

Recurrent Neural Networks: Some Systems-Theoretic Aspects*

Eduardo Sontag
Dept. of Mathematics, Rutgers University
New Brunswick, NJ 08903

sontag@control.rutgers.edu

Abstract

This paper provides an exposition of some recent research regarding system-theoretic aspects of continuous-time recurrent (dynamic) neural networks with sigmoidal activation functions. The class of systems is introduced and discussed, and a result is cited regarding their universal approximation properties. Known characterizations of controllability, observability, and parameter identifiability are reviewed, as well as a result on minimality. Facts regarding the computational power of recurrent nets are also mentioned.

*Supported in part by US Air Force Grant AFOSR-94-0293

1 Introduction

Recurrent nets have been introduced in control, computation, signal processing, optimization, and associative memory applications. Given matrices $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times m}$, $C \in \mathbb{R}^{p \times n}$, as well as a fixed Lipschitz scalar function $\sigma : \mathbb{R} \rightarrow \mathbb{R}$, the *continuous time recurrent network* Σ with *activation function* σ and *weight matrices* (A, B, C) is given by:

$$\frac{dx}{dt}(t) = \bar{\sigma}^{(n)}(Ax(t) + Bu(t)), \quad y(t) = Cx(t), \quad (1)$$

where $\bar{\sigma}^{(n)} : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is the diagonal map

$$\bar{\sigma}^{(n)} : \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \mapsto \begin{pmatrix} \sigma(x_1) \\ \vdots \\ \sigma(x_n) \end{pmatrix}. \quad (2)$$

The terminology of neural networks arises when one thinks of each coordinate x_i of the composite state x as a representation of the internal state of the i th neuron in a set of n interconnected “neurons” or processors. The rate of change of the i th dynamic element is determined by the current state of each other neuron j , either in an inhibitory or excitatory fashion (depending on the sign of the respective “synaptic strength” a_{ij}) as well as by the current values of the coordinates $u_i, i = 1, \dots, m$ of the external input signal u (similarly weighed by the b_{ij} ’s).

The role of the activation or response function σ is to saturate the total rate of change, and is motivated by the simplistic binary “fire or not fire” model of biological neurons. Typically, the function σ is of a “sigmoidal” type as illustrated in Figure 1. Most often in experimental practice as well as theory,

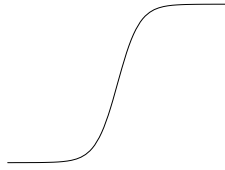


Figure 1: *Sigmoidal activation*

one takes

$$\sigma(x) = \tanh x = \frac{e^x - e^{-x}}{e^x + e^{-x}}.$$

or equivalently, up to translations and change of coordinates, the “standard sigmoid” or “logistic” function $\sigma(x) = 1/(1 + e^{-x})$. Finally, the coordinates of $y(t)$ represent the output of p probes, or measurement devices, each of which provides a weighted average of the current values $x_i(t)$ of the states of the various neurons.

As an illustration, take the system shown in Figure 2.

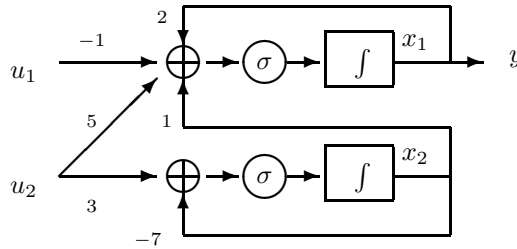


Figure 2: *Example of a two-dimensional, two-input, one-output net*

The equations for this example are

$$\frac{dx}{dt}_1 = \sigma(2x_1 + x_2 - u_1 + 5u_2), \quad \frac{dx}{dt}_2 = \sigma(-7x_2 + 3u_2), \quad y = x_1,$$

or the matrix form in (1) with

$$A = \begin{pmatrix} 2 & 1 \\ 0 & -7 \end{pmatrix}, \quad B = \begin{pmatrix} -1 & 5 \\ 0 & 3 \end{pmatrix}, \quad C = (1 \quad 0).$$

