International Journal of Pure and Applied Mathematics

Volume 32 No. 1 2006, 71-82

LEARNING C^2 AND HÖLDER FUNCTIONS

Duane A. Cooper

Department of Mathematics
Morehouse College
830 Westview Drive SW, Atlanta, GA 30314, USA
e-mail: dcooper@morehouse.edu

Abstract: Considered here is the problem of learning nonlinear mappings drawn from certain classes of functions with uncountable domain and range. The learning model used is that of piecewise linear interpolation on random samples from the domain. In more detail, 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.

The primary results of this article are the derivations of bounds showing that, given ϵ and δ and arbitrary C^2 function $f : [0,1]^k \to \mathbb{R}$,

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

samples from the uniform distribution on $[0,1]^k$ are sufficient to reliably learn f, and that

$$m \geq (\frac{Ck}{4})^{k/2} \cdot (\frac{1}{\epsilon})^{k/2} \cdot \ln \frac{1}{\delta}$$

samples are necessary for reliable learning.

Furthermore, given ϵ and δ and arbitrary Hölder function $f : [0,1]^k \to \mathbb{R}$,

$$m \geq (3C)^{k/\alpha} \cdot k^{k/2} \cdot (\frac{1}{\epsilon})^{k/\alpha} \cdot (\frac{k}{\alpha} \ln(3C) + \frac{k}{2} \ln k + \frac{k}{\alpha} \ln \frac{1}{\epsilon} + \ln \frac{1}{\delta})$$

samples from the uniform distribution on $[0,1]^k$ are sufficient to reliably learn f, and

Received: August 24, 2006

$$m \geq (\frac{C^{1/\alpha}\sqrt{k}}{2})^k \cdot (\frac{1}{\epsilon})^{k/\alpha} \cdot \ln \frac{1}{\delta}$$

samples are necessary for reliable learning.

AMS Subject Classification: 68T05, 26B35
Key Words: function learning, PAC learning, C^2 functions, Hölder functions

1. Introduction

When studying neural networks, it is often important to analyze a system's ability to reliably learn functions. In this article we focus on a network's ability to reliably learn nonlinear mappings having both uncountable domain and range. Such mappings arise in applications involving robotics, machine vision, speech processing, and graphics. For instance, a robot arm learning task can be modeled using a function from the robot's "kinematic space" into its "visual space", i.e., from a closed subset of $\mathbb{R}^3$ into $\mathbb{R}^6$ [4].

A network *learns* a function by approximating its value, typically within some small ϵ, when presented an arbitrary element of the domain [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 [1], [5] is a standard of analysis utilized by learning theorists; part of the PAC model defines learning with the aforementioned parameters ϵ and δ.

Mathematical analysis of learning over uncountable domain and range is necessarily restricted to consideration of particular classes of functions. Attempting to learn, by approximation, totally arbitrary functions would be futile since huge changes in function values could occur over tiny changes in the domain. In this article, as in [2], we restrict our analysis to classes of functions on which such changes in function values are constrained.

To "learn" a function f, we will construct an approximation $\tilde{f}$ in a piecewise linear fashion over a set of random samples of the domain taken during a "training" period. Though we employ random sampling, other researchers — notably, Zabinsky et al [8] — seek to determine optimal locations for the training samples. In this article we restrict our function learning problem to compact domains, in particular to the unit cube in $\mathbb{R}^k$, in order to obtain meaningful results. The approximation $\tilde{f}$ is developed as a piecewise linear interpolation over simplices of $k + 1$ points, determined by Delaunay triangulation of the samples taken in the domain.

Furthermore, our analysis considers functions into $\mathbb{R}$, which lend themselves more easily to analysis. The learning problem is essentially the same as for functions into $\mathbb{R}^l$. For instance, we can learn a function $f : \mathbb{R}^k \to \mathbb{R}^l$ as an approximation within ϵ by approximating f within $\epsilon/\sqrt{l}$ in each of the l dimensions of the codomain.

2. Learning C^2 Functions

First, we consider C^2 function learning, in which consideration is restricted to the class of functions with bounded second derivatives.

In particular, consider a real-valued function $f \in C^2[0,1]^k$, the class of continuous functions with continuous second derivatives on $[0,1]^k$. The partial second derivatives are necessarily bounded on the compact domain, say by $\pm 2C$. Omohundro [6] discusses how f can be approximated by linearly interpolating over a triangulation of sample points from $[0,1]^k$ and shows that the Delaunay triangulation always produces the approximation $\tilde{f}$ which is best in worst-case analysis.

For a simplex $\mathcal{S}$ in $\mathbb{R}^k$ with circumsphere of radius R, he proves that for the worst-case functions, the error on the linear interpolation is at most CR^2. He proceeds to show that, when every Delaunay sphere over the sample points has radius smaller than $\sqrt{\epsilon/C}$, the absolute error $|f(x) - \tilde{f}(x)| < \epsilon$ for every point x within the triangulation.

