

Shattering all sets of k points in “general position” requires $(k - 1)/2$ parameters

Eduardo D. Sontag*

Department of Mathematics

Rutgers University, New Brunswick, NJ 08903[†]

Abstract

For classes of concepts defined by certain classes of analytic functions depending on n parameters, there are nonempty open sets of samples of length $2n + 2$ which cannot be shattered. A slightly weaker result is also proved for piecewise-analytic functions. The special case of neural networks is discussed.

1 Introduction

The generalization capabilities of neural networks are often quantified in terms of the maximal number of possible binary classifications that could be obtained, by means of weight assignments, on any given set of input patterns. Obviously, the larger the number of such potential classifications, the lower the predictability that is possible on the basis of a partial assignment (“loading of training data”) already achieved. Thus, it is of interest to obtain useful upper bounds on this number, and in particular to study the Vapnik-Chervonenkis (VC) dimension, which is the size of the largest set of inputs that can be shattered (arbitrary binary labeling is possible). Recent results (cf. [Koiran and Sontag 1996], and also [Maass 1994] for closely related work and a survey) show that the VC dimension grows at least as fast as the *square* n^2 of the number of adjustable weights n in the net, and this number might grow as fast as n^4 ([Karpinski and Macintyre 1996]). These results are quite pessimistic, since they imply that the number of samples required for reliable generalization, *in the sense of PAC learning*, is very high.

On the other hand, it is conceivable that those sets of input patterns which can be shattered are all in some sense “special” and that if we ask instead, as done in the classical literature

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

[†]E-mail: sontag@hilbert.rutgers.edu

in pattern recognition, for the shattering of *all sets in “general position”* (e.g. Cover’s work on capacity of perceptrons [Cover 1988]), then an upper bound of $O(n)$ might hold. Strong evidence for this possibility was provided by Adam Kowalczyk, who showed in [Kowalczyk 1996] that this indeed happens for the (very special) case of hard-threshold neural networks with fully connected first layer. In this paper, we establish a linear upper bound for arbitrary sigmoidal (as well as threshold) neural nets, proving that in that sense the perceptron results can be recovered in a strong sense (up to a factor of two).

There is potential relevance of our results to variations of PAC learning, when one weakens the requirement that generalization capabilities must hold with respect to all possible input distributions; this will be the subject of future work. The estimate is also useful in the very different context of understanding computational abilities; as an illustration of this fact, we mention that our main result is being employed by Maass in [Maass 1996b] for contrasting the computational power of spiking neurons ([Maass 1996a]) with that of sigmoidal neural networks.

Parametric Classes, Shattering

Assume given two positive integers n and m , and a function

$$\beta : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}.$$

In this context we will call n the *number of parameters*, elements of \mathbb{R}^n parameter vectors (or *weights*), m the *input dimension*, elements of \mathbb{R}^m input vectors, and β a *response* function. We think of $\beta(x, u)$ as a function of the inputs u for each parameter vector x . For each integer k , a k -tuple of the form

$$\vec{u} = (u_1, \dots, u_k) \in (\mathbb{R}^m)^k = \mathbb{R}^{mk}$$

will be called an (ordered) *sample of length k* .

We want to measure how rich the set of functions $\{\beta(x, \cdot), x \in \mathbb{R}^n\}$ is for binary classification purposes. Consider a sample $\vec{u} = (u_1, \dots, u_k)$ of length k . A sequence $\vec{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_k) \in \{-1, 1\}^k$ will be said to be *assignable* to \vec{u} (with respect to the given response function) if there exists some parameter vector x so that

$$\text{sign}(\beta(x, u_i)) = \varepsilon_i, \quad i = 1, \dots, k,$$

where we are defining $\text{sign}(y) = 1$ if $y > 0$ and -1 otherwise. We will say here that a sample \vec{u} of length k is \pm -*shattered* (by β) if, for each $\vec{\varepsilon} \in \{-1, 1\}^k$, either $\vec{\varepsilon}$ or $-\vec{\varepsilon}$ is assignable to \vec{u} .

For each k , let S_k be the subset of $(\mathbb{R}^m)^k = \mathbb{R}^{mk}$, possibly empty, consisting of those samples that are \pm -shattered. We define

$$\mu = \mu_\beta := \sup \left\{ k \geq 1 \mid S_k \text{ is a dense subset of } \mathbb{R}^{mk} \right\}.$$

Our goal is to show that, under reasonable assumptions on β , necessarily $\mu \leq 2n + 1$.