There are many variants of the basic model presented above. First of all, one may consider *discrete time* models, in which the time evolution is described

by a difference instead of a differential equation:

$$x(t+1) = \bar{\sigma}^{(n)}(Ax(t) + Bu(t)), \quad y(t) = Cx(t)$$

or an Euler approximation

$$x(t+1) = x(t) + \bar{\sigma}^{(n)}(Ax(t) + Bu(t)).$$

Second, one may consider systems in continuous time in which the right-hand side of the differential equation has a slightly different form, such as

$$\frac{dx}{dt}(t) = Dx(t) + \bar{\sigma}^{(n)}(Ax(t) + Bu(t)),$$

or

$$\frac{dx}{dt}(t) = A\bar{\sigma}^{(n)}(x(t)) + Bu(t).$$

For instance, Hopfield nets have D a diagonal matrix with negative entries (and A symmetric). The paper [2] showed how, at least for certain problems, it is possible to transform among the different models, in such a way that once that results are obtained for (1), corollaries for the variants are easily obtained. For instance, the transformation $z = Ax + Bu$ takes a recurrent net as studied in this paper into the second model: $\frac{dz}{dt}(t) = A\bar{\sigma}^{(n)}(z(t)) + Bv(t)$, where the new input is $v = \frac{du}{dt}$.

In this paper we restrict attention to the form (1). One advantage of this form is that the linear systems customarily studied in control theory are precisely those nets for which the activation σ is the identity function. This suggests that the above model may be amenable to a theoretical development parallel to linear systems theory (for which see e.g. [11]). Indeed, there are complete characterizations of basic systems theoretic properties such as controllability, observability, minimality, and parameter identifiability. This paper presents a brief survey of some such results. We also review the fact that recurrent nets can approximate arbitrary nonlinear systems (albeit in a restricted fashion). Finally, we discuss the role of recurrent nets as universal models of digital as well as analog computation.

2 System-Theory Results: Statements

We next state several results, which are discussed later in the paper in some more detail (for those results for which a proof is already available in the literature, appropriate citations will be given). For simplicity of exposition, and because that is the most often-used case in applications, we restrict all statements here to the case $\sigma = \tanh$; the later discussion will be done in somewhat more generality.

Approximation Capabilities

Recurrent nets provide universal identification models, in the restricted sense that any system can be simulated by a net, on compact subsets of the state and input-value spaces and finite time intervals. We consider systems $\widehat{\Sigma}$ (cf. [11])

$$\frac{dx}{dt} = f(x, u), \quad y = h(x) \quad (3)$$

with input space \mathbb{R}^m , output space \mathbb{R}^p and state space $\mathbb{R}^{\widehat{n}}$ (the integer \widehat{n} is called the dimension of the system) where $h : \mathbb{R}^{\widehat{n}} \rightarrow \mathbb{R}^p$ is continuous, and $f : \mathbb{R}^{\widehat{n}} \times \mathbb{R}^m \rightarrow \mathbb{R}^{\widehat{n}}$ is continuously differentiable on x for each $u \in \mathbb{R}^m$, with f and f_x jointly continuous on x and u . We assume that solutions $\widehat{x}(t, \widehat{\xi}, u)$, $t \in [0, T]$, exist for the initial value problem $\frac{dx}{dt} = f(x, u)$, $x(0) = \widehat{\xi}$, for each possible input (i.e., locally essentially bounded map $u : [0, T] \rightarrow \mathbb{R}^m$) and each initial state $\widehat{\xi} \in \mathbb{R}^{\widehat{n}}$. Suppose we are given compact subsets $K_1 \subseteq \mathbb{R}^{\widehat{n}}$ and $K_2 \subseteq \mathbb{R}^m$, as well as an $\varepsilon > 0$ and a $T > 0$. We say that the net Σ , with input and output spaces also \mathbb{R}^m and \mathbb{R}^p respectively, *simulates* $\widehat{\Sigma}$ on the sets K_1, K_2 in time T and up to accuracy ε if there exist two differentiable mappings

$$\alpha : \mathbb{R}^n \rightarrow \mathbb{R}^{\widehat{n}} \quad \text{and} \quad \beta : \mathbb{R}^{\widehat{n}} \rightarrow \mathbb{R}^n$$

so that the following property holds: For each $\widehat{x}(0) = \widehat{\xi} \in K_1$ and each $u(\cdot) : [0, T] \rightarrow K_2$,

