tag{6.65}$$

minimization of $\|f\|_{\mathcal{N}}$ is achieved through the minimization of the individual $\|f_j\|_F$, $j = 1, \ldots, m$.

The preceding two remarks explain the validity of the following theorem, a formal proof of which is given elsewhere [1,2,7].

THEOREM 2 Problem 1, expressed by (6.63), has a unique solution, $\hat{f}$, which is the solution of the minimum norm problem

$$\min_{f \in \chi} \|f\|_{\mathcal{N}} \tag{6.66}$$

Each component $\hat{f}_j$ of the solution is the unique vector belonging to the subspace of F spanned by the representers $\xi_{x^i}(\varphi\langle x^i, \cdot\rangle)$ satisfying the interpolating constraints (6.60). This leads to the closed-form expression for $\hat{f}$:

$$\hat{f}(x) = \mathbf{W}^T K(x) \tag{6.67}$$

where $\mathbf{W} = q \times m$ matrix and $K(\cdot)$, a q-dimensional vector, are computed as follows

$$K(x) = (\xi_{x^1}(\varphi(\langle x^1, x\rangle)), \ldots, \xi_{x^q}(\varphi(\langle x^q, x\rangle)))^T \tag{6.68}$$

$$\mathbf{W} = \mathbf{G}^{-1}\mathbf{Y}^T \tag{6.69}$$

where $\mathbf{G}$ is the $q \times q$ matrix with elements

$$\begin{aligned} G_{ij} &= \langle \xi_{x^i}(\varphi(\langle x^i, \cdot\rangle)), \xi_{x^j}(\varphi(\langle x^j, \cdot\rangle))\rangle_F \\ &= \langle \psi(\langle x^i, \cdot\rangle), \psi(\langle x^j, \cdot\rangle)\rangle_F \\ &= \psi(\langle x^i, x^j\rangle) \end{aligned} \tag{6.70}$$

and $\mathbf{Y}$ is the $m \times q$ matrix

$$\mathbf{Y} = (y^1, \ldots, y^q) \tag{6.71}$$

A bound for the residual error ξ is

$$\xi \leq \gamma - YG^{-1}Y^T = \alpha \tag{6.72}$$

Remark 8 Without prior assumption that the solution to the optimization Problem 1 has a neural structure, the optimal solution is in the form of a feedforward two-layer abstract neural network, called by us an *optimal interpolating* (OI) net, depicted in Figure 6.7. In this network the $N \times q$ synaptic weight matrix,[7] X, for the first layer of the network is obtained from the set of input vectors (called *exemplary inputs* or simply *exemplars*)

$$X = (x^1, x^2, \ldots, x^q) \tag{6.73}$$

Therefore, if the input space is E^N, X in (6.73) is an $N \times q$ matrix and X_{ij} is the synaptic weight from the ith input node to the jth node of the first internal layer, as shown in Figure 6.7a. If that space is $L^2(I)$, X_{ij} is a "functional synaptic weight" between the entire x^i signal and the jth node of the first internal layer, as depicted in Figure 6.7b. Functional artificial neural networks, without being called this, were first introduced in [6] and further discussed in [9–10].

Remark 9 The solution provided by Theorem 1 permits one to simultaneously extract optimal structure and the optimal set of parameters that belong to it. The process of this acquisition is called *intelligent learning* vis-a-vis other types of learning based on numerical optimization algorithms like the Amari [36] and Hebbian [37] learning ones. Algorithms for "instantaneous learning" (by obtaining (6.67) with all the exemplars included) and "sequential (adaptive/evolutionary) learning" by (a recursive least-square (RLS) procedure) have been presented and discussed at length by the author in collaboration with Sin [11–13]. Additional comments appear in Section 6.7.3.

Remark 10 The following five comments are of particular relevance:

1. *Point Evaluation Functionals* In the case in which the y^j, $j = 1, \ldots, q$, represent outputs rather than more general observations/specifications described by (6.42), then according to Proposition 2, the representers ξ_x^i are simply

$$\xi_x^i(\varphi(\langle x^i, \cdot \rangle)) = \varphi(\langle x^i, \cdot \rangle) \qquad i = 1, \ldots, q \tag{6.74}$$

[7] Even though we used X to denote the input data space, for convenience we use the same notation for the matrix (6.73), its meaning being clear from the context.

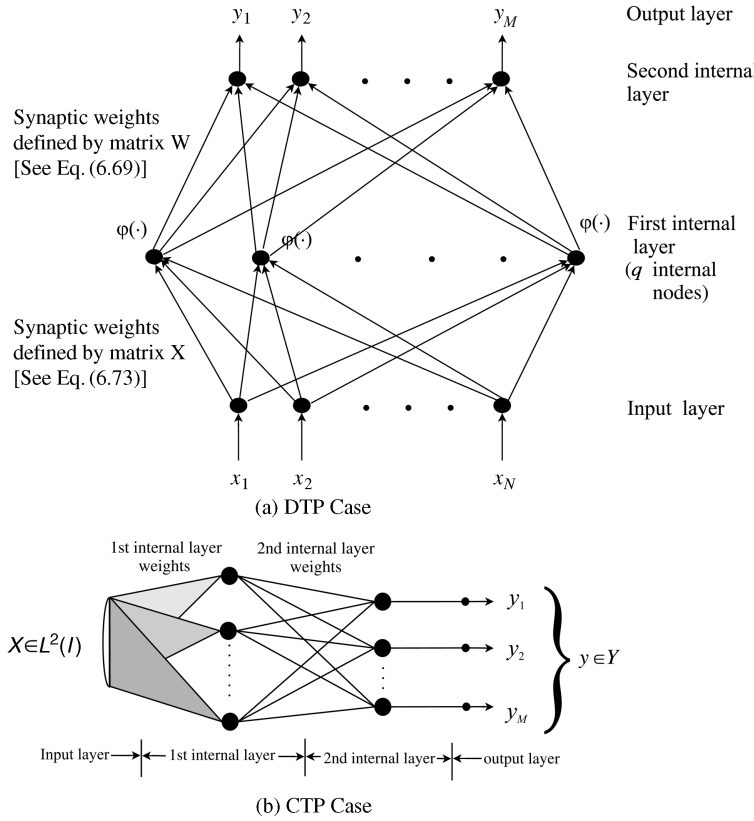


Figure 6.7 Signal-flow graphs for the OI net. (a) DTP case, (b) CTP case.