The organization of this note is as follows. The next section discusses the precise statement of the result, for functions β that are analytic and “definable” as well as the fact that this bound is optimal. The following section provides a proof of the result, which is based on material on analytic functions and especially on new results from the theory of exponentials, which are reviewed briefly in an appendix. The last section shows how to generalize the result to the piecewise analytic definable case, which allows the consideration of neural networks with discontinuous activation functions.

2 Precise Statements and Remarks

We remark first that $2n + 1$ is the *best possible upper bound*, assuming that one wants to prove a result which includes at least all polynomially parameterized classes. To show this we must exhibit, for each integer $n > 0$, a polynomial response function $\beta : \mathbb{R}^n \times \mathbb{R}^{m_n} \rightarrow \mathbb{R}$ with the property that S_{2n+1} is dense in $\mathbb{R}^{(2n+1)m_n}$. In fact, we give, for each n , a β with $m_n = 1$ so that *every* sample of length $2n + 1$ consisting of distinct vectors can be \pm -shattered:

$$\beta((x_1, \dots, x_n), u) := (u - x_1)^2 (u - x_2)^2 \dots (u - x_n)^2 .$$

Take any $\vec{u} \in \mathbb{R}^{2n+1}$ with all u_i distinct; the claim is that this is \pm -shattered. Indeed, pick any $\vec{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_{2n+1}) \in \{-1, 1\}^{2n+1}$. Without loss of generality, we assume that there are $s \leq n$ elements “ -1 ” in the sequence $\vec{\varepsilon}$ (otherwise, we assign $-\vec{\varepsilon}$ instead) and these are $\varepsilon_{i_1}, \dots, \varepsilon_{i_s}$. Let $x_j := u_{i_j}$ for $j = 1, \dots, s$ and pick x_{s+1}, \dots, x_n to be any elements not in the set $\{u_1, \dots, u_{2n+1}\}$. We have that $\beta(x, u) = 0$ for each $u = u_{i_j}$, $j = 1, \dots, s$, and $\beta(x, u) > 0$ for all other $u \in \mathbb{R}$. Thus $\text{sign}(\beta(x, u_i)) = \varepsilon_i$ for all i .

The question we address next is what are “reasonable” assumptions on the class β so that an upper bound like $2n + 1$ holds. It is not difficult to see that merely asking β to be an analytic function is not sufficient even to guarantee finiteness; this is illustrated trivially by the example $\beta(x, u) = \sin(xu)$, which has $n = m = 1$ but for which $\mu_\beta = +\infty$ (this is a consequence of the fact that, if the numbers u_1, \dots, u_k, π are linearly independent over the rationals, then the set of vectors $(\sin(\ell u_1), \dots, \sin(\ell u_k))$, as ℓ ranges over the positive integers, is a dense subset of $[-1, 1]^k$). In fact, it is possible to define a fixed analytic function β , with $m = 1$, for which S_k consists of *all* sequences of k distinct elements of \mathbb{R} , not merely dense in \mathbb{R}^{km} , for all k ; see [Sontag 1992].

Thus, analyticity is in itself not enough to obtain nontrivial bounds. We will assume that β is analytic and *definable*. The Appendix recalls the definition and basic facts about (“exp-RA”)

definable functions—informally, these are functions that can be defined in terms of any first-order logic sentence which is built out of the standard propositional connectives, existential and universal quantification, and which involves rational operations and exponentiation. (Also allowed in the formulas are certain “restricted analytic” functions such as for instance $\arctan(x)$, but $\sin(x)$ is not included.) In particular, and this is of interest in the context of applications to “artificial neural networks,” any response of a “multilayer sigmoidal network” with “activation function” $1/(1 + e^{-x})$ is definable in this sense, because it is obtained by iterative compositions and polynomial combinations involving the activation function, which in turn is constructed by means of rational operations involving exponentiation.

Theorem 1 *If $\beta : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}$ is an analytic definable response, then $\mu_\beta \leq 2n + 1$.*

This result is proved in the next section.

