Bounding VC-Dimension for Neural Networks: Progress and Prospects

Marek Karpinski* (Bonn)

Angus Macintyre** (Oxford)

Abstract. Techniques from differential topology are used to give polynomial bounds for the VC-dimension of sigmoidal neural networks. The bounds are quadratic in w, the dimension of the space of weights. Similar results are obtained for a wide class of Pfaffian activation functions. The obstruction (in differential topology) to improving the bound to an optimal bound $\mathcal{O}(w \log w)$ is discussed, and attention is paid to the role of other parameters involved in the network architecture.

* Dept. of Computer Science, University of Bonn, 53117 Bonn. Research partially supported by the DFG Grant KA 673/4-1, and by ESPRIT BR Grants 7097 and ECUS 030.

** Mathematical Institute, University of Oxford, Oxford OX1 3LB. Research supported in part by a Senior Research Fellowship of the SERC.

1 Introduction

We refer to Macintyre-Sontag [MS93](cf, e.g., also [AB92] and [GJ93]) for all notions required from the theory of neural architectures, and to Hirsch for necessary notions from differential topology [H76]. There [MS93], some general (but profound) results from logic were used to show that for feedforward neural architectures with activation the standard sigmoid $\sigma(y) = 1/1 + e^{-x}$ the VC-dimension is finite. The method applies to a huge variety of other activation functions, and the polynomials involved in computing could be replaced by much more general functions. The method is, however, inadequate to give interesting bounds.

Taking a hint from Goldberg and Jerrum works [GJ93], and a reference to Warren's [W68], we have found a method, not appealing to logic but using rather more differential topology, for giving very good polynomial bounds when the activation function satisfies a special kind of Pfaffian differential equation. The restriction to Pfaffian is because we use a method of Khovanski [K91].

2 Formulation and Main Results

2.1. The method is by no means restricted to neural architectures, and we choose to present it in greater generality. Fix integers k, l and C^∞ (infinitely differentiable) functions $\tau_1, \ldots, \tau_s$ from $\mathbf{R}^{k+\ell}$ to $\mathbf{R}$. Write τ_i as $\tau_i(v_i, \ldots, v_k, y_1, \ldots, y_\ell)$ (or $\tau_i(\bar{v}, \tilde{y})$).

Let $\Phi(\bar{v}, \tilde{y})$ be a Boolean combination of conditions

$$\tau(\bar{v}, \tilde{y}) > 0$$

or

$$\tau(\bar{v}, \tilde{y}) = 0,$$

where τ is one of the τ_i.

Φ defines a set in $\mathbf{R}^{k+\ell}$. For $\tilde{\beta}$ in $\mathbf{R}^\ell$, let $\Phi_{\tilde{\beta}}$ be the set in in $\mathbf{R}^k$ defined by $\Phi(\bar{v}, \tilde{\beta})$. Let $\mathcal{C}_\Phi$ be the family of all $\Phi_{\tilde{\beta}}$ as $\tilde{\beta}$ varies through $\mathbf{R}^\ell$.

We give good bounds of the VC-dimension of $\mathcal{C}_\Phi$, under some assumptions (some are necessary!) about the τ_i. In [KM94] it is shown how the VC-dimension of a neural architecture with k inputs and l weights is a special case of a VC-Dim($\mathcal{C}_\Phi$).

2.2. We maintain the notation of 2.1.
Call a Θ-function from $\mathbf{R}^{k+\ell}$ to $\mathbf{R}$ one of the form $\tau(\bar{\alpha}, \tilde{y})$ for a τ in the list, and some $\bar{\alpha} \in \mathbf{R}^k$. An F-function from $\mathbf{R}^\ell$ to some $\mathbf{R}^r$ $(r \leq \ell)$ is a function

$$\tilde{y} \longmapsto \langle \Theta_1(\tilde{y}), \cdots, \Theta_r(\tilde{y}) \rangle \; = \; F(\tilde{y}) \rangle$$

where the Θ_i are Θ-functions.

