UNIQUENESS OF NETWORK PARAMETERIZATIONS AND FASTER LEARNING

Paul C. Kainen[*], **Věra Kůrková**[†], **Vladik Kreinovich**[‡], **Ongard Sirisengtaksin**[§]

Abstract. Any single-hidden-layer feedforward network based on Gaussian or asymptotically constant odd or even rational non-polynomial activation functions has the same property as such networks based on hyperbolic tangent: input-output function determines weights and biases up to a permutation of the hidden units and sign-flips.

1 Introduction

In recent years, capabilities of feedforward neural networks to approximate arbitrary continuous or measurable functions have been intensively studied. Extending previous results for sigmoidals (e.g., Cybenko (1989) and Hornik et al. (1989)), Mhaskar and Micchelli (1992) and Leshno et al. (1993) showed that the continuous activation functions, which guarantee the universal approximation property for one-hidden-layer networks, are exactly the non-polynomial functions (provided that mild conditions hold). Using more than one hidden layer, one can allow *any* smooth non-linearity, so polynomials of degree at least two also give rise to universal approximation in this multilayer case (Kreinovich, 1991). Hence, theoretically there are many possible activation functions. However, for all of them, the number of hidden units must grow with the required accuracy.

Hecht-Nielsen (1990) proposed studying weight vectors which determine the same input-output functions. His idea was that by choosing a single network parameterization for each of the possible I/O functions, it would be possible to improve the performance of training algorithms. Sussmann (1992) and Chen et al. (1993) have shown that in the case of one-hidden-layer networks with hyperbolic tangent as activation function, the network parameterization is determined uniquely up to a permutation of hidden units and sign flips. Albertini and Sontag (1993) extended this result to infinitely differentiable functions f with the properties $f(0) = 0$, $f'(0) \neq 0$ and $f''(0) = 0$. Kůrková and Kainen (1993) showed that for asymptotically constant activation functions, the problem of uniqueness is independent of the input dimension.

[*] *Industrial Math, 3044 N St., N.W., Washington, D.C. 20007, USA*,

[†] *Institute of Computer Science, Czech Academy of Sciences, P.O. Box 5, 182 07, Prague 8, Czechia*,

[‡] *Computer Science Department, University of Texas at El Paso, El Paso, TX 79968, USA*,

[§] *Department of Computer and Mathematical Sciences, University of Houston-Downtown, Houston, TX 77002, USA*

Moreover, they showed that uniqueness holds when the activation function satisfies two basic properties, being neither "self-affine" nor "affinely recursive".

This paper extends uniqueness results to other activations, like Gaussian and certain rational functions such as $1/(1+x^2)$. For the rational function case, our argument involves analytic continuation. It is further shown that networks with polynomial activation functions allow non-unique parameterization. Section 2 has definitions and tools, main results are in the third section with proofs in section 4.

2 Uniqueness and sign-uniqueness

Any function $s : \mathbf{R} \to \mathbf{R}$ is called an *activation function*. The most widely used activation function is the *logistic sigmoid* $s(y) = 1/(1 + \exp(-y))$ which is affinely equivalent to hyperbolic tangent: $\tanh(y) = 2s(2y - 1)$. Other activation functions which are currently used include polynomials, exponentials, Gaussian and rational. See, e.g., Hecht-Nielsen (1990), Kosko (1992).

For n a positive integer, a function $f : \mathbf{R}^n \to \mathbf{R}$ is *representable by a one-hidden-layer neural network* (with s as activation function) if

$$f(x_1, \ldots, x_n) = \sum_{k=1}^{K} \beta_k s(\sum_{i=1}^{n} w_{ki} x_i + b_k) + t$$

for some positive integer K, and real numbers β_k, w_{ki}, b_k, and t, where $1 \leq k \leq K$ and $1 \leq i \leq n$.

The tuple $w = (K, \beta_1, \ldots, \beta_K, w_{11}, w_{12}, \ldots, w_{1n}, w_{21}, \ldots, w_{K1}, \ldots, w_{Kn}, b_1, \ldots, b_K, t)$, is called the *network parameterization*. For every k from 1 to K, call β_k the *output weight*, $\mathbf{w}_k = (w_{k1}, \ldots, w_{kn})$ the *input weight vector* and b_k the *bias*. The function f is called the *input-output (or I/O) function* of the network.

Two network parameterizations are called *functionally equivalent* if they induce the same I/O function. In particular, simply permuting units in the hidden layer produces functionally equivalent parameterizations. We call this more restrictive notion *interchange equivalence*.