Remark 2.1 In computational learning theory, one defines *shattering* of a sample \vec{u} by the requirement that each $\vec{\varepsilon}$ be assignable, and studies the set S_k^+ consisting of all those samples that are shattered. For our results, it seems more natural to look at \pm -shattering, as we then obtain a best-possible bound. In any case, the two concepts are almost identical. Since $S_k^+ \subseteq S_k$ for all k , the upper bound $2n + 1$ will also apply in particular to $\sup\{k \geq 1 \mid S_k^+ \text{ is a dense subset of } \mathbb{R}^{km}\}$. (Conversely, if \vec{u} can be \pm -shattered, then the new response function with $n + 1$ parameters, and defined for parameters vectors $(x_0, x_1, \dots, x_n) \in \mathbb{R}^{n+1}$ by $\beta_0((x_0, x), u) := 1 - x_0\beta(x, u)$, shatters the same sample \vec{u} .) In learning theory and statistical estimation, the *Vapnik-Chervonenkis (VC) dimension* is defined as the supremum of the integers k such that S_k^+ is *nonempty*; here we are asking questions regarding shattering of arbitrary sets of points in “general position,” a concept which appears in Cover’s and in other classical pattern recognition work. For the best known (obviously larger) upper bounds for VC dimension, for closely related classes of responses, see [Karpinski and Macintyre 1996], and see also [Maass 1994, Koiran and Sontag 1996] for superlinear lower bounds. \square

3 Proof of Main Result

We need to show that if $k \geq 2n + 2$ then S_k is not dense, that is to say, there is some nonempty open subset $Q \subseteq \mathbb{R}^{mk}$ so that $Q \cap S_k = \emptyset$. It will be enough to show this when $k = 2n + 2$. In fact, we will establish a much stronger fact: for each nonempty open subset $W_0 \subseteq \mathbb{R}^{m(n+1)}$, there is a nonempty open subset $Q \subseteq W_0^2$ such that $Q \cap S_{2n+2} = \emptyset$.

For any integer ℓ , we introduce the following subset of $\mathbb{R}^{m\ell}$:

$$A_\ell := \{(u_1, \dots, u_\ell) \mid \exists x \text{ st } \beta(x, \cdot) \not\equiv 0, \beta(x, u_1) = \dots = \beta(x, u_\ell) = 0\}.$$

By Corollary (A.2), it follows that A_ℓ has dimension strictly less than $m\ell$ whenever $\ell = n + 1$. (The content of that Lemma is totally intuitive: for each fixed x , except those for which the map $\beta(x, u)$ is identically zero, the set of u_1 's so that $\beta(x, u_1) = 0$ is of dimension at most $m - 1$ (since this set is the set of zeroes of a nontrivial analytic function); and similarly for each of u_2, \dots, u_ℓ . Thus, for each such x , the Cartesian product of these sets, namely the set of vectors (u_1, \dots, u_ℓ) so that $\beta(x, u_1) = \dots = \beta(x, u_\ell) = 0$, has dimension at most $\ell(m - 1)$. If we now let x range over the whole parameter space \mathbb{R}^n (but not including those x for which $\beta(x, \cdot) \equiv 0$), the set A_ℓ is obtained, as a union of an n -parameter family of sets, each of which is of dimension $\leq \ell(m - 1)$, from which it follows that A_ℓ has dimension at most $n + \ell(m - 1)$, which is less than $m\ell$ provided $\ell = n + 1$. The technicalities have to do only with defining precisely the meaning of “dimension” –this will mean the maximum possible dimension of a submanifold of the set– and establishing the obvious formulas, such as the fact that the union of an n -parameter family of sets of dimension $\leq s$ has dimension $\leq n + s$.)

Let W_0 be any open subset of $\mathbb{R}^{m(n+1)}$. The fact that β is definable implies, by Lemma (A.3), that there are only two possibilities for each set A_ℓ : either (a) it contains some open set or (b) it is nowhere dense. Since when $\ell = n + 1$ this set does not have full dimension, as shown in the previous paragraph, alternative (a) cannot hold, so it must be nowhere dense. That is, its closure has empty interior, which implies in particular that $W_0 \setminus \text{clos } A_{n+1}$ is a nonempty open set. Thus there is some open set V of $W_0 \subseteq \mathbb{R}^{m(n+1)}$ of the special product form $V = V_1 \times \dots \times V_{n+1}$, for some nonempty connected open subsets V_1, \dots, V_{n+1} of \mathbb{R}^m , so that $V \cap A_{n+1} = \emptyset$.

