Neural Networks, Rational Functions and Realization Theory

Uwe Helmke
Department of Mathematics
University of Regensburg
Regensburg 8400 Germany

Robert C. Williamson
Department of Engineering
Australian National University
Canberra 0200 Australia

April 18, 1995

The problem of parametrizing single hidden layer scalar neural networks with continuous activation functions is investigated. A connection is drawn between realization theory for linear dynamical systems, rational functions and neural networks that appears to be new. A result of this connection is a general parametrization of such neural networks in terms of strictly proper rational functions. Some existence and uniqueness results are derived. Jordan decompositions are developed which show how the general form can be expressed in terms of a sum of canonical second order sections. The parametrization may be useful for studying learning algorithms.

1 Introduction

Nonlinearly parametrized representations of functions $\phi\colon \mathbb{R} \to \mathbb{R}$ of the form

$$\phi(x) = \sum_{i=1}^{n} c_i \sigma(x - a_i) \quad x \in \mathbb{R}, \tag{1.1}$$

have attracted considerable attention recently in the neural network literature. Here $\sigma\colon \mathbb{R} \to \mathbb{R}$ is typically a sigmoidal function such as

$$\sigma(x) = \frac{1}{1 + e^{-x}}, \tag{1.2}$$

but other choices than (1.2) are possible and of interest. Sometimes more complex representations such as

$$\phi(x) = \sum_{i=1}^{n} c_i \sigma(b_i x - a_i) \tag{1.3}$$

or even compositions of these are considered. Yet more generally, functions $\phi\colon \mathbb{R}^m \to \mathbb{R}$ are also studied.

Different choices of the "activation function" $\sigma\colon \mathbb{R} \to \mathbb{R}$ correspond to different representation problems. For example, if $\sigma(x) = x^{-1}$ then (1.1) amounts to finding the partial fraction decomposition of a rational function $\phi(x)$. The coefficients a_i and c_i arising in (1.1) here have the interpretation of the poles and residues respectively of $\phi(x)$. In this example it is obvious also that *complex* coefficients $a_i, c_i \in \mathbb{C}$ arise naturally, as a real rational function $\phi\colon \mathbb{R} \to \mathbb{R}$ may well have complex poles and residues.

Another case of interest is where $\sigma(x) = x^d$, $d \in \mathbb{N}$, is a monomial. Then (1.1) is equivalent to finding a decomposition of a polynomial $\phi(x)$ of degree $\leq d$ as a weighted sum of d-th powers of linear polynomials $x - a_i$. This is usually referred to as Waring's problem for binary forms, with early results going back to Sylvester [25] and Gundelfinger (1886). There is also an interesting connection with Hilbert's 17th problem, asking whether a positive polynomial $\phi\colon \mathbb{R}^m \to \mathbb{R}$ can be represented as a sum of squares of polynomials (or rational functions). In fact if the coefficients c_i in (1.1) are all positive and the degree d of ϕ is even, then (1.1) is such a representation of a polynomial as a sum of squares of polynomials.

If $\sigma\colon \mathbb{R} \to \mathbb{R}$ is a sigmoidal function such as (1.2), then functions of the form (1.1) are described by one "hidden layer" neural networks with n hidden layer thresholds a_i and output weights c_i, but with no input weights. The task then is to find, or to "learn", an exact or approximate representation (1.1) of some function. There are now a number of results available describing the "universal approximation properties" of such classes of feedforward neural networks. Nearly all of these [5, 8, 14] are in the form of denseness results, saying that if one takes enough nodes, one can make an arbitrarily good approximation.

Motivated by analogies with model reduction of linear control systems we became interested in finding best approximations of functions by neural network representations (1.1), with an upper bound on the number n of such nodes. Such questions are important in order to estimate the approximation theoretic capabilities of learning algorithms for (1.1). For the special sigmoid function (1.2) an analysis has been presented in [26], where the problem has been shown to be deeply related to classical rational approximation theory. In order to approach such neural network approximation tasks, possibly valid for a large class of activation functions $\sigma\colon \mathbb{R} \to \mathbb{R}$, it becomes important to study the parametrization problem for the class of functions described by (1.1).

It has been shown in [26] that there often exists no best approximation of a function $\Phi\colon \mathbb{R} \to \mathbb{R}$ by functions of the type (1.1). Thus the class of functions (1.1) is not rich enough to guarantee the convergence of general learning algorithms. Also, the different parametrizations of the class (1.1) may well have an impact on the transient behaviour of learning algorithms and are thus worth being studied in more detail. (See [4] for some examples of the effect of different parametrizations on some very simple learning problems.)

1.1 What this paper is about

The purpose of this paper is to explore such parametrization issues regarding (1.1) (and to a lesser extent (1.3)), and in particular to show the close connection these representations have with the standard system-theoretic realization theory for rational functions. The main result of this paper is theorem 5.1. We firstly show how to define a generalization of (1.1) parametrized by (A, b, c), where A is a matrix over a field, and b and c are vectors. (This is made more precise below). The parametrization involves (A, b, c) being used to define a rational function. The generalized σ-representations are then defined in terms of the rational function. Representations (1.1) correspond to the case where A is diagonalizable. In that case, the thresholds a_i and the output weights c_i are interpreted as the poles and residues of the associated rational "transfer function" $c(xI - A)^{-1}b$. This connection allows us to use results available for rational functions in the study of neural-network representations such as (1.1). It will also lead to an understanding of the geometry of the space of functions.

That there is indeed a close connection between representations of the form (1.1) and rational functions was shown in [26] (and previously used in [23]). There it was shown that (1.1) can be written as $e^x r(e^x)$ when $\sigma(\cdot)$ is given by (1.2). The function $r(\cdot)$ is a strictly proper rational function, and the coefficients a_i in (1.1) correspond to the logarithm of the *poles* of $r(\cdot)$, and the c_i coefficients correspond to the *residues* of $r(z)$ at $z = e^{a_i}$.

In the following section we shall show that representations of the form (1.1) in general can be parametrized by rational functions in a single variable ξ. Furthermore, the more general representations (1.3) are shown to be parametrized by rational functions in two variables. They correspond to so-called separable 2–D systems, arising in two dimensional image processing. Then, by using ideas originally from the theory of state-space realizations of linear dynamical systems, we give conditions under which a representation (1.1) exists for a given function $\phi(\cdot)$. More generally, the existence and uniqueness properties of representations $\phi(x) = c\sigma(xI + A)b$ are investigated, where A is an $n \times n$ matrix, b is a n-vector and c is a n-covector. Such representations naturally extend representations (1.1) where $A = \mathrm{diag}(a_1, \ldots, a_n)$ is diagonal.

2 Realizations Relative to a Function

In this section we explore the relationship between sigmoidal representations (1.2) of real analytic functions $\phi\colon \mathbb{I} \to \mathbb{R}$ defined on an interval $\mathbb{I} \subset \mathbb{R}$, real rational functions defined on the complex plane $\mathbb{C}$, and the well established realization theory for linear dynamical systems

$$
\begin{aligned}
\dot{x}(t) &= Ax(t) + bu(t) \\
y(t) &= cx(t) + du(t).
\end{aligned}
$$

For standard textbooks on systems theory and realization theory we refer to [15, 17, 22].

Let $\mathbb{K}$ denote either the field $\mathbb{R}$ of real numbers or the field $\mathbb{C}$ of complex numbers. Let $\Delta \subset \mathbb{C}$ be an open and simply connected subset of the complex plane, containing $\mathbb{I}$, and let $\sigma\colon \Delta \to \mathbb{C}$ be an analytic function defined on Δ. For example, σ may be obtained by an analytic continuation of some sigmoidal function $\sigma\colon \mathbb{R} \to \mathbb{R}$ into the domain of holomorphy of the complex plane.

Let $T\colon \mathbb{V} \to \mathbb{V}$ be a linear operator on a finite-dimensional $\mathbb{K}$-vector space $\mathbb{V}$ such that T has all its eigenvalues in Δ. Let $\Gamma \subset \Delta$ be a simple closed curve, oriented in the counter-clockwise direction, enclosing all the eigenvalues of T in its interior. More generally, Γ may consist of a finite number of simple closed curves Γ_k with interiors Δ'_k such that the union of the domains Δ'_k contains all the eigenvalues of T.

Definition 2.1 *The matrix valued function $\sigma(T)$ is defined as the contour integral*

[18, p.44]

$$\sigma(T) := \frac{1}{2\pi i}\int_\Gamma \sigma(z)\,(zI - T)^{-1}\,\mathrm{d}z. \tag{2.1}$$

Note that for each linear operator $T\colon \mathbb{V} \to \mathbb{V}$, $\sigma(T)\colon \mathbb{V} \to \mathbb{V}$ is again a linear operator on $\mathbb{V}$ that is independent of the choice of Γ.