$$\left\| \widehat{x}(t, \widehat{\xi}, u) - \alpha(x(t, \beta(\widehat{\xi}), u)) \right\| < \varepsilon, \quad \left\| h(\widehat{x}(t, \widehat{\xi}, u)) - C(x(t, \beta(\widehat{\xi}), u)) \right\| < \varepsilon$$

for all $t \in [0, T]$, where $x(t, \xi, u)$ denotes, in general, the unique solution $x : [0, T] \rightarrow \mathbb{R}^n$ of $\frac{dx}{dt} = \vec{\sigma}^{(n)}(Ax + Bu)$ with $x(0) = \xi$, given the (measurable essentially bounded) input function $u : [0, T] \rightarrow \mathbb{R}^m$, and similarly for the second system. From [9] we have:

Theorem 1 *For each system $\widehat{\Sigma}$ and for each K_1, K_2, ε, T as above, there is a net Σ , with $\sigma = \tanh$, which simulates $\widehat{\Sigma}$ on the sets K_1, K_2 in time T and up to accuracy ε .*

Controllability and Observability

Several results from now on assume a certain generic property for the input matrix B , namely that all its rows are nonzero and they are pairwise distinct even after a sign reversal. More precisely, letting $\text{row}_i(Q)$ denote the i th row of a matrix Q , we define, for each pair of positive integers n and m :

$$\mathbf{B}_{n,m} := \{B \in \mathbb{R}^{n \times m}, (\forall i) \text{row}_i(B) \neq 0 \text{ and } (\forall i \neq j) \text{row}_i(B) \neq \pm \text{row}_j(B)\}.$$

(Observe that, for the special but most important case $m = 1$, a vector $b \in \mathbf{B}_{n,1}$ if and only if all its entries are nonzero and have different absolute values.)

We omit the subscripts n, m if they are clear from the context. Since the complement of $\mathbf{B}_{n,m}$ is an algebraic subset, the complement of $\mathbf{B}_{n,m}$ has zero Lebesgue measure and is an open dense subset of $\mathbb{R}^{n \times m}$.

The net (1) is (completely) *controllable* if any state can be steered to any other state, i.e., for each pair of states $\xi, \zeta \in \mathbb{R}^n$, there is some $T \geq 0$ and some input u on $[0, T]$ such that $x(T, \xi, u) = \zeta$. (The output y is irrelevant to this definition.) When $\sigma(x) = x$, that is, for linear systems, controllability is equivalent to the requirement that the matrix pair (A, B) be a reachable pair, i.e. the rank of the $n \times nm$ matrix $(B, AB, \dots, A^{n-1}B)$ must be n . For nets with activation \tanh , we have the following from [12]:

Theorem 2 *Assume that $B \in \mathbf{B}$ and $\sigma = \tanh$. Then the net (1) is controllable.*

The net (1) is *observable* if any two states can be distinguished by input/output experiments, i.e., for each pair of states $\xi, \zeta \in \mathbb{R}^n$, there is some $T \geq 0$ and some input u on $[0, T]$ such that $Cx(T, \xi, u) \neq Cx(T, \zeta, u)$. For linear systems, observability is equivalent to the requirement that the transposed pair (A', C') be a reachable matrix pair. For nets, we have as follows, from [4]. Consider the directed graph G , with node set $\{1, \dots, n\}$, in which there is an edge from i to j whenever $a_{ij} \neq 0$. Let N be the set consisting of those nodes i for which the i th column of C is nonzero. If every node can be reached from N by some path in the graph G , we say that *every variable influences the output*.

