Reprinted from: **Artificial Neural Networks for Speech and Vision**
(Proc. Workshop held at Rutgers, 1992, Edited by Richard J. Mammone)
Chapman and Hall, London, 1993, ISBN 0 412 54850 X, pp. 113-125

UNIQUENESS OF WEIGHTS FOR NEURAL NETWORKS

Francesca Albertini* and Eduardo D. Sontag†
Department of Mathematics
Rutgers University, New Brunswick, NJ 08903
Vincent Maillot‡
Département de Mathématiques et d'Informatique
Ecole Normale Supérieure, 75005 Paris

1 Introduction

In most applications dealing with learning and pattern recognition, neural nets are employed as models whose parameters, or "weights," must be fit to training data. Gradient descent and other algorithms are used in order to minimize an error functional, which penalizes mismatches between the desired outputs and those that a candidate net —with a fixed architecture and varying weights— produces.

There are many numerical issues that arise naturally when using such a design approach, in particular: (i) the possibility of local minima which are not globally optimal, and (ii) the possibility of multiple global minimizers. The first question was dealt with by many different authors —see for instance [5, 13, 14]— and will not reviewed here. Regarding point (ii), observe that there are obvious transformations that leave the behavior of a network invariant, such as interchanges of all incoming and outgoing weights between two neurons, that is the relabeling of neurons, or, for odd activation functions, flipping the signs of all incoming and outgoing weights at any given node. Two networks differing in such a manner give the same error on the training data. When there is a net that fits perfectly the data, all nets differing from it by one of the above transformations also attain the global minimum (zero) of the error functional.

*This research was supported in part by US Air Force Grant AFOSR-91-0343. Currently also at Universita' di Padova, Dipartimento di Matematica, 35100 Padova, Italy. E-mail address: albertin@hilbert.rutgers.edu

†This research was supported in part by US Air Force Grant AFOSR-91-0343. E-mail address: sontag@hilbert.rutgers.edu

‡E-mail address: vmaillot@dmi.ens.fr

A natural question, asked by Hecht-Nielsen in [10], is to what extent are neuron exchanges and sign flips the only transformations that generically occur. If indeed these are the only possible ones, then essentially all the internal structure is uniquely determined by the external behavior of the network. Moreover, the set of invariant transformations is then finite. (One may want to build additional symmetries into a network, in order to increase representational bias, by imposing artificial conditions such as asking that certain weights be equal, which is helpful in designing networks that focus on invariances in the input patterns; see e.g. [7]. This leads to richer transformation groups, but that is a different issue than the one treated here.)

Various conditions can be given which assure that equality of behaviors between two networks implies equality up to neuron relabeling and signs flips. An important consequence in those cases in which the conditions apply is that there is no possible dimensionality reduction in the parameter space, contrary to the situation in classical linear identification, where canonical forms have to be introduced in order to achieve parameter identifiability. (Seen more positively, the parameterizations provided by neural networks have very little redundancy.) In this short expository survey, we sketch various known facts about this issue, including recent results about recurrent nets, and we provide a new and simple proof of a uniqueness result that applies in the single hidden layer case.

2 Single-Hidden Layer Nets

Let $\sigma : \mathbb{R} \to \mathbb{R}$ be any function, and let m, n, p be positive integers. A *single-hidden layer net with m inputs, p outputs, n hidden units, and activation function σ* is specified by a pair of matrices B, C and a pair of vectors β, c_0, where B and C are respectively real matrices of sizes $n \times m$ and $p \times n$, and β and c_0 are respectively real vectors of size n and p. We denote such a net by a 5-tuple

$$\Sigma = \Sigma(B, C, \beta, c_0, \sigma)\,,$$

omitting σ if obvious from the context. In particular, Σ has *no offsets* if $\beta = c_0 = 0$ (the terminology "biases" or "thresholds" is sometimes used instead of offsets).

For simplicity, we will assume from now on that $p = 1$; generalizations to the multiple-output case are not hard but they complicate the notations. Thus, from now on, C is a row n-vector and c_0 is a constant.

