Communications in Applied Analysis **10** (2006), no. 1, 19–27

AN IMPROVED BOUND FOR LEARNING LIPSCHITZ FUNCTIONS

Duane A. Cooper
Department of Mathematics
Morehouse College
Atlanta, GA 30314, USA

Communicated by D. Bainov

ABSTRACT: Revisited here is the problem of learning a nonlinear mapping with uncountable domain and range. The learning model used is that of piecewise linear interpolation on random samples from the domain. More specifically, a network *learns* a function by approximating its value, typically within some small ϵ, when presented an arbitrary element of the domain. For *reliable* learning, the network should accurately return the function's value with high probability, typically higher than $1-\delta$ for some small δ.

With new analytic results, we show that, given ϵ and δ and an arbitrary Lipschitz function $f : [0,1]^k \to \mathbb{R}$,

$$m \geq (2M\sqrt{k})^k \cdot (\frac{1}{\epsilon})^k \cdot (k \ln 2M\sqrt{k} + k \ln \frac{1}{\epsilon} + \ln \frac{1}{\delta})$$

samples from the uniform distribution on $[0,1]^k$ are sufficient to reliably learn f. The Delaunay triangulation technique is applied in order to keep simplex sizes small. The sufficient condition is an improvement of approximately $(\frac{2}{3})^k$ over the previously known upper bound.

AMS (MOS) Subject Classification. 68T05, 26A16

1. INTRODUCTION

We investigate in this article a network's ability to learn nonlinear mappings having both uncountable domain and range, as arises in robotics and other applications. A network *learns* a function by approximating its value, typically within some small ϵ, when presented an arbitrary element of the domain Poggio and Girosi [7]. For *reliable* learning, the network should accurately return the function's value with high probability, typically higher than $1-\delta$, for some small δ. The Probably Approximately Correct (PAC) model of learning Cucker and Smale [3], Valiant [9] is a standard of

Received May 19, 2005

analysis utilized by learning theorists; part of the PAC model defines learning with the aforementioned parameters ϵ and δ.

By necessity, analysis of learning over uncountable domain and range is restricted to consideration of specific classes of functions. Attempting to learn totally arbitrary functions by approximation would be futile since huge changes in function values could occur over tiny changes in the domain. Restricting the analysis to the class of Lipschitz functions is attractive in that the Lipschitz condition, which merely bounds the rate of change, is perhaps the weakest restriction we can impose to determine significant conditions and bounds on the capabilities of a network to learn functions Williamson and Paice [10].

Beyond this problem's mathematical aspects, statistical questions arise from the random sampling of the function which is part of the learning process. Furthermore, computing concerns must also be addressed since a desired property of neural models is computational efficiency.

2. WORST-CASE LIPSCHITZ FUNCTIONS

Consider the problem of learning functions over $\mathbb{R}^k$ by approximating in a piecewise linear fashion over a set of random input-output pairs. In particular, consider the class of Lipschitz functions, whose members have bounds on their rates of change.

Function $f : \mathbb{R}^k \to \mathbb{R}^l$ satisfies the *Lipschitz condition* with *Lipschitz constant* $M > 0$ if for every $\mathbf{x}, \mathbf{y} \in \mathbb{R}^k$ we have $|f(\mathbf{x}) - f(\mathbf{y})| \leq M \cdot |\mathbf{x} - \mathbf{y}|$. Learning occurs in the following sense: for some arbitrary member f of the class of Lipschitz functions, a "training period" occurs, in which a number of samples from the domain are taken and the values of f are given or ascertained. From this sampling, we construct an approximation $\tilde{f}$ which, it is hoped, reliably returns values sufficiently close to the actual value of f. In this work, $\tilde{f}$ will be developed as a piecewise linear interpolation over simplices of $k + 1$ points, determined by a triangulation of the samples taken in the domain.

In the following work, consider real-valued functions $f : \mathbb{R}^k \to \mathbb{R}$. These functions lend themselves more easily to analysis, and the learning problem is essentially the same. For instance, we can learn a function into $\mathbb{R}^l$ as an approximation within ϵ by approximating the function within $\epsilon/\sqrt{l}$ in each of the l dimensions of the range.

