

Critical Points for Least-Squares Problems Involving Certain Analytic Functions, with Applications to Sigmoidal Nets

*Eduardo D. Sontag**

Department of Mathematics

Rutgers University, New Brunswick, NJ 08903[†]

Abstract

This paper deals with nonlinear least-squares problems involving the fitting to data of parameterized analytic functions. For generic regression data, a general result establishes the countability, and under stronger assumptions finiteness, of the set of functions giving rise to critical points of the quadratic loss function. In the special case of what are usually called “single-hidden layer neural networks,” which are built upon the standard sigmoidal activation $\tanh(x)$ (or equivalently $(1 + e^{-x})^{-1}$), a rough upper bound for this cardinality is provided as well.

1 Introduction

A very typical problem concerning function approximation and regression with so-called artificial neural networks, especially in applications dealing with learning and pattern recognition, is as follows. There is given a specification of a wiring diagram (a labeled graph) that stipulates how information flows from node to node (nodes being typically called “neurons”), and, for each such node, there is a rule that restricts the particular type of combination (linear, polynomial, and so forth) of the incoming signals that will be used as input to the node. These signals arrive from other nodes as well as from external sources. In addition, a transfer function (“activation”) is specified for each node; this function indicates what computation is performed by that node on its input in order to produce the output computed by the respective node. One of the nodes acts as a designated “output node,” and its output represents the response of the whole network to the external inputs. Once such an architecture has been defined, it remains to set the numerical values of the constants appearing as “weights” or “parameters” (such as the coefficients of linear combinations or polynomials); for each choice of these parameters, a particular function of inputs is computed. The values of parameters are often obtained by minimization of a quadratic loss function which measures the goodness of fit to a given set of numerical data.

By far the most common model in experimental work is that in which affine combinations are performed at the input of each internal node, each of which then computes an application

¹Supported in part by US Air Force Grant AFOSR-94-0293

²E-mail: sontag@hilbert.rutgers.edu

³To appear in *Advances in Computational Mathematics*, Special Issue on Neural Networks

of the “standard sigmoid” $\tanh(x) = (e^x - e^{-x})/(e^x + e^{-x})$ —or equivalently, if a range of $\{0, 1\}$ is preferred, then up to a rescaling and translation $(1 + e^{-x})^{-1}$. The output node then performs a final affine combination of the outputs of the internal nodes. These are “single hidden layer networks,” which compute functions of the following type:

$$\beta(x, u) = c_0 + \sum_{i=1}^K c_i \tanh(A_i u + b_i) .$$

The inputs u are vectors in R^m . The integer K (the “number of hidden units” in neural network terminology) is assumed to be fixed. The $K(m + 2) + 1$ parameters of the network (summarized by the vector “ x ”), namely the scalars c_0, \dots, c_K and b_1, \dots, b_K , and the m -row vectors A_1, \dots, A_K , are thought of as variables that will be tuned so as to make $\beta(x, u_i) \approx y_i$ when given a set of inputs and corresponding target outputs. There are portions of the parameter space that give rise to degeneracies. For instance, if one coefficient c_i ($i \neq 0$) vanishes, then the loss function β is independent of the values of the corresponding A_i and b_i . If some $A_i = 0$ then the corresponding term is constant and can be absorbed into c_0 . If for some pair $i \neq j$ it is the case that $A_i = A_j$ and $b_i = b_j$, then the terms corresponding to i and j can be combined, and only the sum $c_i + c_j$ is relevant, resulting also in a loss of dimensionality. Similarly, since \tanh is an odd function, if $(A_i, b_i) = -(A_j, b_j)$ then terms can be combined as well. Thus a natural parameter space is the set \mathbb{X} consisting of all the b_i ’s, c_i ’s, and A_{ij} ’s for which these exceptional situations do not occur.

Assume given a training or regression data set (“labeled sample”)

$$(u, y) = \left((u_1, \dots, u_N), (y_1, \dots, y_N) \right)$$

where we interpret the u_i ’s as input vectors (“regressors” in statistical terms) and the scalars y_i ’s as targets or response vectors desired for the respective u_i ’s. The regression problem is that of minimizing (typically by means of steepest descent or other local search techniques) the quadratic loss

$$E^{(u,y)}(x) := \frac{1}{2} \sum_{i=1}^N \left(\beta(x, u_i) - y_i \right)^2$$

over \mathbb{X} . It has been often remarked that, even for extremely simple cases (such as $K=1$ and supposing that the inputs are binary vectors) there arise critical points associated to non-global local minima, and thus the study of the set of critical points of $E^{(u,y)}$ has been frequently put forward as a research topic; see [3, 4, 8, 14, 17]. In this context it has also been observed many times that —as with other least-squares problems— pathological behavior will depend heavily on the training sets not being in “general position” in appropriate senses of probability or topological density (cf. [4, 8, 14]). In this paper, a combination of techniques from [1, 11, 16, 19] —dealing with reconstruction of parameters from the functional form, the need for generic data (u, y) with large enough N , and the use of certain tools from analytic geometry and from model theory in logic— is used in order to obtain several characterizations of the critical set.

One of the main results given in this paper (Corollary 6.4) is that the set of critical points is finite, and in particular less than $2^{8(NK)^2}$ (assuming that there are enough samples to make the problem not underdetermined, specifically that $N \geq 2K(m + 2) + 3$, and for generic regression data). If the number of samples scales linearly on the number of nodes K , and assuming a constant input dimension, an upper bound of the type

$$2^{cK^4}$$

results. (A lower bound of the type $2^{e'K \log K}$ also holds, due to symmetries in the problem: any exchange among the K terms in the sum preserves β .) The finiteness results (not the above bound) can be generalized to more general “neural networks” and in fact many of the intermediate results apply equally well to completely general least-squares problems involving analytic functions.

The paper is organized as follows. Section 2 presents the basic terminology. Section 3 provides a result showing that analytically parameterized classes of functions can be identified generically on the basis of just $2r + 1$ samples, if r is the number of free parameters. This part of the paper depends on basic facts about real-analytic functions discussed in Appendix A. Section 4 studies critical points for least-squares error criteria; this part of the paper relies upon elementary differential topology (Morse theory). Section 5 combines the results of Sections 3 and 4. It establishes, for generic analytic problems, the countability of the set of functions giving rise to critical parameter values. A refinement shows that this set is in fact finite, provided that the parametric class of functions be definable logically in terms of the exponential and certain other special analytic functions; this is shown on the basis of recent work in logic, dealing with “o-minimal logical theories,” and discussed Appendix B. Finally, Section 6 specializes to single hidden layer networks, where one can use results on identifiability of parameters in order to obtain finiteness of the set of critical parameters. At this point, a Khovanskii-type estimate gives immediately the bound mentioned above. Though not strictly related to the previous results, we include in Section 7 some observations regarding approximate interpolation problems for analytically parameterized families; we do so because the basic ideas are close, and the techniques used are essentially the same as those employed in the least-squares problem.

2 Parametric Classes of Functions

Most of the technical results to be given in this note depend only upon the fact that the output of a neural network is simply a joint function of external inputs and parameters. At this level, we simply study regression problems for parametrized families of functions. Only towards the end do we need to specialize to some cases in which the precise form of the parameterization (“internal” structure) is relevant.

The main technical hypothesis that we make is that the functional form is analytic on inputs and parameters. (“Analytic” means real-analytic: an analytic function on an open subset of \mathbb{R}^l is one which admits a convergent power series representation, locally around each point of its domain. See Appendix A for some basic facts about analytic functions.) Analyticity is essential if one wants to obtain results in the form stated here. Relaxing to simply differentiability or piecewise differentiability leads to far weaker conclusions.