Theorem 3 *Assume that $B \in \mathbf{B}$ and $\sigma = \tanh$. Then the net (1) is observable if and only if every variable influences the output and $\text{rank}(A', C') = n$.*

Identifiability of Parameters and Minimality

A natural question is as follows. Assume that we do manage to find a net Σ which matches *exactly* the complete i/o behavior of an observed input/output behavior. Can we then say something regarding the relation between the internal structure of the object generating the behavior (in control theoretic terminology, the “plant”) and the equations defining Σ ? We now state two results which address this question. The first one says that if the plant happened to be itself a net $\widehat{\Sigma}$, and if both nets satisfy the generic observability condition just given, then Σ and $\widehat{\Sigma}$ must be identical (up to a possible relabeling and sign change of variables). The second deals with the general case in which the plant is a more arbitrary dynamical system $\widehat{\Sigma}$. In this case, again provided Σ is observable, we can conclude that $\widehat{\Sigma}$ must be larger than Σ , in the sense that there is a natural map from a subset of the state space of $\widehat{\Sigma}$ onto that of our model Σ , compatible with the dynamics of both systems; this means in particular that Σ is a minimal model.

For any net Σ , any input function $u : [0, T] \rightarrow \mathbb{R}^m$, and any initial state $\xi \in \mathbb{R}^n$, we consider the ensuing output function $y(t, \xi, u) := Cx(t, \xi, u)$. Two

initialized nets (Σ, ξ) and $(\widehat{\Sigma}, \widehat{\xi})$ with same input and output spaces are *i/o equivalent* if $y(\cdot, \xi, u) = \widehat{y}(\cdot, \widehat{\xi}, u)$ for all inputs u , where \widehat{y} indicates the output function associated to $\widehat{\Sigma}$.

A particular change of variables possible for nets is as follows. Take any sequence

$$\varepsilon = (\varepsilon_1, \dots, \varepsilon_n) \in \{-1, 1\}^n$$

and any permutation $\pi : \{1, \dots, n\} \rightarrow \{1, \dots, n\}$. Consider the new state \widehat{x} whose coordinates $\widehat{x}_{\pi(i)} := \varepsilon_i x_i$ are obtained by exchanging the x_i 's and possibly (if $\varepsilon_i = -1$) inverting signs. This means that $\widehat{x} = Tx$, with

$$T = \text{diag}(\varepsilon_1, \dots, \varepsilon_n)(e_{\pi(1)}, \dots, e_{\pi(n)}),$$

where e_i is the i th canonical basis vector. If σ is an odd function, then $T\bar{\sigma}^{(n)}(v) = \bar{\sigma}^{(n)}(Tv)$ for all $v \in \mathbb{R}^n$. Thus, the new state \widehat{x} satisfies the equations

$$\dot{\widehat{x}} = \bar{\sigma}^{(n)}(\widehat{A}\widehat{x} + \widehat{B}u), \quad y = \widehat{C}\widehat{x},$$

with

$$\widehat{A} = TAT^{-1}, \quad \widehat{B} = TB, \quad \widehat{C} = CT^{-1}. \quad (4)$$

If $\xi \in \mathbb{R}^n$, let $\widehat{\xi} := T\xi$. Any initialized net $(\widehat{\Sigma}, \widehat{\xi})$ obtained in this fashion is said to be *sign-permutation equivalent* to (Σ, ξ) . It is easy to see that sign-permutation equivalent nets are also i/o equivalent. We have the following converse from [3]:

Theorem 4 *Assume that Σ and $\widehat{\Sigma}$ are two observable nets with $\sigma = \tanh$ and $B, \widehat{B} \in \mathbf{B}$. Then, Σ and $\widehat{\Sigma}$ are sign-permutation equivalent if and only if they are i/o equivalent.*

For the next result, we consider systems $\widehat{\Sigma}$ as in (3), except that we now also ask that h and $f : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^n$ be (real-)analytic on x . More generally, we allow the state space \widehat{X} to be any paracompact (real-)analytic connected manifold, with $h : \widehat{X} \rightarrow \mathbb{R}^p$ analytic, and a continuous mapping $f : \widehat{X} \times \mathbb{R}^m \rightarrow T\widehat{X}$ such that $\pi(f(x, u)) = x$, where $\pi : T\widehat{X} \rightarrow \widehat{X}$ is the tangent bundle projection, so that $f(x, u)$ is analytic on x and f_x continuous on $\widehat{X} \times \mathbb{R}^m$. For technical reasons, we assume completeness: for each function $u : [a, b] \rightarrow \mathbb{R}^m$ with $0 \in [a, b]$, and each $\widehat{\xi} \in \widehat{X}$, there is a solution of $\frac{dx}{dt} = f(x, u)$, $x(0) = \widehat{\xi}$, defined for all $t \in [a, b]$. As before, we may consider the outputs $\widehat{y}(t, \widehat{\xi}, u) = h(\widehat{x}(t, \widehat{\xi}, u))$, and we call two initialized systems i/o equivalent if these the outputs coincide for all possible inputs.

