Correction to "Lower Bounds on the VC-Dimension of Smoothly Parametrized Function Classes"

Wee Sun Lee
Dept. of Systems Engineering,
RSISE, Aust. Nat. University,
Canberra, ACT 0200, Australia.
WeeSun.Lee@anu.edu.au

Peter L. Bartlett
Dept. of Systems Engineering,
RSISE, Aust. Nat. University,
Canberra, ACT 0200, Australia.
Peter.Bartlett@anu.edu.au

Robert C. Williamson
Department of Engineering,
Australian National University,
Canberra, ACT 0200, Australia.
Bob.Williamson@anu.edu.au

January 3, 1996

Abstract

The paper [3] gives lower bounds on the VC-dimension of various smoothly parametrized function classes. The results were proven by showing a relationship between the uniqueness of decision boundaries and the VC-dimension of smoothly parametrized function classes. The proof is incorrect; there is no such relationship under the conditions stated in [3]. For the case of neural networks with tanh activation functions, we give an alternative proof of a lower bound for the VC dimension proportional to the number of parameters which holds even when the magnitude of the parameters is restricted to be arbitrary small.

Theorem 10 from [3] which was used to prove lower bounds on various neural network classes is incorrect.

Theorem 10 ([3]) *Let A be an open subset of $\mathbb{R}^m$, X be an open subset of $\mathbb{R}^n$ and $f : A \times X \to \mathbb{R}$ be a continuously differentiable function (in all of its arguments). Let $F := \{f(a, \cdot) : a \in A\}$. If there exists a k-dimensional manifold $M \subset A$ which has unique decision boundaries, then $\mathrm{VCdim}(F_\theta) \geq k$, where F_θ is the thresholded function class formed from F.*

As counterexamples, observe that $y = (a - x)^2$ has unique decision boundaries but VC dimension zero. Similarly $y = (x - a)(x - b)^2$ has unique decision boundaries but VC dimension 1.

In [3], Theorem 10 was used to prove a lower bound for the VC-dimension of neural networks with tanh activation functions. We can prove this lower bound more directly using Lemma 9 from [3].

Lemma 9 ([3]) *Let A be an open subset of $\mathbb{R}^m$, X be an open subset of $\mathbb{R}^n$ and $f : A \times X \to \mathbb{R}$ be a continuously differentiable function (in all of its arguments). Let $F := \{f(a, \cdot) : a \in A\}$. Let φ be any diffeomorphism from an open subset of A to a subset of $\mathbb{R}^m$. For $x \in \mathbb{R}^m$, denote $\psi_x = f_x \circ \varphi^{-1}$, where $f_x = f(\cdot, x)$. If $\mathrm{VCdim}(F_\theta) < k$, then for every $a \in A$, for every $x \in X^m$, $f_x(a) = 0 \Rightarrow \mathrm{rank}(D\psi_x(b)) < k$, where $b = \varphi(a)$.*

Definition 1 *A* two layer feedforward network *with an n-dimensional input $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, k hidden units with σ activation function, and weights $w = (v_{10}, \ldots, v_{kn}, w_0, \ldots, w_k) \in \mathbb{R}^W$ (where $W = kn + 2k + 1$) is a function $f : \mathbb{R}^W \times \mathbb{R}^n \to \mathbb{R}$ given by*

$$f(w, x) = w_0 + \sum_{i=1}^{k} w_i \ \sigma(v_i \cdot x + v_{i0}),$$

where $v_i = (v_{i1}, \ldots, v_{in})$ and $v_i \cdot x = \sum_{j=1}^{n} v_{ij} x_j$. The weights $w_0, v_{i0}, i = 1, \ldots, k$ are called the offsets.

First we give some conditions on the activation function σ for which our result below hold. These conditions are the similar to that used in [1] for proving the independence property. The function tanh satisfies these conditions.