Definition 2.1 An *architecture* $\mathcal{A} = (\beta_{\mathcal{A}}, \mathbb{X}, \mathbb{U})$ is specified by two analytic manifolds \mathbb{X} and \mathbb{U} , of dimensions respectively r (the *number of parameters*) and m (the *input dimension*), and an analytic function

$$\beta_{\mathcal{A}} : \mathbb{X} \times \mathbb{U} \rightarrow \mathbb{R}$$

called the *behavior* of \mathcal{A} . □

The function computed by the architecture \mathcal{A} corresponding to a given parameter vector $x_0 \in \mathbb{X}$ is by definition the function

$$\beta_{\mathcal{A}}(x_0, \cdot) : \mathbb{U} \rightarrow \mathbb{R}.$$

The *class of functions computed by \mathcal{A}* is defined as the set of functions

$$\{\beta_{\mathcal{A}}(x_0, \cdot) : \mathbb{U} \rightarrow \mathbb{R}, x_0 \in \mathbb{X}\}.$$

When \mathcal{A} is clear from the context, we drop the subscript and write simply β .

Many of the results to be given will hold in general, but those involving finiteness claims will only be proved in the special cases when the function β is also (exponential and/or restricted analytic) *definable*. These are functions which can be expressed in terms of logical operations involving exponentials (on all of \mathbb{R}), as well as other analytic functions but restricted to bounded domains; see Appendix B for details. Similarly, *definable sets* are those defined in terms of such operations. A *definable architecture* is one for which \mathbb{X} is a definable submanifold of \mathbb{R}^r , \mathbb{U} is a definable submanifold of \mathbb{R}^m , and $\beta_{\mathcal{A}}$ is a definable function.

Remark 2.2 In particular, any “neural network” made up of linear (or polynomial) interconnections, and employing either the activation \tanh or the activation \arctan , gives rise to a definable architecture. \square

3 Minimal Sample Sizes

Let \mathcal{A} be an architecture, and let \mathbb{U}_0 be a subset of the input set \mathbb{U} . Two parameters x_1 and x_2 are said to be *indistinguishable modulo \mathbb{U}_0* , and we write

$$x_1 \underset{\mathbb{U}_0}{\sim} x_2,$$

if

$$\beta(x_1, u) = \beta(x_2, u)$$

for all $u \in \mathbb{U}_0$. If this property holds with $\mathbb{U}_0 = \mathbb{U}$, we write $x_1 \sim x_2$ and simply say that x_1 and x_2 are indistinguishable; this means that $\beta(x_1, u) = \beta(x_2, u)$ for all $u \in \mathbb{U}$, that is, the behavior of the architecture is the same, for all possible external inputs, whether the parameter is x_1 or x_2 .

Given a parameter $x_0 \in \mathbb{X}$, a *distinguishing subset* for it is a subset \mathbb{U}_0 of \mathbb{U} such that, for every $x \in \mathbb{X}$,

$$x_0 \underset{\mathbb{U}_0}{\sim} x \Rightarrow x_0 \sim x.$$

That is to say, if two parameters give rise to different functions, then they can be distinguished on the basis of these inputs. The *distinguishing dimension* $\mathfrak{D}(\mathcal{A})$ is the smallest integer κ (possibly infinite) with the property that for each $x_0 \in \mathbb{X}$ there is some distinguishing subset of size κ .

The set \mathbb{U}_0 is a *universal distinguishing set* if it is a distinguishing subset with respect to all possible $x_0 \in \mathbb{X}$. That is, for such a set \mathbb{U}_0 , the relation “ $\underset{\mathbb{U}_0}{\sim}$ ” is the same as simply “ \sim .” Equivalently, for a finite subset $\mathbb{U}_0 = \{u_1, \dots, u_s\}$, this means that the following mapping, which maps parameters into the vector of outputs corresponding to inputs in \mathbb{U}_0 :

$$\mathbb{X} \rightarrow \mathbb{R}^s : x \mapsto \begin{pmatrix} \beta(x, u_1) \\ \vdots \\ \beta(x, u_s) \end{pmatrix} \quad (1)$$

is one-to-one from the quotient set \mathbb{X}/\sim into \mathbb{R}^s . The *universal distinguishing dimension* $\text{UD}(\mathcal{A})$ is the smallest integer κ (possibly infinite) with the property that there is some universal distinguishing subset of size κ . Clearly $\text{D}(\mathcal{A}) \leq \text{UD}(\mathcal{A})$.

Remark 3.1 Similar concepts arise in different areas. In control theory (see e.g. [15], Chapter 5), one studies the possibility of separating internal states (corresponding to the parameters in the current context) on the basis of input/output experiments. In computational learning theory, there is an analogous concept of “teaching dimension” —see e.g. [7]— to model the smallest cardinality of a set of inputs that allows a teacher to uniquely specify the particular function being “taught” among all other functions of interest. Related notions appear also in automata theory and sequential machines (cf. [5]), though in both the cases of computational learning theory and automata, the emphasis is on discrete sets and combinatorics, as opposed to analytic parameterizations. \square

The next result provides a simple upper bound on the size needed for (universal) distinguishing subsets. Moreover, the result shows that in a precise sense, almost every subset of this minimal cardinality (or, therefore, of any larger cardinality as well) has the desired property.

By abuse of terminology, we’ll say that a family \mathcal{Z} of k -element subsets of \mathbb{U} is (finitely) analytically thin if the set of vectors $(u_1, \dots, u_k) \in \mathbb{U}^k$ so that $\{u_1, \dots, u_k\} \in \mathcal{Z}$ is (finitely) analytically thin (cf. Appendix A).

Theorem 1 *Assume that \mathcal{A} is an architecture for which \mathbb{U} is connected. Then,*

$$\text{D}(\mathcal{A}) \leq r + 1 \quad \text{and} \quad \text{UD}(\mathcal{A}) \leq 2r + 1 .$$

Moreover, the set of universal distinguishing subsets of size $2r + 1$, and, for each $x_0 \in \mathbb{X}$, the set of distinguishing subsets for x_0 of size $r + 1$, have analytically thin complements. If in addition \mathcal{A} is definable, then these statements hold with “finitely analytically thin” instead of “analytically thin.”

Proof. Fix a parameter $x_0 \in \mathbb{X}$. As a first step, we characterize the distinguishing subsets of size $r + 1$ for the parameter x_0 . Consider the set of parameters that can be distinguished from x_0 :

$$\mathcal{W}_0 := \{x \in \mathbb{X} \mid x \not\sim x_0\}$$

and, for each element $x \in \mathcal{W}_0$, the set of inputs that do not distinguish x from x_0 :

$$\mathcal{B}(x) := \left\{ u \mid u \in \mathbb{U} \text{ and } x \underset{u}{\sim} x_0 \right\} .$$

For each such $x \in \mathcal{W}_0$, $\mathcal{B}(x)$ is a semianalytic subset of \mathbb{U} of dimension at most $m - 1$, since it is the set where the nonzero analytic function $\beta(x, u) - \beta(x_0, u)$ vanishes and \mathbb{U} is connected. Thus, the following subset of \mathbb{U}^{r+1} :

$$\mathcal{T}(x) = \{(u_1, \dots, u_{r+1}) \mid u_i \in \mathcal{B}(x) \forall i = 1, \dots, r + 1\} = \prod_{i=1}^{r+1} \mathcal{B}(x)$$

has dimension at most $(m - 1)(r + 1)$ (Proposition A.2, Part 3 in Appendix A).

Next, consider the following σ -analytic subset of $\mathcal{W}_0 \times \mathbb{U}^{r+1}$:

$$\mathcal{G} := \{(x, u_1, \dots, u_{r+1}) \mid x \in \mathcal{W}_0, u_i \in \mathcal{B}(x) \forall i = 1, \dots, r+1\}.$$

Let $\pi_1 : \mathcal{W}_0 \times \mathbb{U}^{r+1} \rightarrow \mathcal{W}_0$ be the projection on the \mathcal{W}_0 factor. For each $x \in \mathcal{W}_0$, $\pi_1^{-1}(x) \cap \mathcal{G} = \mathcal{T}(x)$ has dimension at most $(m-1)(r+1)$. Applying then Proposition A.2, Part 2, it follows that

$$\dim \mathcal{G} \leq r + (m-1)(r+1) = m(r+1) - 1.$$

Finally, consider the projection π_2 of \mathcal{G} on the \mathbb{U}^{r+1} coordinates. Its image is exactly the set B consisting of those vectors (u_1, \dots, u_{r+1}) which give rise to *non* distinguishing sets $\mathbb{U}_0 = \{u_1, \dots, u_{r+1}\}$ for x_0 . As projections cannot increase dimension (by Proposition A.2, Part 1, applied with $f = \pi_2$), the set B must have dimension at most $m(r+1) - 1$, which allows us to conclude the first part of Theorem 1.