Theorem 5 *Assume that the initialized analytic $(\widehat{\Sigma}, \widehat{\xi})$ and the observable initialized net (Σ, ξ) with $\sigma = \tanh$ and $B \in \mathbf{B}$ are i/o equivalent. Then, there is an analytic submanifold X_0 of \widehat{X} and an analytic onto mapping $\Pi : X_0 \rightarrow \mathbb{R}^n$, such that $h(q) = C\Pi(q)$ for all $q \in X_0$ and, for each input $u : [0, T] \rightarrow \mathbb{R}^m$ and each $t \in [0, T]$, $\widehat{x}(t, \widehat{\xi}, u) \in X_0$ and $\Pi(\widehat{x}(t, \widehat{\xi}, u)) = x(t, \xi, u)$. In particular, the dimension of Σ is minimal among all systems i/o equivalent to it.*

3 System-Theory Results: Discussion

Theorems 1, 2, 3, and 4 hold for activations σ more general than \tanh , as we discuss next. (In every case, in addition to the conditions stated, one assumes that $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ is locally Lipschitz, so that solutions of the evolution equations are defined at least locally.)

Approximation

Theorem 1 is from [9]. It is proved for any σ that has the following *spanning property*: the linear span of the functions $\sigma(as + b)$, with $a, b \in \mathbb{R}$, that is, the set of all finite linear combinations

$$\sum_i c_i \sigma(a_i s + b_i),$$

restricted to any finite interval $[\alpha, \beta] \subset \mathbb{R}$, constitute a dense subset of $C^0[\alpha, \beta]$, the set of continuous functions on $[\alpha, \beta]$ endowed with the metric of uniform convergence.

Not every function has the spanning property. For instance, if σ is a polynomial of degree k then the above span is the set of all polynomials of degree $\leq k$, hence it forms a closed subspace and cannot be dense. This turns out to be the only exception: [6] shows that any locally Riemann integrable σ (i.e., any function which is continuous except at most in a set of measure zero, and bounded on each compact) has the spanning property *if and only if it is not a polynomial*.

Controllability

Theorem 2 is from [12]. It is proved for σ odd and with the properties that there exists $\lim_{s \rightarrow +\infty} \sigma(s) = \sigma_\infty > 0$, $\sigma(s) < \sigma_\infty$ for all $s \in \mathbb{R}$, and, for each $a, b \in \mathbb{R}$ with $a > 1$,

$$\lim_{s \rightarrow +\infty} \frac{\sigma_\infty - \sigma(as + b)}{\sigma_\infty - \sigma(s)} = 0.$$

This latter asymptotic property is essential; for instance, the sigmoid arctan does not satisfy it, and in fact the Theorem is false for $\sigma = \arctan$.

The proof of the Theorem is based on establishing that the positive cone generated by the vector fields

$$\{\vec{\sigma}^{(n)}(Ax + Bu), u \in \mathbb{R}^m\}$$

equals the tangent space at each point x of the state space, which provides local controllability at each state.

Observe that there are no assumptions on A . In fact, the condition that $B \in \mathbf{B}$ is necessary in the following sense: if $B \in \mathbb{R}^{n \times m}$ is so that for all

$A \in \mathbb{R}^{n \times n}$ the system (1) is controllable, then $B \in \mathbf{B}$; however, for a specific A it may very well happen that the net is controllable even if $B \notin \mathbf{B}$.

A related fact is that “forward accessibility” (the reachable set from each state has nonempty interior) holds for every net as in Theorem 2, provided that σ has the “IP property” to be discussed below. This result had been earlier shown in the paper [1] (which dealt mainly with accessibility for the much harder discrete-time case). It is an immediate consequence of the fact that, when the IP property holds, the *linear span* of $\{\vec{\sigma}^{(n)}(Ax + Bu), u \in \mathbb{R}^m\}$ equals the tangent space at each point x .