Let $\vec{\sigma}_n : \mathbb{R}^n \to \mathbb{R}^n$ indicate the application of σ to each coordinate of an n-vector:

$$\vec{\sigma}_n(x_1, \ldots, x_n) \;=\; (\sigma(x_1), \ldots, \sigma(x_n))\,.$$

(We will drop the subscript as long as its value is clear from the context.) The *behavior* of Σ is defined to be the map

$$\mathrm{beh}_\Sigma \;:\; \mathbb{R}^m \to \mathbb{R} \;:\; u \mapsto C\vec{\sigma}(Bu + \beta) + c_0\,.$$

In other words, the behavior of a network is a composition of the type $f \circ \vec{\sigma} \circ g$, where f and g are affine maps. Given two networks Σ and $\hat{\Sigma}$, we say that they are (input/output) *equivalent* and denote

$$\Sigma \sim \hat{\Sigma}\,,$$

if $\mathrm{beh}_\Sigma = \mathrm{beh}_{\hat{\Sigma}}$ (equality of functions). The question to be studied, then, is: when does $\Sigma \sim \hat{\Sigma}$ imply $\Sigma = \hat{\Sigma}$?

Consider first the case in which σ is the identity. In that case, $\mathrm{beh}_\Sigma = CBu + (C\beta + c_0)$, and we see that any two nets Σ giving rise to the same products CB and $C\beta + c_0$ have the same behavior. Assume that $\Sigma \sim \hat{\Sigma}$. Under suitable minimality conditions (B, $\hat{B}$ of full row rank and C, $\hat{C}$ of full column rank), there must exist an invertible matrix T such that $\hat{C} = CT$, $\hat{B} = T^{-1}B$, and $\hat{c}_0 = C(\beta - T\hat{\beta}) + c_0$. Conversely, for any given Σ, any such T, and any $\hat{\beta}$, the above formulas define a $\hat{\Sigma}$ which is equivalent to the given one. Thus, uniqueness is very far from being satisfied. The same argument applies if σ is any linear map. Observe, as the same fact will be needed later, that without minimality assumptions nothing at all can be concluded; for instance, if $B, \hat{B}, \beta, \hat{\beta}, c_0, \hat{c}_0$ all vanish, one has $\Sigma \sim \hat{\Sigma}$ but there need be no relation among C and $\hat{C}$.

One might at first think that nonlinear maps σ provide uniqueness up to finitely many symmetries. But it is easy to see that far more is needed. For instance, such a property cannot hold for polynomials, nor for periodic functions, nor for the exponential function (see below). Thus one is led to the search for easily verifiable conditions on the mapping σ which imply the desired property. We formalize what is needed:

Definition 2.1 The function σ satisfies the *independence property* ("**IP**" from now on) if, for every positive integer l, nonzero real numbers $b_1, \ldots, b_l$, and real numbers $\beta_1, \ldots, \beta_l$ for which the pairs (b_i, β_i), $i = 1, \ldots, l$ satisfy

$$(b_i, \beta_i) \;\neq\; \pm(b_j, \beta_j) \;\; \forall i \neq j\;,$$

it must hold that the functions

$$1\,,\ \sigma(b_1x+\beta_1)\,,\ \dots\,,\ \sigma(b_lx+\beta_l)$$

are linearly independent. The function σ satisfies the *weak* independence property ("**WIP**") if the above linear independence property is only required to hold for all pairs with $\beta_i = 0$, $i = 1, \dots, l$. □

Observe that the independence condition is:

$$c_0 + \sum_{i=1}^{l} c_i\sigma(b_ix+\beta_i) \;=\; 0 \;\; \forall x \in \mathbb{R} \;\;\Rightarrow\;\; c_0 = c_1 = \ldots = c_l = 0\,.$$

This is the property needed for the desired uniqueness results, as we discuss next.

Recall that some sort of nontriviality hypothesis is needed. Let $\Sigma(B, C, \beta, c_0, \sigma)$ be given, and denote by B_i the transpose of the ith row of the matrix B and by c_i and β_i the ith entries of C and β respectively. With these notations, $\mathrm{beh}_\Sigma(u) = c_0 + \sum_{i=1}^n c_i\sigma(B_iu+\beta_i)$. As in [15], we say that Σ is *irreducible* if the following properties hold:

1. $c_i \neq 0$ for each $i = 1, \dots, n$.