Denote by S some closed simplex having vertices $\mathbf{s_0}, \ldots, \mathbf{s_k} \in \mathbb{R}^k$. It turns out that the graph over S of any Lipschitz function f must lie entirely within the intersection of the spaces outside of the $k + 1$ right circular cones of slope M in $\mathbb{R}^{k+1} = \mathbb{R}^k \times \mathbb{R}^1$ having vertices $(\mathbf{s_0}, f(\mathbf{s_0})), \ldots, (\mathbf{s_k}, f(\mathbf{s_k}))$. These cones consist of all points $(\mathbf{x}, y)$ which satisfy

$$(y - f(\mathbf{s_j}))^2 = M^2 \cdot |\mathbf{x} - \mathbf{s_j}|^2,$$

for $j = 0, \ldots, k$. Thus we want to show, for $\mathbf{x} \in S$, that a Lipschitz f must satisfy

$$(f(\mathbf{x}) - f(\mathbf{s_j}))^2 \leq M^2 \cdot |\mathbf{x} - \mathbf{s_j}|^2,$$

for $j = 0, \ldots, k$. We establish this known result below in a manner consistent for our purposes in the following lemma, showing the lower and upper boundaries of this region to be worst-case functions.

Lemma 1. *Let S be a simplex in $\mathbb{R}^k$, and let the values $f(\mathbf{s_0}), \ldots, f(\mathbf{s_k})$ be known at the simplex vertices for Lipschitz function f. Then f is bounded entirely on S – below and above, respectively – by the worst-case functions*

$$f_*(\mathbf{x}) = \max_j (f(\mathbf{s_j}) - M \cdot |\mathbf{x} - \mathbf{s_j}|)$$

and

$$f^*(\mathbf{x}) = \min_j (f(\mathbf{s_j}) + M \cdot |\mathbf{x} - \mathbf{s_j}|).$$

Proof. Without loss of generality, we consider f_*. To establish f_* as a worst-case Lipschitz function, we need to establish (i) that f_* indeed satisfies the Lipschitz condition, and (ii) that f cannot assume any value below the surface of f_*, lest the Lipschitz condition be violated.

Let $\mathbf{x}, \mathbf{y} \in S$. Assume, w.l.o.g., that $f_*(\mathbf{x}) \geq f_*(\mathbf{y})$. Then:

$$\begin{aligned} |f_*(\mathbf{x}) - f_*(\mathbf{y})| &= |\max_j (f(\mathbf{s_j}) - M \cdot |\mathbf{x} - \mathbf{s_j}|) - \max_j (f(\mathbf{s_j}) - M \cdot |\mathbf{y} - \mathbf{s_j}|)| \\ &= (f(\mathbf{s_{j_0}}) - M \cdot |\mathbf{x} - \mathbf{s_{j_0}}|) - \max_j (f(\mathbf{s_j}) - M \cdot |\mathbf{y} - \mathbf{s_j}|) \\ &\quad \text{(for some } j_0 \in \{0, \ldots, k\}) \\ &\leq (f(\mathbf{s_{j_0}}) - M \cdot |\mathbf{x} - \mathbf{s_{j_0}}|) - (f(\mathbf{s_{j_0}}) - M \cdot |\mathbf{y} - \mathbf{s_{j_0}}|) \\ &\leq M \cdot |\mathbf{x} - \mathbf{y}|. \end{aligned}$$

Hence, f_* indeed satisfies the Lipschitz condition.