Now let $Q \subseteq W_0^2 \subseteq \mathbb{R}^{2m(n+1)}$ be the open set defined as follows:

$$Q := V \times V.$$

The claim is that $Q \cap S_{2n+2} = \emptyset$. To establish this fact, we take an arbitrary element

$$(u_1, \dots, u_{n+1}, v_1, \dots, v_{n+1})$$

of Q and we show that it cannot be \pm -shattered, that is, there is some sequence $\vec{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_{2n+2}) \in \{-1, 1\}^{2n+2}$ for which neither $\vec{\varepsilon}$ nor $-\vec{\varepsilon}$ is assignable. Consider the following sequence:

$$\vec{\varepsilon} := \underbrace{(1, \dots, 1)}_{n+1}, \underbrace{(-1, \dots, -1)}_{n+1}.$$

Assume that there would exist some parameter vector x so that this vector is assigned:

$$\beta(x, u_i) > 0, i = 1, \dots, n + 1, \quad \beta(x, v_i) \leq 0, i = 1, \dots, n + 1.$$

Observe that, in particular, it holds that $\beta(x, \cdot) \not\equiv 0$ for this x . For each $i = 1, \dots, n + 1$, let

$$\gamma_i : [0, 1] \rightarrow V_i \subseteq \mathbb{R}^m$$

be a continuous function so that

$$\gamma_i(0) = u_i \quad \text{and} \quad \gamma_i(1) = v_i.$$

(recall that V_i is connected). Since $\beta(x, \gamma_i(t))$ is a continuous function of t , we conclude that for each i there is some t_i so that $\beta(x, \gamma_i(t_i)) = 0$. Writing $w_i := \gamma_i(t_i)$, this says that $\vec{w} := (w_1, \dots, w_{n+1}) \in A_{n+1}$, which contradicts the fact that $\vec{w} \in V_1 \times \dots \times V_{n+1} = V$ and that V was picked so that it is disjoint from A_{n+1} . Thus $\vec{\varepsilon}$ is not assignable. If $-\vec{\varepsilon}$ would be assignable, we would have an x so that $\beta(x, u_i) \leq 0$ for $i = 1, \dots, n+1$ and $\beta(x, v_i) > 0$ for $i = 1, \dots, n+1$, and the same argument leads once more to a contradiction.

Remark 3.1 Observe that the definability property is used when proving that A_{n+1} cannot be dense. If instead one would pick for instance $\beta(x, u) = \sin(xu)$, the property is false: for this example, A_2 is the union of all the lines in \mathbb{R}^2 that pass through $(0, 0)$ and have a rational slope (plus the y -axis), so A_2 is in this case a dense set with empty interior. \square

Remark 3.2 As remarked during the proof, a stronger result is established. It is shown, in particular, that for each non empty open subset $W \subseteq \mathbb{R}^m$, there is a nonempty open subset $Q \subseteq W^{2n+2}$ such that $Q \cap S_{2n+2} = \emptyset$. (Apply with $W_0 := W^{n+1}$.) \square

4 Discontinuous Responses

The main theorem requires that the response function β be analytic as well as definable. It is possible, however, to extend its validity to some other classes of responses, including those obtained by considering neural networks which use ‘‘Heaviside’’ activation functions

$$H(v) = \begin{cases} 0 & \text{if } v \leq 0 \\ 1 & \text{if } v > 0 \end{cases}$$

as well as sigmoidal activations. Such responses are a particular case of a piecewise analytic definable response, meaning a response that is described by a (possibly different) definable analytic function in each piece of a partition of the space of parameters and inputs; the partition in turn is assumed to be determined by the signs of a finite number of definable analytic functions. We now make this concept precise.

The map $\beta : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}$ is a *piecewise analytic definable* response if there is some integer p and there exist $p + 2^p$ analytic definable functions

$$\phi_1, \dots, \phi_p : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}$$

and

$$\{\psi_\alpha : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}, \alpha \in \{0, 1\}^p\}$$

such that the following property holds: for each $(x, u) \in \mathbb{R}^n \times \mathbb{R}^m$, letting

$$\alpha(x, u) := (H(\phi_1(x, u)), \dots, H(\phi_p(x, u)))$$

then

$$\beta(x, u) = \psi_{\alpha(x, u)}(x, u).$$

(That is to say, in each of the components of the partition of the (x, u) space determined by the possible signs of the functions ϕ_i , β is expressed by an appropriate ψ_α .)