If we now make the substitution $T := xI + A$ for $x \in \mathbb{C}$ and $A\colon \mathbb{V} \to \mathbb{V}$ $\mathbb{K}$-linear, then

$$\sigma(xI + A) = \frac{1}{2\pi i}\int_\Gamma \sigma(z)\,((z-x)I - A)^{-1}\,\mathrm{d}z$$

becomes a function of the complex variable x, at least as long as Γ contains all the eigenvalues of $xI + A$. Using the change of variables $\xi := z - x$ we obtain

$$\sigma(xI + A) = \frac{1}{2\pi i}\int_{\Gamma'} \sigma(x+\xi)\,(\xi I - A)^{-1}\,\mathrm{d}\xi \tag{2.2}$$

where $\Gamma' = \Gamma - x \subset \Delta$ encircles all the eigenvalues of A.

Given an arbitrary vector $b \in \mathbb{V}$ and a linear functional $c\colon \mathbb{V} \to \mathbb{K}$ we then achieve the representation

$$\boxed{c\sigma(xI + A)b = \frac{1}{2\pi i}\int_\Gamma \sigma(x+\xi)\,c(\xi I - A)^{-1}b\,\mathrm{d}\xi.} \tag{2.3}$$

Note that in (2.3) the simple closed curve $\Gamma \subset \mathbb{C}$ is arbitrary, as long as it satisfies the two conditions

$$\Gamma \text{ encircles all the eigenvalues of } A \tag{2.4}$$

$$x + \Gamma = \{x + \xi \mid \xi \in \Gamma\} \subset \Delta. \tag{2.5}$$

We will take (2.3) to be the *definition* of $c\sigma(xI + A)b$.

Let $\phi\colon \mathbb{I} \to \mathbb{R}$ be a real analytic function in a single variable $x \in \mathbb{I}$, defined on an interval $\mathbb{I} \subset \mathbb{R}$.

Definition 2.2 *A quadruple (A, b, c, d) is called a finite-dimensional σ*-realization *of $\phi\colon \mathbb{I} \to \mathbb{R}$ over a field of constants $\mathbb{K}$ if for all $x \in \mathbb{I}$*

$$\phi(x) = c\sigma(xI + A)b + d \tag{2.6}$$

holds, where the right hand side is given by (2.3) and Γ is assumed to satisfy the *conditions (2.4)–(2.5). Here* $d \in \mathbb{K}$*,* $b \in \mathbb{V}$*, and* $A\colon \mathbb{V} \to \mathbb{V}$*,* $c\colon \mathbb{V} \to \mathbb{K}$ *are* $\mathbb{K}$*-linear maps and* $\mathbb{V}$ *is a finite dimensional* $\mathbb{K}$*-vector space. If* $d \equiv 0$*, we will sometimes just write of a* σ*-realization* (A, b, c)*.*

Definition 2.3 *The* dimension *(or degree) of a* σ*-realization is* $\dim_{\mathbb{K}} \mathbb{V}$*. The* σ-degree *of* ϕ*, denoted* $\delta_\sigma(\phi)$*, is the minimal dimension of all* σ*-realizations of* ϕ*. A* minimal σ-realization *is a* σ*-realization of minimal dimension* $\delta_\sigma(\phi)$*.*

The above definition of a σ-realization is a rather straightforward extension of the familiar system-theoretic notion of a realization of a transfer function. In this paper we will address the following specific questions concerning σ-realizations.

Q1 What are the existence and uniqueness properties of σ-realizations?

Q2 How can one characterize minimal σ-realizations?

Q3 How can one compute $\delta_\sigma(\phi)$?

Q4 Given ϕ, when does there exist a σ-realization (A, b, c, d) with A diagonalizable over $\mathbb{K}$, and what is the minimal dimension of such a realization?

Examples of $\sigma(\cdot)$

Important examples of activation functions $\sigma\colon \mathbb{R} \to \mathbb{R}$ are:

1. $\boxed{\sigma(x) = x^{-1}}$
   In this case σ-realizations are just the standard realizations of analytic functions or formal power series in systems theory. Kalman's realization theorem [17] solves questions 1–3 for rational functions and, in fact, for arbitrary formal power series $\phi(x)$.

2. $\boxed{\sigma(x) = x^d,\ d \in \mathbb{N}}$
   This is known as Waring's problem for binary forms (being a generalization of the number theoretic question bearing that name [12]). A function $\phi(x)$ admits a σ-realization (A, b, c) if and only if it is a polynomial of degree $\leq d$. Over $\mathbb{C}$ and over $\mathbb{R}$ with d even, the problem has been solved by Helmke [13].

3. $\boxed{\sigma(x) = (1 + e^{-x})^{-1}}$
   We refer to this as the standard sigmoid case. This function $\sigma(x)$ is widely used in feedforward neural networks. Other cases of interest include bump functions such as e.g. $\sigma(x) = e^{-x^2}$.

Remarks

1. $\mathbb{K}$ is usually either the field $\mathbb{R}$ of real numbers or the field $\mathbb{C}$ of complex numbers. Even if $\sigma(x)$ and $\phi(x)$ are considered to be *real* valued functions, *complex* realizations may be of interest; i.e. where $\mathbb{V}$ is a $\mathbb{C}$-vector space, $b \in \mathbb{V}$ and $A\colon \mathbb{V} \to \mathbb{V}$, $c\colon \mathbb{V} \to \mathbb{C}$ are $\mathbb{C}$-linear maps. This topic will be of importance, for example, if realizations are sought where A is a diagonal matrix (see section 6.1 below). In fact, if A is diagonalizable with complex eigenvalues then the diagonalization of A leads a finite sum sigmoidal representation of $\phi(x)$. Such complex realizations may thus yield finite sum representations using complex coefficients, even in cases where exact finite sum representations with real coefficients do not exist. This is one reason why complex realizations may prove to be important.

2. The standard realizations arising in systems theory have $\sigma(x) = x^{-1}$. In this case, a function $\phi(x)$ has a finite dimensional σ-realization if and only if $\phi(x)$ is rational with $\phi(\infty) \neq \infty$ [17].

2.1 Generalizations: Realizations Related to $\sum_{i=1}^{n} c_i\sigma(b_i x - a_i)$

In (2.3), the strictly proper "transfer function"

$$g(\xi) := c(\xi I - A)^{-1} b$$

can be extended to an arbitrary rational function (and thus in particular to a polynomial). To see this we consider the integral transformation

$$\boxed{\phi(x) = \frac{1}{2\pi i}\int_\Gamma \sigma(x+\xi)\, g(\xi)\, \mathrm{d}\xi} \tag{2.7}$$

for an arbitrary rational function $g(\xi) = p(\xi)/q(\xi) \in \mathbb{K}(\xi)$. Clearly (2.7) makes sense for any simple closed integration path Γ which encircles the poles of $g(\xi)$ such that $x + \Gamma \subset \Delta$ for all $x \in \mathbb{I}$. Moreover, by the residue theorem,

$$\frac{1}{2\pi i}\int_\Gamma \sigma(x+\xi)\, g(\xi)\, \mathrm{d}\xi = \sum_{g(z)=\infty} \operatorname*{res}_{\xi=z} \left[\sigma(x+\xi)\, g(\xi)\right]. \tag{2.8}$$

When all the poles of $g(\xi)$ are distinct, (2.8) is just a weighted sum of residues, where the sum is over all finite poles of $g(\xi)$. We note in passing that formula (2.8) is reminiscent of Gaussian quadrature formulae; see, e.g. [9, 20].

More generally one can consider σ-realizations (E, A, b, c, d), where $E, A\colon \mathbb{V} \to \mathbb{V}$, $c\colon \mathbb{V} \to \mathbb{K}$ are $\mathbb{K}$-linear maps satisfying $\det(xE - A) \not\equiv 0$, $b \in \mathbb{V}$, $d \in \mathbb{K}$ such that

$$\phi(x) = c\sigma(xE + A)b + d \tag{2.9}$$

holds for all $x \in \mathbb{I}$. Here we define

$$\boxed{c\sigma(xE + A)b := \frac{1}{2\pi i}\int_\Gamma \sigma(\xi)\, c(\xi I - xE - A)^{-1} b \,\mathrm{d}\xi} \tag{2.10}$$

where, for $x \in \mathbb{I}$ fixed, Γ encircles all the roots ξ of $\det(\xi I - xE - A) = 0$ such that $x + \Gamma \subset \Delta$ for all $x \in \mathbb{I}$. Note that the "transfer function" $g(x, \xi) = c(\xi I - xE - A)^{-1}b$ is a rational function in two variables (ξ, x). While such rational functions arise naturally in applications, such as two dimensional image processing, the system theoretic interpretation of such transfer functions in the context of neural networks is left as an open question.

Of course *commuting σ-realizations* (E, A, b, c, d) with

$$EA = AE \tag{2.11}$$

are of obvious interest.

Lemma 2.4 *1. Let $S\colon \mathbb{V} \to \mathbb{V}$ be an invertible $\mathbb{K}$-linear map. Then (E, A, b, c, d) is a σ-realization of ϕ if and only if*

$$(SES^{-1}, SAS^{-1}, SB, cS^{-1}, d)$$

is a σ-realization of ϕ. In particular, for $E = I_n$ (the $n \times n$ identity), if (A, b, c, d) is a σ-realization of ϕ then $(SAS^{-1}, Sb, cS^{-1}, d)$ is a σ-realization of ϕ for all invertible transformations $S\colon \mathbb{V} \to \mathbb{V}$.