A network parameterization is *reduced* if the following two conditions are true: for every k, $\beta_k \neq 0$ and $\mathbf{w}_k \neq \mathbf{0}$; different units in a hidden layer have different parameterizations (i.e., if $k \neq k'$, then $\mathbf{w}_k \neq \mathbf{w}_{k'}$). The reader can easily verify the following:

Proposition 1. *Every network parameterization is functionally equivalent to a reduced network parameterization.*

A network parameterizaton is *sign-reduced* if for every hidden unit the first non-zero entry of the input weight vector is positive. Recall that a function f is *odd (resp. even)* if $f(-t) = -f(t)$ (resp. $f(-t) = f(t)$). It is easy to check the following:

Proposition 2. *If the activation function is odd or even, then every network parameterization is functionally equivalent to a sign-reduced network parameterization.*

An activation function has the *uniqueness property* if for the set of reduced network parameterizations, functional equivalence and interchange equivalence coincide. If the activation is also either even or odd, then it has the *sign-uniqueness property* if for the set of sign-reduced network parameterizations, functional and interchange equivalence coincide.

An activation function s that has finite limits at $+\infty$ and $-\infty$ is called *asymptotically constant*. The following result was proved by Kůrková and Kainen (1993) using a careful analysis of the affine geometry.

Theorem 1. *If an activation function is asymptotically constant and does not have the uniqueness (sign-uniqueness) property, then uniqueness (sign-uniqueness) is violated by single-input networks.*

Recall that the Gaussian function is $\exp(-x^2)$ and that a rational function is the ratio of two polynomials. Note that a rational function f/g is asymptotically constant if and only if $deg(f) \leq deg(g)$.

3 Main results

Intuitively, an activation function has the uniqueness property if its graph cannot be expressed as a sum of copies of itself which have been shifted and rescaled horizontally and vertically. But proving uniqueness or non-uniqueness for specific functions may require powerful techniques.

Theorem 2. *The Gaussian function has the sign-uniqueness property.*

Theorem 3. *An even or odd, asymptotically constant, non-polynomial rational activation function has the sign-uniqueness property.*

These results show that important activation functions do have the sign-uniqueness property. However, there are other natural activation functions, like the ramp sigmoid, which do not have sign-uniqueness. If ρ is the ramp sigmoid, defined by $\rho(t) = t$ for $t \in [-\frac{1}{2}, \frac{1}{2}]$ and constant outside the interval, then the reader can check that $\rho(t) = \frac{1}{2}(\rho(2t - \frac{1}{2}) + \rho(2t + \frac{1}{2}))$ so sign-uniqueness is violated.

The exponential function $s(y) = \exp(y)$ also does not have the uniqueness property since there are two functionally equivalent but not interchange equivalent single input, single hidden unit networks, with the exponential activation function. For example, take both to have input weight 1, where the first has bias 1 and output weight 1 while the second has bias 0 and output weight e. Similarly, the reader can check that standard trigonometric formulas show that sine and cosine don't have the sign-uniqueness property.

The same result holds for polynomials.

Theorem 4. *Polynomial functions do not have the uniqueness property; even or odd polynomial functions do not have the sign-uniqueness property.*

Uniqueness enables easy description of "canonical" parameterizations since any network parameterization is plainly interchange equivalent to one with the parameter vectors in lexicographic order; see, e.g., Kůrková and Kainen (1993). This is important for learning because the search need only consider the canonical network parameterizations. Algorithms based on gradient descent cannot be restricted to such syntactically defined subsets as lexicographically ordered vectors, but other learning methods (e.g., genetic) may be able to take advantage of the reduced set of network parameterizations.

Fast learning is not the only possible criterion for choosing an activation function. For example, polynomial activations do not satisfy the uniqueness property. However, once the weights are chosen, the resulting network uses only addition and multiplication to compute y from the x_i and thus could be faster than a standard network that uses the logistic sigmoid $1/(1+\exp(-y))$. So if we are interested in computational complexity of the resulting neural network and not in the learning time, then polynomial units might be acceptable in spite of their non-uniqueness. A general optimization approach to choosing an activation function was described by Kreinovich and Quintana (1991).

4 Proofs

Proof of Theorem 2.

Since Gaussian is asymptotically constant, by Theorem 1, if sign-uniqueness is violated, it must be violated by single-input networks. Thus, by Proposition 2, we need to prove that whenever for all x, the following equality holds for two sign-reduced parameterizations w, w'

$$\sum_{k=1}^{K} \beta_k s(w_{k1}x + b_k) + t = \sum_{k=1}^{K'} \beta'_k s(w'_{k1}x + b'_k) + t', \tag{1}$$

then K must equal K' and w' is obtained from w by a permutation.