Definition 2 *The function σ is admissible if it satisfies the following conditions. It is a real-analytic function, and it extends to a function $\sigma: \mathbb{C} \to \mathbb{C}$ analytic on a subset $\mathcal{D} \subset \mathbb{C}$ of the form*

$$\mathcal{D} = \{z \in \mathbb{C}: |\mathrm{Im}\ z| \le \lambda\} \backslash \{z_0, \overline{z_0}\}$$

where $\mathrm{Im}\ z_0 = \lambda$ and $\infty > \lambda > 0$. Here $\overline{z_0}$ is the complex conjugate of z_0 and z_0 and $\overline{z_0}$ are poles of σ.

We want to use Lemma 9 to find a lower bound for the VC-dimension of a two layer feedforward network with admissible activation functions. First we show that the following matrix of derivatives with respect to first layer weights is nonsingular for almost all sets of kn inputs $S = \{x^1, \ldots, x^{kn}\}$. (The technique used is similar to that in [2].) Define

$$D(S) = \begin{bmatrix} \sigma'(v_1 \cdot x^1 + v_{10}) w_1 x_1^1 & \sigma'(v_1 \cdot x^2 + v_{10}) w_1 x_1^2 & \cdots & \sigma'(v_1 \cdot x^{kn} + v_{10}) w_1 x_1^{kn} \\ \sigma'(v_1 \cdot x^1 + v_{10}) w_1 x_2^1 & \cdots & & \vdots \\ \vdots & & & \\ \sigma'(v_1 \cdot x^1 + v_{10}) w_1 x_n^1 & & & \\ \vdots & & & \\ \sigma'(v_k \cdot x^1 + v_{k0}) w_k x_1^1 & & & \\ \vdots & & & \\ \sigma'(v_k \cdot x^1 + v_{k0}) w_k x_n^1 \cdots & & & \end{bmatrix} \qquad (1)$$

Let $\omega_{i,j} = v_i \cdot x^j + v_{i0}$. Define a diagonal matrix $Z(S) = \mathrm{diag}((\omega_{1,1} - p_{1,1})^m, \ldots, (\omega_{1,n} - p_{1,n})^m, (\omega_{2,n+1} - p_{2,n+1})^m, \ldots, (\omega_{k,kn} - p_{k,kn})^m)$ where m is the multiplicity of the pole of $\sigma'(\cdot)$ at $p_{i,j}$ and $p_{i,j} \in \{z_0, \overline{z_0}\}$ for each component of $Z(S)$.

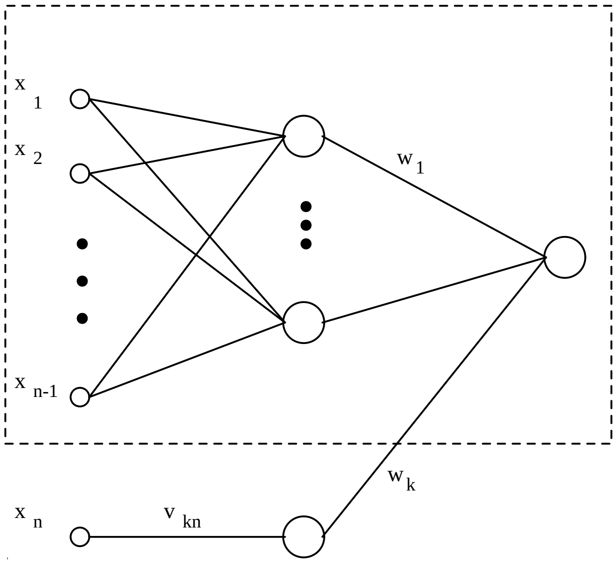

Figure 1: Network with weights deleted

We will show that there is a value of S such that $D(S)Z(S)$ is nonsingular which implies that $D(S)$ is nonsingular. Each component of $D(S)$ and $Z(S)$ is an analytic function of S on the set $\mathcal{D}^{kn}$ hence the determinant is also an analytic function of S. This means that the determinant is nonzero for almost all values of S in $\mathcal{D}^{kn}$ and hence also for almost all values of S when each member of S is from $\mathbb{R}^n$.