For the second part of the proof, assume $\exists \mathbf{x_0} \in S$ such that $f(\mathbf{x_0}) < f_*(\mathbf{x_0})$. Then, for at least one $j_0 \in \{0, \ldots, k\}$, $f(\mathbf{x_0}) < f(\mathbf{s_{j_0}}) - M \cdot |\mathbf{x_0} - \mathbf{s_{j_0}}|$. Therefore,

$$\begin{aligned} |f(\mathbf{x_0}) - f(\mathbf{s_{j_0}})| &= f(\mathbf{s_{j_0}}) - f(\mathbf{x_0}) \\ &> M \cdot |\mathbf{x_0} - \mathbf{s_{j_0}}|, \end{aligned}$$

violating the Lipschitz condition. □

3. A WORST-CASE ERROR BOUND ON A SIMPLEX

Next we establish a bound on the error that can result from linearly interpolating over a simplex.

Lemma 2. *Let S be a simplex in $\mathbb{R}^k$, and let the values $f(\mathbf{s_0}), \ldots, f(\mathbf{s_k})$ be known at the simplex vertices for Lipschitz function f. Let $\tilde{f} : \mathbb{R}^k \to \mathbb{R}$ be the affine function whose graph is the hyperplane through the points $(\mathbf{s_0}, f(\mathbf{s_0})), \ldots, (\mathbf{s_k}, f(\mathbf{s_k}))$. Then $\forall \mathbf{x} \in S$,*

$$|f_*(\mathbf{x}) - \tilde{f}(\mathbf{x})| \leq 2MR$$

and, similarly,

$$|f^*(\mathbf{x}) - \tilde{f}(\mathbf{x})| \leq 2MR,$$

where M is the Lipschitz constant and R is the radius of the sphere in $\mathbb{R}^k$ through $\mathbf{s_0}, \ldots, \mathbf{s_k}$.

Proof. Since relabeling and translation of our simplex will not change the essence of this problem, assume that $f(\mathbf{s_0}) \leq f(\mathbf{s_j})$, for $j = 1, \ldots, k$, and that $(\mathbf{s_0}, f(\mathbf{s_0})) = (\mathbf{0}, 0)$.

The distance between $f_*(\mathbf{x})$ and $\tilde{f}(\mathbf{x})$ is greatest when $(\mathbf{x}, f_*(\mathbf{x}))$ occurs at the lower point of intersection of the aforementioned cones satisfying

$$(y - f(\mathbf{s_j}))^2 = M^2 \cdot |\mathbf{x} - \mathbf{s_j}|^2.$$

Otherwise, for any $\mathbf{x} \in S$, the gradient of f_* at a point $(\mathbf{x}, f_*(\mathbf{x}))$ is M, following the path of one of the cones, whereas the gradient of $\tilde{f}$ at a point $(\mathbf{x}, \tilde{f}(\mathbf{x}))$ is less than M by virtue of $\tilde{f}(\mathbf{x})$ being determined as the linear interpolant of the values of Lipschitz function f on S. Thus, in this case where $(\mathbf{x}, f_*(\mathbf{x}))$ does not occur at the point of intersection, it would be possible to move $\mathbf{x}$ in a direction that increases the distance between $f_*(\mathbf{x})$ and $\tilde{f}(\mathbf{x})$.

Now consider worst-case function f_*. For $\mathbf{x} \in S$, $\mathbf{x}$ must lie within R of at least one of the vertices of S, where R is the radius of the circumsphere of S. In particular here, since the minimum value of Lipschitz f on the simplex S occurs at $\mathbf{s_0} = \mathbf{0}$, $\mathbf{x}$ must lie within R of $\mathbf{0}$. Therefore, $f_*(\mathbf{x}) \geq -M \cdot R$ and $\tilde{f}(\mathbf{x}) \leq M \cdot R$. Consequently, $|f_*(\mathbf{x}) - \tilde{f}(\mathbf{x})| \leq 2MR$.

Similarly, $|f^*(\mathbf{x}) - \tilde{f}(\mathbf{x})| \leq 2MR$. □

4. A SUFFICIENT NUMBER OF SAMPLES FOR RELIABLE LIPSCHITZ FUNCTION LEARNING

Now, in order to establish meaningful results, we restrict consideration of our function learning problem to compact domains. We will work in the unit cube in $\mathbb{R}^k$.