2. If $A = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ is diagonal, then for each x,

$$\sigma(xI + A) = \operatorname{diag}(\sigma(x + \lambda_1), \ldots, \sigma(x + \lambda_n)). \tag{2.12}$$

Proof The proof is an immediate consequence of definition (2.3), once it is noticed that

$$cS^{-1}\sigma(xSES^{-1} + SAS^{-1})Sb = c\sigma(xE + A)b$$

holds for all $x \in \mathbb{I}$. ∎

Remark In the case $\sigma(x) = (1 + e^{-x})^{-1}$ it is of interest to compare the representation (2.7) with the representation used in [26], namely $\phi(x) = e^x r(e^x)$, where $r(\cdot)$ is a strictly proper rational function with all its poles in $\mathbb{C} - [0, \infty)$. (Such a representation was also used in [23].) It is easily verified (see [26] for details), that a representation (1.1) is equivalent to $\phi(x) = e^x r(e^x)$, with

$$r(z) = \sum_{i=1}^{n} \frac{c_i}{z + e^{a_i}}.$$

Let $g(\cdot)$ be a strictly proper rational function and Γ an arc encircling the poles of $g(\cdot)$. Equating the two representations we have

$$\begin{aligned} e^x r(e^x) &= \frac{1}{2\pi i} \int_\Gamma \sigma(z + x)\, g(z)\, \mathrm{d}z \\ &= \frac{1}{2\pi i} \int_\Gamma \frac{e^x}{e^x + e^{-z}}\, g(z)\, \mathrm{d}z. \end{aligned}$$

Let $y = e^x$ and so $x = \log y$. Then

$$r(y) = \frac{1}{2\pi i} \int_\Gamma \frac{1}{y + e^{-z}}\, g(z)\, \mathrm{d}z.$$

Let z_k, $k = 1, \ldots, n$, denote the poles of $g(z)$. Then by the residue theorem,

$$r(y) = \sum_{k=1}^{n} \operatorname*{res}_{z = z_k} \frac{g(z)}{y + e^{-z}}.$$

Thus if $g(z)$ has a pole at $z = \beta$, $r(z)$ has a pole at $z = -e^{-\beta}$. Specifically, if $g(z) = c(zI - A)^{-1} b$, then

$$r(z) = c(zI + e^{-A})^{-1} b. \tag{2.13}$$

3 Existence of σ-realizations

We now consider the question of existence of σ-realizations. To set the stage, we consider the systems theory case $\sigma(x) = x^{-1}$ first. Assume we are given a formal power series

$$\phi(x) = \sum_{i=0}^{N} \frac{\phi_i}{i!} x^i, \quad N \le \infty, \tag{3.1}$$

and that (A, b, c) is a σ-realization in the sense of definition 2.2. The Taylor expansion of $c(xI + A)^{-1}b$ at 0 is (for A nonsingular)

$$c(xI + A)^{-1}b = \sum_{i=0}^{\infty} (-1)^i c A^{-(i+1)} b x^i. \tag{3.2}$$

Thus

$$\frac{\phi_i}{i!} = (-1)^i c A^{-(i+1)} b, \qquad i = 0, \ldots, N. \tag{3.3}$$

if and only if the expansions of (3.1) and (3.2) coincide up to order N. Observe [17] that

$$\begin{aligned} & \phi(x) = c(xI + A)^{-1}b \text{ and } \dim \mathbb{V} < \infty \\ \iff & \\ & \phi(x) \text{ is rational with } \phi(\infty) = 0. \end{aligned}$$

The possibility of solving (3.3) is now easily seen as follows. Let $\mathbb{V} = \mathbb{R}^{N+1} = \mathrm{Map}(\{0, \ldots, N\}, \mathbb{R})$ be the finite or infinite $(N+1)$-fold product space of $\mathbb{R}$. (Here $\mathrm{Map}(X, Y)$ denotes the set of all maps from X to Y.) If N is finite let

$$\begin{aligned} A^{-1} &= -\begin{bmatrix} 0 & \cdots & 0 & 1 \\ 1 & \cdots & 0 & 0 \\ & \ddots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 \end{bmatrix} \in \mathbb{R}^{(N+1)\times(N+1)}, \\ b &= (1\ 0\ \cdots\ 0)^T \in \mathbb{V}, \quad c = \left(\frac{\phi_N}{N!},\ \phi_0,\ \phi_1,\ \frac{\phi_2}{2!},\ \ldots,\ \frac{\phi_{N-1}}{(N-1)!} \right). \end{aligned} \tag{3.4}$$

For $N = \infty$ we may take, by an abuse of notation, $A^{-1} \colon \mathbb{R}^{\mathbb{N}} \to \mathbb{R}^{\mathbb{N}}$ as the shift operator

$$\begin{aligned} & A^{-1} \colon \mathbb{R}^{\mathbb{N}} \to \mathbb{R}^{\mathbb{N}} \\ & A^{-1} \colon (x_0, x_1, \ldots) \mapsto -(0, x_0, x_1, \ldots) \\ \text{and} \quad & b = (1, 0, \ldots), \quad c = (0, \phi_0, \phi_1, \frac{\phi_2}{2!}, \ldots). \end{aligned} \tag{3.5}$$

Note that there may be finite-dimensional realizations as well.

We will now consider more general $\sigma(\cdot)$. We then have

Definition 3.1 *Let $\sigma(x)$ be a meromorphic function and let*

$$\phi(x) = \sum_{i=0}^{N} \frac{\phi_i}{i!} x^i, \quad N \leq \infty$$

*be a, possibly finite, formal power series in the variable x. A triple (A, b, c) is called an N-*th order σ-realization of $\phi(x)$ *if the following conditions are satisfied:*

(a) $\sigma(x)$ is analytic on a neighborhood of the spectrum of A in $\mathbb{C}$.

(b) The power series expansions of $\phi(x)$ and $c\sigma(xI + A)b$ at $x = 0$ coincide up to order N.

By condition (a) the i-th derivative of the function $x \mapsto \sigma(xI+A)$ at $x = 0$ exists and is equal to $\sigma^{(i)}(A)$, where $\sigma^{(i)}(x)$ denotes the i-th derivative of the function $\sigma(x)$. Therefore condition (b) is equivalent to

$$\phi_i = c\sigma^{(i)}(A)b \text{ for } i = 0, \ldots, N. \tag{3.6}$$

Observe that condition (a) is satisfied for $\sigma(x) = x^{-1}$ if and only if A is invertible. Moreover, since the standard sigmoid $\sigma(z) = (1+e^{-z})^{-1}$ is analytic for any complex variable $z \in \mathbb{C}$ with $|\Im z| < \pi$, condition (a) is satisfied for the standard sigmoid if A has all its eigenvalues in $\{z \in \mathbb{C} : \ |\Im z| < 2\pi\}$.

If $\sigma(x) = x^{-1}$, and $A \in \mathbb{C}^{n \times n}$ is invertible, then $\sigma^{(i)}(-A) = i!(A^{-1})^{i+1}$. Consequently, in this case, a triple (A, b, c) is an N-th order σ-realization if and only if the partial realization condition (3.3) is satisfied. Using the terminology of linear systems theory we may thus say that (A, b, c), with A invertible, is an N-th order σ-realization for $\sigma(x) = x^{-1}$ if and only if $(F, g, h) := (-A^{-1}, A^{-1}b, c)$ is a partial realization of $(\phi_0, \phi_1, \ldots, \frac{\phi_N}{N!})$, i.e. if and only if

$$\frac{\phi_i}{i!} = hF^i g$$

holds for $i = 0, \ldots, N$.

The existence part of the realization question Q1 can now be restated as

Q5 Given a meromorphic function $\sigma(x)$ and a sequence of real numbers $(\phi_0, \ldots, \phi_N)$, does there exist an N-th order σ-realization (A, b, c) with

$$\phi_i = c\sigma^{(i)}(A)b, \quad i = 0, \ldots, N? \tag{3.7}$$

Thus the realization question Q1 is essentially an interpolation question (Loewner interpolation [2, 7]), which extends the partial realization task from linear system theory, where $\sigma(x) = x^{-1}$ to more general sigmoid functions $\sigma(x)$.