Corollary 4.1 If β is piecewise analytic definable, then $\mu_\beta \leq 2n + 3$.

The proof will be based on the construction of an analytic definable response

$$\beta' : \mathbb{R}^{n+1} \times \mathbb{R}^m \rightarrow \mathbb{R}$$

with the following property: if a given sequence $\vec{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_k) \in \{-1, 1\}^k$ is assignable to a sample $\vec{u} = (u_1, \dots, u_k)$, with respect to the response β , then it is also assignable to the same sample with respect to the response β' . This means, in particular, that if a sample \vec{u} is \pm -shattered with respect to β then it is also \pm -shattered with respect to β' , so we have that $\mu_\beta \leq \mu_{\beta'}$. By the main theorem we know that $\mu_{\beta'} \leq 2(n + 1) + 1$, so the desired conclusion follows. We now show how to construct β' .

As a preliminary step, we establish a simple fact regarding the approximation of piecewise constant functions by sigmoidal combinations. We first introduce some additional notations. Given any vector $r = (r_1, \dots, r_p) \in \mathbb{R}^p$, we let $\gamma(r) := (H(r_1), \dots, H(r_p))$. Consider the function

$$\theta : \mathbb{R}^p \times \mathbb{R}^{2^p} \rightarrow \mathbb{R} : (r, s) \mapsto s_{\gamma(r)}$$

(where we are indexing the coordinates of $s \in \mathbb{R}^{2^p}$ by vectors $\alpha = (\alpha_1, \dots, \alpha_p) \in \{0, 1\}^p$). Now let $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ be the function $\sigma(v) = (1 + e^{-v})^{-1}$ and, denoting $\sigma_1 := \sigma$, $\sigma_0 := 1 - \sigma$, consider, for each positive real number ρ , the (definable and analytic) function:

$$\chi_\rho : \mathbb{R}^p \times \mathbb{R}^{2^p} \rightarrow \mathbb{R} : (r, s) \mapsto \sum_{\alpha \in \{0, 1\}^p} \sigma_{\alpha_1}(\rho^2 r_1 - \rho) \dots \sigma_{\alpha_p}(\rho^2 r_p - \rho) s_\alpha. \quad (1)$$

Lemma 4.2 For each pair $(r, s) \in \mathbb{R}^p \times \mathbb{R}^{2^p}$,

$$\lim_{\rho \rightarrow \infty} \chi_\rho(r, s) = \theta(r, s) \quad (2)$$

and

$$\theta(r, s) = 0 \Rightarrow \lim_{\rho \rightarrow \infty} (1 + \rho^2) \chi_\rho(r, s) = 0. \quad (3)$$

Proof. Observe that we can rewrite θ as follows:

$$\theta(r, s) = \sum_{\alpha \in \{0,1\}^p} H_{\alpha_1}(r_1) \dots H_{\alpha_p}(r_p) s_\alpha$$

where $H_1 := H$, $H_0 := 1 - H$. Thus the first conclusion of the lemma is an immediate consequence of the fact that $\lim_{\rho \rightarrow \infty} \sigma(\rho^2 r - \rho) = H(r)$ for each $r \in \mathbb{R}$. Now assume that (r, s) is so that $\theta(r, s) = s_{\gamma(r)} = 0$. This means that the term corresponding to the index $\alpha = \gamma(r)$ does not appear in the sum (1) defining $\chi_\rho(r, s)$. Fix any other index α ; we claim that

$$\lim_{\rho \rightarrow \infty} (1 + \rho^2) \sigma_{\alpha_1}(\rho^2 r_1 - \rho) \dots \sigma_{\alpha_p}(\rho^2 r_p - \rho) = 0,$$

(from which the second conclusion will follow). Indeed, since $\alpha \neq \gamma(r)$, there must exist some $j \in \{1, \dots, p\}$ for which $\alpha_j \neq H(r_j)$; fix one such j . If $\alpha_j = 1$ then $H(r_j) = 0$ means that $r_j \leq 0$, and therefore $(1 + \rho^2) \sigma(\rho^2 r_j - \rho) \rightarrow 0$ as $\rho \rightarrow \infty$; since the other terms are bounded, it follows that their product converges to zero. If instead $\alpha_j = 0$ then $H(r_j) = 1$ means that $r_j > 0$, and thus $(1 + \rho^2)(1 - \sigma(\rho^2 r_j - \rho)) \rightarrow 0$ as $\rho \rightarrow \infty$; again since the other terms are bounded, the product goes to zero. \blacksquare

Equation (3) implies in particular that, when $\theta(r, s) = 0$,

$$\chi_\rho(r, s) - \frac{1}{1 + \rho^2} < 0$$

for all ρ large enough. Together with Equation (2), this means that

$$\text{sign} \left(\chi_\rho(r, s) - \frac{1}{1 + \rho^2} \right) = \text{sign} \theta(r, s)$$

as $\rho \rightarrow \infty$, for each $(r, s) \in \mathbb{R}^p \times \mathbb{R}^{2^p}$.