Pick an S such that $v_i \cdot x^j + v_{i0} = p_{i,j}$ for each $p_{i,j}$ which appears in a component of $Z(S)$ and for all other i, j, $v_i \cdot x^j + v_{i,0} \in \mathcal{D}$ and is not a pole of $\sigma'(\cdot)$. This can be done as long as no two hidden units are identical. (For each hidden unit i, we can choose its n corresponding x^j's so that the activations are at the same pole and avoid the hyperplanes defining the poles of the other hidden units. If we choose the components of each x^j small enough, the activation will remain in $\mathcal{D}$.) All the singularities in $D(S)Z(S)$ are removable and the determinant can be made an analytic function of S. With the singularities removed, the matrix $D(S)Z(S)$ assumes a block diagonal form as the entries where $v_i \cdot x^j + v_{i,0}$ is not a pole become zero. To complete the proof, we need that each block in the block diagonal matrix has full rank. To do that we need only choose the x^j's within each block such that they are linearly independent and avoid a finite number of hyperplanes which define the poles for the other hidden units. This is always possible since the input dimension is n and each block is of size $n \times n$.

We can use this result to prove the following modification of Theorem 13 in [3]. The bound obtained is $(k-1)(n-1)$ instead of $(k-1)(n+1)+1$.

Theorem 3 *Let F be the class of two layer feedforward networks with k hidden units with tanh activation, input space $X := \{(x_1, \ldots, x_n) \in \mathbb{R}^n : |x_i| < C\}$ (where C is a constant greater than zero), and $k(n+2)+1$ weights restricted to an open set which includes the origin. Then $\mathrm{VCdim}(F_\theta) \geq (k-1)(n-1)$.*

Proof. Delete all the weights from the nth input and all the weights connected to the kth hidden unit except the weight connecting the two of them as shown in Figure 1. Let the smaller network (with weights deleted) be

$$\begin{aligned} f(w,x) &= w_0 + \sum_{i=1}^{k-1} w_i \ \tanh(v_i' \cdot x' + v_{i0}) \\ &\quad + w_k \ \tanh(v_{kn} x_n) \\ &= g(w', x') + w_k \ \tanh(v_{kn} x_n) \end{aligned}$$

where $w' = (v_{10}, \ldots, v_{k-1,n-1}, w_0, \ldots, w_{k-1})$, $v_i' = (v_{i1}, \ldots, v_{i,n-1})$ and $x' = (x_1, \ldots, x_{n-1})$. We will also fix w_k and v_{kn} to be constants.

Now, $g(w', x')$ is a network with $n-1$ inputs and $k-1$ hidden units (the network within the box in Figure 1). We have shown that for almost all sets of $(k-1)(n-1)$ real inputs, the rank of the matrix (1) is at least $(k-1)(n-1)$ as long as no two hidden units have the same set of weights. If we choose the outputs for $g(w', \cdot)$ in the range of $h : x_n \mapsto -w_k \tanh(v_{kn} x_n)$, the inputs in $g(w', \cdot)$, can be made to belong to the boundary by choosing x_n appropriately. Hence inputs such that the matrix in Lemma 9 have rank $(k-1)(n-1)$ exist (almost all sets of $n-1$ dimensional inputs of size $(k-1)(n-1)$ that map $g(w', \cdot)$ to the range of $h : x_n \mapsto -w_k \tanh(v_{kn} x_n)$, and the corresponding x_n's will have this property). Lemma 9 then gives the desired result. □

Acknowledgements

We would like to thank Adam Kowalczyk, Yossi Erlich and Dan Chazan for pointing out the error.

References

[1] F. Albertini, E. D. Sontag, and V. Maillot. Uniqueness of weights for neural networks. In R. Mammone, editor, *Artificial Neural Networks for Speech and Vision*, pages 115–125. Chapman and Hall, London, 1993.

[2] P. L. Bartlett and S. Dasgupta. Exponential convergence of a gradient descent algorithm for a class of recurrent neural networks. In *Proceedings of the 38th Midwest Symposium on Circuits and Systems*, 1995.

[3] W. S. Lee, P. L. Bartlett, and R. C. Williamson. Lower bounds on the VC-dimension of smoothly parametrized function classes. *Neural Computaion*, 7(5):1040–1053, 1995.