and (6.68) and (6.70) take the form

$$K(x) = (\varphi(\langle x^1, x\rangle), \ldots, \varphi(\langle x^q, x\rangle))^T \tag{6.75}$$

$$G_{ij} = \varphi\langle x^i, x^j\rangle \tag{6.76}$$

2. *Rationale for Shifted Sigmoids Model Representation* One can interpret the model (6.67) in terms of shifted sigmoids of the type considered in Section (6.4). For simplicity in presentation, we will consider the case in which $m = 1$ (single output), use the abbreviated notation for the representers in (6.59) and (6.68).

$$\xi_{x^i}(\varphi\langle x^i, x\rangle) = \psi(\langle x^i, x\rangle) \tag{6.77}$$

and denote by r_i the shift in the ith sigmoid.

The following cases are of interest:

(a) *Exponential Sigmoid* (*6.47*) In this case, shifts correspond to the scaling of coefficients of unshifted sigmoids. The shifts are automatically taken into account by our procedure according to:

$$\begin{aligned}\exp(\lambda_0(\langle x^i, x\rangle - r_i)) &= \exp(-\lambda_0 r_i)\exp(\lambda_0\langle x^i, x\rangle) \\ &= A_i \exp(\lambda_0\langle x^i, x\rangle) \qquad i = 1, \ldots, q \end{aligned} \tag{6.78}$$

(b) *Other Sigmoids* ((*6.50*), (*6.51*), (*6.54*), (*6.55*)) In this case, sigmoid shifts can be interpreted and taken into account in one of the following two ways.

(**i**) They can result from an offset x^0 in the input signal x, i.e.,

$$\psi(\langle x^i, x\rangle - r_i) = \psi(\langle x^i, x - x^0\rangle) \tag{6.79}$$

where

$$r_i = \langle x^i, x^0\rangle, \qquad i = 1, \ldots, q \tag{6.80}$$

(**ii**) More fundamentally, one would like to represent the model, f, as a sum of abstract power series (Taylor series) around p points $\tilde{x}^k \in X$, $k = 1, \ldots, p$, rather than a single power series around the origin (McLaurin series), as we have done thus far. Typically, the points $\tilde{x}^k$ could be cluster centers of the set of exemplars. One would then construct a new space, F, as a direct sum of spaces, F_k, i.e.,

$$F = F_1 \oplus F_2 \oplus \cdots \oplus F_p \tag{6.81}$$

each F_k being constructed as was F before, except that the space, X, for F_k would be centered at $\tilde{x}^k$.

From such a development it follows that (with $m = 1$)

$$\begin{aligned}\hat{f}(x) &= \sum_{k=1}^{p}\sum_{i=1}^{q} \tilde{w}_{ik}\psi(\langle x^i - \tilde{x}^k, x - \tilde{x}^k\rangle) \\ &= \sum_{k=1}^{p}\sum_{i=1}^{q} \tilde{w}_{ik}\psi(\langle v^{ik}, x\rangle - r_{ik}) \end{aligned} \tag{6.82}$$

where

$$v^{ik} = x^i - \tilde{x}^k \tag{6.83}$$

$$r_{ik} = \langle v^{ik}, \tilde{x}^k\rangle \tag{6.84}$$

Thus, according to our formulation, one would determine $\tilde{x}^k$ from the training data, and obtain the parameters v^{ik} and r^{ik} from (6.83) and (6.84). By

arranging the subscripts ik lexicographically, we relabel them in terms of a single indexing variable $l = 1, \ldots, L$, where $L = pq$. With this notational agreement, we rewrite (6.82) using the subscript j to cover the case of multiple outputs corresponding to $j = 1, \ldots, m$, in the form

$$\hat{f}_j(x) = \sum_{l=1}^{L} w_{lj} \psi(\langle v^l, x \rangle - r_l) \tag{6.85}$$

Here w_{lj} denotes the weight $\tilde{w}_l = \tilde{w}_{ik}$ for the jth output. By substituting (6.85) in (6.69), taking the preceding notational agreement into account, the desired optimal model (6.67) based on shifted sigmoids can be obtained.

Comment Barron and others (see e.g., [50]) have studied the property of approximation of a known function by a linear combination of shifted sigmoids. Their problem is clearly different from the one we have considered to be best, approximating an unknown function, f, based on the training data, an uncertainty model for f, and an appropriate space where f may reside. Sigmoids appear as a possible consequence rather than a cause according to our formulation.

3. *Rationale for Radial Basis Functions Model Representations* Another popular scheme for modeling artificial neural systems is that based on the RBF [38]. We now show how this scheme fits our model (6.67). Assume that the space, F, consists of VS on $X \times X$, where[8] $X = E^N$. Then using the exponential reproducing kernel (6.37) and the functional (6.42) with $\alpha_n = (-1)^n$, we have for (6.43)

$$\psi(\langle I, x \times x \rangle) = \exp(-\lambda_0 \|x\|^2) \tag{6.86}$$

where I denotes the $N \times N$ diagonal matrix

$$I = \mathrm{Diag}(1, \ldots, 1) \tag{6.87}$$

Constructing now a new space, F, in a way analogous to (6.81), we are led to

$$\hat{f}(x) = \sum_{k=1}^{p} \sum_{i=1}^{q} \tilde{w}_{ik} \exp(-\lambda_0 \|x - \tilde{x}^k\|^2) \tag{6.88}$$

where $\tilde{x}^k$, $k = 1, \ldots, p$, are the vectors around which the VS expansions for F_k, $k = 1, \ldots, p$ occur. The coefficients $\tilde{w}_{ik}$, $k = 1, \ldots, p$, are obtained by interpolating the expression (6.88) at the exemplars using the formulas (6.68)–(6.71) with appropriate interpretation of the notation.