2. $B_i \neq 0$ for each $i = 1, \dots, n$.

3. $(B_i, \beta_i) \neq \pm(B_j, \beta_j)$ for all $i \neq j$.

Given $\Sigma(B, C, \beta, c_0, \sigma)$, a *sign-flip* operation consists of simultaneously reversing the signs of c_i, B_i, and β_i, for some i. A *node-permutation* consists of interchanging (c_i, B_i, β_i) with (c_j, B_j, β_j), for some i, j. Given two nets Σ and $\hat{\Sigma}$, we say that they are *equivalent* if $n = \hat{n}$ and (B, C, β, c_0) can be transformed into $(\hat{B}, \hat{C}, \hat{\beta}, \hat{c}_0)$ by means of a finite number of sign-flips and node-permutations. Of course, equivalent nets have the same behavior (since σ has been assumed to be odd). The next simple remark establishes the connection between the concepts just introduced. The proof is adapted from [15]. We assume for simplicity that the function σ is odd, but it is easy to generalize this in various ways.

Lemma 2.2 Let σ be odd and satisfy property **IP**. Assume that Σ and $\hat{\Sigma}$ are both irreducible, and $\Sigma \sim \hat{\Sigma}$. Then, Σ and $\hat{\Sigma}$ are equivalent. If σ only satisfies **WIP**, the same statement is true for nets with no offsets.

Proof. Assume that Σ and $\hat{\Sigma}$ are as in the statement, so

$$C\vec{\sigma}(Bu+\beta)+c_0 \;=\; \hat{C}\vec{\sigma}(\hat{B}u+\hat{\beta})+\hat{c}_0 \quad \text{for all } u \in \mathbb{R}^m\,. \tag{1}$$

Pick any fixed vector $\bar{u} \in \mathbb{R}^m$ such that:

- $B_i\bar{u} \neq 0$ and $\hat{B}_i\bar{u} \neq 0$ for all $i = 1, \ldots, n$,
- $(B_i\bar{u}, \beta_i) \neq \pm(B_j\bar{u}, \beta_j)$ and $(\hat{B}_i\bar{u}, \hat{\beta}_i) \neq \pm(\hat{B}_j\bar{u}, \hat{\beta}_j)$ for all $i \neq j$.

Such vectors exist, because we only need to avoid the union of the hyperplanes in $\mathbb{R}^m$ determined by each of the equations: $B_iu = 0$, $(B_i+B_j)u = 0$ for each i, j for which $\beta_i = -\beta_j$, $(B_i - B_j)u = 0$ for each i, j for which $\beta_i = \beta_j$, and the corresponding ones for $\hat{\Sigma}$.

In particular, we may consider elements $u \in \mathbb{R}^m$ of the form $u = \bar{u}x$ in Equation (1). With the notations $b_i = B_i\bar{u}$ and $\hat{b}_i = \hat{B}_i\bar{u}$, we obtain the identity

$$(c_0 - \hat{c}_0) + \sum_{i=1}^{n} c_i\sigma(b_ix+\beta_i) - \sum_{i=1}^{n} \hat{c}_i\sigma(\hat{b}_ix+\hat{\beta}_i) \;=\; 0 \quad \text{for all } x \in \mathbb{R}\,.$$

If the functions 1, $b_ix+\beta_i$, $\hat{b}_ix+\hat{\beta}_i$, $i = 1, \ldots, n$ are linearly independent, then all $c_i = 0$, contradicting irreducibility. Since property **IP** holds (or **WIP**, in the case of nets with no offsets, for which all $\beta_i = \hat{\beta}_i = 0$), the only way in which linear independence can fail is if some b_i or $\hat{b}_i$ vanishes, which cannot be the case because of the choice of $\bar{u}$, or —since also by construction $(b_i, \beta_i) \neq \pm(b_j, \beta_j)$ and similarly for the (b_i, β_i)'s— if $(b_i, \beta_i) = \pm(\hat{b}_j, \hat{\beta}_j)$ for some i, j. Thus, using that σ is odd, we may relabel indices, apply if necessary a sign-flip and collect these two terms, there resulting an equation:

$$(c_0 - \hat{c}_0) + (c_1 - \varepsilon\hat{c}_1)\sigma(b_1x+\beta_1) + \sum_{i=2}^{n} c_i\sigma(b_ix+\beta_i) - \sum_{i=2}^{n} \hat{c}_i\sigma(\hat{b}_ix+\hat{\beta}_i) \;=\; 0$$

with $\varepsilon = \pm1$, where now no pair (b_i, β_i) or $(\hat{b}_i, \hat{\beta}_i)$ equals $\pm(b_1, \beta_1)$. We may iterate this argument until all terms have been collected, which leads to an equation such as

$$(c_0 - \hat{c}_0) + \sum_{i=1}^{n}(c_i - \varepsilon_i\hat{c}_i)\sigma(b_ix+\beta_i) \;=\; 0\,.$$

Once more using property **IP**, this implies that $c_0 = \hat{c}_0$ and $c_i = \varepsilon_i\hat{c}_i$ for all i, completing the proof. ∎

Remark 2.3 For infinitely differentiable σ, there is a slightly different argument that can be used in the above proof, instead of the choice of a direction $\bar{u}$, but which makes the stronger assumption that all derivatives of σ satisfy **IP** or **WIP**. This argument was given in [8], which dealt with projection-pursuit algorithms in statistics, an area closely related to neural networks; we sketch the idea next. Again, we need to reduce an equation such as (1) to the scalar case. To do this, we apply a sequence of partial derivation operators $w_k \cdot \nabla$, where each w_k is chosen so as to kill one direction at a time among the vectors B_i, while the rest of the directions provide a nonzero inner product. After this procedure, there results a scalar linear dependence involving a derivative of σ instead of σ itself. □

Our goal is then to explore easily verifiable and weak conditions for **IP** and **WIP** to hold.

2.1 The Property WIP

Characterizing **WIP** is especially easy, and very classical: for odd analytic functions σ, property **WIP** holds if and only if σ is not a polynomial.

Lemma 2.4 If σ is a polynomial, **WIP** does not hold. Conversely, if σ is odd and infinitely differentiable, and if there are an infinite number of nonzero derivatives $q_k = \sigma^{(k)}(0)$, then σ satisfies property **IP**.

Proof. If σ is a polynomial of degree r, the functions $\sigma(b_i x)$ are all polynomials of degree r, and hence are linearly dependent, for any choice of distinct and positive numbers b_i, $i = 1, \ldots, r+2$. For the converse, we need to see that $c_0 + \sum_{i=1}^{l} c_i \sigma(b_i x) \equiv 0$ implies that $c_0 = c_1 = \ldots = c_l = 0$, assuming that all the b_i are nonzero and have different absolute values. Since σ is odd, $\sigma(0) = 0$, so $c_0 = 0$. Furthermore, we may assume after sign-flips if necessary, that all $b_i > 0$. Taking derivatives of various orders, and evaluating at $x = 0$, one obtains $q_k \sum_{i=1}^{l} c_i b_i^k = 0$ for all k. Let $C = (c_1, \ldots, c_l)$. Picking l nonzero derivatives q_{k_j}, $j = 1, \ldots, l$, there results that $CM = 0$, where M is the generalized Vandermonde matrix with entries $b_i^{k_j}$. It is a classical fact that such a matrix is nonsingular (Descartes' rule of signs), so $C = 0$ as desired. ∎

Thus the conditions in Lemma 2.2 are satisfied for many interesting nonlinearities, for the case of nets with no offset. Nets with no offsets appear naturally in signal processing and control applications, as there it is often the case that one requires that the zero input signal causes no effect, corresponding to equilibrium initial states for a controller or filter.