Let w, w' be sign-reduced parameterizations satisfying (1). Since the parameterizations are sign-reduced, for all k, $w_{k1} \neq 0$ and $w'_{k1} \neq 0$. Hence, taking the limit as $x \to \infty$, we get $t = t'$ and thus

$$\sum_{k=1}^{K} \beta_k \exp(-(w_{k1}x + b_k)^2) = \sum_{k=1}^{K'} \beta'_k \exp(-(w'_{k1}x + b'_k)^2). \tag{2}$$

By reordering the hidden units and, if necessary, interchanging the two sides of (1), without loss of generality, we can assume that the following hold: (i) for all $k, 1 \leq k \leq K$, $w_{11} \leq w_{k1}$ and if $w_{11} = w_{k1}$, then $b_1 < b_k$ (ii) for all $k, 1 \leq k \leq K'$, $w_{11} \leq w'_{k1}$ and if $w_{11} = w'_{k1}$, then $b_1 < b'_k$.

Next we prove that there exists $k, 1 \leq k \leq K'$, for which $w'_{k1} = w_{11}$ and $b'_k = b_1$. We use the fundamental properties of the exponential function.

Multiply both sides of (2) by $\exp((w_{11}x + b_1)^2)$, getting

$$\sum_{k=1}^{K} \beta_k \exp(-(w_{k1}x+b_k)^2+(w_{11}x+b_1)^2) = \sum_{k=1}^{K'} \beta'_k \exp(-(w'_{k1}x+b'_k)^2+(w_{11}x+b_1)^2). \quad (3)$$

As $x \to \infty$, since $\exp(0) = 1$, the first term of the left-hand side of (3) tends to $\beta_1 \neq 0$, while an easy calculation, using (i), shows that the other terms all tend to 0 so the limit of the left-hand side of (3) is non-zero.

If on the the right-hand side of (3) there were no k with $w'_{k1} = w_{11}$ and $b'_k = b_1$, then, using (ii), all the terms on this side would tend to 0 as $x \to \infty$, and so such k exists.

The limit of the corresponding term is β'_k so $\beta_1 = \beta'_k$. Therefore, we can subtract the corresponding terms from both sides of (2) resulting in a similar equality but with $K - 1$ units in the left-hand side. Repeating the same procedure shows that $K = K'$ and the two parameterizations w and w' correspond up to a permutation. □

Proof of Theorem 3.

Let w, w' be sign-reduced parameterizations satisfying (1). By Theorem 1 and Proposition 2, it suffices (as before) to show that $K = K'$ and that w' is obtained from w by a permutation. We prove this by contradiction.

If there is no such permutation, then one can rewrite (1) in the form

$$\sum_{j=0}^{J} c_j s(a_j x + b_j) + c = 0 \quad (4)$$

where at least one $a_j > 0$ (since s is odd or even) and $c_j \neq 0$ $(0 \leq j \leq J)$ and c. In this equation, all the pairs (a_j, b_j) are different.

Similarly to the proof of Theorem 2, take j so that a_j is the smallest of the positive numbers a_j and if there are several such j, choose the (unique) integer with b_j smallest possible. By permuting terms in (4), we may assume without loss of generality that $j = 0$.

Hence,

$$s(a_0 x + b_0) = \sum_{j=1}^{J} C_j s(a_j x + b_j) + C, \quad (5)$$

where $C_j = c_j/(-c_0)$ and $C = c/(-c_0)$.

To simplify this equation further, put $y = a_0 x + b_0$, $A_j = a_j/a_0$ and $B_j = b_j - a_j b_0/a_0$. Then reinterpreting (5),

$$s(y) = \sum_{j=1}^{J} C_j s(A_j y + B_j) + C. \quad (6)$$

Since $0 < a_0 \leq a_j$, $A_j = a_j/a_0 \geq 1$. Further, if $A_j = 1$, then $B_j > 0$.

Since $s(y)$ is a rational activation function, it can be extended to a complex rational function $s(z)$. Such a rational function of a complex variable is representable as a finite sum (see, e.g., [Flanigan 1983]):

$$s(z) = P(z) + \sum_{k,l} \frac{c_{kl}}{(z - z_k)^l}. \tag{7}$$

where the z_k are complex numbers, called *poles*, and $P(z)$ is a polynomial. The poles can be characterized as complex values z for which $s(z) = \infty$. Further, if z is a pole, so is z^* (the complex conjugate).

It is now easy to check that since s is an even or odd activation function, $s(z)$ must have a pole with positive imaginary part. That is, there exists z_k such that $\Im z_k > 0$.

¿From all such poles, let us find a pole with the largest possible value of $\Im z_k$. If there are several such poles, pick one among them for which the real part $\Re z_k$ is the largest possible; in other words, a pole for which the tuple $(\Im z_k, \Re z_k)$ is the largest possible in the sense of the lexicographic ordering. Denote this pole by z_p.