Note that, as indicated under (b)(ii), the $\tilde{x}^k$, $k = 1, \ldots, p$, constitute a set of fiducial points extracted from the training (such as cluster centers or some

[8] X Here denotes the input space.

exemplars themselves) that best represent the structure of the training set for the purpose just explained.

4. *The deF Dimension* The bound α on the residual error (6.72) enables us to define a criterion that we denote by $deF(\alpha)$, which measures an *intrinsic dimensionality* of the OI net, in terms of the minimum number of neurons in its first layer to achieve correct classification of all exemplars. Specifically, we define $deF(\alpha)$ as the minimum number of exemplary pairs (x^i, y^i) needed to keep the uncertainty error in (6.72) below a prescribed α.
5. *Optimal Solution with Noisy Data* If the output data are noisy, i.e.,

$$y_j^i = z_j^i + v_j^i, \qquad i = 1, \ldots, q, \qquad j = 1, \ldots, m$$

where z_j^i, $i = 1, \ldots, q$, constitute the nonnoisy component of the data vector y_j, and v_j^i, $i = 1, \ldots, q$, are the corresponding components of an additive WGN vector v_j with zero-mean and covariance

$$R_j = \text{Diag}(\rho_{1j}, \ldots, \rho_{qj}) \tag{6.89}$$

there are a number of ways one can formulate the approximation problem. One of the simplest ways is to note that the optimal solution lies in the span of $\psi(x^i, \cdot)$, $i = 1, \ldots, q$, i.e.,

$$f_j(x) = \sum_{i=1}^{q} w_{ij} \psi(x^i, x) \tag{6.90}$$

and obtain the desired model as the solution to the penalized optimization problem

$$\min \left\{ \beta \| f_j(\cdot) \|_F^2 + \sum_{i=1}^{q} \rho_{ij}^{-1} (\langle \psi_i(\cdot), f_j(\cdot) \rangle_F - y_j^i)^2 \right\} \tag{6.91}$$

where β is a positive constant to be chosen by the modeler. A small value of β expresses fidelity to the observed data at the expense of smoothness.

Substituting (6.90) and (6.70)–(6.71) in (6.91), we get

$$\beta w_j^T G w_j + \sum_{i=1}^{q} \rho_{ij}^{-1} \left(\sum_{n=1}^{q} G_{in} w_{nj} - y_j^i \right)^2 \tag{6.92}$$

where

$$w_j = (w_{1j}, \ldots, w_{qj})^T \tag{6.93}$$

By differentiating (6.92) partially with respect to each w_{ij} and setting the result equal to zero, we obtain

$$\hat{w}_j = (\beta I + R_j^{-1}G)^{-1}R_j^{-1}y_j \tag{6.94}$$

and

$$\hat{f}_j(x) = y_j^T R_j^{-1}(\beta I + R_j^{-1}G)^{-1}K(x), \qquad j = 1, \ldots, m \tag{6.95}$$

We call the two-layer neural network represented by (6.95) an OS network.

6.6 OPTIMAL COMBINED STRUCTURAL AND PARAMETRIC MODELING OF NONLINEAR DYNAMICAL SYSTEMS IN $\mathcal{N}$

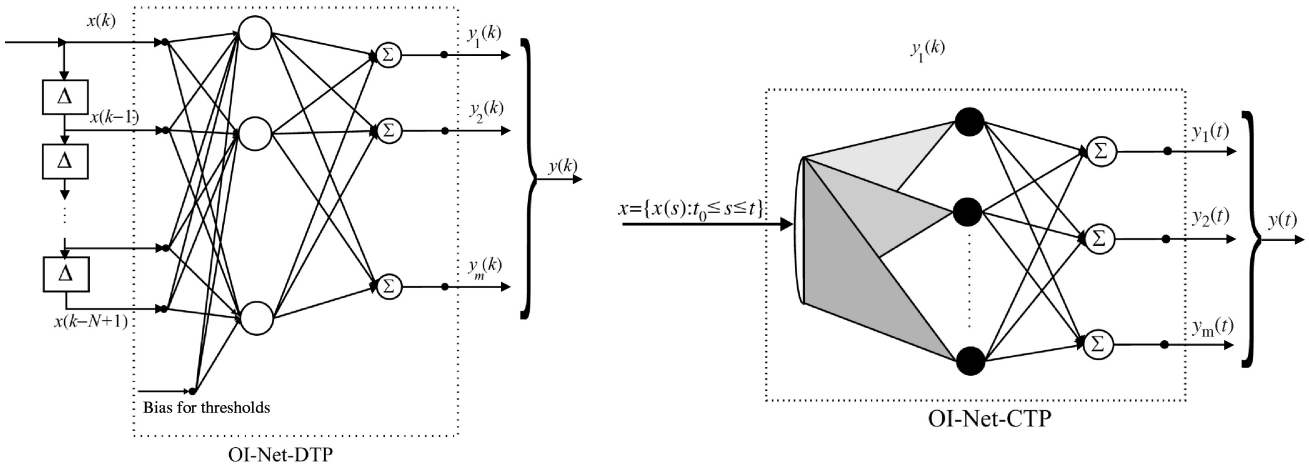
Based on the developments of the preceding section, it is now possible to obtain optimal structural and parametric realizations of the three classes of generic models described in Section 6.2. For this purpose, in each case one picks each block described by a nonlinear functional such as f or g in $\mathcal{N}$ as a feedforward artificial neural system that may constitute the desired solution, or, after combination with other blocks realized in a similar manner, may lead to an overall system that is a recurrent or state-space realization. These procedures are explained graphically for the DTP and CTP cases for feedforward, recurrent, and state-space models in Figures 6.8–6.10, and are clear enough so as not to require further discussion.

Depending on the application, the generic two-layer artificial neural system modules of OI or OS nets can be assembled to produce larger and more complex models. Each module may have an additional Winner–Take–All layer in tandem with it for the purpose of decision as explained in the following section.