Even stronger results can be proved if constraints are imposed on the matrix B. For instance, one may require that the successive rows of B have the form $(d_1, \dots, d_k, 0, \dots, 0)$, $(0, d_1, \dots, d_k, 0, \dots, 0)$, $(0, 0, d_1, \dots, d_k, 0, \dots, 0)$, ... Such a constraint is natural if one is dealing with a composition

$$f \circ \sigma \circ g \,,$$

where f, g are finite impulse response filters, and the inputs u are thought of as time signals. (The d_i's are the coefficients of the filter g; the coefficients defining f are the entries of C.) If any d_i is nonzero, then all rows of B are nonzero, and it holds automatically that $B_i \neq \pm B_j$ for all $i \neq j$; essentially, due to the regularity of B, one is dealing here with a case closer to that of one neuron ($n = 1$) than general n. The uniqueness result in this context is essentially what was proved in [3]. (Actually, [3] treated more general time-invariant linear systems than FIR filters, as well as a continuous-time version, and in [4] the authors generalized their work to other structures containing one scalar nonlinearity. That work was in turn motivated by the older work [11] and [9] which dealt with interconnections of linear systems and memory free nonlinearities.)

2.2 The Property IP

It appears to be harder to obtain elegant characterizations of the stronger property **IP**. For obvious examples of functions not satisfying **IP**, take $\sigma(x) = e^x$, any periodic function, or any polynomial. One case is relatively simple: the one concerning dependence equations in which all $b_i = 1$ in the definition of the property **IP**. Now the only condition left is that the elements β_i must be all distinct. Given an equation

$$c_0 + \sum_{i=1}^{l} c_i \sigma(x + \beta_i) \;\equiv\; 0 \,, \qquad (2)$$

taking Fourier transforms results in the desired conclusion that all c_i must vanish, as long as σ is not identically zero (a.e.). Of course, many functions do not admit Fourier transforms, but observe that any linear combination of functions satisfying a nontrivial identity as above again satisfies a similar identity (with larger l). Thus, for instance, if there is a nonzero linear combination of translates of σ which is integrable, then σ itself cannot satisfy such an equation. An example is any "squashing" function (σ is measurable, nondecreasing, and bounded), in which case $\sigma(x+1)-\sigma(x-1)$ is in L^1.

The most interesting case, for neural network applications, is $\sigma(x) = \tanh(x)$, or equivalently after a linear transformation, the standard sigmoid $\frac{1}{1+e^{-x}}$. (It is more convenient to work with $\tanh(x)$, as it is odd.) For this function σ, consider first again the case when $b_i = 1$ and equation (2). From this equation, with a change of variables $z = e^{-2x}$ we obtain

$$\hat{c}_0 + \sum_{i=1}^{l} \frac{\hat{c}_i}{q_i + z} \equiv 0 \tag{3}$$

where $\hat{c}_0 = c_0 - \sum c_i$, $\hat{c}_i = 2c_i q_i$, and $q_i = e^{2\beta_i}$. Taking the limit as $z \to +\infty$, we have that $\hat{c}_0 = 0$. We may consider the identity (3) over the complexes (analytic continuation), and take residues at the various $z = -q_i$; from here one concludes the desired linear independence.

In place of the residue argument, we may instead use a formula due to Cauchy, which shows the stronger fact that from the values of the right-hand side of (3) at any l points $z_1, \ldots, z_l$ one can retrieve the c_i's uniquely: let M be the matrix with entries $M_{ij} = \frac{1}{q_i + z_j}$, then ([6], Section 11.3):

$$\det M = \frac{\prod_{i>j}(z_i - z_j)(q_i - q_j)}{\prod_{i,j}(q_i + z_j)} \neq 0 \, .$$

Thus $(\hat{c}_1, \ldots, \hat{c}_l)M = 0$ implies once more that all $\hat{c}_i = 0$.

The reduction of questions about tanh-nets to questions about rational functions, by means of the transformation $z = e^{-2x}$, formed the basis of the approach taken in [13] to study local minima of gradient descent; see also the recent work [16], which carries this much further into deeper questions of approximation theory. A similar reduction can be done whenever the b_i are rational numbers, but the above proof works only under the assumption that all the b_i are equal. However, for the particular function $\sigma = \tanh$, the full property **IP**, with no further restrictions on the b_i's, was established by Sussmann in [15], using a very different argument. We wish to show now that a residue type of argument works in general, and in the process we extend considerably the class of functions to which it applies.