We now prove Corollary 4.1. Let $\beta, \phi_1, \dots, \phi_p$, and the ψ_α 's be as in the definition of piecewise analytic definable response. Denote $\Phi(x, u) := (\phi_1(x, u), \dots, \phi_p(x, u))$ and let $\Psi(x, u) \in \mathbb{R}^{2^p}$ be the vector whose α th component is $\psi_\alpha(x, u)$. Then $\beta(x, u) = \theta(\Phi(x, u), \Psi(x, u))$. We define

$$\beta'((\rho, x), u) := \chi_\rho(\Phi(x, u), \Psi(x, u)) - \frac{1}{1 + \rho^2}.$$

Thus

$$\text{sign} \beta'((\rho, x), u) = \text{sign} \beta(x, u)$$

as $\rho \rightarrow \infty$, for each $(x, u) \in \mathbb{R}^n \times \mathbb{R}^m$. Now assume that the sequence $\vec{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_k) \in \{-1, 1\}^k$ is assignable to the sample $\vec{u} = (u_1, \dots, u_k)$ with respect to the response β . This means that there is some parameter vector $x \in \mathbb{R}^n$ so that $\text{sign}(\beta(x, u_i)) = \varepsilon_i$ for each $i = 1, \dots, k$. For large enough ρ , then, also $\text{sign}(\beta'((\rho, x), u_i)) = \varepsilon_i$ for all i , so we conclude that $\vec{\varepsilon}$ is also assignable to the same sample with respect to the response β' .

Remark 4.3 The bound $2n + 3$ can be reduced to $2n + 1$ provided that the classes of functions $\phi_i(x, \cdot)$ and $\psi(x, \cdot)$ are closed under scaling and translation (as is the case commonly when dealing with artificial neural networks, where these are affine functions and the parameters x correspond to the coefficients of linear combinations (weights) and constant term (bias)). In that case, the parameter ρ can be absorbed into the remaining parameters. \square

A Appendix

We first review some elementary facts regarding real-analytic geometry. If M is any (analytic, second-countable) manifold, by a σ -analytic subset Z of M we mean one which can be decomposed into a countable union of embedded analytic submanifolds; for such a set Z , we define $\dim Z$ as the largest dimension of a submanifold appearing in such a union. (The dimension is well-defined, in the sense that it doesn't depend on the particular decomposition.)

Consider now any analytic map $\pi : M \rightarrow N$ between two manifolds. For each $y \in N$, the "fiber" $M_y = \{x \in M \mid \pi(x) = y\}$ provides an example of a σ -analytic subset (and, if M is connected, either $M_y = M$ or $\dim M_y < \dim M$). Now take any σ -analytic subset Z of M . Then the image $\pi(Z)$ is a σ -analytic subset of N , having $\dim \pi(Z) \leq \dim Z$, and the following inequality holds:

$$\dim Z \leq \dim \pi(Z) + \max_{y \in \pi(Z)} \dim [M_y \cap Z] . \quad (4)$$

(This inequality is a simple consequence of stratification theory; it is proved in the Appendix of [Sontag 1996]. Note, incidentally, that a strict inequality may hold, as illustrated by the case $M = \mathbb{R}^2$, $N = \mathbb{R}$, $\pi =$ projection on first factor, and $Z =$ the union of the x and the y axes.)

We prove the following fact, which is basically Theorem 1 in [Sontag 1996]. (That theorem could also be applied more directly, but the slight generalization given here should be useful for other purposes.)