A generic feedforward $2n$ layer net called **OMNI** net is shown in Figure 6.11. If the modules are OS nets, the resultant multilayer net is called **OSMAN**. Both OMNI and OSMAN are multilayer perceptrons, the structure and parameters of which are obtained using the framework developed in this chapter. Algorithms extending the procedure (6.67)–(6.71) to this case are described in [8].

Through various interconnections of OI and/or OS modules, a nonlinear dynamical system of any required complexity can be modeled, including, of course, systems with feedback. As an example, Figure 6.12 illustrates a proposed OMNI net implementation of a cortical column of the primary visual cortex of a cat. The connectionist structure was obtained by Bolz, Gilbert, and Wisel [63,64] from pharmacological experiments. Figure 6.12 shows the six OI nets that are used to implement the six layers (regions) in the column and the interconnections.

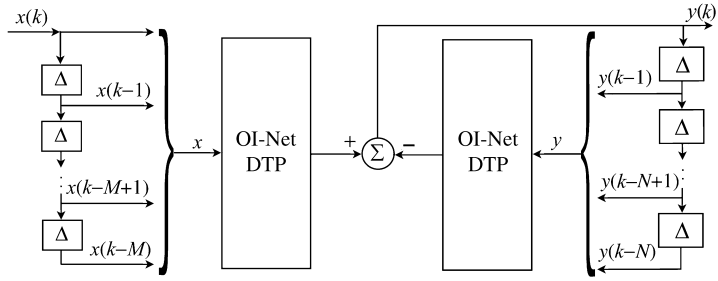
For the Wiener–Bose (WB) model, we have proposed a modification [10,65] using an OI net (Fig. 6.13). In this approximation, the linear (Laguerre) part of the WB model is preserved and provides ''functional'' synaptic weights to the first layer. The activations for this layer are provided by the Hermite part of the WB model, and the second layer of the model is linear and consists of a single neuron.



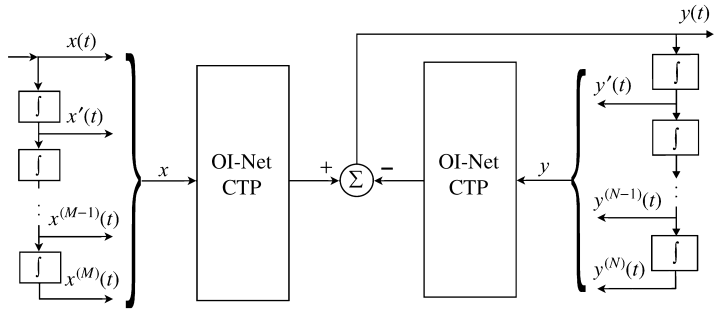
(a) DTP Case

(b) CTP Case

Figure 6.8 Feedforward models for single-input/multiple-output dynamical systems.



(a) DTP Case

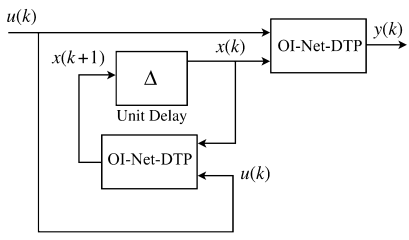


(b) CTP Case

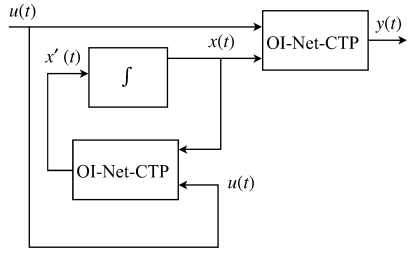
Figure 6.9 Recurrent models for single-input/single-output dynamical systems.

6.7 COMPUTATIONALLY INTELLIGENT (CI) SYSTEMS

The concepts and methodology developed thus far in the present chapter can be of value in the identification and realization of computationally intelligent (CI) systems. In what follows, we present a characterization and follow-up to the previous developments oriented toward this goal.



(a) DTP Case



(b) CTP Case

Figure 6.10 State-space models for multiple-input/multiple-output dynamical systems.

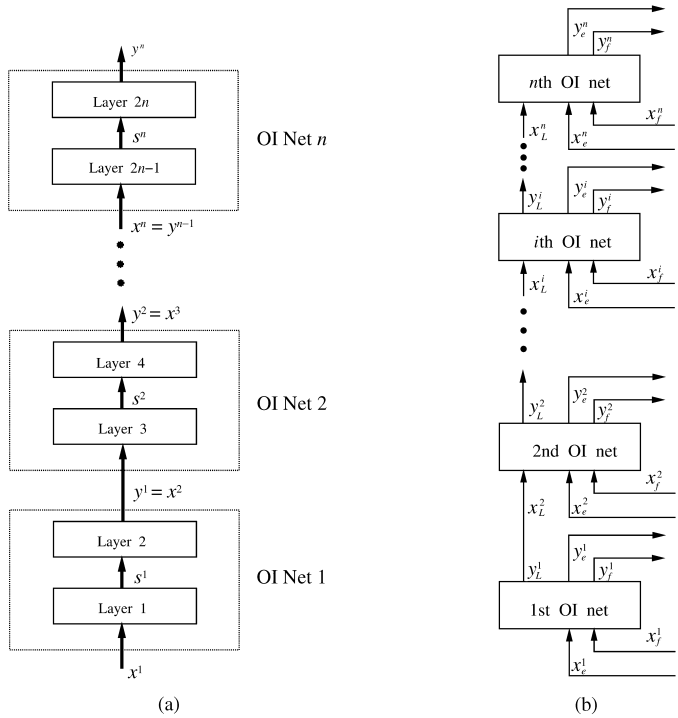


Figure 6.11 A diagram of a general $2n$-layer OMNI net with feedforward, feedback, and external connections at every two-layer level (the figure depicts the most general situation).