If the architecture is definable, then the set B is definable, so from Fact B.3 and the above dimension count it follows that B must be finitely analytically thin.

We now show the existence of universal distinguishing sets of cardinality $2r+1$ and that almost all sets of that cardinality are universal distinguishing sets. For this, consider the architecture

$$\mathcal{A}' = (\beta', \mathbb{X} \times \mathbb{X}, \mathbb{U})$$

where

$$\beta'((x_1, x_2), u) := \beta(x_1, u) - \beta(x_2, u).$$

Fix any arbitrary x_0 , and consider the problem of finding a distinguishing set for (x_0, x_0) with respect to the architecture \mathcal{A}' . Any such set is also a universal distinguishing set for the original \mathcal{A} . As the parameter space is now of dimension $2r$, the conclusion is obtained. ■

Remark 3.2 In the smooth (rather than analytic) case a local result is possible: there is a dense open subset of \mathbb{X} , and an open covering of this set, so that on each subset V of this cover, some set of r inputs serves as a universal distinguishing set with respect to parameters on V . This is easy to prove via an argument using the rank theorem. The global versions are obviously false, however. This is illustrated by the following example: let $\mathbb{X} = \mathbb{U} = (0, \infty)$ and $\beta(x, u) := \gamma(x - u)$, where γ is a smooth map which is nonzero on $(-\infty, 0)$ and zero elsewhere. Then every two parameters are distinguishable (if $x \neq y$ then picking $u := (x + y)/2$ results in $\beta(x, u) \neq 0 = \beta(y, u)$). But there is no finite universal distinguishing set (even though $r = 1$): given any bounded $\mathbb{U}_0 \subseteq \mathbb{U}$, pick $x, y \geq \sup\{u \mid u \in \mathbb{U}_0\}$; then $\beta(x, u) = 0 = \beta(y, u)$ for all $u \in \mathbb{U}_0$. □

4 Outputs at Critical Points of Error Function

Given a differentiable function E among two differentiable manifolds, and an element x in its domain, $(E)_* [x]$ denotes the differential of E at the point x . In local coordinates this is just the Jacobian, or, for real-valued functions E the gradient, evaluated at x .

Throughout this section, f is a fixed analytic mapping

$$f : \mathbb{X} \rightarrow \mathbb{R}^N$$

where \mathbb{X} is an analytic manifold (typically, an open subset of some Euclidean space) and N is some positive integer.

For each fixed $y \in \mathbb{R}^N$, we consider the function

$$E_y : \mathbb{X} \rightarrow \mathbb{R} : x \mapsto \frac{1}{2} \|f(x) - y\|^2 \quad (2)$$

as well as the set of critical points of E_y :

$$M_y := \{x \mid (E_y)_* [x] = 0\} \quad (3)$$

and its image under f :

$$S_y := f(M_y) . \quad (4)$$

Note that, for each fixed y , M_y is a semianalytic subset of \mathbb{X} (it can be characterized through the vanishing of analytic functions) and thus S_y is what we call a σ -analytic subset of \mathbb{R}^N (cf. Appendix A), for each fixed y . We will be interested in knowing when S_y is a countable set, or equivalently, when it has zero dimension. (In this paper, by “countable” set we mean denumerable or finite.)

Remark 4.1 In least-squares problems, the variables x represent parameters to be fit to data specified by the target vector y , and one attempts to minimize E_y in order to find a best fit. The local extrema of E_y are in particular points in the sets $E_y(M_y)$, that is to say critical values of E_y . For any fixed $y \in \mathbb{R}^N$, just from the smoothness of the function E_y one knows that this set has measure zero (Sard’s Theorem). Since in addition E_y is analytic, this set of critical values is countable, because it has measure zero and is a σ -analytic subset of \mathbb{R} . (The set of critical values is not necessarily discrete, however, as illustrated by $f(x) = e^{-x} \sin(x)$ with $\mathbb{X} = \mathbb{R}$, $N = 1$, $y = 0$.) However, here we are not interested in the image of M_y under E_y , but rather in its image under f . This latter image may fail to be countable, at least for certain target values y (example: $f(x) = (\cos(x), \sin(x))$, $\mathbb{X} = \mathbb{R}$, $N = 2$, $y = (0, 0)$). The next result shows that such a situation holds only exceptionally. \square

From now on, we say that a property holds *generically* for points y in a manifold M if the set where this property fails to hold is included in an analytically thin subset of M , that is, in a countable union of submanifolds of strictly smaller dimension. (In particular, the set where the property fails has measure zero, and also is of the first category. Furthermore, a countable intersection of generic subsets is again generic.)

The following transversality fact will be essential to the further results:

Proposition 4.2 Generically for $y \in \mathbb{R}^N$, $\dim S_y = 0$.

Proof. We start by observing that there is a covering of \mathbb{X} by countably many embedded submanifolds \mathbb{X}_i with the property that each restriction $f|_{\mathbb{X}_i}$ has constant rank differential. This can be proved by induction on the dimension r of \mathbb{X} , as follows.

The case $r = 0$ (that is, when \mathbb{X} consists of a countable union of points) is clear. Assume that we proved the existence of such a covering for the case of maps on manifolds of dimension $r - 1$. Assume now that \mathbb{X} has dimension r . Without loss of generality, we assume that \mathbb{X} is connected; if this were not the case, we could start by decomposing \mathbb{X} into its —at most

countably many— connected components; the existence of a good covering for each component then implies the existence of a covering for the original set. Let q be the largest possible rank of the differential of f . Let \mathbb{X}^{q-1} be the set of points in \mathbb{X} where the rank is less or equal to $q - 1$. This is a proper closed semianalytic subset of \mathbb{X} . Since \mathbb{X} is connected, \mathbb{X}^{q-1} is analytically thin in \mathbb{X} . Thus, by Fact A.1, it can be written as a (disjoint) union of embedded analytic submanifolds, $\mathbb{X}^{q-1} = \bigcup\{M_j, j \in J\}$, where J is countable. Each M_j has dimension at most $q - 1$, so by inductive hypothesis, for each j there is a family of submanifolds $\{M_{jk}, k \in K_j\}$, K_j countable, which cover M_j and so that each $f|_{M_{jk}}$ has constant rank. Then the family consisting of $\mathbb{X} \setminus \mathbb{X}^{q-1}$ together with all the M_{jk} provide the desired covering of the original space \mathbb{X} . (If desired, the same proof can be used to provide a partition.)

Next we remark that we may assume, in addition, that the restrictions $f|_{\mathbb{X}_i}$ are submersions onto embedded submanifolds of \mathbb{R}^N . That is, there are embedded submanifolds \mathbb{Z}_i of \mathbb{R}^N so that $f(\mathbb{X}_i) = \mathbb{Z}_i$ for each i , and so that the (constant) rank of the differential of $f|_{\mathbb{X}_i}$ equals the dimension of \mathbb{Z}_i , for each i . Indeed, pick any i . Locally, by the Rank Theorem, about each $x \in \mathbb{X}_i$ there is a neighborhood of x in \mathbb{X}_i so that f restricted to this neighborhood defines a submersion into the image. Covering in this way each \mathbb{X}_i , and picking countable subcoverings (Lindelöf property), one obtains the desired conclusion.

Fix any $y \in \mathbb{R}^N$ and, for each index i , let E_y^i be the restriction of E_y to \mathbb{X}_i . Let M_y^i be the set of critical points of E_y^i and consider the respective images $S_y^i := f(M_y^i)$. If $(E_y)_*[x] = 0$ and $x \in \mathbb{X}_i$ then, since E_y^i factors as $E_y \circ \psi_i$, where ψ_i is the inclusion of \mathbb{X}_i in \mathbb{X} , also

$$(E_y^i)_*[x] = 0.$$

Thus M_y is contained in the union of the sets M_y^i . So S_y is contained in the union of the sets S_y^i . Assume that, for each i , it is known that S_y^i is countable whenever y does not belong to the analytically thin subset Q_i of \mathbb{R}^N . It then follows that S_y is countable if y is not in $Q = \bigcup Q_i$, which again is analytically thin, and hence the desired conclusion holds. Thus we reduced the problem to establishing the result for each \mathbb{X}_i .