By Sard's Theorem, the p in $\mathbf{R}$ such that $F^{-1}(p)$ is either $\emptyset$ or an $l-r$ manifold, have measure 1.

Our assumption on $\tau_1, \ldots, \tau_s$ is that there is a bound B, independent of the F-function F (and $r \leq \ell$) such that for $p \in \mathbf{R}$ $F^{-1}(p)$ has $\leq B$ connected components. This is true, for example, for the Θ corresponding to sigmoidal neural networks [KM94]. Our general result, based essentially on Warren's ideas [W68], is:

Theorem 1 VC-Dim$(\mathcal{C}_\Phi) \leq 2 \log\ B\ +\ 16\ l\ (\log\ s + 1)$.

Note: Goldberg and Jerrum [GJ93] have a result of this type when the τ are polynomial and apply it to get a bound of order $l \log l$ for the VC-dimension of neural netwoks with semi-algebraic activation functions. B was estimated via a result of Warren [W68] (or Milnor [M64] can be used) but Warren's method was irrelevant in [GJ93]. $l \log l$ appears not from $\log B$, but from the $l \log s$, by crude estimates.

2.1 Application to Neural Nets

In this case $\Phi(\bar{v}, \tilde{y})$ is

$$\tau(\bar{v}, \tilde{y}) > 0,$$

where τ is composed out of polynomials and activation functions according to the structure of the underlying graph. The problem then is to bound $\log B$ explicitly. For sigmoidal activation functions, one can appeal to a method, and results, of Khovanski [K91].

In order to avoid a clash of notation (cf. [M93]), we now write w for ℓ.
w is the dimension of the weight space. The other relevant parameters are:

i) d = a bound for the degrees of all polynomials involved;
ii) m = number of computation nodes.

Then:

Theorem 2 (For the neural network $\mathcal{A}$ with the standard sigmoid activation function σ).

$$\begin{aligned} VC - Dim(\mathcal{A}) &\leq (mw)(mw-1) \\ &+ 6mw \log w \\ &+ 8mw \log(2md+1) \\ &+ 16w \log(2md+1). \end{aligned}$$

Note: The only troublesome term is $(mw)(mw-1)$. Its presence is easily traced to the term $2^{q(q-1)/2}$ in Khovanskii's basic estimate [K91], where q is the number of exponentials involved in a problem.

For $w = 1$, we can use another method, going back to Hardy [H12],allowing us to replace $mw\ (mw-1)\ \ (= m(m-1))$ by a term linear in m.

For the proof of Theorem 2, see [KM94].

3 Generalization and Prospects

3.1. Theorem 2, with a quadratic dominant term, does not depend too strongly on the form of the activation function. To see this, one can appeal to Khovanskii's book [K91](p. 91). There an argument is given generalizing that used earlier for sets defined by conditions

$$p(\tilde{y}, e^{\Lambda_1(\tilde{y})}, \cdots, e^{\Lambda_q(\tilde{y})}) = 0,$$

p polynomial, Λ_i's linear.

One can now replace the $e^{\Lambda_i(\tilde{y})}$ by q many functions occurring in a Pfaffian chain of length $\leq q$. So we can extend Theorem 2 to Pfaffian activation functions, but now we have to take into account the degrees of the polynomials occurring in the Pfaffian chain. Let D be a bound for these degrees. The only alteration in Theorem 2 is the replacement of the terms $\log(2md+1)$ by $\log(2md+D)$.

There is a more remarkable further generalization. There is an obvious way to consider networks with multivariate activation functions. If these are Pfaffian, we still get a quadratic dominant term. We will elaborate this in a future publication.

3.2. How to remove the quadratic term? One sees easily that it occurs because Khovanski [K91] removes one exponential at a time in his basic inductive method. We are hard at work on a method for removing all at once, and we expect to replace $mw(mw-1)$ by mw. □

References

[AB92] M. Anthony, N. Biggs, Computational Learning Theory: An Introduction, Cambridge University Press, 1992.