6.7.1 CI Systems as Mixed Systems

Computationally intelligence usually involves functions that perform decisions regarding simple or complex events present in the data acquired by some sensing system. For this reason, CI systems are best modeled as *mixed continuous/discrete systems*: *continuous* with regard to the computation of a score (or event membership value in the case of a fuzzy CI system[9] [66–68]) on the basis of which a decision is made; and *discrete* in the representation of the event that needs to be detected, classified, or interpreted.

In the case of detection/classification, that is, of mapping a sensed vector x into one of m hypotheses $H_1, \ldots, H_m$ of the occurrence of m possible events, the jth output, $j = 1, \ldots, m$ of an m-output, CI system will be 1 or 0, depending on whether H_j is or is not true. In the case of interpretation, the outputs will be appropriate arrays of 1s and 0s, corresponding to graphs expressing the various interpretations in a given language.

[9] Due to space limitations, we do not discuss applications with appropriate modifications and interpretations to fuzzy neural networks and systems.

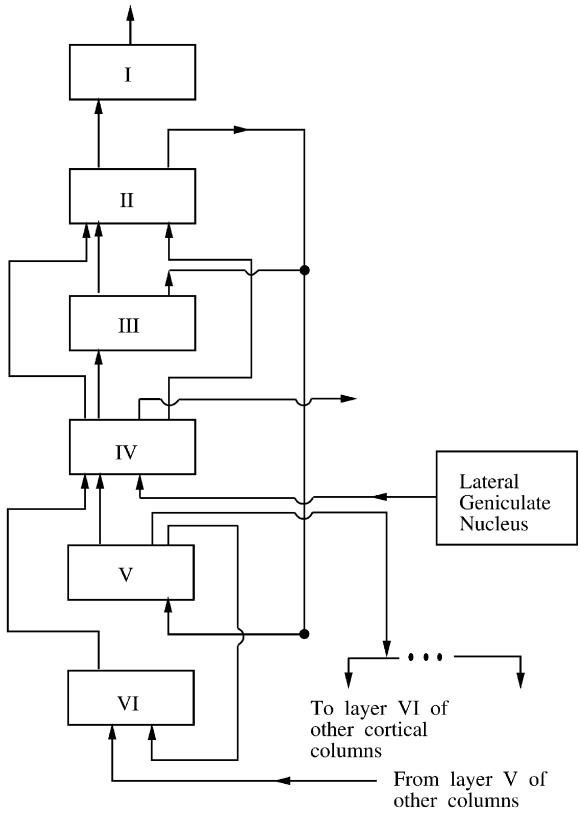


Figure 6.12 Proposed OMNI model for primary visual cortex of a cat.

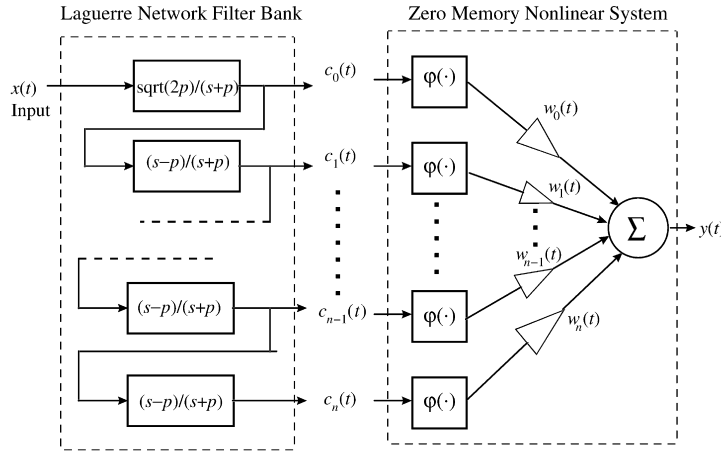


Figure 6.13 Robust best approximation to the Wiener model. Coefficients $w_i(t)$ may be generated recursively using the matrix inversion lemma. (*Note*: sqrt$(\cdot)$ denotes the square root operator, and the blocks on the left side denote transfer functions.)

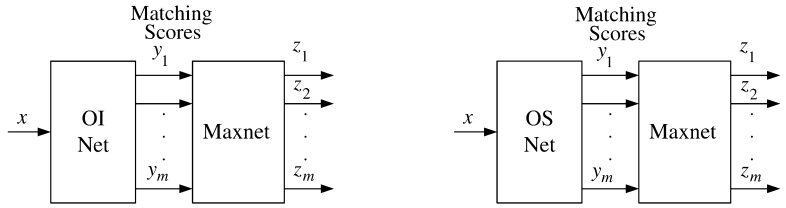


Figure 6.14 Mixed net modules for CI systems.

6.7.2 MOI and MOS Nets

At a basic level, the scheme just described can be incorporated into a MOI net, which is a 3-layer net consisting of an OI net (defined previously) in tandem with a Maxnet that incorporates a Winner-Take-All (WTA) decision based on the scores provided by the OI net. As depicted in Figure 6.14, the OI net maps the input data vector x into the scores y_i, $i = 1, \ldots, m$ provided by its m inputs. These in turn feed into the Maxnet, which selects the output of the OI net with the highest score (see [2]). The Maxnet is a m-input/m-output system, for which the input is a score vector $y = (y_1, \ldots, y_m)^T$, and the output is a binary vector $z = (z_1, \ldots, z_m)^T$, the components of z being all zero, except one, say z_k, corresponding to the component y_k of y having the highest value (score). A (MOS) net may be defined and used in a similar manner (see Fig. 6.14b). Appropriate combinations of MOI and MOS nets can be utilized to identify or simulate complex decision systems or realize new ones.

6.7.3 Intelligent Learning as Combined Adaptation and Evolution

The approach presented here lends itself naturally to a process of *combined adaptation and evolution*. This is what we call *intelligent learning* vis-a-vis conventional learning. Conventional learning is implemented as a strictly numerical optimization algorithm like the error back-propagation algorithm.