Observability

Theorem 3 is from [4]. It is proved for every σ that satisfies the *independence property* (IP). This property is basically a dual to the spanning property. For odd σ , it states that translates and dilations of σ must be linearly independent: for any positive integer l , any l -tuple of distinct pairs (a_i, b_i) with $a_i > 0$, the functions $1, \sigma(a_1s + b_1), \dots, \sigma(a_ls + b_l)$ are linearly independent, i.e.,

$$c_0 + \sum_{i=1}^l c_i \sigma(a_i s + b_i) \equiv 0 \Rightarrow c_0 = c_1 = \dots = c_l = 0.$$

(A variation of the property, more interesting for non-odd σ , asks linear independence of pairs (a_i, b_i) with $a_i \neq 0$ but now requiring also $(a_i, b_i) \neq -(a_j, b_j)$ for all $i \neq j$.)

A simple sufficient condition can be used to show that many maps, including \tanh and \arctan , satisfy the IP property (cf. [4]): it is enough that σ admit an extension as a complex analytic function $\sigma : \mathbb{C} \rightarrow \mathbb{C}$ defined on a subset of the form

$$\{|\operatorname{Im}z| \leq \lambda\} \setminus \{z_0, \bar{z}_0\}$$

for some $\lambda > 0$, where $\operatorname{Im}z_0 = \lambda$ and z_0 and \bar{z}_0 are singularities. Another way of establishing the IP property is by an asymptotic analysis of σ , in the spirit as in the statement given above for the controllability property; this was the approach taken in [14]. For instance, cf. [5], σ has the IP property if it is continuously differentiable, $\sigma(s)/\sigma'(s)$ is defined and has constant sign for all s large, and:

$$\lim_{s \rightarrow +\infty} \frac{\sigma(s)}{\sigma'(s)} = 0.$$

As remarked in [5], this establishes the IP property whenever $\sigma(s) = q(s)e^{p(s)}$, and p, q are polynomials with $\deg p \geq 2$. Even weaker conditions from [5] are to require that for each $b > 0$, $\sigma(s + b)/\sigma(s)$ be defined and bounded for all sufficiently large s , and

$$\sigma(s + b)/\sigma(s) \rightarrow 0 \text{ as } s \rightarrow +\infty,$$

or that the same property hold for $1/\sigma$.

The condition that every variable affects the output can be equivalently stated in terms of invariant subspaces. This provides an elegant connection to the case of linear systems, since for the latter observability means that there is no nonzero A -invariant subspace of the kernel of C . To be precise, the condition means that there cannot exist any nonzero subspace of $\ker C$ which is invariant under A and also under all θ_i , $i \in \{1, \dots, n\}$, where

$$\theta_i : \mathbb{R}^n \rightarrow \mathbb{R}^n$$

is the projection on the i th axis, i.e., $\theta_i e_j = \delta_{ij} e_i$. (We let δ_{ij} be the Kronecker delta and e_i the i th canonical basis vector.)

Parameter Identifiability

Theorem 4 is from [3]. It is proved there for every σ that is odd and satisfies the IP property. Thus it holds as well for any σ for which any of the sufficient conditions stated above are verified.

Minimality

Theorem 5 does not appear to have been mentioned in the literature. It is an easy consequence of the uniqueness theorem for minimal realizations, as we describe next.

The restriction of the dynamics of $\widehat{\Sigma}$ to the orbit X_0 passing through the initial state $\widehat{\xi}$ provides an initialized system $(\widehat{\Sigma}_0, \widehat{\xi}_0)$ which is orbit-minimal in the sense of [13] and is again i/o equivalent to (Σ, ξ) . One may then apply Theorem 1 in [13] to conclude that there is also an initialized analytic system $(\widehat{\Sigma}', \widehat{\xi}')$ with state space X' , i/o equivalent to (Σ, ξ) and minimal in the sense of [13], and an analytic onto mapping

$$\Pi_0 : X_0 \rightarrow X'$$

such that $h(q) = h'(\Pi_0(q))$ for all $q \in X_0$ and, for each input $u : [0, T] \rightarrow \mathbb{R}^m$ and each $t \in [0, T]$, $\widehat{x}(t, \widehat{\xi}, u) \in X_0$ and (with the obvious notations)

$$\Pi_0(\widehat{x}(t, \widehat{\xi}, u)) = \widehat{x}'(t, \xi, u).$$

(The statement of Theorem 1 in [13] is somewhat weaker than this, but the proof actually shows the claimed facts.) Next, Theorem 5 in [13], applied to the two minimal systems $(\widehat{\Sigma}', \widehat{\xi}')$ and (Σ, ξ) provides an isomorphism Π_1 , which composed with Π_0 provides the mapping desired for Theorem 5 in this paper.