Lemma A.1 Let $\mathbb{X}, \mathbb{U}_1, \dots, \mathbb{U}_\ell$ be (analytic, second countable) manifolds, with dimensions n, m_1, \dots, m_ℓ respectively, and each \mathbb{U}_i connected. Assume given ℓ analytic maps

$$\beta_i : \mathbb{X} \times \mathbb{U}_1 \times \dots \times \mathbb{U}_i \rightarrow \mathbb{R}, \quad i = 1, \dots, \ell$$

so that the following property holds for each $i = 1, \dots, \ell$:

$$\forall \bar{x} \in \mathbb{X}, \forall \bar{u}_1 \in \mathbb{U}_1, \dots, \forall \bar{u}_{i-1} \in \mathbb{U}_{i-1}, \exists u \in \mathbb{U}_i \text{ so that } \beta_i(\bar{x}, \bar{u}_1, \dots, \bar{u}_{i-1}, u) \neq 0. \quad (5)$$

Let

$$\mathcal{G}_\ell := \{(x, u_1, \dots, u_\ell) \in \mathbb{X} \times \mathbb{U}_1 \times \dots \times \mathbb{U}_\ell \mid \beta_1(x, u_1) = \dots = \beta_\ell(x, u_1, \dots, u_\ell) = 0\} .$$

Then this is a σ -analytic set with

$$\dim \mathcal{G}_\ell \leq n + \sum_{j=1}^{\ell} m_j - \ell .$$

Proof. Define $\mathcal{G}_i := \{(x, u_1, \dots, u_i) \in \mathbb{X} \times \mathbb{U}_1 \times \dots \times \mathbb{U}_i \mid \beta_1(x, u_1) = \dots = \beta_i(x, u_1, \dots, u_i) = 0\}$ for each $i = 1, \dots, \ell$, and $\mathcal{G}_0 := \mathbb{X}$; we prove by induction on i that $\dim \mathcal{G}_i \leq n + \sum_{j=1}^i m_j - i$.

For $i = 0$ the result holds, so we now assume it has been proved for i and show the case $i + 1$.
Let

$$\pi : \mathbb{X} \times \mathbb{U}_1 \times \dots \times \mathbb{U}_{i+1} \rightarrow \mathbb{X} \times \mathbb{U}_1 \times \dots \times \mathbb{U}_i$$

be the projection on the first $1 + i$ factors. Write $Z := \mathcal{G}_{i+1}$. Note that $\pi(Z) \subseteq \mathcal{G}_i$ by definition of these sets, so in particular it holds by inductive hypothesis that

$$\dim \pi(Z) \leq n + \sum_{j=1}^i m_j - i. \quad (6)$$

For each fixed $y = (\bar{x}, \bar{u}_1, \dots, \bar{u}_i) \in \pi(Z)$, $\pi^{-1}(y) \cap Z = \{\bar{x}\} \times \{\bar{u}_1\} \times \dots \times \{\bar{u}_i\} \times Q$, where

$$Q_{(\bar{x}, \bar{u}_1, \dots, \bar{u}_i)} := \{u \in \mathbb{U}_{i+1} \mid \beta_{i+1}(\bar{x}, \bar{u}_1, \dots, \bar{u}_i, u) = 0\}.$$

Property (5) implies that $\dim Q_{(\bar{x}, \bar{u}_1, \dots, \bar{u}_i)} \leq m_{i+1} - 1$. Thus from Equation (4) and using (6) we conclude that

$$\dim Z \leq n + \sum_{j=1}^i m_j - i + m_{i+1} - 1 = n + \sum_{j=1}^{i+1} m_j - (i + 1)$$

and the induction is complete. ■

Corollary A.2 Let $\beta : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}$ be analytic, and consider for any integer ℓ the set

$$A_\ell := \{(u_1, \dots, u_\ell) \mid \exists x \text{ st } \beta(x, \cdot) \not\equiv 0, \beta(x, u_1) = \dots = \beta(x, u_\ell) = 0\}.$$

Then, this is a σ -analytic set of dimension at most $n + \ell(m - 1)$.

Proof. Let $\mathbb{X} = \{x \in \mathbb{R}^n \mid \beta(x, \cdot) \not\equiv 0\}$; this is an open subset of \mathbb{R}^n and hence a submanifold. (If $\mathbb{X} = \emptyset$, there is nothing to prove.) Apply Lemma (A.1) with $\beta_i(x, u_1, \dots, u_i) := \beta(x, u_i)$ (and all $\mathbb{U}_i = \mathbb{U}$); observe that property (5) holds by definition of \mathbb{X} . Then $\dim \mathcal{G}_\ell \leq n + \ell(m - 1)$, and the same bound applies to its projection A_ℓ on the \mathbb{U} components. ■

Finally, we review several facts about exp-RA definable functions. This discussion is based on recent work in model theory due to Gabrielov, Van den Dries, Wilkie, and many others; see [Macintyre and Sontag 1993, Sontag 1996] for more details and extensive bibliographical references.