The underlying training procedure in our approach is explained for the MOI net shown in Figure 6.15. In this figure, the first layer of the net consists of neurons representative of the exemplars, the properties of which are to be retained by the net. These exemplars are called *prototypes*. Thus the synaptic weights connecting to the ith neuron of this first layer are the components of x^i, assuming that x^i is an exemplar that has been retained as a prototype.

The second-layer neurons correspond to the m hypotheses/events, into one of which the input vector is to be classified and the synaptic weights for this layer are calculated using (6.67)–(6.71).

The training process begins by inserting the first neuron in the first layer of the net with its synaptic weights consisting of the components of x^1 and the second-layer weights being calculated so as to enable its output vector to be the vector, y^1, associated with the first exemplary pair. As this process continues, and a new

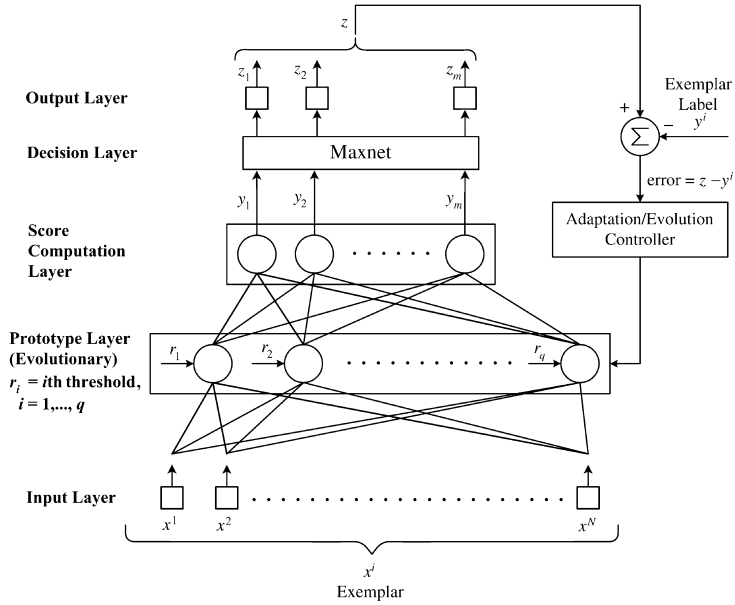


Figure 6.15 Intelligent learning: combined adaptation/evolution process.

exemplar, say x^k, is presented to the net, it gives rise to an output y that the Maxnet maps into a vector, z. This z is then compared to the correct exemplary output vector,[10] y^k. If they agree, the exemplar, x^k, is laid aside and the next exemplar is presented to the net so that the process can continue as before. If z and y^k do not agree, the error signal $z - y^k$ from the subtractor shown in the figure is applied to an adaptation–evolution controller that adds a neuron to the first layer corresponding to x^k as an additional prototype and adjusts the synaptic weights of the second layer accordingly. This process of prototype addition corresponds to evolution, and the one of adjusting the weights of the second layer to adaptation. This entire procedure is based on an algorithm described by Sin and de Figueiredo [13].

The preceding training procedure converges, classifying all exemplars correctly after recycling through the training set a few (typically two or three) times. During this recycling, exemplars that have been laid aside are put back in the training set after each cycle to make sure that all the exemplars have been taken into account.

Note that the number of exemplars retained as prototypes (number of neurons in the first layer) depend on the order in which they have been presented to the net. From the author's experience, a very small number close to $deF(\alpha)$ results when the prototypes are close to the boundaries of the decision region. For a number of examples, see [13] and references therein.

[10] As indicated under 6.7.1, for any given pair $(x^i, y^i), i = 1, \ldots, q$, the components $y^i_j, j = l, \ldots, m$, are binary and $y^i_j = l$ if and only if x^i belongs to H_j.

6.7.4 Discovery

Our formulation allows one to model the process of ''discovery'' by a neural system, which we will explain, for simplicity, in terms of an MOI net. After the training of a neural system is completed according to the preceding section, some vectors in the new data being received may not classify properly. If a significant number of such outliers appear, the cluster-detection-and-labeling (CDL) network recently developed by us [21] can be applied to such a set of outliers. The clusters obtained by the CDL network can then be considered to constitute training sets for corresponding new classes, and additional neurons added to the second layer of the net by training it with the exemplars from these clusters. As a consequence, we can say that the net discovered those new classes from its experience with the new data. This encapsulates the concept of *discovery* by an artificial or natural neural system, according to our formulation.

6.8 CONCLUDING REMARKS

The framework presented here is especially useful in applications where the dynamics of the generation, processing, and/or delivery of data is nonlinear. Applications of this approach have been made to a number of problems, including nonlinear adaptive time-series prediction and nonlinear equalization in communication channels, sonar signal analysis and detection, and neuroscience [6,12,13,15,17, 20–23]. Limitations in space did not permit us to discuss them here. The potential of this technology is enormous for further applications in many fields, including wireline, cable, fiber, and wireless communications, automated manufacturing, and medical diagnostics and treatment.

REFERENCES

1. R. J. P. de Figueiredo. ''A New Nonlinear Functional Analytic Framework for Modeling Artificial Neural Networks.'' In *Proceedings of the IEEE International Symposium on Circuits and Systems*, New Orleans, LA, May 1990, pp. 723–726.
2. R. J. P. de Figueiredo. ''An Optimal Matching-Score Net for Pattern Classification.'' In *Proceedings of 1990 International Joint Conference on Neural Networks, IJCNN-90*, San Diego, CA, June 1990.
3. V. Volterra. *Theory of Functionals and of Integral and Integro-Differential Equations*. Dover, New York, 1959.
4. R. J. P. de Figueiredo and T. A. W. Dwyer III. ''A Best Approximation Framework and Implementation for Simulation of Large-Scale Nonlinear Systems.'' *IEEE Trans. Circuits Syst.*, vol. CAS-27, no. 11, 1005–1014, 1980.
5. R. J. P. de Figueiredo. ''A Generalized Fock Space Framework for Nonlinear System and Signal Analysis.'' *IEEE Trans. Circuits Syst.*, vol. CAS-30, no. 9, 637–647, 1983.
6. L. V. Zyla and R. J. P. de Figueiredo. ''Nonlinear System Identification Based on a Fock Space Framework.'' *SIAM J. Control Optim.*, 931–939, Nov. 1983.