Let $\gamma_\ell = cA^\ell b,\ \ \ell \in \mathbb{N}_0$, and let

$$F = \begin{bmatrix} \sigma_0 & \sigma_1 & \sigma_2 & \cdots \\ \sigma_1 & \sigma_2 & \sigma_3 & \cdots \\ \sigma_2 & \sigma_3 & \sigma_4 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} = (\sigma_{i+j})_{i,j=0}^{\infty}. \tag{3.8}$$

Write

$$[\gamma] = \begin{bmatrix} \gamma_0 \\ \gamma_1 \\ \gamma_2/2! \\ \gamma_3/3! \\ \vdots \end{bmatrix}, \quad \text{and} \quad [\phi] = \begin{bmatrix} \phi_0 \\ \phi_1 \\ \phi_2 \\ \vdots \end{bmatrix}. \tag{3.9}$$

Then (3.6) (for $N = \infty$) can formally be written as

$$[\phi] = F \cdot [\gamma]. \tag{3.10}$$

Of course, any meaningful interpretation of (3.10) requires that the infinite sums $\sum_{j=0}^{\infty} \frac{\sigma_{i+j}}{j!}\gamma_j$, $i \in \mathbb{N}_0$, exist. This happens, for example, if $\sum_{j=0}^{\infty} \sigma_{i+j}^2 < \infty$, $i \in \mathbb{N}_0$ and $\sum_{j=0}^{\infty} \left(\frac{\gamma_j}{j!}\right)^2 < \infty$ exist. We have already seen that every finite or infinite sequence $[\gamma]$ has a realization (A, b, c). Thus we obtain

Corollary 3.2 *Suppose that the infinite sums $\sum_{j=0}^{\infty} \frac{\sigma_{i+j}}{j!}\gamma_j$, $i \in \mathbb{N}_0$, exist. A function $\phi(x)$ admits a σ-realization if and only if $[\phi] \in \mathrm{image}(F)$.*

Corollary 3.3 *Suppose that the infinite sums $\sum_{j=0}^{\infty} \frac{\sigma_{i+j}}{j!}\gamma_j$, $i \in \mathbb{N}_0$, exist. Let*

$$H = (\gamma_{i+j})_{i,j=0}^{\infty}. \tag{3.11}$$

There exists a finite dimensional σ-realization of $\phi(x)$ if and only if $[\phi] = F[\gamma]$ with rank $H < \infty$. *In this case* $\delta_\sigma(\phi) = \operatorname{rank} H$.

Proof This follows immediately from Kronecker's theorem and systems realization theory; see [15, 17]. ∎

Remark It would be desirable to have an explicit characterization of when F is invertible or surjective via Hankel operator theory for analytic functions. Also it would be nice to know how to compute the inverse of F; i.e. $[\gamma] = F^{-1}[\phi]$. Furthermore, how can one characterize pairs of analytic functions (σ, ϕ) such that $(\gamma_{i+j})_{i,j=0}^{\infty}$ has finite rank? When $\operatorname{rank} H < \infty$, then $\sum_{i=0}^{\infty} \gamma_i x^{-i-1}$ is rational and thus $\gamma_i = cA^i b$, $i \in \mathbb{N}_0$ for (A, b, c) is finite dimensional.

4 Uniqueness

In this section we consider the uniqueness of the representation (2.3). We require several definitions first.

Definition 4.1 *A system* $\{g_1, \ldots, g_n\}$ *of continuous functions* $g_i \colon \mathbb{I} \to \mathbb{R}$*, defined on an interval* $\mathbb{I} \subset \mathbb{R}$*, is said to be* linearly independent*, if for every* $c_1, \ldots, c_n \in \mathbb{R}$ *with* $\sum_{i=1}^{n} c_i g_i(x) = 0$ *for all* $x \in \mathbb{I}$*, then* $c_1 = \cdots = c_n = 0$.

Remark The linear independency condition is implied by the stronger condition that

$$\det \begin{bmatrix} g_1(x_1) & \cdots & g_1(x_n) \\ \vdots & & \vdots \\ g_n(x_1) & \cdots & g_n(x_n) \end{bmatrix} \neq 0$$

for all distinct $(x_i)_{i=1}^{n}$ in $\mathbb{I}$. Equivalently, if $\sum_{i=1}^{n} c_i g_i(x)$ has n distinct roots in $\mathbb{I}$, then $c_1 = \cdots = c_n = 0$.

Definition 4.2 *A subset* A *of* $\mathbb{C}$ *is called* self-conjugate *if* $a \in A$ *implies* $\overline{a} \in A$.

Let $\sigma \colon \mathbb{R} \to \mathbb{R}$ be a continuous function and define $\sigma_{z_i}^{(j)}(x) := \sigma^{(j)}(x + z_i)$. Let

$$\kappa := (\kappa_1, \ldots, \kappa_m) \;\; \text{where} \;\; \sum_{j=1}^{m} \kappa_j = n, \;\; \kappa_j \in \mathbb{N}, \;\; \kappa_j \geq 1, \;\; j = 1, \ldots, m$$

denote a combination of n of size m. For a given combination $\kappa = (\kappa_1, \dots, \kappa_m)$ of n, let $I := \{1, \dots, m\}$ and let $J_i := \{1, \dots, \kappa_i\}$. Let $Z_m := \{z_1, \dots, z_m\}$ and let

$$\sigma(\kappa, Z_m) := \{\sigma_{z_i}^{(j-1)} \colon i \in I, \, j \in J_i\}. \tag{4.1}$$

Definition 4.3 *If for all $m \leq n$, for all combinations $\kappa = (\kappa_1, \dots, \kappa_m)$ of n of size m, and for any self-conjugate set Z_m of distinct points, the class of functions $\sigma(\kappa, Z_m)$ are linearly independent, then σ is said to be* l.i. (linearly independent) generating of order *n.*

A condition similar to this was defined in [1]. They considered a uniqueness condition on $\sigma(\cdot)$ for representations of the form $\sum_{i=1}^{n} c_i \sigma(b_i x - a_i)$ (IP) and $\sum_{i=1}^{n} c_i \sigma(b_i x)$ (WIP).

Theorem 4.4 (Uniqueness) *Let $\sigma \colon \mathbb{R} \to \mathbb{R}$ be l.i. generating of order at least $2n$ on $\mathbb{I}$ and let (A, b, c) and $(\tilde{A}, \tilde{b}, \tilde{c})$ be minimal σ-realizations of order n of functions ϕ and $\tilde{\phi}$ respectively. Then the following equivalence holds*

$$\begin{array}{c} c\sigma(xI + A)b = \tilde{c}\sigma(xI + \tilde{A})\tilde{b} \quad \forall x \in \mathbb{I} \\ \Longleftrightarrow \\ c(\xi I - A)^{-1} b = \tilde{c}(\xi I - \tilde{A})^{-1} \tilde{b} \quad \forall \xi \in \mathbb{R}. \end{array} \tag{4.2}$$

Conversely, if (4.2) holds for almost all order n triples (A, b, c), $(\tilde{A}, \tilde{b}, \tilde{c})$, then $\sigma \colon \mathbb{R} \to \mathbb{R}$ is l.i. generating on $\mathbb{I}$ of order $\geq n$.

Proof

(Equivalence) By hypothesis (A, b, c) and $(\tilde{A}, \tilde{b}, \tilde{c})$ are minimal realizations of ϕ and $\tilde{\phi}$ (of order n). If $\phi = \tilde{\phi}$, then

$$c\sigma(xI + A)b = \tilde{c}\sigma(xI + \tilde{A})\tilde{b}.$$

Thus for

$$g(\xi) := c(\xi I - A)^{-1} b$$

and

$$\tilde{g}(\xi) := \tilde{c}(\xi I - \tilde{A})^{-1} \tilde{b}$$

we have

$$\frac{1}{2\pi i}\int_\Gamma \sigma(x+\xi)h(\xi)\,\mathrm{d}\xi = 0 \quad \forall x \in \mathbb{I} \tag{4.3}$$

with $h(\xi) := g(\xi) - \tilde{g}(\xi)$, having degree $\leq 2n$. By the residue theorem, condition (4.3) is equivalent to

$$\sum_{h(z_i)=\infty} \operatorname*{res}_{z=z_i}\left[\sigma(x+z)h(z)\right] = 0, \tag{4.4}$$

where the sum is over all distinct poles $\{z_1, \ldots, z_m\}$ of $h(z)$ $(m \leq 2n)$. Write the partial fraction expansion of $h(z)$ as

$$h(z) = \sum_{i=1}^{m}\sum_{j=1}^{\kappa_i} \frac{h_{ij}}{(z-z_i)^j}, \qquad \sum_{i=1}^{m} \kappa_i = 2n. \tag{4.5}$$

(It is always of the form (4.5) since $h(z)$ is the difference of two strictly proper rational functions and hence is itself strictly proper.) Recall the general formula for residues of a product of an analytic and a meromorphic function. For the case of one pole (possibly of multiplicity greater than one) we have:

$$\operatorname*{res}_{z=z_0}\left[f(z)\left(\sum_{j=1}^{k}\frac{a_j}{(z-z_0)^j}\right)\right] = \sum_{j=1}^{k}\frac{a_j}{(j-1)!}f^{(j-1)}(z_0)$$

(see for example [21]). Thus we can express (4.4) as

$$\sum_{i\in I}\sum_{j\in J_i}\frac{h_{ij}}{(j-1)!}\sigma^{(j-1)}(x+z_i) = 0. \tag{4.6}$$

Since σ is l.i. generating of order at least $2n$, (4.6) implies that $h_{ij} = 0$, $i \in I$, $j \in J_i$. Hence $h(z) = 0$ and so $g(z) = \tilde{g}(z)$ as required.