From the previous considerations, it is sufficient to treat the case in which f maps submersively onto an embedded submanifold \mathbb{Z} of \mathbb{R}^N . We assume from now on that this is the case. For each $y \in \mathbb{R}^N$, let

$$H_y : \mathbb{Z} \rightarrow \mathbb{R} : z \mapsto \frac{1}{2} \|z - y\|^2$$

so that $E_y = H_y \circ f$. If $z = f(x)$ and $(E_y)_*[x] = 0$ then

$$(H_y)_*[f(x)] \circ (f)_*[x]$$

is zero. Since f is a submersion, that is, $(f)_*[x]$ is an onto map from $T_x\mathbb{X}$ to $T_{f(x)}\mathbb{Z}$, this means that $(H_y)_*[f(x)] = 0$. In other words, the elements of S_y are precisely the critical points of H_y as a map on the submanifold \mathbb{Z} .

Let Q be the set of those $y \in \mathbb{R}^N$ for which the map H_y is a Morse map, that is, it is so that all of its critical points are nondegenerate (Hessian is nonsingular). It is well-known, and easy to prove via a parametric Sard theorem, that the complement of Q has Lebesgue measure zero. The proof is based on the fact that H_y has no degenerate critical points if and only if y is not a focal point for \mathbb{Z} , that is, it cannot be written as a “focus” of a set of nearby points, or more precisely, the endpoint map $(x, v) \mapsto x + v$ from the normal bundle of \mathbb{Z} to

\mathbb{R}^N is an isomorphism. See [12], §6, or [13], Section 9.6 for details. Notice that $\mathbb{R}^N \setminus Q$ is a σ -analytic set, because it is the projection into the y variables of the semianalytic set of pairs (x, y) for which the differential and the Hessian of H_y both vanish. Being a σ -analytic subset and having measure zero, $\mathbb{R}^N \setminus Q$ must be analytically thin. For each $y \in Q$, all critical points are nondegenerate, hence isolated; thus there can only be a countable number of them. This completes the proof of Proposition 4.2. \blacksquare

5 Extremal Parameters

We now assume given an architecture $\mathcal{A} = (\beta_{\mathcal{A}}, \mathbb{X}, \mathbb{U})$. Let N be a positive integer. A *regression data sequence* of size N is by definition a pair of sequences

$$(u, y) = \left((u_1, \dots, u_N), (y_1, \dots, y_N) \right) \in \mathbb{U}^N \times \mathbb{R}^N.$$

For each regression data sequence, we consider the quadratic loss function

$$E^{(u,y)} : \mathbb{X} \rightarrow \mathbb{R}$$

defined by the formula

$$E^{(u,y)}(x) := \frac{1}{2} \sum_{i=1}^N \left(\beta(x, u_i) - y_i \right)^2.$$

In other words, using the notations of Section 4, if we denote

$$f_u : \mathbb{X} \rightarrow \mathbb{R}^N : x \mapsto \begin{pmatrix} \beta(x, u_1) \\ \vdots \\ \beta(x, u_N) \end{pmatrix}$$

then $E^{(u,y)}(x) = E_y(x)$, where y is as above the vector with components y_1, \dots, y_N and E^y is understood as the error function with respect to the function f_u .

We are interested in studying the set of critical points of the map $E^{(u,y)}$. More precisely, since indistinguishable parameters give rise to the same behavior, we look for an upper bound on the number of equivalence classes that may give rise to critical values of the error function.

A *class of parameters* \mathcal{C} will mean an equivalence class under \sim , using the notations in Section 3. A class \mathcal{C} will be said to be *critical*, with respect to a given regression data sequence (u, y) , if there is any parameter $x_0 \in \mathcal{C}$ for which $\left(E^{(u,y)} \right)_* [x_0] = 0$. We let $\rho(u, y)$ be the number of critical classes with respect to (u, y) .

For each (u, y) , we may consider the sets $S_{(u,y)}$ and $M_{(u,y)}$ equal respectively to the sets S_y and M_y in Section 5 when applied to the map f_u . A class \mathcal{C} is critical if and only if the image $f_u(\mathcal{C})$ is in $S_{(u,y)}$. For any $\{u_1, \dots, u_N\}$ and any class \mathcal{C} , the image $f_u(\mathcal{C})$ consists of just one point. Thus

$$\rho(u, y) \geq \text{card } S_{(u,y)}.$$

If, in addition, $\{u_1, \dots, u_N\}$ happens to be a universal distinguishing set for \mathcal{A} , then $f_u(x) = f_u(x')$ if and only if $x \sim x'$. So in that case

$$\rho(u, y) = \text{card } S_{(u,y)}.$$

The next result says that (provided the data is “overdetermined” enough) the number of critical classes is, in a generic sense, countable. Recall that r is the dimension of the parameter space \mathbb{X} .

Theorem 2 *Assume that $N \geq 2r + 1$. Then, generically in u , for generic y there are only countably many critical classes.*

Proof. Pick any sequence $u = (u_1, \dots, u_N)$ so that $\{u_1, \dots, u_N\}$ is a universal distinguishing set for \mathcal{A} . By Theorem 1, there are such sets, and all sequences are like this except for those in an analytically thin set B . As remarked above, $\rho(u, y) = \text{card } S_{(u,y)}$. Combined with Proposition 4.2, this gives that, for all y except those in an analytically thin set B_u , the set of critical classes is countable. ■

Remark 5.1 Let F be the complement of the set of regression data sequences of size N for which there are countably many critical classes. With the notations in the above proof, F is contained in the (σ -analytic, and therefore measurable) set

$$\{(u, y) \mid u \in B \text{ or } y \in B_u\},$$

so by Fubini’s theorem it has zero measure. □

5.1 The Definable Case

From now on, assume that \mathcal{A} is a definable architecture. Consider the following formula $\Phi(z, u, y)$ over the language L :

$$(\exists x \in \mathbb{X}) (\beta(x, u_i) = z_i, i = 1, \dots, N) \text{ and } (E^{(u,y)})_* [x] = 0.$$

The set $S_{(u,y)}$ obtained as the image of the set of critical points of $E^{(u,y)}$ under f_u is precisely the set of points z defined by the formula $\Phi_{(u,y)}$ obtained from Φ by fixing the variables (u, y) . Thus, by Fact B.2, the number of connected components of $S_{(u,y)}$ is bounded by some fixed integer κ (which depends only on the architecture and will be fixed from now on). In particular, if $S_{(u,y)}$ happens to be a countable set, then it must be a finite set of cardinality at most κ . Recall that for universal distinguishing sets this cardinality is the same as $\rho(u, y)$.

Suppose that $N \geq 2r + 1$. Let G be the set of regression data sequences of size N for which $\rho(u, y) \leq \kappa$. *This set is definable.* Indeed, it is $\mathcal{S}(\Phi)$, where $\Phi(u, y)$ is the formula that states that there exist κ vectors z_1, \dots, z_κ in \mathbb{R}^N with the property that, if x is a critical point of $E^{(u,y)}$ then one of the equalities $f_u(x) = z_1, \dots, f_u(x) = z_\kappa$ holds. On the other hand, by Remark 5.1, we know that the complement F of G , which is also definable, has measure zero. It follows from Fact B.3 that F is finitely analytically thin in the sense of the appendixes (i.e., a finite union of embedded submanifolds of positive codimension). We summarize as follows.

Theorem 3 *Assume that $N \geq 2r + 1$ and that the architecture is definable. Then there is some integer κ and a finitely analytically thin subset $F \subseteq \mathbb{U}^N \times \mathbb{R}^N$ so that, for each regression data sequence of size N which is not in F , the number of critical classes is at most κ .* ■

6 The Single-Hidden Layer Network Case

In this section we specialize the results to the case of single-hidden layer networks. Let $\theta : \mathbb{R} \rightarrow \mathbb{R}$ be a given function, to be called from now on an *activation*. For simplicity we'll assume that θ is an *odd* function ($\theta(-x) = -\theta(x)$), but results can be generalized in obvious ways at the cost of somewhat more notational complication.