We can now offer the following extension, stating and proving a sufficient number of samples for reliable function learning in this case.

Theorem 1. *Let $f \in C^2[0,1]^k$; denote by $2C$ a bound on the partial second derivatives of f. Given ϵ and δ ($\epsilon > 0$, $0 < \delta < 1$), if*

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

or more 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 $s = \sqrt{\epsilon/(kC)}$; 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 $R < \sqrt{\epsilon/C}$ and piecewise linear

interpolation over Delaunay simplices will guarantee approximation error $|f - \tilde{f}| \leq CR^2 < C(\sqrt{\frac{\epsilon}{C}})^2 = \epsilon$.

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

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

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

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

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 Delaunay method yields 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

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

A similar result exists for functions over the more general cube $[0, \rho]^k$ as follows.

Corollary 2. *Let $f \in C^2[0,1]^k$. Denote by $2C$ a bound on the partial second derivatives of f. If*

$$m = (\rho\sqrt{kC})^k \cdot (\frac{1}{\epsilon})^{k/2} \cdot (k \ln(\rho\sqrt{kC}) + \frac{k}{2} \ln \frac{1}{\epsilon} + \ln \frac{1}{\delta})$$

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.

Here, as with Lipschitz functions, learning is efficient since

$$m = O((\frac{1}{\epsilon})^{k/2} \cdot (\ln \frac{1}{\epsilon} + \ln \frac{1}{\delta})),$$

polynomial in $1/\epsilon$ and $1/\delta$.

3. Learning Hölder Functions

Having examined Lipschitz function learning in [2], it makes sense to attempt to extend the previous results to learning on the class of Hölder functions.

A function $f : \mathbb{R}^k \to \mathbb{R}^l$ satisfies the *Hölder condition* with *Hölder constant* $C > 0$ and *order* α, $0 < \alpha \leq 1$, if for every $\mathbf{x}, \mathbf{y} \in \mathbb{R}^k$ $|f(\mathbf{x}) - f(\mathbf{y})| \leq C \cdot |\mathbf{x} - \mathbf{y}|^\alpha$. Hölder functions and Lipschitz functions are interrelated as follows. A Lipschitz function satisfies the Hölder condition with $\alpha = 1$. On the other hand, a Hölder function under a metric d satisfies the Lipschitz condition under the metric d^α, see [3].

3.1. Worst-Case Hölder Functions

As before, we wish to learn a function f by means of a piecewise linear approximation $\tilde{f}$. We again restrict our consideration to real-valued functions $f : \mathbb{R}^k \to \mathbb{R}$ to simplify the analysis, though the results can be extended to apply to functions into $\mathbb{R}^l$.

Let us denote by $\mathcal{S}$ some closed simplex having vertices $\mathbf{s_0}, \ldots, \mathbf{s_k} \in \mathbb{R}^k$. It happens that the graph over $\mathcal{S}$ of any Hölder function f must lie entirely within an intersection of the spaces bounded by $k+1$ surfaces in $\mathbb{R}^{k+1} = \mathbb{R}^k \times \mathbb{R}^1$ governed by the following: for $\mathbf{x} \in \mathcal{S}$, f must satisfy

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

for $j = 0, \ldots, k$. We establish this below by showing the lower and upper boundaries of this region to be worst-case functions.

Lemma 3. *Let $\mathcal{S}$ 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 Hölder function f. Then f is bounded entirely on $\mathcal{S}$—below and above, respectively—by the worst-case functions*

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

and

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

Proof. Without loss of generality, we consider f_*. To establish f_* as a worst-case Hölder function, we need to establish (i) that f_* indeed satisfies the

Hölder condition, and (ii) that f cannot assume any value below the surface of f_*, lest the Hölder condition be violated.