In fact, a stronger result holds as well, namely, if the orbit X_0 equals the whole space X and if $\widehat{\Sigma}$ is observable, then Π is a diffeomorphism.

4 Computational Power

We close with a mention of results regarding computational capabilities of recurrent networks, seen from the point of view of classical formal language theory. The papers [7, 8] considered discrete-time networks with the “semilinear” or “saturated linearity” activation

$$\pi(x) = \begin{cases} -1 & \text{if } x \leq -1 \\ 1 & \text{if } x \geq 1 \\ x & \text{otherwise.} \end{cases}$$

It is assumed, for simplicity (but not changing the results in any substantial way) that there are just one input and one output channel ($m = p = 1$). The cited papers established that with rational weights recurrent networks are computationally equivalent, up to polynomial time, to Turing machines, and with real weights to a large class of “analog computers”. (With no time constraints, all possible binary functions, recursive or not, are “computable” in exponential time by real-weight machines.)

Formally, we say that a pair consisting of a recurrent network Σ and an initial state $\xi \in \mathbb{R}^n$ is *admissible* if, for every input of the special form

$$u(\cdot) = \alpha_1, \dots, \alpha_k, 0, 0, \dots, \quad (5)$$

where each $\alpha_i = \pm 1$ and $1 \leq k < \infty$, the output that results with $x(0) = \xi$ is either $y \equiv 0$ or y is a sequence of the form

$$y(\cdot) = \underbrace{0, 0, \dots, 0}_s, \beta_1, \dots, \beta_l, 0, 0, \dots, \quad (6)$$

where each $\beta_i = \pm 1$ and $1 \leq l < \infty$. A *rational* (Σ, ξ) is one for which the matrices defining Σ , and ξ , all have rational entries. (In that case, for rational inputs all ensuing states and outputs remain rational.) Given an admissible (Σ, ξ) , there is an associated partial function

$$\phi : \{-1, 1\}^+ \rightarrow \{-1, 1\}^+,$$

where $\{-1, 1\}^+$ is the free semigroup in the two symbols ± 1 , given as follows: for each sequence

$$w = \alpha_1, \dots, \alpha_k,$$

consider the input in Equation (5) and its corresponding output, which is either identically zero or has the form in Equation (6). If $y \equiv 0$, then $\phi(w)$ is undefined; otherwise, if Equation (6) holds, then $\phi(w)$ is defined as the sequence β_1, \dots, β_l . In the latter case, we say that the response to the input sequence w was computed *in time* $s+l$. If ϕ is obtained in this form, the (partial) function ϕ is said to be *realized* by the initialized network (Σ, ξ) . It is shown in [7] that any partial function $\phi : \{-1, 1\}^+ \rightarrow \{-1, 1\}^+$ can be realized by some admissible pair, and ϕ can be realized by some rational admissible pair if and only if ϕ is a partial recursive function.

Constraints in computational time are of course more interesting. Restricting for simplicity to language recognition, the results can be summarized as follows. If $\phi(w)$ is defined for all inputs and if there is a function on positive integers $T : \mathbb{N} \rightarrow \mathbb{N}$ so that the response to each sequence w is computed in time at most $T(|w|)$, where $|\alpha_1, \dots, \alpha_k| = k$, then (Σ, ξ) is said to *compute in time* T . If ϕ is everywhere defined and