We'll say that (A, b, c) is an (m, K) *triple* if $A \in \mathbb{R}^{K \times m}$, $b \in \mathbb{R}^K$, and $c \in \mathbb{R}^{K+1}$, and use A_i and b_i , $i = 1, \dots, K$, to denote the i th rows of A and b respectively, and c_i , $i = 0, \dots, K$, for the rows of c . The triple is *irreducible* if the following properties hold:

$$\begin{aligned} c_i &\neq 0 & \text{for } i = 1, \dots, K \\ A_i &\neq 0 & \text{for } i = 1, \dots, K \\ (A_i, b_i) &\neq \pm (A_j, b_j) & \text{for } i, j = 1, \dots, K, i \neq j. \end{aligned}$$

We now define “single-hidden layer neural networks” (with the obvious nonredundancy constraints).

Definition 6.1 An (m, K) *irreducible architecture with activation θ* is an architecture $\mathcal{A} = (\beta, \mathbb{X}, \mathbb{U})$ of the following form:

- The input set $\mathbb{U} = \mathbb{R}^m$.
- With $r = K(m + 2) + 1$, and writing the elements of the Euclidean space \mathbb{R}^r as triples $x = (A, b, c)$, the parameter set \mathbb{X} is the subset of \mathbb{R}^r consisting of irreducible triples.
- The behavior β has the following form:

$$\beta(x, u) = c_0 + \sum_{i=1}^K c_i \theta(A_i u + b_i) . \quad \square$$

The activation θ is said to satisfy the *property IP* (“independence property”) if, for each positive integer l , any positive real numbers a_1, \dots, a_l , and any real numbers b_1, \dots, b_l such that

$$(a_i, b_i) \neq (a_j, b_j) \quad \forall i \neq j ,$$

the set of dilated and translated functions $\mathbb{R} \rightarrow \mathbb{R}$

$$\{1, \theta(a_1 s + b_1), \dots, \theta(a_l s + b_l)\}$$

is linearly independent.

The function $\theta = \tanh$, the standard sigmoid used in the neural networks experimental literature, satisfies IP, as shown in [19]. A simple proof of this fact, as well as an extension to far more general θ , is given in [1], from which we cite the following sufficient condition for θ to satisfy IP:

Fact 6.2 Assume that θ extends to an analytic function defined on some subset $D \subseteq \mathbb{C}$ of the form:

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

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

This condition encompasses many, or perhaps most, examples of interest in neural networks. Observe that if θ has a meromorphic extension which has a unique pole of minimal positive real part, then it satisfies the above hypotheses. This includes many rational functions as well as $\tanh(s)$. Another useful example that satisfies the above sufficient condition, and hence also property IP, is $\arctan(s)$.

Lemma 6.3 Let $\mathcal{A} = (\beta, \mathbb{X}, \mathbb{U})$ be an (m, K) irreducible architecture with an activation θ which satisfies property IP. Then, each equivalence class \mathcal{C} under \sim has cardinality exactly $2^K K!$.

Proof. It is shown in [1] (in the same manner as done for the corresponding result for \tanh in [19]) that two parameter vectors are equivalent, $(A, b, c) \sim (A', b', c')$, if and only if (A', b', c') can be obtained from (A, b, c) by some permutation of the rows $i = 1, \dots, K$ of each of A, b, c , and/or a sign reversal in each row, for some subset of these rows. ■

A parameter $x = (A, b, c)$ will be said to be *critical*, with respect to a given regression data sequence (u, y) , if $\left(E^{(u,y)}\right)_* [x] = 0$, that is, if $x \in M_{(u,y)}$.

Theorem 4 Let $\mathcal{A} = (\beta, \mathbb{X}, \mathbb{U})$ be an (m, K) irreducible architecture with activation θ . Assume that:

- $N \geq 2r + 1$;
- θ is definable; and
- θ satisfies IP.

Then, there is some integer $\rho = \rho_{\theta, r, N}$ and a finitely analytically thin subset $F \subseteq \mathbb{U}^N \times \mathbb{R}^N$ so that, for each regression data sequence of size N which is not in F , the number of critical parameters is at most ρ .

Proof. Theorem 3 showed that there are κ and F so that, for each regression data sequence of size N and not in F , the number of critical classes is at most κ . Lemma 6.3 provides a uniform bound on the cardinality of classes. Thus the conclusion holds with $\rho = 2^K K! \kappa$. ■

Observe that Theorem 4 applies, in particular, to the choices $\theta = \tanh$ and $\theta = \arctan$.

6.1 Explicit Estimates for \tanh

We now specialize to the case when $\theta = \tanh$, in which case we can use explicit estimates derived from Khovanskii's theory of sparse and exponential polynomials. The objective is to estimate the cardinality of $M_{(u,y)}$, the set of critical parameters. To do so, we need to count the equations defining the partial derivatives and to analyze the complexity of these equations. A potential difficulty in applying Khovanskii's techniques is that in general one must first reduce the problem to one dealing with submanifolds of Euclidean spaces defined by exponential polynomials. Thus we assume that a regression data sequence (u, y) is given, satisfying the genericity assumptions of Theorem 4, so that the set $M_{(u,y)}$ is already known to be finite (and hence a manifold). Next we compute explicitly the partial derivatives of $E^{(u,y)}$, as follows. As before, we write the parameter vector in the form $x = (A, b, c)$.

We start by considering the set of equations

$$(z_{ij} + 1) \left(1 + e^{L_{ij}}\right) = 2, \quad i = 1, \dots, K, \quad j = 1, \dots, N, \quad (5)$$

where, for each i, j , we are denoting

$$L_{ij} = -2(A_i u_j + b_i),$$

which is a linear function of the parameters. Note that Equation (5) is equivalent to

$$z_{ij} = \tanh(A_i u_j + b_i), \quad i = 1, \dots, K, \quad j = 1, \dots, N.$$

For simplicity in displays, we use the following notation, for each $j = 1, \dots, N$:

$$\rho_j := c_0 + \sum_{i=1}^K c_i z_{ij} - y_j$$

(this represents the value $\beta(x, u_j)$ of the output, corresponding to the parameters x and an input vector u_j). The derivatives with respect to the variables c_μ provide the equations

$$\sum_{j=1}^N \rho_j z_{\mu j} = 0, \quad \mu = 1, \dots, K \quad (6)$$

and, for the case $\mu = 0$:

$$\sum_{j=1}^N \rho_j = 0. \quad (7)$$

The derivatives with respect to the variables b_μ provide the equations

$$\sum_{j=1}^N \rho_j \left(1 - z_{\mu j}^2\right) = 0, \quad \mu = 1, \dots, K \quad (8)$$

where we used the fact that $\tanh'(s) = 1 - \tanh(s)^2$ and we cancelled the factor c_μ (which is necessarily nonzero in the irreducible case). Finally, derivatives with respect to the entries of A give the equations

$$\sum_{j=1}^N \rho_j \left(1 - z_{\mu j}^2\right) u_{lj} = 0, \quad \mu = 1, \dots, K, \quad l = 1, \dots, m \quad (9)$$

(again dropping the factor c_μ).

Irreducibility of the triple $x = (A, b, c)$ is equivalent to the solvability of the following set of equations:

$$c_\mu \tilde{c}_\mu = 1, \quad \mu = 1, \dots, K, \quad (10)$$

$$\|A_\mu\|^2 \tilde{a}_\mu = 1, \quad \mu = 1, \dots, K, \quad (11)$$

$$\|(A_i + A_j, b_i + b_j)\|^2 \tilde{d}_{ij} = 1, \quad i = 1, \dots, K-1, \quad j = i+1, \dots, K, \quad (12)$$

and

$$\|(A_i - A_j, b_i - b_j)\|^2 \tilde{e}_{ij} = 1, \quad i = 1, \dots, K-1, \quad j = i+1, \dots, K, \quad (13)$$

where the \tilde{c}_μ , \tilde{a}_μ , \tilde{d}_{ij} , and \tilde{e}_{ij} represent a set of additional variables.