Suppose we wish a system to have the ability to learn, by piecewise linear interpolation, an arbitrary Lipschitz function $f : [0,1]^k \to \mathbb{R}$. We can achieve this by the following deterministic procedure.

Tesselate. $[0,1]^k$ by small cubes of side $s = \epsilon/(2M\sqrt{k})$, and *select* one point—a *sample* point—from each small cube.

Construct. simplices by the Delaunay triangulation method. The Delaunay method yields simplices with the property that the circumsphere (the *Delaunay sphere*) of any simplex contains none of the sample points in its interior; such a triangulation always exists Omohundro [5], Preparata and Shamos [8].

Determine. $\tilde{f}$ by linearly interpolating between values of f over the Delaunay simplices. The property that Delaunay spheres contain no other samples keeps the radii of these circumspheres small; in the tesselation described here, the radius of any Delaunay sphere satisfies $R < \epsilon/2M$. Hence (by Lemma 2.), piecewise linear interpolation over the Delaunay simplices guarantees approximation error $|f - \tilde{f}| \leq 2MR < 2M(\frac{\epsilon}{2M}) = \epsilon$.

This suffices as a deterministic method, but PAC-learning calls for random sampling on the domain. Thus, instead, we need to determine a number m of random sample points from the uniform distribution on $[0,1]^k$ such that, with high probability, every cube of the tesselation is likely to contain a sample. Then we proceed as in the deterministic case, constructing Delaunay simplices from the samples and linearly interpolating over them.

Theorem 3. *Let f be a Lipschitz function with Lipschitz constant M, $f : [0,1]^k \to \mathbb{R}$. Given ϵ and δ ($\epsilon > 0$, $0 < \delta < 1$), if*

$$m \geq (2M\sqrt{k})^k \cdot (\frac{1}{\epsilon})^k \cdot (k \ln(2M\sqrt{k}) + k \ln \frac{1}{\epsilon} + \ln \frac{1}{\delta})$$

samples are taken from the uniform distribution on $[0,1]^k$, then $P[|f - \tilde{f}| < \epsilon] > 1 - \delta$, where $\tilde{f}$, defined on the convex hull of the sample points, is the piecewise linear approximation over simplices constructed by the Delaunay triangulation method.

Proof. Tesselate $[0,1]^k$ by cubes of side $s = \epsilon/(2M\sqrt{k})$; let $n = (\frac{1}{s})^k$, the number of small cubes in the tesselation. Choose m samples from the uniform distribution on $[0,1]^k$. If each small cube contains at least one sample point, then any Delaunay sphere will have radius $R < \epsilon/2M$ and, by Lemma 2., piecewise linear interpolation over Delaunay simplices will guarantee approximation error $|f - \tilde{f}| \leq 2MR < 2M(\frac{\epsilon}{2M}) = \epsilon$.

Denote by E the event that, after m samples, at least one cube is empty, and denote by E_i the event that a specific cube i is empty. Then

$$\begin{aligned} P[E] &\leq \sum_{i=1}^{n} P[E_i] \\ &= n \cdot P[E_1] \\ &= n(1 - \frac{1}{n})^m \\ &< n(e^{-1/n})^m. \end{aligned}$$

Here, let us write $m = n \ln n + nb$, where $b \in \mathbb{R}$. Now

$$\begin{aligned} P[E] &< n(e^{-1/n})^m \\ &= n(e^{-1/n})^{n \ln n + nb} \\ &= n(e^{-\ln n - b}) \\ &= n(\frac{1}{n})(e^{-b}) \\ &= e^{-b}. \end{aligned}$$

We have $e^{-b} \leq \delta$ whenever $-b \leq \ln \delta$, that is, whenever $b \geq \ln \frac{1}{\delta}$. So the probability that some cube is unsampled after m samples of the unit cube is small—$P[E] < \delta$—whenever $b \geq \ln \frac{1}{\delta}$.