$$\phi : \{-1, 1\}^+ \rightarrow \{-1, 1\},$$

that is, the length of the output is always one, one can think of ϕ as the characteristic function of a subset L of $\{-1, 1\}^+$, that is, a *language* over the alphabet $\{-1, 1\}$. Given $T : \mathbb{N} \rightarrow \mathbb{N}$, the language L is *recognizable in time* T if the corresponding characteristic function is, for some admissible pair that computes in time T . It can be proved that languages recognizable in polynomial time by rational admissible pairs are exactly those in the class P of polynomial-time recursive languages. Using real weights, a new class, “analog P,” arises. This class can be characterized as the class of all languages recognizable by arbitrary nonlinear (but Lipschitz-continuous) dynamical systems, see [7] for details. The class analog P strictly contains P, and it turns out to coincide with a class already studied in computer science, namely the languages recognized in polynomial time by Turing machines which consult oracles, where the oracles are sparse sets. This gives a precise characterization of the power of recurrent nets in terms of a known complexity class. The following table summarizes the results just discussed:

Weights	Capability	Polytime
integer	regular	regular
rational	recursive	(usual) P
real	arbitrary	analog P

5 Some Remarks

It would be quite interesting to have complete characterizations of controllability in the case when the matrix B does not belong to \mathbf{B} . It is easy to see that the block matrix $[A, B]$ must be in \mathbf{B} (defined for sizes n by $n + m$), but useful necessary and sufficient conditions are unknown.

We have avoided discussion of system-theoretic issues for discrete-time networks. Approximation, observability, and identifiability results are known for the discrete time case, and most are similar to those for continuous time (see the respective references). The controllability case is still open, though partial characterizations are known (see [1]).

References

- [1] Albertini, F., and P. Dai Pra, "Forward accessibility for recurrent neural networks," *IEEE Trans. Automat. Control* **40** (1995): 1962-1968
- [2] Albertini, F., and E.D. Sontag, "For neural networks, function determines form," *Neural Networks* **6**(1993): 975-990.
- [3] Albertini, F., and E.D. Sontag, "Uniqueness of weights for recurrent nets," *Systems and Networks: Math Theory and Applics*, Proc. MTNS '93, Vol. 2, Akademie Verlag, Regensburg, pp. 599-602. Extended version: http://www.math.rutgers.edu/sontag/FTP_DIR/93mtns-nn-extended.ps.gz
- [4] Albertini, F., and E.D. Sontag, "State observability in recurrent neural networks," *Systems & Control Letters* **22**(1994): 235-244.
- [5] Hautus, M., "A set of IP-functions," unpublished manuscript, Eindhoven University, August 1993.
- [6] Leshno, M., V.Ya. Lin, A. Pinkus, and S. Schocken, "Multilayer feedforward networks with a non-polynomial activation function can approximate any function," *Neural Networks* **6**(1993): 861-867.
- [7] Siegelmann, H.T., and E.D. Sontag, "Analog computation, neural networks, and circuits," *Theor. Comp. Sci.* **131**(1994): 331-360.
- [8] Siegelmann, H.T., and E.D. Sontag, "On the computational power of neural nets," *J. Comp. Syst. Sci.* **50**(1995): 132-150.
- [9] Sontag, E.D., "Neural nets as systems models and controllers," in *Proc. Seventh Yale Workshop on Adaptive and Learning Systems*, pp. 73-79, Yale University, 1992.
- [10] Sontag, E.D., "Neural networks for control," in *Essays on Control: Perspectives in the Theory and its Applications* (H.L. Trentelman and J.C. Willems, eds.), Birkhauser, Boston, 1993, pp. 339-380.
- [11] Sontag, E.D., *Mathematical Control Theory: Deterministic Finite Dimensional Systems*, Springer, New York, 1990.
- [12] Sontag, E.D., and H.J. Sussmann, "Complete controllability of continuous-time recurrent neural networks," *Systems and Control Letters* **30**(1997): 177-183.
- [13] Sussmann, H.J., "Existence and uniqueness of minimal realizations of nonlinear systems," *Math. Sys. Theory* **10**(1977): 263-284.
- [14] Sussmann, H.J., "Uniqueness of the weights for minimal feedforward nets with a given input-output map," *Neural Networks* **5**(1992): 589-593.
- [15] Zbikowski, R., "Lie algebra of recurrent neural networks and identifiability," *Proc. Amer. Auto. Control Conf.*, San Francisco, 1993, pp.2900-2901.