7. R. J. P. de Figueiredo. "Mathematical Foundations of Optimal Interpolative Neural Networks." In E. Houstis and J. R. Rice, (eds.), *Artificial Intelligence, Expert Systems, and Symbolic Computing*, pp. 303–319, Elsevier, Amsterdam, 1992.
8. R. J. P. de Figueiredo. "An Optimal Multilayer Neural Interpolating (OMNI) Net in a Generalized Fock Space Setting." *Proceedings of 1992 International Joint Conference on Neural Networks, IJCNN-92*, vol. 1, Baltimore, MD, 1992, pp. 111–118.
9. R. W. Newcomb and R. J. P. de Figueiredo. "A Multi-Input Multi-Output Functional Artificial Neural Network." *J. Intell. Fuzzy Syst.*, vol. 4, no. 3, 207–213, 1996.
10. R. J. P. de Figueiredo. "Optimal Interpolative and Smoothing Functional Artificial Neural Networks (FANNs) Based on a Generalized Fock Space Framework." *Circuits, Syst., Signal Process.*, vol. 17, no. 2, 271–287, 1998.
11. S. K. Sin and R. J. P. de Figueiredo. "An Incremental Fine Adjustment Algorithm for the Design of Optimal Interpolating Neural Networks." *Int. J. Pattern Recogn. Artif. Intell.*, Nov. 1991.
12. S. K. Sin and R. J. P. de Figueiredo. "An Evolution Oriented Learning Algorithm for the Optimal Interpolative Net." *IEEE Trans. Neural Networks*, vol. 3, no. 2, 315–323, 1992.
13. S. K. Sin and R. J. P. de Figueiredo. "Efficient Learning Procedures for Optimal Interpolating Networks." *Neural Networks*, vol. 6, 99–113, 1993.
14. R. J. P. de Figueiredo and G. Chen. *Nonlinear Feedback Control Systems: An Operator Theory Approach*. Academic Press, New York, 1993.
15. A. Maccato and R. J. P. de Figueiredo. "Structured Neural Network Topologies with Application to Acoustic Transients." *Proceedings of the 1990 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Albuquerque, NM, April 1990, pp. 877–880.
16. A. Maccato and R. J. P. de Figueiredo. "A Neural Network Based Framework for Classification of Oceanic Acoustic Signals." *Proceedings of the 1989 OCEANS Conference*, Seattle, WA, Sept. 18, 1989.
17. R. J. P. de Figueiredo and T. A. W. Dwyer III. "Approximation-Theoretic Methods for Nonlinear Deconvolution and Inversion." *Inf. Sci.*, vol. 31, 209–220, 1983.
18. R. J. P. de Figueiredo. "Implications and Applications of Kolmogorov's superposition theorem." *IEEE Trans. Automatic Control*, vol. AC-25, no. 5, 1227–1231, 1980.
19. R. J. P. de Figueiredo. "Generalized Nonlinear Functional and Operator Splines in Fock Spaces." In Ward Cheney, (ed.), *Approximation Theory*, pp. 937–944, Academic Press, New York, 1980.
20. T. Eltoft and R. J. P. de Figueiredo. "A DCT-Based D-FANN for Nonlinear Adaptive Time Series Prediction." *IEEE Trans. Circuits Syst.*, Part II, 1131, 2000.
21. T. Eltoft and R. J. P. de Figueiredo. "A New Neural Network for Cluster-Detection and-Labelling." *IEEE Trans. Neural Networks*, vol. 9, 1021–1035, 1998.
22. R. J. P. de Figueiredo and T. Eltoft. "Pattern Classification of Non-Sparse Data Using Optimal Interpolative Nets." *Neurocomputing*, vol. 10, no. 4, 385–403, 1996.
23. R. J. P. de Figueiredo, W. R. Shankle, A. Maccato, M. B. Dick, P. Y. Mundkur, I. Mena, and C. W. Cotman. "Neural-Network-Based Classification of Cognitively Normal, Demented, Alzheimer's Disease and Vascular Dementia from Brain SPECT Image Data." In *Proceedings of the National Academy of Sciences USA*, vol. 92, June 1995, pp. 5530–5534.
24. A. Speis and R. J. P. de Figueiredo. "A Generalized Fock Space Framework for Nonlinear System Identification for Random Input-Output Data." *UCI Int. Rep.*, to be submitted for publication.