The probability that the interpolation over our chosen simplices is a "bad" approximation is at most the probability that some cube is unsampled, so $P[|f - \tilde{f}| \geq \epsilon] \leq P[E] < \delta$ (and equivalently $P[|f - \tilde{f}| < \epsilon] > 1 - \delta$) whenever $b \geq \ln \frac{1}{\delta}$, namely whenever the number of samples is

$$\begin{aligned} m &= n \ln n + nb \\ &= n(\ln n + b) \\ &\geq (\frac{1}{s})^k (\ln(\frac{1}{s})^k + \ln \frac{1}{\delta}) \\ &= (\frac{2M\sqrt{k}}{\epsilon})^k (\ln(\frac{2M\sqrt{k}}{\epsilon})^k + \ln \frac{1}{\delta}) \\ &= (2M\sqrt{k})^k \cdot (\frac{1}{\epsilon})^k \cdot (k \ln(2M\sqrt{k}) + k \ln \frac{1}{\epsilon} + \ln \frac{1}{\delta}). \qquad \square \end{aligned}$$

Similarly, we establish a result for functions over the more general cube $[0, \rho]^k$ as follows.

Corollary 4. *Let f be Lipschitz with constant M, $f : [0, 1]^k \to \mathbb{R}$. If*

$$m \geq (2\rho M\sqrt{k})^k \cdot (\frac{1}{\epsilon})^k \cdot (k \ln 2\rho M\sqrt{k} + k \ln \frac{1}{\epsilon} + \ln \frac{1}{\delta})$$

samples are taken from the uniform distribution on $[0, \rho]^k$, then $P[|f - \tilde{f}| < \epsilon] > 1 - \delta$.

The proof of the corollary follows the proof of the theorem, except here we need $n = (\rho/s)^k$ small cubes to tesselate our new domain.

Note for the theorem and its corollary that the values k, M, and ρ are all predetermined quantities for any given application. Hence, the sufficient number of samples m is polynomial in $1/\epsilon$ and $1/\delta$; in particular,

$$m = O((\frac{1}{\epsilon})^k \cdot (\ln \frac{1}{\epsilon} + \ln \frac{1}{\delta})). \tag{1}$$

Thus, we say that Lipschitz functions are *efficiently* or *feasibly learnable* Natarajan [4], Valiant [9].

The number of samples sufficient for reliable learning is improved roughly by a factor of $(\frac{2}{3})^k$ from that proved in Cooper [2]. This improvement is based on the stronger bound of $2MR$, rather than $3MR$, in Lemma 2.

5. USE OF THE DELAUNAY TRIANGULATION METHOD

The preceding results were inspired by the related work on function learning by Steve Omohundro in Omohundro [5], in which consideration is restricted to the class of functions with bounded second derivatives. Omohundro's central result in Omohundro [5] is the proof that, over the class of real-valued functions with partial second derivatives bounded by $2C$, Delaunay triangulation of the known sample points in the domain yields the piecewise linear approximation with the smallest maximum error at each point over the functions of that class. This follows from the determination that, for a simplex S with vertices $\mathbf{s_0}, \ldots, \mathbf{s_k}$, the worst-case functions – those for which linear interpolation has greatest error – are the paraboloids in $\mathbb{R}^{k+1}$ through the points $(\mathbf{s_0}, f(\mathbf{s_0})), \ldots, (\mathbf{s_k}, f(\mathbf{s_k}))$.

Although the analogous proposition is not true for the broader class of Lipschitz functions, as demonstrated in Cooper [2], the Delaunay method remains a satisfactory means of determining simplices from the domain samples because it does not allow for the long, narrow simplices which could be troublesome. And aside from providing the desired small simplices, the Delaunay triangulation is attractive computationally as our method of partitioning the input space. Because the Delaunay simplex in which a domain element lies can be determined locally in logarithmic expected time Omohundro [6], it is not necessary to store the entirety of the piecewise linear approximation function $\tilde{f}$ in memory; instead, the appropriate piece can be determined as necessary.