We now consider the set of equations (5) to (13) as a set of simultaneous constraints on the variables

$$z_{ij}, c_i, b_i, A_{ij}, \tilde{c}_i, \tilde{a}_i, \tilde{d}_{ij}, \tilde{e}_{ij}.$$

Let $\tilde{M}_{(u,y)}$ be the subset of \mathbb{R}^ν defined by these equations, where

$$\nu = K(3 + N + m + K).$$

There is a one-to-one correspondence between the set $M_{(u,y)}$ and the set $\tilde{M}_{(u,y)}$, since Equations (10) to (13) provide unique values for the “tilde variables” in terms of the original variables. The advantage of working with the extended set $\tilde{M}_{(u,y)}$ is that this is a subset of Euclidean space defined by a finite set of equations, each of them a polynomial on the variables or on the exponentials of the linear functions L_{ij} . Since the set $\tilde{M}_{(u,y)}$ is known to be finite, we can apply the estimates provided for precisely such equations by Khovanskii in [9], page 91. (A priori, these estimates can be applied to any set of equations provided that the set of solutions is already known to define a submanifold.) Let $q = NK$, the number of distinct linear functions appearing in the exponentials. Then Khovanskii’s estimates (in his notations, “ k ” is zero) gives the following upper bound for the cardinality of $\tilde{M}_{(u,y)}$:

$$2^{\frac{q(q-1)}{2}} \delta^\nu (\nu\delta)^q \tag{14}$$

where δ is the maximum degree of the equations. Note that $\delta = 4$, which is achieved by the terms of the type $c_i z_{ij} z_{\mu_j}^2$ which appear in Equation (8). Note that if $N \geq 2r + 1$, then $N > m + 3$, so we have an upper bound estimate as follows.

Corollary 6.4 Let $\mathcal{A} = (\beta, \mathbb{X}, \mathbb{U})$ be an (m, K) irreducible architecture with activation \tanh . Assume that $N \geq 2r + 1$. Then, there is a finitely analytically thin subset $F \subseteq \mathbb{U}^N \times \mathbb{R}^N$ so that, for each regression data sequence of size N which is not in F , the number of critical parameters is at most $2^{8(NK)^2}$. \square

7 Interpolation Capabilities

In the context of solving least-squares problems, it seems of interest to ask how many parameters are necessary in order to be able to obtain an arbitrarily small error on a given number of samples. We formalize this question as follows.

A sequence of elements (u_1, \dots, u_N) will be said to be *I-shattered* by the architecture \mathcal{A} if for every possible sequence of target values (y_1, \dots, y_N) ,

$$\inf_{x \in \mathbb{X}} E^{(u,y)}(x) = 0.$$

Thus, shattering in this sense means that all possible values can be approximately obtained. This property may be too restrictive, for instance if β is a bounded function (as happens for neural networks if parameters in the output layer are required to be small). A weaker requirement is that, for some $\varepsilon > 0$, the sequence be ε -*I-shattered* by \mathcal{A} , meaning that one requires this property only for all those sequences of target values y_1, \dots, y_N for which $|y_i| < \varepsilon$ for all i .

Observe that I-shattering amounts to asking that the mapping in Equation (1) has a dense image, and ε -I-shattering is the same as the requirement that the image of this map intersect $(-\varepsilon, \varepsilon)^N$ densely.

The *interpolation dimension* $\text{ID}(\mathcal{A})$ is the supremum (possibly infinite) of the integers N for which there is an $\varepsilon > 0$ and some sequence of length N that can be ε -I-shattered by \mathcal{A} . (Note that if one would define $\text{ID}(\mathcal{A})$ using I-shattering rather than ε -I-shatterings the dimension would be no greater; thus the upper bound to be given below holds in that case as well.)

We next show that a parameter count provides the right upper bound. This fact is not true in general, and the assumption of definability is essential; such a result is in general false, even for networks obtained from analytic activations that qualitatively look very much like tanh (strictly increasing, limits at $\pm\infty$, etc); see [16] for such counterexamples.

Theorem 5 *For every definable architecture \mathcal{A} , $\text{ID}(\mathcal{A}) \leq r$.*

Proof. If a sequence (u_1, \dots, u_N) can be ε -I-shattered, then the image of the map in Equation (1) intersects $(-\varepsilon, \varepsilon)^N$ densely, for some $\varepsilon > 0$. By Corollary B.4 in Appendix B this image, being a definable set, must have nonempty interior. But the map is analytic, so then Sard’s Theorem implies that its differential must have full rank N at some point. In particular, it must then be the case that $N \leq r$, establishing the result. ■

Remark 7.1 Note that the inequality $\text{ID}(\mathcal{A}) \leq r$ is trivial in the case of *bounded* parameters, assuming only that $\beta_{\mathcal{A}}$ is smooth. That is, if one takes any class of functions of the type $\{\beta_{\mathcal{A}}(x_0, \cdot), \|x_0\| \leq \gamma\}$, then the image of the map (1) (with domain $\|x\| \leq \gamma$) is compact, hence closed. Thus the image cannot intersect $(-\varepsilon, \varepsilon)^N$ densely unless it contains all of $(-\varepsilon, \varepsilon)^N$. Now Sard’s theorem again provides the conclusion. □

Remark 7.2 As an example, take an $(1, K)$ architecture with activation $\theta = \tanh$. The above result says that $\text{ID}(\mathcal{A}) \leq 3K + 1$. This fact had also been proved, for this very special case, by an ad-hoc argument in [16], where it was also shown that $\text{ID}(\mathcal{A}) \geq 2k - 1$. Determining in this example the precise value of $\text{ID}(\mathcal{A})$ in the interval $2K - 1, \dots, 3K + 1$ would seem to be an open question. □

Remark 7.3 Observe that it is possible for $\text{ID}(\mathcal{A})$ to be far smaller than r . For instance, for “neural nets” consisting of a string of linearly ordered nodes, (see Fig. 1) i.e., $\beta(x, u)$ is an

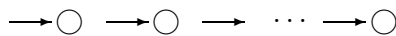


Figure 1: Linearly Ordered Nodes

iteration of functions $\theta(a_1\theta(a_2\theta(\dots(a_k u + b_k)\dots) + b_2) + b_1)$, $r = 2(k - 1)$, but $\text{ID}(\mathcal{A}) = 2$, independently of the number of nodes, since all the functions $\beta_{\mathcal{A}}(x, \cdot) : \mathbb{R} \rightarrow \mathbb{R}$ are necessarily monotone. □

A Some Facts about Real-Analytic Functions

In this appendix we describe various elementary facts regarding (real-)analytic functions in a manner suitable for reference in the main text.

Let M be any analytic manifold of dimension l . (In this paper, “manifold” always means second-countable manifold.) Recall that an embedded submanifold Z of M , of dimension q , is a connected subset which, locally around each of its points and up to analytic diffeomorphisms, looks like a “slice” $\{(x_1, \dots, x_l) \mid x_{q+1} = \dots = x_l = 0\}$. When $q = l$, this is just an open set.

Assume now that the submanifold Z has positive codimension, that is, $q \leq l - 1$. From the definition, it follows that for some open subset $M_0 \subseteq M$, Z is a closed subset of M_0 (in the relative topology). Observe that Z is nowhere dense, that is, its closure has empty interior. (That is, if U is any open subset of M , then $Z \cap U$ cannot be dense in U : if U does not intersect M_0 then this is clear; otherwise $U \cap M_0$ is a nonempty open set and we may assume without loss of generality that $U \subseteq M_0$; then $U = (U \setminus Z) \cup (U \cap Z)$, so either $U \setminus Z$ is a nonempty open set, and we are again done, or $U \cap Z = U$, but in this latter case Z would contain an open set and hence could not have positive codimension.) Also, such a Z has measure zero.