Theorem 1 *Assume that σ is a real-analytic function, and it extends to an analytic function $\sigma : \mathbb{C} \to \mathbb{C}$ defined on a subset $D \subseteq \mathbb{C}$ of the form:*

$$D = \{z \mid |\mathrm{Im}\, z| \leq \lambda\} \setminus \{z_0, \bar{z}_0\}$$

for some $\lambda > 0$, where $\operatorname{Im} z_0 = \lambda$ and z_0 and $\bar{z}_0$ are singularities, that is, there is a sequence $z_n \to z_0$ so that $|\sigma(z_n)| \to \infty$, and similarly for $\bar{z}_0$. Then, σ satisfies property **IP***.*

Proof. Assume that

$$c_0 + c_1\sigma(b_1 z + \beta_1) + \ldots + c_r\sigma(b_r z + \beta_r) \;\equiv\; 0 \tag{4}$$

is an equation of linear dependence, with r as small as possible. Thus $c_i \neq 0$ for all $i = 1, \ldots, r$. Without loss of generality, we may assume that $|b_1| \geq b_i$ for all $i > 1$. After a change of variables $b_1 z + \beta_1 \to z$, we have that $(b_1, \beta_1) = (1, 0)$, $(b_i, \beta_i) \neq \pm(1, 0)$ for all $i \geq 2$, and $b_i \leq 1$ for all i. Thus, by the assumptions on singularities, $b_i z_0 + \beta_i$ is not a singularity of σ, for all $i \geq 2$. Dividing the expression in (4) by $\sigma(z)$ and taking limits as $z \to z_0$, we obtain $c_1 = 0$, a contradiction. ∎

A typical example of a σ satisfying the hypotheses of the theorem is that of a σ having a meromorphic extension which has a unique pole of minimal positive real part. Most rational functions satisfy this, as well as the main example in neural networks research, $\sigma(x) = \tanh(x)$. In this case, the set of poles is the set $\{(k\pi/2)i\,,\ k \text{ odd}\}$ and one can take $z_0 = (\pi/2)i$. Another interesting example for neural nets is that of $\arctan(x)$. In this case, integrating $\frac{1}{1+z^2}$, one can find a branch defined on the complement of $\{\operatorname{Re} z = 0, |\operatorname{Im} z| \geq 1\}$.

Remark 2.5 C. Fefferman (personal communication) has recently been able to extend this argument to "multiple hidden layer" nets, for the special case of the nonlinearity $\tanh(x)$. □

3 Recurrent Nets

A *recurrent net with m inputs, p outputs, dimension n, and activation function σ* is specified by a triple of matrices A, B, C and a pair of vectors β, c_0, where A, B, and C are respectively real matrices of sizes $n \times n$, $n \times m$ and $p \times n$, and β and c_0 are respectively real vectors of size m and p. We use the notation

$$\Sigma = \Sigma(A, B, C, \beta, c_0, \sigma)\,,$$

again omitting σ if obvious from the context. Because this is natural in control applications, and since the only results to be described are for that case, we assume that Σ has *no offsets*, i.e. $\beta = c_0 = 0$, and write just $\Sigma(A, B, C, \sigma)$.

We will interpret the above data (A, B, C) as defining a controlled and observed dynamical system evolving in $\mathbb{R}^n$ (in the standard sense of control theory; see e.g. [12]) by means of a differential equation

$$\dot{x} \; = \; \vec{\sigma}\,(Ax + Bu) \;,\; y \; = \; Cx \tag{5}$$

in continuous-time (dot indicates time derivative), or a difference equation

$$x^+ \; = \; \vec{\sigma}\,(Ax + Bu) \;,\; y \; = \; Cx \tag{6}$$

in discrete-time ("+" indicates a unit time shift). Other systems models are possible; for instance, "Hopfield nets" have dynamics of the form

$$\dot{x} \; = \; -Dx \, + \, \vec{\sigma}\,(Ax + Bu) \tag{7}$$

(with D a diagonal matrix and often A symmetric); results analogous to those to be described can be obtained for these more general models as well.