[AS93] M. Anthony, J. Shawe-Taylor, A Result of Vapnik with Applications, Discrete Applied Math. **47** (1993), pp. 207–217.

[BT90] A. Borodin, P. Tiwari, On the Decidability of Sparse Univariate Polynomial Interpolation, Proc. 22nd ACM STOC (1990), pp. 535–545.

[D92] L. van den Dries, Tame Topology and 0-minimal Structures, preprint, University of Illinois, Urbana, 1992; to appear as a book.

[DMM94] L. van den Dries, A.Macintyre and D.Marker, The Elementary Theory of Restricted Analytic Fields with Exponentation, Annuals of Mathematics **140** (1994), pp 183-205.

[GJ93] P.Goldberg and M.Jerrum, Bounding the Vapnik Chervonenkis Dimension of Concept Classes Parametrized by Real Numbers. Machine Learning, 1994 (to appear). A preliminary version appeared in Proc. 6th ACM Workshop on Computational Learning Theory, pp. 361–369, 1993.

[H12] G.H. Hardy, Properties of Logarithmic-Exponential Functions, Proc. London Math. Soc. 10 (1912), pp. 54–90.

[H92] D. Haussler, Decision Theoretic Generalizations of the PAC Model for Neural Nets and other Learning Applications, Information an Computation **100**, (1992), pp. 78–150.

[HKP91] J. Hertz, A. Krogh and R. G. Palmer, Introduction to the Theory of Neural Computation, Addison-Wesley, 1991.

[H76] M. W. Hirsch, Differential Topology, Springer-Verlag, 1976.

[KM94] M. Karpinski and A. Macintyre, Quadratic Bounds for VC Dimension at Sigmoidal Neural Networks, Research Report No. 85116-CS, Universität Bonn, 1994; to be submitted.

[KW93] M.Karpinski and T.Werther, VC Dimension and Uniform Learnability of Sparse Polynomials and Rational Functions, SIAM J. Computing **22** (1993), pp 1276–1285.

[K91] A.G.Khovanski, Fewnomials, American Mathematical Society, Providence, R.I., 1991.

[KPS86] J.Knight, A.Pillay and C.Steinhorn, Definable Sets and Ordered Structures II, Trans. American Mathematical Society **295** (1986), pp.593-605.

[L92] M.C.Laskowsky, Vapnik-Chervonenkis Classes od Definable Sets, J.London Math. Society **45** (1992), pp 377–384.

[M93] W.Maass, On the Complexity of Learning on Feedforward Neural Nets, in Proc. EATCS Advanced School on Computational Learning and Cryptography, Vietri sul Mare, 1993.

[MSS91] W. Maass, G. Schnitger and E. D. Sontag, On the Computational Power of Sigmoidal versus Boolean Threshold Circuits, Proc. 32nd IEEE FOCS (1991), pp. 767–776.

[MS93] A.J.Macintyre and E.D.Sontag, Finiteness results for Sigmoidal Neural Networks, Proc. 25th ACM STOC (1993), pp.325–334.

[M64] J.Milnor, On the Betti Numbers of Real Varieties, Proc. of the American Mathematical Society **15** (1964), pp 275–280.

[M65] J.Milnor, Topology from the Differentiable Viewpoint, Univ.Press, Virginia, 1965.

[S-T94] J. Shawe-Taylor, Sample Sizes for Sigmoidal Neural Networks, Preprint, University of London, 1994.

[TV94] G. Turan and F. Vatan, On the Computation of Boolean Functions by Analog Circuits of Bounded Fan-in, Proc. 35th IEEE FOCS (1994), pp. 553–564.

[W68] H.E.Warren, Lower Bounds for Approximation by Non-linear Manifolds, Trans. of the AMS **133** (1968), pp. 167–178.

[W94] A.J.Wilkie, Model Completeness Results of Restricted Pfaffian Functions and the Exponential Function, to appear in Journal of the AMS, 1994.