For simplicity, if Z is a countable union of embedded analytic submanifolds of M of dimension $\leq q$, and q is the smallest such integer, we say that Z is a σ -analytic subset of M and call q the *dimension* of Z . (It is not hard to verify that the dimension is well-defined, in the sense that it doesn’t depend on the particular union of countably many submanifolds being used.) The σ -analytic subset Z of M will be said to be *analytically thin* if $\dim Z < \dim M$. Such a set has zero measure and is of the first category (a countable union of nowhere dense sets), as remarked above. Conversely, if a σ -analytic subset and it has measure zero then it must be analytically thin (otherwise it contains a submanifold of full dimension, that is, an open subset of M). We also use the following terminology: a subset Z of M is *finitely analytically thin* if it is a finite union of embedded analytic submanifolds of positive codimension.

In studying analytic sets and mappings, it is useful to introduce the notion of a semianalytic subset Z of an analytic manifold M . This is a set Z so that, for each $z \in M$, there is some neighborhood U of z so that $Z \cap U$ is in the Boolean algebra generated by a finite family of subsets of the form $\{f_j(x) > 0\}$, for some analytic functions $f_j : U \rightarrow \mathbb{R}$, $j = 1, \dots, J_z$. It is easy to see from the definition of embedded submanifold that if $Q \subseteq M$ is such a submanifold, then Q is a countable union of compact semianalytic subsets. It follows that every σ -analytic subset of M is a countable union of compact semianalytic subsets. Conversely, every semianalytic subset is a countable union of embedded submanifolds; see for instance Property 8(e) in [18]. In conclusion, being a σ -analytic subset is the same as being a countable union of semianalytic subsets (or of compact semianalytic subsets). We will use the following very special consequence of general stratification theorems (see for example Theorem 9.2 in [18], as well as [2]):

Fact A.1 Let M and N be analytic manifolds and $f : M \rightarrow N$ an analytic mapping. Assume that Z is a compact semianalytic subset of M . Then there is a partition of N into a countable union of connected analytic embedded submanifolds Q_j , and a partition of Z into a countable union of connected analytic embedded submanifolds so that each such submanifold is diffeomorphic to $\mathbb{R}^{n_j} \times Q_j$, for integers $n_j \geq 0$ and suitable indices $j \in J_Z$, and on each such set the mapping f is (up to the same diffeomorphism) the projection $\mathbb{R}^{n_j} \times Q_j \rightarrow Q_j$. \square

Observe that it also follows from Fact A.1 that the image $f(Z)$ is a σ -analytic subset of N (since it is a union of a subfamily of the Q_j ’s). Note that if Z is a σ -analytic subset of M ,

then the above discussion shows that $\dim Z = q$ if and only if Z can be written as a countable union of embedded submanifolds in such a way that the maximum of the dimensions of the submanifolds is q .

The next statements amount to saying that naive parameter counts are well-justified when dealing with analytic mappings.

Proposition A.2 Assume that M , N , and M_i , $i = 1, \dots, k$, are analytic manifolds. Let $f : M \rightarrow N$ be an analytic mapping. Then:

1. If Z is a σ -analytic subset of M , then $f(Z)$ is a σ -analytic subset of N , and $\dim f(Z) \leq \dim Z$.
2. For all $Z \subseteq M$,

$$\dim Z \leq \dim f(Z) + \max_{y \in N} [\dim f^{-1}(y) \cap Z] .$$
3. If Z_i is analytically thin in M_i , for $i = 1, \dots, k$, then $Z = Z_1 \times \dots \times Z_k$ is analytically thin in $M_1 \times \dots \times M_k$ and

$$Z = Z_1 \times \dots \times Z_k \subseteq M_1 \times \dots \times M_k$$

satisfies $\dim Z = \sum_i \dim Z_i$.

Proof. In order to calculate the dimension of $f(Z)$, it is enough, by the above considerations, to do this when Z is a relatively compact semianalytic subset, and thus the dimension inequality follows by Fact A.1.

To prove the statement about fibres $f^{-1}(y)$, we proceed as follows. Note that each such fibre is semianalytic, so its dimension is well-defined. Write Z as a countable union of compact semianalytic subsets Z_i ; then

$$\max_y [\dim f^{-1}(y) \cap Z] = \max_{y,i} \dim [f^{-1}(y) \cap Z_i] .$$

Fix any i , and apply Fact A.1 with Z_i instead of Z . Thus $q = \max_y \dim [f^{-1}(y) \cap Z_i]$ is the largest of the n_j 's, while $\dim Z_i$ is at most $q + t$, $t =$ largest dimension of the Q_j 's, $j \in J_Z$, and $f(Z)$ has dimension t . This shows that

$$\dim Z_i \leq \dim f(Z) + \max_{y \in N} [\dim f^{-1}(y) \cap Z_i] \leq \dim f(Z) + \max_{y \in N} [\dim f^{-1}(y) \cap Z]$$

from which, since $\dim Z = \max_i \dim Z_i$, the conclusion follows.

Finally, to prove that $\dim Z_1 \times Z_2 = \dim Z_1 + \dim Z_2$, simply note that $Z_1 \times Z_2$ equals a union of the type $Z_1^j \times Z_2^k$, for countable coverings by submanifolds for each of Z_1 and Z_2 respectively, and dimensions add as they should for submanifolds. \blacksquare

B Some Facts about Order-Minimality

Here we summarize certain recent facts from model theory used in order to prove the results given in the text.

Pick any positive integer l , and a cube $C = [-k, k]^l$ in \mathbb{R}^l . Assume that g is a real-valued function which is (real-)analytic in a neighborhood of C . By the θ -restriction of g to C we will mean the function $f : \mathbb{R}^l \rightarrow \mathbb{R}$ which equals 0 outside C and equals g on C . A *restricted analytic* (RA) *function* is any function obtained in this manner. Below we formally state what it means for a function to be (“exp-RA”) definable —informally, these are functions that can be defined in terms of a first-order logic sentence involving the standard propositional connectives, existential and universal quantification, algebraic operations, and symbols for the exponential function as well as all RA functions. Of course, \tanh is definable, since $y = \tanh(x)$ if and only if $(y + 1)(1 + e^x) = 2$. Any RA function is in particular definable. The function $\arctan(x)$ is also definable, since $y = \arctan(x)$ if and only if $-\pi/2 < y < \pi/2$ and $\text{SIN}(y) = x \text{COS}(y)$, where SIN and COS denote the restrictions of \sin and \cos to $[-\pi/2, \pi/2]$. Compositions such as $\arctan(\exp(\exp(x)))$ are also allowed. (However, the function $\sin(x)$ is *not* definable.)

Formally, consider the structure

$$L = (\mathbb{R}, +, \cdot, <, 0, 1, \exp, \{f, f \in \text{RA}\}),$$

and the corresponding language for the real numbers with addition, multiplication, and order, as well as one function symbol for real exponentiation and one for each restricted analytic function. The set of (first order) formulas over L is the set of all well-formed logical expressions obtained by using propositional connectives, real numbers as constants, the operations of addition and multiplication, the relations $<$ and $=$, and \exp and restricted analytic functions as functions; quantification is allowed over variables. (This is an example of a formula $\Phi(x, y)$ over L :

$$\forall z \left[e^{7z^2e^y} - \pi xz \geq \arctan(e^x) \right].$$

We write $\Phi(x, y)$ to indicate the fact that the only free —i.e., non-quantified— variables in the formula are x and y .) Each such formula will be interpreted over the real numbers, that is, all variables are assumed to take real values. Thus all quantifiers are implicitly assumed to be over \mathbb{R} . Given a formula Φ with free variables x_1, \dots, x_l , we write $\mathcal{S}(\Phi)$ for the subset of \mathbb{R}^l that it defines. A *definable* set is a set of the form $\mathcal{S}(\Phi)$, for some first order formula Φ over the language L . For instance, the above $\Phi(x, y)$ gives rise to:

$$\mathcal{S}(\Phi) = \left\{ (x, y) \in \mathbb{R}^2 \mid (\forall z \in \mathbb{R}) \left[e^{7z^2e^y} - \pi xz \geq \arctan(e^x) \right] \right\}.$$

Similarly, the truth of a formula Φ with no free variables is defined as the truth of the statement obtained when quantifying over the reals. By abuse of notation, when giving such a formula, we will also allow other symbols, such as “ $-$ ” or “ \geq ” which could be in turn defined on the basis of the above primitives, or even symbols for any set already known to be definable. A (*exp*-RA) *definable function* is a function $f : M \rightarrow N$ whose graph is a definable set in the above sense, and where N and M are definable subsets of two spaces \mathbb{R}^{l_1} and \mathbb{R}^{l_2} respectively.