(Converse) Assume σ is not l.i. generating of order at least n. That is, for some $m \leq n$ there is a combination $\kappa = (\kappa_1, \ldots, \kappa_m)$ of n of size m and a self-conjugate set of points $Z_m = \{z_1, \ldots, z_m\}$ of size m such that for some $l \in I = \{1, \ldots, m\}$, and some $j_l \in J_l = \{1, \ldots, \kappa_l\}$ there is a sequence of coefficients (α_{ij}) such that

$$\sigma^{(j_l-1)}(x+z_l) = \sum\sum_{(i,j)\in\mathcal{I}(l)}\frac{\alpha_{ij}}{(j-1)!}\sigma^{(j-1)}(x+z_i) \tag{4.7}$$

where

$$\mathcal{I}(l) := \{(i,j)\colon \ (i \in I - \{l\},\, j \in J_i) \text{ or } (i = l,\, j \in J_l - \{j_l\})\}.$$

Now

$$\begin{aligned} c\sigma(xI+A)b &= \sum_{g(z_i)=\infty} \operatorname*{res}_{z=z_i} [\sigma(x+z)g(z)] \\ &= \sum_{i\in I}\sum_{j\in J_i} \frac{g_{ij}\sigma^{(j-1)}(x+z_i)}{(j-1)!} \end{aligned} \tag{4.8}$$

where g_{ij} are the coefficients of the partial fraction expansion of $g(z)$:

$$g(z) = \sum_{i\in I}\sum_{j\in J_i} \frac{g_{ij}}{(z-z_i)^j}. \tag{4.9}$$

By (4.7) and (4.8) we can then write

$$\begin{aligned} c\sigma(xI+A)b &= \sum\sum_{(i,j)\in\mathcal{I}(l)} \frac{g_{ij}\sigma^{(j-1)}(x+z_i)}{(j-1)!} + \frac{g_{l,j_l}\sigma^{(j_l-1)}(x+z_l)}{(j_l-1)!} \\ &= \sum\sum_{(i,j)\in\mathcal{I}(l)} \frac{g_{ij}\sigma^{(j-1)}(x+z_i)}{(j-1)!} + \sum\sum_{(i,j)\in\mathcal{I}(l)} \frac{g_{l,j_l}\alpha_{ij}\sigma^{(j-1)}(x+z_i)}{(j-1)!} \\ &= \sum\sum_{(i,j)\in\mathcal{I}(l)} \lambda_{ij}\sigma^{(j-1)}(x+z_i) \end{aligned}$$

where

$$\lambda_{ij} := \frac{g_{ij} + g_{l,j_l}\alpha_{ij}}{(j-1)!}, \quad (i,j) \in \mathcal{I}(l). \tag{4.10}$$

Now $|\mathcal{I}(l)| = n-1$ but there are n independent parameters g_{ij}, $i \in I$, $j \in J_i$. Thus given λ_{ij} and α_{ij}, g_{l,j_l} can be chosen arbitrarily (with the g_{ij} given by (4.10)) such that λ_{ij} remains invariant. Hence $c\sigma(xI+A)b$ is unchanged. (In fact g_{l,j_l} generates a linear subspace.) But $g(\xi)$ changes and, for different choices of g_{l,j_l}, is not necessarily equal to $\tilde{g}(\xi)$. ∎

The following result gives examples of activation functions $\sigma\colon \mathbb{R} \to \mathbb{R}$ which are l.i. generating.

Lemma 4.5 *Let $d \in \mathbb{N}_0$. Then*

1) *The function $\sigma(x) = x^{-d}$ is l.i. generating of arbitrary order.*

2) *The monomial $\sigma(x) = x^d$ is l.i. generating of order $d+1$.*

3) The function e^{-x^2} is l.i. generating of arbitrary order.

Proof For $\sigma(x) = x^{-d}$ we have $\sigma^{(j)}(x) = (-1)^j \prod_{l=0}^{j-1}(d+l)x^{-d-j}$. Let $z_1, \ldots, z_m \in \mathbb{C}$ be distinct complex numbers and let $\kappa = (\kappa_1, \ldots, \kappa_m)$ be a combination of n of size m. Then

$$\sum_{i=1}^{m} \sum_{j=0}^{\kappa_i - 1} c_{ij} \sigma^{(j)}(x + z_i) = \sum_{i=1}^{m} \sum_{j=0}^{\kappa_i - 1} c_{ij} \frac{\tilde{c}_{ij}}{(x + z_i)^{d+j}},$$

with $\tilde{c}_{ij} = (-1)^j c_{ij} \prod_{l=0}^{j-1}(d + l)x^{-d-j}$, is a strictly proper rational function of degree $\leq n + m(d-1)$. Therefore it vanishes identically on a nontrivial interval $\mathbb{I}$ if and only if the coefficients c_{ij} are all zero. This proves 1).

To prove 2) let us assume for simplicity that $(\kappa_1, \ldots, \kappa_m) = (1, \ldots, 1)$, $m = d+1$. The general case is treated analogously. Thus suppose there are $c_1, \ldots, c_{d+1} \in \mathbb{R}$ and $d + 1$ distinct complex numbers $z_1, \ldots, z_{d+1}$ given with $\sum_{i=1}^{d+1} c_i \sigma(x + z_i)$ identically zero on the interval $\mathbb{I} \subset \mathbb{R}$. Thus for all $x \in \mathbb{I}$

$$0 = \sum_{i=1}^{n} c_i \sigma(x + z_i) = \sum_{i=1}^{d+1} c_i (x + z_i)^d = \sum_{j=0}^{d} \left(\sum_{i=1}^{d+1} c_i \binom{d}{j} z_i^{d-j} \right) x^j.$$

Thus for $l = 0, \ldots, d$, $\sum_{i=1}^{d+1} c_i z_i^l = 0$. Equivalently,

$$\begin{bmatrix} 1 & \cdots & 1 \\ z_1 & \cdots & z_{d+1} \\ \vdots & & \vdots \\ z_1^d & \cdots & z_{d+1}^d \end{bmatrix} \begin{bmatrix} c_1 \\ \vdots \\ c_{d+1} \end{bmatrix} = 0.$$

Since the Vandermonde matrix is invertible (by assumption that the $z_1, \ldots, z_{d+1}$ are distinct) it follows that $c_1 = \cdots = c_{d+1} = 0$. Thus x^d is l.i. generating of order $d+1$.

To prove 3) assume there are $c_1, \ldots, c_n$ such that for distinct complex numbers $z_1, \ldots, z_n$ we have

$$\sum_{i=1}^{n} c_i e^{-(x-z_i)^2} = e^{-x^2} \sum_{i=1}^{n} c_i e^{-z_i^2} e^{-2z_i x} = 0$$

for all $x \in \mathbb{I}$. Then by the identity theorem for analytic functions,

$$\sum_{i=1}^{n} \tilde{c}_i e^{-2z_i x} = 0 \tag{4.11}$$

for all $x \in \mathbb{R}$ with $\tilde{c}_i = c_i e^{-z_i^2}$. By differentiating (4.11) $n-1$ times and evaluating at $x = 0$ we obtain

$$\sum_{i=1}^{n} \tilde{c}_i(-2z_i)^l = 0, \qquad l = 0, \ldots, n-1.$$

Thus, using the invertibility of the Vandermonde matrix,

$$\begin{bmatrix} 1 & \cdots & 1 \\ z_1 & \cdots & z_n \\ \vdots & & \vdots \\ z_1^{n-1} & \cdots & z_n^{n-1} \end{bmatrix} \begin{bmatrix} \tilde{c}_1 \\ \vdots \\ \tilde{c}_n \end{bmatrix} = 0$$

implies that $\tilde{c}_1 = \cdots = \tilde{c}_n = 0$ and hence $c_1 = \cdots = c_n = 0$. Thus e^{-x^2} is l.i. generating of order n. Since n is arbitrary, the result follows. The general case can be proved using a residue type argument similar to the proof of theorem 1 in [1]. ∎

Remark A simple example of a σ which is not l.i. generating of order ≥ 2 is $\sigma(x) = e^x$. In fact, in this case $\sigma(x + z_j) = c_j \sigma(x + z_i)$ for $c_j = e^{z_j - z_i}$, $j = 2, \ldots, n$.

Remark Similarly, the standard sigmoid function $\sigma(x) = (1+e^{-x})^{-1}$ is *not* l.i. generating of any order ≥ 2. In fact, by the periodicity of the complex exponential function we have $\sigma(x+2\pi i) = \sigma(x-2\pi i)$, $i = \sqrt{-1}$, for all x. Thus the l.i. condition fails for $Z_2 = \{2\pi i, -2\pi i\}$.

In particular, the above uniqueness result fails for the standard sigmoid case. In order to cover this case we need a further definition.