6. A NECESSARY NUMBER OF SAMPLES FOR RELIABLE LIPSCHITZ FUNCTION LEARNING

Theorem 3 establishes an upper bound on the number of random samples with which we can reliably learn any Lipschitz function $f : [0,1]^k \to \mathbb{R}$ by piecewise linear interpolation. In addition to this upper bound, we establish a lower bound on the number of random samples required to reliably learn any such function (see Cooper [2]).

We accomplish this by determining a bound on m for which, if fewer than m random samples are selected, we can show the existence of a Lipschitz function f and some $\mathbf{x_0}$ in its domain such that we cannot reliably learn by piecewise linear approximation on the samples. Specifically, the result states that for a set of m sample points taken from the uniform distribution on $[0,1]^k$ and for given M, ϵ, and δ with $M > 0$, $0 < \epsilon < M/2$, and $0 < \delta < 1$, if

$$m < (\frac{M}{2}\sqrt{k})^k \cdot (\frac{1}{\epsilon})^k \cdot \ln \frac{1}{\delta},$$

then there exists a Lipschitz function f with Lipschitz constant M which cannot be reliably learned in $[\frac{\epsilon}{M}, 1-\frac{\epsilon}{M}]^k$ by any piecewise linear approximation $\tilde{f}$ over simplices on the sample points because, for some $\mathbf{x_0} \in [\frac{\epsilon}{M}, 1-\frac{\epsilon}{M}]^k$, $P[|f(\mathbf{x_0}) - \tilde{f}(\mathbf{x_0})| \geq \epsilon] \geq \delta$.

7. LEARNING ON RELATED FUNCTION CLASSES

Results like those presented here are established in Cooper [1] for learning C^2 functions and Hölder functions on compact subsets of $\mathbb{R}^k$.

Learning C^2 functions is efficient, as is the case with Lipschitz functions, as a polynomial in $1/\epsilon$ and $1/\delta$, with $O((\frac{1}{\epsilon})^{k/2} \cdot (\ln \frac{1}{\epsilon} + \ln \frac{1}{\delta}))$ samples sufficient and $\Omega((\frac{1}{\epsilon})^{k/2} \cdot \ln \frac{1}{\delta})$ samples necessary for reliable learning.

Similarly, Hölder functions of order α can be learned with sampling polynomial in $1/\epsilon$ and $1/\delta$. Here, $O((\frac{1}{\epsilon})^{k/\alpha} \cdot (\ln \frac{1}{\epsilon} + \ln \frac{1}{\delta}))$ samples are sufficient and $\Omega((\frac{1}{\epsilon})^{k/\alpha} \cdot \ln \frac{1}{\delta})$ samples are necessary for reliable learning. However, when the order α is near zero, these bounds become troublesome.

REFERENCES

[1] D.A. Cooper, *Probably Approximately Correct Learning on the Class of Lipschitz Functions.* PhD thesis, University of California, Berkeley, 1993.

[2] D.A. Cooper, Learning Lipschitz functions, *International Journal of Computer Mathematics*, **59** (1995), 15-26.

[3] F. Cucker and S. Smale, On the mathematical foundations of learning, *Bulletin of the American Mathematical Society*, **39** (2002), 1-49.

[4] B.K. Natarajan, On learning sets and functions, *Machine Learning*, **4** (1989), 67-97.

[5] S.M. Omohundro, The Delaunay triangulation and function learning, *Technical Report* TR-90-001, International Computer Science Institute (1990).

[6] S.M. Omohundro, Geometric learning algorithms, *Physica D*, **42** (1990), 307-321.

[7] T. Poggio and F. Girosi, Networks for approximation and learning, *Proceedings of the IEEE*, **78** (1990), 1481-1497.

[8] F.P. Preparata and M.I. Shamos, *Computational Geometry: An Introduction*, New York, Springer-Verlag, 1985.

[9] L.G. Valiant, A theory of the learnable, *Communications of the ACM*, **27** (1984), 1134-1142.

[10] R.C. Williamson and A.D.B. Paice, The number of nodes required in a feed-forward neural network for functional representation, Department of Systems Engineering, The Australian National University, Canberra (1990).