When the exponential is left out, definable sets are precisely those called “finitely sub-analytic” in [20]. Restricted analytic functions were introduced in [22]. (The definition in that reference is slightly different from the one we gave in the previous section: it assumes that the functions g have a convergent power series representation valid on all of the cube

$C = [-k, k]^l$, but a standard compactness argument shows that the two definitions are equivalent.) Gabrielov showed in [6] that the theory of real numbers with restricted analytic functions is model-complete, which means that every formula is equivalent to one that involves only existential quantification. (We do not give the precise statements here, as they are not needed for explaining the further material.) In a recent major development, Wilkie showed in [25, 26] that using exponentiation (but now leaving out the RA functions), model-completeness obtains as well. Finally, in [22] and [23], it was shown that the full theory (RA as well as exponentials) is model-complete, and hence order-minimal:

Fact B.1 ([22], Theorem 6.9, and [23]) The theory of L is *order-minimal*, that is, for each formula Φ having just one free variable, $\mathcal{S}(\Phi)$ is a subset of \mathbb{R} consisting of a finite union of intervals (possibly unbounded or just points). \square

The terminology reflects that such finite unions are the smallest Boolean algebra of subsets that can be defined using order. The forthcoming book [21] by van den Dries deals in detail with order-minimal theories. Sets definable (in any dimension) for order-minimal theories admit finite cell decompositions into topological submanifolds. In particular, this applies to parametric versions. Decompositions can be obtained which are uniform on parameters, and in particular the number of connected components is uniformly bounded. To be more precise, assume given a formula $\Phi(\Lambda, x)$, where Λ denotes a set of p variables and x denotes a set of q variables, for some integers p and q . For each fixed $\lambda \in \mathbb{R}^p$, we may consider the formula $\Phi_\lambda(x) = \Phi(\lambda, x)$ on the free variables x , and the respective definable set. It then follows from the general theory (see [10, 21]):

Fact B.2 Given a formula Φ as above, there is an integer κ so that for all $\lambda \in \mathbb{R}^p$, the number of connected components of $\mathcal{S}(\Phi_\lambda)$ is at most κ . \square

When dealing as here with a language whose primitives stand for analytic functions, the cell decomposition results can be stated in a stronger fashion. By [22], Theorem 8.8, one knows that each definable subset is a finite union of what are called in that paper analytic cells, each of which is definable and definably-isomorphic to an Euclidean space. The definition of analytic cell in that paper implies that each such cell is an embedded analytic submanifold. Thus one also has the following result:

Fact B.3 Let S be a definable subset of \mathbb{R}^q . Then, either S contains an open subset or it is finitely analytically thin. \square

Observe that a function such as $\sin(x)$ (seen as a function of $x \in \mathbb{R}$) is not definable, so there is no contradiction with the fact that its set of zeroes is not finitely analytically thin. (The zero set is of course analytically thin, consistent with the analyticity of $\sin(x)$.)

Note that a *finitely* analytically thin subset is nowhere dense (as it is a finite union of nowhere dense subsets). So this follows from Fact B.3:

Corollary B.4 If S is a definable subset of \mathbb{R}^q , then either it has nonempty interior or it is nowhere dense.

Acknowledgments

There is some overlap between the material presented here, in particular that dealing with distinguishing and interpolation dimensions, and parts of the conference paper [11]. The author wishes to thank his coauthor in that paper, Angus MacIntyre, for his permission to use that material in this paper. Many thanks are also due to Lou van den Dries for making available a draft of part of his book on o-minimal theories as well as providing many references.

References

- [1] Albertini, F., E.D. Sontag, and V. Maillot, “Uniqueness of weights for neural networks,” in *Artificial Neural Networks with Applications in Speech and Vision* (R. Mammone, ed.), Chapman and Hall, London, 1993, pp. 115-125.
- [2] Bierstone, E., and Pierre D. Milman, “Semianalytic and subanalytic sets,” *Inst. Hautes Études Sci. Publ. Math.* **67**(1988): 5-42.
- [3] Blum, E.K., “Approximation of Boolean functions by sigmoidal networks: Part I: XOR and other two-variable functions,” *Neural Computation* **1**(1989): 532-540.
- [4] M. Brady, R. Raghavan and J. Slawny, “Backpropagation fails to separate where perceptrons succeed,” *IEEE Trans. Circuits and Systems* **36**(1989): 665-674.
- [5] Conway, J.B., *Regular Algebra and Finite Machines*, Chapman and Hall, London, 1971.
- [6] Gabrielov, A., “Projections of semi-analytic sets,” *Functional Anal. Appl.*, **2**(1968): 282-291.
- [7] Goldman, S.A., and M.J. Kearns, “On the complexity of teaching,” *Proc. Forth ACM Workshop on Computational Learning Theory*, July 1991, pp. 303-314.
- [8] Gori, M., and A. Tesi, “On the problem of local minima in back-propagation,” Tech. Report RT-DSI 6/90, Univ. di Firenze, April 1990.
- [9] Khovanskii, A.G., *Fewnomials*, American Mathematical Society, Providence, R.I., 1991.
- [10] Knight, J., A. Pillay, and C. Steinhorn, “Definable sets in ordered structures, II” *Trans. Amer. Math. Soc.* **295** (1986): 593-605.
- [11] Macintyre, A., and E.D. Sontag, “Finiteness results for sigmoidal ‘neural’ networks,” in *Proc. 25th Annual Symp. Theory Computing*, San Diego, May 1993, pp. 325-334.
- [12] Milnor, J.W., *Morse Theory* Princeton University Press, 1963.
- [13] Palais, R., and C-I. Terng, *Critical Point Theory and Submanifold Geometry*, Springer-Verlag, Berlin, New York, 1988.
- [14] Poston, T., C-N Lee, Y-J Choie, and Y. Kwon, “Local minima and backpropagation,” in *Int. Joint Conf. Neural Networks*, Seattle, IEEE Press, 1991, pp. 173-176.
- [15] Sontag, E.D., *Mathematical Control Theory: Deterministic Finite Dimensional Systems*, Springer, New York, 1990.

- [16] Sontag, E.D., “Feedforward nets for interpolation and classification,” *J. Comp. Syst. Sci.* **45**(1992): 20-48.
- [17] 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.
- [18] Sussmann, H.J., “Real analytic desingularization and subanalytic sets: An elementary approach,” *Trans. Amer. Math. Soc.* **317**(1990): 417-461.
- [19] Sussmann, H.J., “Uniqueness of the weights for minimal feedforward nets with a given input-output map,” *Neural Networks* **5**(1992): 589-593.
- [20] van den Dries, L., “A generalization of the Tarski-Seidenberg theorem, and some nondefinability results,” *Bull. AMS* **15**(1986): 189-193.
- [21] van den Dries, L., “Tame topology and 0-minimal structures”, preprint, University of Illinois, Urbana, 1991-2.
- [22] van den Dries, L., and C. Miller, “On the real exponential field with restricted analytic functions,” *Israel J. Math.* **85** (1994): 19-56.
- [23] van den Dries, L., A. Macintyre, and D. Marker, “The elementary theory of restricted analytic fields with exponentiation,” *Annals of Math.* **140** (1994): 183-205.
- [24] Williamson, R.C., and U. Helmke, “Approximation Theoretic Results for Neural Networks,” in *Proceedings of the Australian Conference on Neural Networks*, 1992, pp. 217-222. (Also, “Existence and uniqueness results for neural network approximations,” *IEEE Transactions on Neural Networks* **6**(1995): 2-13.)
- [25] Wilkie, A.J., “Some model completeness results for expansions of the ordered field of reals by Pfaffian functions,” preprint, Oxford, 1991, submitted.
- [26] Wilkie, A.J., “Smooth 0-minimal theories and the model completeness of the real exponential field,” preprint, Oxford, 1991, submitted.