Definition 4.6 *Let $\Omega = \overline{\Omega} \subset \mathbb{C}$ be a self-conjugate subset of $\mathbb{C}$. A function $\sigma \colon \mathbb{R} \to \mathbb{R}$ is said to be* l.i. generating of order n on Ω, *if for all $m \leq n$, for all combinations $\kappa = (\kappa_1, \ldots, \kappa_m)$ of n of size m, and for any self-conjugate subset $Z_m \subset \Omega$ of distinct points of Ω, $\sigma(\kappa, Z_m)$ consists of linearly independent functions.*

Of course for $\Omega = \mathbb{C}$, this definition coincides with definition 4.3. We have the following extension of theorem 4.4. The proof is completely analogous to that of theorem 4.4 and is thus omitted.

Theorem 4.7 (Local Uniqueness) *Let $\sigma\colon \mathbb{R} \to \mathbb{R}$ be analytic and let $\Omega \subset \mathbb{C}$ be a self-conjugate subset contained in the domain of holomorphy of σ. Let $\mathbb{I}$ be a nontrivial subinterval of $\Omega \cap \mathbb{R}$. Suppose $\sigma\colon \mathbb{R} \to \mathbb{R}$ is l.i. generating on Ω of order at least $2n$, $n \in \mathbb{N}$. Then for any two minimal σ-realizations (A, b, c) and $(\tilde{A}, \tilde{b}, \tilde{c})$ of orders at most n the following equivalence holds:*

$$
\begin{aligned}
& c\sigma(xI + A)b = \tilde{c}\sigma(xI + \tilde{A})\tilde{b} \quad \forall x \in \mathbb{I} \\
\Longleftrightarrow \quad & \\
& c(\xi I - A)^{-1}b = \tilde{c}(\xi I - \tilde{A})^{-1}\tilde{b} \quad \forall \xi \in \mathbb{R}.
\end{aligned} \tag{4.12}
$$

The next result shows that the standard sigmoid function is l.i. generating on a suitable subset Ω.

Lemma 4.8 *Let $\Omega := \{z \in \mathbb{C}\colon |\Im z| < \pi\}$. Then the standard sigmoid function $\sigma(x) = (1 + e^{-x})^{-1}$ is l.i. generating on Ω of arbitrary order.*

Proof Let distinct points $z_1, \ldots, z_n \in \Omega$ and $c_1, \ldots, c_n \in \mathbb{R}$ be given such that $\sum_{i=1}^{n} c_i\sigma(x + z_i) = 0$. The general case is treated similarly, using the fact that $\sigma'(x) = \sigma(x)(1 - \sigma(x))$, so that all higher order derivatives of $\sigma(x)$ can be expressed as polynomials in $\sigma(x)$. Thus for all $x \in \mathbb{I} \subset \Omega$

$$
0 = \sum_{i=1}^{n} c_i\sigma(x + z_i) = \sum_{i=1}^{n} \frac{c_i e^x}{e^x + e^{-z_i}} = e^x r(e^x)
$$

with $r(z) = \sum_{i=1}^{n} \frac{c_i}{z + e^{-z_i}}$. Thus $r(z) = 0$ vanishes identically. As $z_1, \ldots, z_n \in \Omega$ are pairwise distinct and as $z \mapsto e^z$ is injective on Ω, the poles $-e^{-z_i}$, $i = 1, \ldots, n$, of $r(z)$ are pairwise distinct. Thus $r(z)$ is of degree n and identically zero. Therefore $c_1 = \cdots = c_n = 0$ and the result follows. ∎

Albertini et al [1] showed $\sigma(x) = \frac{1}{1+e^{-x}}$ satisfied their uniqueness condition IP using both a residue argument (as above) and directly using Cauchy's formula for $\det(\frac{1}{a_i + b_j})_{i,j}$ (see [6, lemma 11.3.1]).

5 Main Result

As a consequence of the uniqueness theorems 4.4 and 4.7 we can now state our main result on the existence of minimal σ-realizations of a function $\phi(x)$. It extends a parallel result for standard transfer function realizations, where $\sigma(x) = x^{-1}$.

Theorem 5.1 (Realization) *Let $\Omega \subset \mathbb{C}$ be a self-conjugate subset, contained in the domain of holomorphy of a real meromorphic function $\sigma\colon \mathbb{R} \to \mathbb{R}$. Suppose σ is l.i. generating on Ω of order at least $2n$ and assume $\phi(x)$ has a finite dimensional realization (A, b, c) of dimension at most n such that A has all its eigenvalues in Ω.*

1. *There exists a minimal σ-realization (A_1, b_1, c_1) of $\phi(x)$ of degree $\delta_\sigma(\phi) \leq \dim(A, b, c)$. Furthermore, there exists an invertible matrix S such that*

$$SAS^{-1} = \begin{bmatrix} A_1 & A_2 \\ 0 & A_3 \end{bmatrix}, \quad Sb = \begin{bmatrix} b_1 \\ 0 \end{bmatrix}, \quad cS^{-1} = [c_1, c_2] \tag{5.1}$$

   holds.

2. *If (A_1, b_1, c_1) and (A_1', b_1', c_1') are minimal σ-realizations of $\phi(x)$ such that the eigenvalues of A_1 and A_1' are contained in Ω, then there exists a unique invertible matrix S such that*

$$(A_1', b_1', c_1') = (SA_1S^{-1}, Sb_1, c_1S^{-1}) \tag{5.2}$$

   holds.

3. *A σ-realization (A, b, c) is minimal if and only if (A, b, c) is controllable and observable; i.e. if and only if (A, b, c) satisfies the generic rank conditions*

$$\operatorname{rank}(b, Ab, \ldots, A^{n-1}b) = n$$
$$\operatorname{rank}\begin{bmatrix} c \\ cA \\ \vdots \\ cA^{n-1} \end{bmatrix} = n$$

   for $A \in \mathbb{K}^{n\times n}$, $b \in \mathbb{K}^n$, $c^T \in \mathbb{K}^n$.

Proof The existence of minimal σ-realizations with $\delta_\sigma(\phi) \leq \dim(A, b, c)$ is trivial: just pick any σ-realization (A_1, b_1, c_1), with eigenvalues of A_1 contained in Ω, from the set of all σ-realizations of $\phi(x)$ with smallest possible dimension $\leq \dim(A, b, c)$.

Let (A_1, b_1, c_1) be a minimal σ-realization of $\phi(x)$ with eigenvalues of A_1 contained in Ω. By the uniqueness theorem 4.7,

$$c_1(\xi I - A_1)^{-1} b_1 = c(\xi I - A)^{-1} b \quad \text{for all } \xi.$$

Thus (5.1) follows from the Kalman decomposition; see [15]. Moreover, statements 2 and 3 follow immediately from Kalman's realization theorem for strictly proper rational transfer functions [15]. ∎

Remark The use of the terms "observable" and "controllable" is solely for formal correspondence with standard systems theory. There are no dynamical systems actually under consideration here.

Remark For any σ-realization (A, b, c) of the form

$$A = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}, \; b = \begin{bmatrix} b_1 \\ 0 \end{bmatrix}, \; c = [c_1, c_2],$$

we have

$$\sigma(A) = \begin{bmatrix} \sigma(A_{11}) & * \\ 0 & \sigma(A_{22}) \end{bmatrix}$$

and thus $c\sigma(xI + A)b = c_1 \sigma(xI + A_{11}) b_1$. Thus transformations of the above kind always *reduce* the dimension of a σ-realization.

The following results are immediate consequences of theorem 5.1 and lemma 4.5.

Corollary 5.2 *Let $\phi(x) = g^{(k)}(x)$ be a k-th derivative of a rational function $g(x)$. Then*

1. *$\phi(x) = c(xI - A)^{-k-1} b$ for $(A, b, c) \in \mathbb{K}^{n \times n} \times \mathbb{K}^n \times \mathbb{K}^{1 \times n}$.*
2. *If (A, b, c) and $(\tilde{A}, \tilde{b}, \tilde{c})$ are minimal realizations such that for all $x \in \mathbb{I}$*

$$\phi(x) = c(xI - A)^{-k-1} b = \tilde{c}(xI - \tilde{A})^{-k-1} \tilde{b}$$

then there exists a unique invertible matrix S such that

$$(\tilde{A}, \tilde{b}, \tilde{c}) = (SAS^{-1}, Sb, cS^{-1}).$$

3. *A realization* $(A,b,c) \in \mathbb{K}^{n\times n} \times \mathbb{K}^n \times \mathbb{K}^{1\times n}$ *satisfying* $\phi(x) = c(xI - A)^{-k-1}b$ *is of minimal size* n *if and only if* (A,b,c) *is controllable and observable.*

Proof By lemma 4.5 the function $\sigma(x) = x^{-k}$ is l.i. generating of all orders. Applying theorem 5.1 to $\sigma(x) = x^{-k}$ and $\Omega = \mathbb{C} - \{0\}$ completes the proof. ∎

The next result is a special case of a more general result appearing in [13].

Corollary 5.3 *Let* $\phi(x)$ *be a polynomial of degree* $\leq 2n$. *Suppose*

$$\begin{aligned}\phi(x) &= \sum_{i=1}^{s} c_i(x-a_i)^{2n} \\ &= \sum_{i=1}^{s} c_i'(x-a_i')^{2n}\end{aligned}$$

are two sums of $2n$*th power representations of* $\phi(x)$ *of minimal length* s *satisfying* $s < n$. *Then*

$$(a_i', c_i') = (a_{\pi(i)}, c_{\pi(i)}), \qquad i = 1, \ldots, s$$

for a permutation $\pi\colon \{1,\ldots,s\} \to \{1,\ldots,s\}$.