Depending on the interpretation (5) or (6), one defines an appropriate behavior beh_Σ, mapping suitable spaces of input functions into spaces of output functions, again in the standard sense of control theory (see [12]). For instance, in continuous time, one proceeds as follows: For any measurable essentially bounded $u(\cdot) : [0, T] \to \mathbb{R}^m$, we denote by $\phi(t, \xi, u)$ the solution at time t of (5) with initial state $x(0) = \xi$; this is defined at least on a small enough interval $[0, \varepsilon)$, $\varepsilon > 0$. (The maps σ of interest in neural network theory tend to be mostly globally Lipschitz, in which case $\varepsilon = T$.) For each input, we let $\text{beh}_\Sigma(u)$ be the output function corresponding to the initial state $x(0) = 0$, that is,

$$\text{beh}_\Sigma(u)(t) \; := \; C(\phi(t, 0, u)) \;,$$

defined at least on some interval $[0, \varepsilon)$. Given recurrent nets Σ and $\hat{\Sigma}$ (necessarily with the same numbers of input and output channels, i.e. with $p = \hat{p}$ and $m = \hat{m}$), we again say that Σ and $\hat{\Sigma}$ are equivalent (in discrete or continuous time, depending on the context) if it holds that $\text{beh}_\Sigma = \text{beh}_{\hat{\Sigma}}$; as before, we denote $\Sigma \sim \hat{\Sigma}$ (To be more precise, in continuous-time, we require that for each u the domains of definitions of $\text{beh}_\Sigma(u)$ and $\text{beh}_{\hat{\Sigma}}(u)$ coincide, and their values be equal for all t in the common domain.)

Next we summarize the main results from [1] and [2] on weight uniqueness for recurrent networks.

3.1 Continuous-Time

We assume from now on that σ is infinitely differentiable, and that it satisfies the following assumptions:

$$\sigma(0) = 0\,, \;\; \sigma^{'}(0) \neq 0\,, \;\; \sigma^{''}(0) = 0\,, \;\; \sigma^{(q)}(0) \neq 0 \text{ for some } q > 2\,. \qquad (*)$$

We let $\mathcal{S}(n,m,p)$ denote the set of all recurrent nets $\Sigma(A,B,C,\sigma)$ with fixed n,m,p. Two nets Σ and $\hat{\Sigma}$ in $\mathcal{S}(n,m,p)$ are *sign-permutation equivalent* if there exists a nonsingular matrix T such that

$$T^{-1}AT = \hat{A}\,, \;\; T^{-1}B = \hat{B}\,, \;\; CT = \hat{C}$$

and T has the special form:

$$T = PD\,,$$

where P is a permutation matrix and $D = \text{diag}(\lambda_1, \ldots, \lambda_n)$, with each $\lambda_i = \pm 1$. The nets Σ and $\hat{\Sigma}$ are just *permutation equivalent* if the above holds with $D = I$, that is, T is a permutation matrix.

Let $\mathbf{B}^{n,m}$ be the class of $n \times m$ real matrices B for which:

1. $b_{i,j} \neq 0$ for all i,j,

2. for each $i \neq j$, there exists some k such that $|b_{i,k}| \neq |b_{j,k}|$.

For any choice of positive integers n,m,p, we denote by $S^c_{n,m,p}$ the set of all triples of matrices (A,B,C), $A \in R^{n\times n}$, $B \in R^{n\times m}$, $C \in R^{p\times n}$ which are "canonical" (observable and controllable, as in [12], section 5.5). This is a generic set of triples, in the sense that the entries of the ones that do not satisfy the property are zeroes of certain nontrivial multivariable polynomials.

Finally, we let:

$$\tilde{\mathcal{S}}(n,m,p) \;=\; \left\{ \Sigma(A,B,C,\sigma) \;\middle|\; B \in \mathbf{B}^{n,m} \text{ and } (A,B,C) \in S^c_{n,m,p} \right\}\,.$$

Then, in [1], the following result was proved:

Theorem 2 *Assume that σ is odd and satisfies property (*). For any two $\Sigma, \hat{\Sigma}$, $\Sigma \sim \hat{\Sigma}$ if and only if Σ and $\hat{\Sigma}$ are sign-permutation equivalent.*

If we simply modify the definition of $\mathbf{B}^{n,m}$ to consist now of matrices for which 1. holds and 2. is replaced by:

2'. for each $i \neq j$, there exists some k such that $b_{i,k} \neq b_{j,k}$,

and defining $\tilde{\mathcal{S}}$ as above, but with this new $\mathbf{B}^{n,m}$, the above reference also proved:

Theorem 3 *Assume that σ is not odd and satisfies property (*). For any two $\Sigma, \hat{\Sigma}$, $\Sigma \sim \hat{\Sigma}$ if and only if Σ and $\hat{\Sigma}$ are permutation equivalent.*

The paper [1] explains how in fact the assumption that both nets have the same activation function σ is basically redundant, as the equality of activation functions can be derived from the equality of behaviors. Many more results are given there, for other continuous-time models.

3.2 Discrete-Time

Similar results hold for discrete-time recurrent nets. These are treated in detail in [2]. Proofs are technically different than in the continuous case, but the results are analogous. More precisely, we assume that σ not only satisfies (*), but also the following extra condition, which appeared above in the context of single-hidden layer nets with no offsets: $\sigma^{(k)}(0) \neq 0$ for infinitely many integers k. Then the same theorems as before hold, provided that we redefine:

$$\hat{S}(n,m,p) \;=\; \left\{(A,B,C) \,\middle|\, (A,B,C) \in \tilde{\mathcal{S}}(n,m,p),\, c_{ij} \neq 0 \;\forall\; i,j\right\} \;.$$

References

[1] Albertini, F., and E.D. Sontag, "For neural networks, function determines form," *Neural Networks*, to appear. Summary in: "For neural networks, function determines form," *Proc. IEEE Conf. Decision and Control, Tucson, Dec. 1992*, IEEE Publications, 1992, pp. 26-31.

[2] Albertini, F., and E.D. Sontag, "Identifiability of discrete-time neural networks," Preprint, Sept. 1992.

[3] Boyd, S., and L.O. Chua, "Uniqueness of a basic nonlinear structure" *IEEE Trans. Circuits and Systems* **CAS-30**(1983): 648-651.

[4] Boyd, S., and L.O. Chua, "Uniqueness of circuits and systems containing one nonlinearity," *IEEE Trans. Automatic Control* **AC-30**(1985): 674-681.

[5] Brady, M., R. Raghavan and J. Slawny, "Backpropagation fails to separate where perceptrons succeed," *IEEE Trans. Circuits and Systems* **CAS-36**(1989): 665-674.

[6] Davis, Philip J., *Interpolation and Approximation*, Blaisdell, New York, 1963.

[7] Denker, J., Schwartz, D., Wittner, B., Solla, S.A., Howard, R., Jackel, L., and Hopfield, J., "Automatic learning, rule extraction and generalization," *Complex Systems*, **1**(1987): 877-922.

[8] Diaconis, P., and M. Shahshahani, "On nonlinear functions of linear combinations," *SIAM J. Sci. Stat. Comput.* **5**(1984): 175-191.

[9] Harper,T.R., and W.J. Rugh, "Structural features of factorable Volterra systems," *IEEE Trans. Automatic Control* **AC-21**(1976): 822-832.

[10] Hecht-Nielsen, R., "Theory of the backpropagation neural network," in *Proceedings of the Int. Joint Conf. on Neural Networks, Washington, 1989*, IEEE Publications, NY, 593–605.

[11] Smith, W.W., and W.J. Rugh, "On the structure of a class of nonlinear systems," *IEEE Trans. Automatic Control* **AC-19**(1974): 701-706.

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

[13] Sontag, E.D., and H.J. Sussmann, "Backpropagation can give rise to spurious local minima even for networks without hidden layers," *Complex Systems* **3**(1989): 91-106.

[14] Sontag, E.D., and H.J. Sussmann, "Backpropagation separates where perceptrons do," *Neural Networks*, **4**(1991): 243-249.

[15] Sussmann, H.J., "Uniqueness of the weights for minimal feedforward nets with a given input-output map," *Neural Networks* **5**(1992): 589-593.

[16] Williamson, R.C., and U. Helmke, "Existence and uniqueness results for neural network approximations," preprint, Sept 1992.