By uniqueness of analytic continuation, $s(z)$ satisfies

$$s(z) = \sum_{j=1}^{J} C_j s(A_j z + B_j) + C. \tag{8}$$

By (8) and the characterization of poles, there exists $j, 1 \leq j \leq J$, such that $A_j z_p + B_j$ is also a pole, say z_q. If $A_j > 1$, then $\Im z_q > \Im z_p$ while if $A_j = 1$, then $B_j > 0$ so $\Re z_q > \Re z_p$. This contradicts the maximality of z_p with respect to the lexicographic order. □

Proof of Theorem 4.

Let us denote the degree of the polynomial $P(y)$ by d. Then, $P(y) = a_0 + a_1 y + \ldots + a_d y^d$ for some a_j. Let us take an arbitrary positive real number α, and form $\Delta P(y) = P(y+\alpha) - P(y)$. Substituting the above expression for $P(y)$ into the formula for $\Delta P(y)$, we can easily see that terms of power d cancel each other, and therefore, $\Delta P(y)$ is a polynomial of degree $\leq d-1$. If we apply the same operation Δ to this polynomial $\Delta P(y)$, we will get a polynomial $\Delta^2 P(y) = \Delta P(y+\alpha) - \Delta P(y)$ whose degree is $\leq d-2$. After repeating this procedure d times, we get $\Delta^d P(y) =$const, and thus $\Delta^{d+1} P(y) = 0$.

One can easily check by induction that

$$\Delta^j P(y) = \sum_{k=0}^{j} (-1)^k C_k^j P(y + (j-k)\alpha),$$

where C_k^j denotes binomial coefficients. In particular,

$$\Delta^{d+1} P(y) = \sum_{k=0}^{d+1} (-1)^k C_k^{d+1} P(y + (d+1-k)\alpha).$$

Therefore, from $\Delta^{d+1}P(y) = 0$, one can conclude that

$$-P(y) = \sum_{k=1}^{d+1}(-1)^k C_k^{d+1} P(y + (d+1-k)\alpha).$$

Hence, uniqueness is violated for $n = 1$, $w = (K = 1, \beta_1 = -1, w_{11} = 1, b_1 = 0, t = 0)$, and $w' = (K' = d+1, \beta'_k = (-1)^k C_{d+1-k}^{d+1}, w'_{k1} = 1, b'_k = (d+1-k)\alpha, t' = 0)$. □

Acknowledgments

This research was supported by NSF grant No. CDA-9015006, and a Research Opportunity Award (for O.S.).

References

Albertini F., Sontag E.D. (1993). For neural networks, functions determines form, *Neural Networks* **6**(7), 975-990.

Chen A.M., Lu, H., Hecht-Nielsen R. (1993). On the geometry of feedforward neural network error spaces. *Neural Computation* **5** (6).

Cybenko G. (1989). Approximation by superpositions of a sigmoidal function, *Mathematics of Control, Signals and Systems* **2**, 303–314.

Flanigan F. J. (1983). *Complex variables: harmonic and analytic functions*, Dover, N.Y..

Hecht-Nielsen R. (1990). *Neurocomputing.* Addison-Wesley, Reading, MA.

Hecht-Nielsen R. (1990). On the algebraic structure of feedforward network weight spaces. In *Advanced Neural Computers* (pp.129 –135), Elsevier.

Hornik K., Stinchcombe M., White H. (1989). Multilayer feedforward networks are universal approximators, *Neural Networks* **2**, 359–366.

Kosko B. (1992). *Neural networks and fuzzy systems.* Prentice Hall, Englewood Cliffs, NJ.

Kreinovich V. (1991). Arbitrary nonlinearity is sufficient to present all functions by neural networks: a theorem. *Neural Networks* **4**, 381–383.

Kreinovich V., Quintana C. (1991). Neural networks: what non-linearity to choose? In *Proceedings of the 4th University of New Brunswick Artificial Intelligence Workshop*(pp. 627–637). Fredericton, N.B., Canada.

Kreinovich V., Sirisaengtaksin O. (1992). 3-layer neural networks are universal approximators for functionals and for control strategies, University of Texas at El Paso, Computer Science Department, Technical Report UTEP-CS-92-27.

Kůrková V., Kainen P. C. (1993). Functionally equivalent feedforward neural networks, *Neural Computation* (in press).

Leshno M., Lin V., Pinkus A., Schocken S. (1993). Multilayer feedforward networks with a non-polynomial activation function can approximate any function, *Neural Networks* **6**(6), 861–867.

Mhaskar H.N., Micchelli C.A. (1992). Approximation by superposition of sigmoidal and radialbasis functions, *Advances in Applied Mathematics* **13**, 350–373.

Sussmann H.J. (1992). Uniqueness of the weights for minimal feedforward nets with a given input-output map, *Neural Networks* **5**, 589–594.