25. N. Wiener. *Nonlinear Problems in Random Theory.* MIT Press, Cambridge, MA, 1958.
26. A. G. Bose. "A Theory of Nonlinear Systems." Tech. Rep. No. 309, Research Laboratory of Electronics, MIT, Cambridge, MA, 1956.
27. Y. W. Lee and M. Schetzen. "Measurement of the Wiener Kernels of a Nonlinear System by Cross-Correlation." *Int. J. Control*, vol. 2, 237–254, 1965.
28. P. Z. Marmarelis and V. Z. Marmarelis. *Analysis of Physiological Systems*. Plenum Press, New York, 1978.
29. M. Schetzen. *The Volterra and Wiener Theories on Non-Linear Systems*. Wiley/Interscience, New York, 1980.
30. W. J. Rugh. *Nonlinear System Theory*: *The Volterra/Wiener Approach*. Johns Hopkins University Press, Baltimore, MD, 1981.
31. L. Ljung and T. Söoderström. *Theory and Practice of Recursive Identification*. MIT Press, Cambridge, MA, 1983.
32. L. Ljung. *System Identification*: *Theory for the User*. Prentice Hall, Englewood Cliffs, NJ, 1987.
33. D. E. Rumelhart, J. L. McClelland, and the PDP Research Group. *Parallel Distributed Processing*: *Explorations in the Microstructure of Cognition*, vol. 1, *Foundations*. MIT Press, Cambridge, MA, 1986.
34. B. Widrow and S. D. Stearns. *Adaptive Signal Processing*. Prentice Hall, Englewood Cliffs, NJ, 1985.
35. G. A. Carpenter and S. Grossberg. "Neural Dynamics of Category Learning and Recognition: Attention Memory Consolidation and Amnesia." In J. Davis, R. Newburgh, and E. Wegman (eds.), *Brain Structure*, *Learning and Memory*, AAAS Symposium Series, Washington, DC, 1986.
36. S. I. Amari. "Mathematical Foundations of Neurocomputing." *Proc. IEEE*, vol. 78, no. 9, 1443–1463, 1990.
37. J. M. Zurada. *Introduction to Artificial Neural Systems*. West, St. Paul, MN, 1992.
38. V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 1995.
39. V. N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.
40. A. Dingankar and I. W. Sandberg, "On Error Bounds for Neural Network Approximation." In *Proceedings of the IEEE International Symposium on Circuits and Systems*, *ISCAS-95*, vol. 1, 1995, pp. 490–492.
41. A. T. Dingankar and I. W. Sandberg. "Tensor Product Neural Networks and Approximation of Dynamical Systems." In *Proceedings of the IEEE International Symposium on Circuits and Systems*, *ISCAS-96*, vol. 3, 1996, pp. 353–356.
42. J. C. Principe, N. R. Euliano, and W. C. Lefebre. *Neural and Adaptive Systems*. Wiley, New York, 2000.
43. S. Haykin. *Adaptive Filter Theory*, 3rd ed., Prentice Hall, Upper Saddle River, NJ, 1996.
44. W. G. Knecht. "Nonlinear Noise Filtering and Beam-Forming Using the Perception and Its Volterra Approximation." *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 2, no. 1, 55–62, 1994.
45. P. Werbos. "Generalization of Backpropagation with Application to a Recurrent Gas Market Model." *Neural Networks*, vol. 1, 339–356, 1998.
46. A. Weigend, B. Huberman, and D. Rumelhart. "Predicting the Future: A Connectionist Approach." *Int. J. Neural Syst.*, vol. 7, no. 3–4, 403–430, 1990.

47. S. G. Tzafestas. *Computational Intelligence in Systems and Control Design and Applications*. Kluwer, Dordrecht, The Netherlands, 1999.
48. A. Gammerman. *Computational Learning and Probabilistic Reasoning*. Wiley, Chichester, UK, 1996.
49. G. Cybenko, "Approximation by Superpositions of a Sigmoidal Function." *Math. Control, Signals, Syst.*, vol. 2, no. 4, 303–314, 1989.
50. A. R. Barron. "Universal Approximation Bounds for Superpositions of a Sigmoidal Function." *IEEE Trans. Inform. Theory*, vol. 39, issue 3, 930–945, 1993.
51. P. A. Regalia. *Adaptive IIR Filtering in Signal Processing and Control*. Dekker, New York, 1995.
52. V. Fock. *Konfigurationsraum und Zweite Quantelung*. *Z. Phys*., vol. 75, 622–647, 1932.
53. F. Berezin. *The Method of Second Quantization*. Academic Press, New York, 1966.
54. B. Simon. *The $P(\Phi)_2$ Euclidean* (*Quantum*) *Field Theory.* Princeton University Press, Princeton, NJ, 1974.
55. V. Bargmann. "On a Hilbert Space of Analytic Functions and on Associated Integral Transform," Part I. *Commun. Pure Appl. Math.*, vol. 14, 187–214, 1961; also Part II. vol. 20, 1–101, 1967.
56. F. Beatrous and J. Burbea. *Dissertations Mathematica*. Warszawa, 1989.
57. M. Beals, C. Fefferman, and R. Grossman. "Strictly Pseudoconvex Domains in $\mathcal{C}^n$." *Bull*. (*New Ser.*) *Am. Math. Soc.*, vol. 8, no. 2, 1983.
58. T. W. A. Dwyer III. "Holomorphic Representation of Tempered Distributions and Weighted Fock Spaces." In L. Nachbin, (ed.), *Analyse Fonctionelle et Aplications*, Hermann, Paris, 1975.
59. F. Treves. *Topological Vector Spaces, Distributions and Kernels*. Academic Press, New York, 1967.
60. E. Hille and R. S. Phillips. "Functional Analysis and Semigroups." *Am. Math. Soc. Colloq.*, Publ. 31, New York, 1957.
61. N. Aronsjahn. "Theory of Reproducing Kernels." *Am. Math. Soc. Trans.*, vol. 68, 337–404, 1950.
62. F. B. Hildebrand. *Introduction to Numerical Analysis*. McGraw-Hill, New York, 1974.
63. J. Bolz, C. D. Gilbert, and T. N. Wisel. "Pharmacological Analysis of Cortical Circuitry." *Trends Neurosci.*, vol. 12, no. 8, 292–296, 1989.
64. J. C. Eccles and O. Creutzfeld (eds.). "The Principles of Design and Operation of the Brain." In *Proceedings of the Study Week Organized by the Pontifical Academy of Sciences, Cassia Pius IV, Vatican City*, Springer-Verlag, New York, 1990.
65. R. J. P. de Figueiredo. "Beyond Volterra and Wiener: Some New Results And Open Problems in Nonlinear Circuits and Systems." In *Proceedings of the IEEE Midwest Symposium on Circuits and Systems*, 1998, pp. 124–127.
66. L. A. Zadeh. "Outline of a New Approach to the Analysis of Complex Systems and Decision Processes." *IEEE Trans. Syst., Man Cybern.*, vol. 3, no. 1, 28–44, 1973.
67. L. A. Zadeh. In J. E. Hayes, D. Michie, and L. I. Mikulich, (eds.), *A Theory of Approximate Reasoning, in Machine Intelligence*, vol. 9, pp. 149–194. Elsevier, New York, 1979.
68. G. J. Klir, U. H. St. Clair, and B. Yuan. *Fuzzy Set Theory*: *Foundations and Applications*. Prentice Hall, Upper Saddle River, NJ, 1997.