A *restricted analytic (RA) function* $\mathbb{R}^q \rightarrow \mathbb{R}$ is any function which is real analytic on some compact subset C of \mathbb{R}^q and is zero outside C (examples: the functions SIN and COS which equal sin and cos on $[-\pi/2, \pi/2]$ and are zero outside this interval). 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. (Arbitrary real constants are allowed.) An (*exp-RA*) *definable set* is any subset of \mathbb{R}^k defined by a first-order formula over L with k free variables. A *definable function* is a function $\beta : M \rightarrow N$ whose graph is a definable set in this sense, and where N and M are definable subsets of two spaces \mathbb{R}^{l_1} and \mathbb{R}^{l_2} respectively. For example, any function obtained by compositions of rational operations and taking exponentials is definable, such as $1/(1 + e^{-x})$; also definable is for instance the function $\arctan(x)$, since its graph is characterized by the formula “ $y = \arctan(x)$ iff $-\pi/2 < y < \pi/2$ and $\sin(y) = x \cos(y)$ ”. Notice that any set obtained from a function β by logical operations (such as the set A_ℓ in Corollary (A.2)) is definable provided that β is definable. The following fact is a nontrivial consequence of the work in logic cited above; see precise references in [Sontag 1996]:

Lemma A.3 Let S be a definable subset of \mathbb{R}^q , for some q . Then, either S contains an open subset, or it is a finite union of connected embedded submanifolds of \mathbb{R}^q (and in particular is nowhere dense). □

References

- [Cover 1988] Cover, T.M., “Capacity problems for linear machines”, in *Pattern Recognition*, L. Kanal ed., *Thompson Book Co.*, 1988, pp. 283-289.
- [Karpinski and Macintyre 1996] Karpinski, M., and A. Macintyre, “Polynomial bounds for VC dimension of sigmoidal and general Pfaffian neural networks,” *J. Computer Systems Sciences* (1996), to appear. (Summarized version: “Polynomial bounds for VC dimension of sigmoidal neural networks,” in Proc. 27th ACM Symposium on Theory of Computing, 1995, pp. 200-208.)
- [Koiran and Sontag 1996] Koiran, P., and E.D. Sontag, “Neural networks with quadratic VC dimension,” *J. Computer Systems Sciences* (1996), to appear. (Summarized version in *Advances in Neural Information Processing Systems 8* (NIPS95) (D.S. Touretzky, M.C. Moser, and M.E. Hasselmo, eds.), MIT Press, Cambridge, MA, 1996, pp. 197-203.)
- [Kowalczyk 1996] Kowalczyk, A., “Estimates of storage capacity of multilayer perceptron with threshold logic units,” *Neural Networks* (1996), to appear. (Preliminary version appeared as “Counting function theorem for multi-layer networks,” in *Advances in Neural Information Processing Systems 6*, Cowan, J.D., Tesauro, G., and Alspector, J., editors, Morgan Kaufman, 1994, pp 375-382.)
- [Maass 1994] Maass, M., “Perspectives of current research about the complexity of learning in neural nets,” in *Theoretical Advances in Neural Computation and Learning*, V.P. Roychowdhury, K.Y. Siu, and A. Orlicsky, editors, Kluwer, Boston, 1994, pp. 295-336.
- [Maass 1996a] Maass, W., “Lower bounds for the computational power of networks of spiking neurons,” *Neural Computation* **8**(1996): 1-40.
- [Maass 1996b] Maass, W., “The third generation of neural network models,” in *Proc. 7th Australian Conference on Neural Networks 1996*, Canberra, to appear.
- [Macintyre and Sontag 1993] 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.
- [Sontag 1992] Sontag, E.D., “Feedforward nets for interpolation and classification,” *J. Comp. Syst. Sci.* **45**(1992): 20-48.
- [Sontag 1996] Sontag, E.D., “Critical points for least-squares problems involving certain analytic functions, with applications to sigmoidal nets,” *Advances in Computational Mathematics* (Special Issue on Neural Networks) (1996), to appear.