Proof By lemma 4.5 the function $\sigma(x) = x^{2n}$ is l.i. generating of order $2n+1$. Then (A,b,c) and (A',b',c') defined by

$$A = \operatorname{diag}(a_1,\ldots,a_s), \quad b = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}, \quad c = (c_1,\ldots,c_s)$$

$$A' = \operatorname{diag}(a_1',\ldots,a_s'), \quad b' = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}, \quad c' = (c_1',\ldots,c_s')$$

are σ-realizations of $\phi(x)$ of order $s < n$. By minimality of s, (A,b,c) and (A',b',c') are controllable and observable. Applying theorem 5.1 to $\sigma(x) = x^{2n}$, $\Omega = \mathbb{C}$, yields

$$(A',b',c') = (SAS^{-1}, Sb, cS^{-1})$$

for a unique invertible matrix S. Since $SAS^{-1} = A'$ and both A and A' are diagonal, S must be a permutation matrix. The result follows. ∎

As a final consequence we obtain another proof of the following result (c.f. [24]which contains a more general result)

Corollary 5.4 *Let* $\sigma(x) = (1 + e^{-x})^{-1}$ *and*

$$\phi(x) = \sum_{i=1}^{n} c_i \sigma(x - a_i) = \sum_{i=1}^{n} c_i' \sigma(x - a_i')$$

be two minimal length σ*-representations with*

$$|\Im a_i| < \pi, \quad |\Im a_i'| < \pi, \quad i = 1, \ldots, n.$$

Then

$$(a_i', c_i') = (a_{\pi(i)}, c_{\pi(i)})$$

for a unique permutation $\pi\colon \{1, \ldots, n\} \to \{1, \ldots, n\}$. *In particular, minimal length representations (1.1) with real coefficients* a_i *and* c_i *are unique up to a permutation of the summands.*

Proof By lemma 4.8, $\sigma(x) = (1+e^{-x})^{-1}$ is l.i. generating on $\Omega = \{z \in \mathbb{C}\colon |\Im z| < \pi\}$ of arbitrary order. By minimality of the representations $a_i \neq a_j$ and $a_i' \neq a_j'$ for $i \neq j$. Let (A, b, c) and (A', b', c') be defined by

$$A = \operatorname{diag}(-\log a_1, \ldots, -\log a_n), \quad b = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}, \quad c = (c_1, \ldots, c_n),$$

$$A' = \operatorname{diag}(-\log a_1', \ldots, -\log a_n'), \quad b' = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}, \quad c' = (c_1', \ldots, c_n'),$$

using the standard branch of the complex logarithm function. Then (A, b, c) and (A', b', c') are controllable and observable σ-realizations of $\phi(x)$. By theorem 5.1 applied to $\sigma(x) = (1 + e^{-x})^{-1}$ and $\Omega = \{z \in \mathbb{C}\colon |\Im z| < \pi\}$ the result follows as in the previous proof.

∎

6 Jordan Canonical Forms

We now explore the connection between the Jordan canonical form for minimal realizations (A, b, c) and σ-realizations. Consider the partial fraction decomposition of a transfer function $g(\xi) = c(\xi I - A)^{-1}b$:

$$g(\xi) = \sum_{i=1}^{m} \sum_{j=1}^{\kappa_i} \frac{c_{ij}}{(\xi - z_i)^j}. \tag{6.1}$$

Then

$$\begin{aligned}
\phi(x) &= \frac{1}{2\pi i} \int_\Gamma \sigma(x+\xi) g(\xi) \, \mathrm{d}\xi \\
&= \sum_{i=1}^{m} \sum_{j=1}^{\kappa_i} c_{ij} \operatorname*{res}_{\xi = z_i} \left(\frac{\sigma(x+\xi)}{(\xi - z_i)^j} \right) \\
&= \sum_{i=1}^{m} \sum_{j=1}^{\kappa_i} \frac{c_{ij}}{(j-1)!} \sigma^{(j-1)}(x + z_i) \qquad (6.2)
\end{aligned}$$

By the realization theorem, any minimal σ-realization (A, b, c) of $\phi(x)$ satisfying (2.7) is a controllable and observable realization of the transfer function $g(\xi)$. Thus by the Jordan control canonical form [15], (A, b, c) is similar to $(A_J, b_J, c_J) = (SAS^{-1}, Sb, cS^{-1})$ with

$$A_J = \operatorname{diag}(z_1 I + N_1, \ldots, z_m I + N_m), \tag{6.3}$$

$$N_i = \begin{bmatrix} 0 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 \\ 0 & \cdots & 0 & 0 \end{bmatrix} \in \mathbb{R}^{\kappa_i \times \kappa_i}, \ i = 1, \ldots, m, \tag{6.4}$$

$$b_J = \begin{bmatrix} e_{\kappa_1} \\ e_{\kappa_2} \\ \vdots \\ e_{\kappa_m} \end{bmatrix}, \quad e_i = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \in \mathbb{R}^{\kappa_i}, \tag{6.5}$$

$$c_J = [c_{11}, \ldots, c_{1\kappa_1}, \ldots, c_{m1}, \ldots, c_{m\kappa_m}] \in \mathbb{K}^n, \tag{6.6}$$

where κ_i are the multiplicities of the poles z_i of $g(\xi)$, and c_{ij} are the coefficients from the partial fraction expansion (6.1). Thus for a σ-realization (A, b, c) of $\phi(x)$

in Jordan control canonical form (6.3)–(6.5),

$$c\sigma(xI+A)b = \sum_{i=1}^{m}\sum_{j=1}^{\kappa_i}\frac{c_{ij}}{(j-1)!}\sigma^{(j-1)}(x+z_i). \tag{6.7}$$

In particular, in the generic case where (A, b, c) has distinct possibly complex poles, then

$$c\sigma(xI+A)b = \sum_{i=1}^{n} c_i\sigma(x+z_i).$$

Remark When ϕ is real, the poles and residues always appear in complex conjugate pairs. Thus, by analogy with the systems theory case, it is always possible (in the generic case of no repeated poles) to write

$$c\sigma(xI+A)b = \sum_{z_i\in\mathbb{R}} c_i\sigma(x+z_i) \;+\; \sum_{z_i\notin\mathbb{R}} \sigma(c_i^R, c_i^I, z_i^R, z_i^I; x)$$

where $z_i \in \operatorname{spect} A$, $c_i^R, c_i^I, z_i^R, z_i^I$ are the real and imaginary parts of the residues and poles of $c(\xi I - A)^{-1}b$, and

$$\sigma(c^R, c^I, z^R, z^I; x) := (c^R + ic^I)\sigma(x + (z^R + iz^I)) \;+\; (c^R - ic^I)\sigma(x + (z^R - iz^I))$$

is the *canonical second-order building block.* Thus it is simple to cope with complex poles in σ-realizations using only real parameters.

As an example, consider $\sigma(x) = (1+e^{-x})^{-1}$. For simplicity write $\alpha = c^R$, $\beta = c^I$, $\gamma = z^R$ and $\delta = z^I$. We then obtain

$$\sigma(\alpha,\beta,\gamma,\delta;x) = \frac{2\,\alpha + 2\,\alpha\, e^{-x-\gamma}\cos(\delta) - 2\,\beta\, e^{-x-\gamma}\sin(\delta)}{1 + 2\, e^{-x-\gamma}\cos(\delta) + e^{-2\,x-2\,\gamma}}$$

The non-uniqueness of the parametrization (due to the periodicity of e^z) is apparent explicitly in the $\cos(\delta)$ and $\sin(\delta)$ terms.

Another example is $\sigma(x) = e^{-x^2}$. In this case,

$$\sigma(\alpha,\beta,\gamma,\delta;x) = 2\, e^{-x^2-2\,x\gamma-\gamma^2+\delta^2}(\alpha\,\cos(2\,x\,\delta + 2\,\gamma\,\delta) \;+\; \beta\,\sin(2\,x\,\delta + 2\,\gamma\,\delta)).$$

Remark Complex parametrizations for standard neural networks have been discussed in [10, 19]. The motivation there was to be able to use neural networks with complex inputs.

6.1 Diagonalizable σ-Realizations

A σ-realization (A, b, c) of $\phi(x)$ is called $\mathbb{K}$-*diagonalizable* if the operator $A\colon \mathbb{V} \to \mathbb{V}$ is diagonalizable over $\mathbb{K}$. Since the set of degree n rational functions $r(x) \in \mathbb{C}(x)$ with distinct poles is dense in the set of all rational functions of degree n, $\mathbb{C}$-diagonalizability of a σ-realization (A, b, c) is generic property.

Over $\mathbb{R}$, a necessary and sufficient condition for diagonalizable σ-realizations is that the poles of the associated rational function $c(xI - A)^{-1}b$ are on the real axis and simple. Certainly this is not a generic property. Note that a *sufficient* condition for a real rational function $g(x) = \sum_{i=1}^{\infty} g_i x^{-i}$ of degree n to have a $\mathbb{R}$-diagonalizable realization is that the $n \times n$-Hankel $(g_{i+j-1})_{i,j=1}^{n}$ is positive definite. In this case also the residues of the partial fraction decomposition of $g(x)$ are positive. The following result provides an answer to question 4.

Theorem 6.1 *Let $\sigma\colon \mathbb{R} \to \mathbb{R}$ be analytic, possibly only on an interval $\mathbb{I} \subset \mathbb{R}$, and let $\Omega \supset \mathbb{I}$ be a self-conjugate subset of $\mathbb{C}$ contained in the domain of holomorphy of σ. Assume that σ is l.i. generating on Ω of order at least $2n$. The set of functions $\phi\colon \mathbb{R} \to \mathbb{R}$ with σ-degree $\delta_\sigma(\phi) = n$ and which admit a diagonalizable σ-realization over $\mathbb{R}$ of length n has the structure of an analytic manifold of dimension $2n$. It has exactly 2^n connected components. Each such $\phi(x)$ with σ-degree $\delta_\sigma(\phi)$ has a decomposition*

$$\phi(x) = \sum_{i=1}^{n} c_i \sigma(x - a_i), \tag{6.8}$$

with real numbers $c_i \neq 0$, $a_i \in \Omega$, $a_i \neq a_j$ for $i \neq j$. The different connected components are characterized by $a_1 < \cdots < a_n$, and $\operatorname{sign}(c_i) = \varepsilon_i$, $\varepsilon_i \in \{-1, 1\}$, $i = 1, \ldots, n$. Moreover, the decomposition (6.8) is unique in the sense that for real numbers $a_1, \ldots, a_n \in \Omega$, $c_1, \ldots, c_n$ and $a'_1, \ldots, a'_n \in \Omega$, $c'_1, \ldots, c'_n$ satisfying (6.8), then

$$a'_i = a_{\pi(i)}, \ c'_i = c_{\pi(i)}, \quad i = 1, \ldots, n,$$

for a permutation $\pi\colon \{1, \ldots, n\} \to \{1, \ldots, n\}$.

Proof Let $\Gamma_n \subset \Omega^n \times \mathbb{R}^n$ denote the open subset of $\mathbb{R}^{2n}$ defined by

$$\Gamma_n := \{(a_1, \ldots, a_n; c_1, \ldots, c_n) \in \Omega^n \times (\mathbb{R} - \{0\})^n : \ a_1 < \cdots < a_n, \ a_i \in \Omega \cap \mathbb{R}, \ c_i \neq 0\}.$$

For any $(a, c) \in \Gamma_n$ let $g_{ac}(x)$ be the rational function defined by

$$g_{ac}(x) := \sum_{i=1}^{n} \frac{c_i}{x - a_i}.$$

Let $\phi_{ac}(x)$ be defined by

$$\phi_{ac}(x) := \frac{1}{2\pi i} \int_{\Gamma} \sigma(x + \xi) g_{ac}(x) \, \mathrm{d}\xi,$$

where Γ is a sufficiently large arc containing all the poles of $g_{ac}(x)$, $(a, c) \in \Gamma_n$. Then the map

$$(a, c) \mapsto \phi_{ac}(x)$$

is, by theorem 4.7, an injective map on Γ_n. The image $\mathcal{M}_n$ is exactly the class of functions described by (6.8). Γ_n is a smooth analytic manifold of dimension $2n$ with 2^n connected components characterized by $\mathrm{sign}(c_i) = \varepsilon_i$, $\varepsilon_i \in \{-1, 1\}$, $i = 1, \ldots, n$. Endow $\mathcal{M}_n$ with the unique structure of an analytic manifold such that $(a, c) \mapsto \phi_{ac}$ is an analytic diffeomorphism. This completes the proof. ∎

7 Conclusions and Related work

We have drawn a connection between the realization theory for linear dynamical systems and neural network representations. This is an exciting connection because it opens the way for the application of some of the machinery of realization theory to neural networks. A number of new open problems arise also. For example there is the problem of partial realizations [11, 16]. We are currently exploring the application of the theory of Padé approximants and continued fractions to the neural networks considered here.

After this paper was substantially completed the authors became aware of the work of Barrar and Loeb [3]. They have considered parametrizations like (2.7) for general nonlinear families. We have considered a slightly more specific case, and we have obtained results not obtainable in their setup.

Finally let us point out that the ability to parametrize general neural network representations in different ways could have a profound effect on learning algorithms: simply by performing gradient descent in the different parameter spaces is expected to offer different behaviour.

Acknowledgements

This work was supported by the Australian Research Council, the Australian Telecommunications and Electronics Research Board, and the Boeing Commercial Aircraft Company (thanks to John Moore).

References

[1] F. Albertini, E. D. Sontag and V. Maillot, Uniqueness of Weights for Neural Networks, in *Artificial Neural Networks for Speech and Vision*, R. Mammone, ed., Chapman and Hall, London, 1993, pp. 115–125.

[2] A. C. Antoulas and B. D. O. Anderson, On the Scalar Rational Interpolation Problem, *IMA Journal of Mathematical Control and Information*, **3** (1986), pp. 61–88.

[3] R. B. Barrar and H. L. Loeb, On Extended Varisolvent Families, *Journal D'Analyse Mathématique*, **26** (1973), pp. 243–254.

[4] K. L. Blackmore, R. C. Williamson and I. M. Y. Mareels, Local Minima and Attractors at Infinity of Gradient Descent Learning Algorithms, to appear in *Journal of Mathematical Systems, Estimation and Control*, 1995.

[5] G. Cybenko, Approximation by Superpositions of a Sigmoidal Function, *Mathematics of Control, Signals, and Systems*, **2** (1989), pp. 303–314.

[6] P. J. Davis, *Interpolation and Approximation*, Dover, New York, 1975.

[7] W. F. Donoghue, Jr, *Monotone Matrix Functions and Analytic Continuation*, Springer-Verlag, Berlin, 1974.

[8] K. -I. Funahashi, On the Approximate Realization of Continuous Mappings by Neural Networks, *Neural Networks*, **2** (1989), pp. 183–192.

[9] W. Gautschi, A Survey of Gauss-Christoffel Quadrature Formulae, in *E.B. Christoffel, The Influence of his work on Mathematics and the Physical Sciences*, P. Butzer and F. Fehér, eds., Birkhäuser, Basel, 1981, pp. 72–147.

[10] G. M. Georgiou and C. Koutsougeras, Complex Domain Backpropagation, *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, **39** (1992), pp. 330–334.

[11] W. B. Gragg and A. Lindquist, On the Partial Realization Problem, *Linear Algebra and its Applications*, **50** (1983), pp. 277–319.

[12] G. H. Hardy and E. M. Wright, *An Introduction to the Theory of Numbers*, Oxford University Press, Oxford, 1979.

[13] U. Helmke, Waring's problem for binary forms, *Journal of Pure and Applied Algebra*, **80** (1992), pp. 29–45.

[14] K. Hornik, M. Stinchcombe and H. White, Multilayer Feedforward Networks are Universal Approximators, *Neural Networks*, **2** (1989), pp. 359–366.

[15] T. Kailath, *Linear Systems*, Prentice-Hall, Englewood Cliffs, 1980.

[16] R. E. Kalman, On Partial Realizations, Transfer Functions, and Canonical Forms, *Acta Polytechnica Scandinavica*, **31** (1979), pp. 9–32.

[17] R. E. Kalman, P. L. Falb and M. A. Arbib, *Topics in Mathematical System Theory*, McGraw-Hill, New York, 1969.

[18] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, 1966.

[19] H. Leung and S. Haykin, The Complex Backpropagation Algorithm, *IEEE Transactions on Signal Processing*, **39** (1991), pp. 2101–2104.

[20] C. Martin and M. Stamp, A Note on the Error in Gaussian Quadrature, Preprint, 1992.

[21] R. A. Silverman, *Introductory Complex Analysis*, Dover, New York, 1972.

[22] E. D. Sontag, *Mathematical Control Theory: Deterministic Finite Dimensional Systems*, Springer-Verlag, New York, 1990.

[23] E. D. Sontag and H. J. Sussmann, Backpropagation can Give Rise to Spurious Local Minima Even for Networks Without Hidden Layers, *Complex Systems*, **3** (1989), pp. 91–106.

[24] H. J. Sussmann, Uniqueness of Weights for Minimal Feedforward Nets with a Given Input-Output Map, *Neural Networks*, **5** (1992), pp. 589–593.

[25] J. J. Sylvester, An Essay on Canonical Forms, Supplement to a Sketch of a Memoir on Elimination, in *Collected Mathematical Paper I, paper 34*, 1851.

[26] R. C. Williamson and U. Helmke, Existence and Uniqueness Results for Neural Network Approximations, To appear, IEEE Transactions on Neural Networks, 1994.