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

$$\begin{aligned}
|f_*(\mathbf{x}) - f_*(\mathbf{y})| &= |\max_j(f(\mathbf{s_j}) - C \cdot |\mathbf{x} - \mathbf{s_j}|^\alpha) - \max_j(f(\mathbf{s_j}) - C \cdot |\mathbf{y} - \mathbf{s_j}|^\alpha)| \\
&\leq (f(\mathbf{s_{j_0}}) - C \cdot |\mathbf{x} - \mathbf{s_{j_0}}|^\alpha) - (f(\mathbf{s_{j_0}}) - C \cdot |\mathbf{y} - \mathbf{s_{j_0}}|^\alpha) \\
&\qquad [\text{for some } j_0 \in \{0, \ldots, k\}] \\
&= C \cdot (|\mathbf{y} - \mathbf{s_{j_0}}|^\alpha - |\mathbf{x} - \mathbf{s_{j_0}}|^\alpha) \leq C \cdot ((|\mathbf{y} - \mathbf{x}| + |\mathbf{x} - \mathbf{s_{j_0}}|)^\alpha - |\mathbf{x} - \mathbf{s_{j_0}}|^\alpha) \\
&\leq C \cdot \left(\frac{|\mathbf{y} - \mathbf{x}| + |\mathbf{x} - \mathbf{s_{j_0}}|}{(|\mathbf{y} - \mathbf{x}| + |\mathbf{x} - \mathbf{s_{j_0}}|)^{1-\alpha}} - \frac{|\mathbf{x} - \mathbf{s_{j_0}}|}{(|\mathbf{y} - \mathbf{x}| + |\mathbf{x} - \mathbf{s_{j_0}}|)^{1-\alpha}}\right) \\
&= C \cdot \frac{|\mathbf{y} - \mathbf{x}|}{(|\mathbf{y} - \mathbf{x}| + |\mathbf{x} - \mathbf{s_{j_0}}|)^{1-\alpha}} \leq C \cdot \frac{|\mathbf{y} - \mathbf{x}|}{|\mathbf{y} - \mathbf{x}|^{1-\alpha}} = C \cdot |\mathbf{x} - \mathbf{y}|^\alpha.
\end{aligned}$$

Hence, f_* indeed satisfies the Hölder condition.

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

$$|f(\mathbf{x_0}) - f(\mathbf{s_{j_0}})| = f(\mathbf{s_{j_0}}) - f(\mathbf{x_0}) > C \cdot |\mathbf{x_0} - \mathbf{s_{j_0}}|^\alpha,$$

violating the Hölder condition. □

3.2. 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 4. *Let $\mathcal{S}$ 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 Hölder 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 $\forall \mathbf{x} \in \mathcal{S}$,*

$$|f_*(\mathbf{x}) - \tilde{f}(\mathbf{x})| \leq 3CR^\alpha$$

and, similarly,

$$|f^*(\mathbf{x}) - \tilde{f}(\mathbf{x})| \leq 3CR^\alpha,$$

where C is the Hölder constant, α is the order, 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, let us assume that $f(\mathbf{s_0}) \leq f(\mathbf{s_j})$, for $j = 1, \ldots, k$, and that $(\mathbf{s_0}, f(\mathbf{s_0})) = (\mathbf{0}, 0)$.

Consider first our linear interpolation $\tilde{f}$. For $\mathbf{x} \in \mathcal{S}$, the farthest $\mathbf{x}$ can be from $\mathbf{s_0}$ is the diameter $2R$ of the circumsphere of $\mathcal{S}$. Thus:

$$\tilde{f}(\mathbf{x}) \leq \max_j(f(\mathbf{s_j})) = f(\mathbf{s_{j_0}}) \quad [\text{for some } j_0] \leq C \cdot |\mathbf{s_{j_0}}|^\alpha \leq C \cdot (2R)^\alpha.$$

Now consider worst-case function f_*. For $\mathbf{x} \in \mathcal{S}$, $\mathbf{x}$ must lie within radius R of at least one of the vertices of $\mathcal{S}$. Thus:

$$\begin{aligned} f_*(\mathbf{x}) = \max_j(f(\mathbf{s_j}) - C \cdot |\mathbf{x} - \mathbf{s_j}|^\alpha) \geq f(\mathbf{s_{j_0}}) - C \cdot R^\alpha \quad [\text{where } |\mathbf{x} - \mathbf{s_{j_0}}| \leq R] \\ \geq 0 - C \cdot R^\alpha = -C \cdot R^\alpha. \end{aligned}$$

Consequently,

$$|f_*(\mathbf{x}) - \tilde{f}(\mathbf{x})| \leq C \cdot (2R)^\alpha + C \cdot R^\alpha = (1 + 2^\alpha) \cdot CR^\alpha \leq 3CR^\alpha.$$

A similar proof shows that $|f^*(\mathbf{x}) - \tilde{f}(\mathbf{x})| \leq 3CR^\alpha$. □

3.3. Sufficient Samples for Reliable Hölder Function Learning

Suppose we wish a system to have the ability to learn, by piecewise linear interpolation, an arbitrary Hölder function $f : [0,1]^k \to \mathbb{R}$. We need to determine a number m of random sample points from the uniform distribution on $[0,1]^k$ such that, with high probability, approximation error is small.

Theorem 5. *Let f be a Hölder function with Hölder constant C and order α, $f : [0,1]^k \to \mathbb{R}$. Given ϵ and δ ($\epsilon > 0$, $0 < \delta < 1$), if*

$$m \geq (3C)^{k/\alpha} \cdot k^{k/2} \cdot (\frac{1}{\epsilon})^{k/\alpha} \cdot (\frac{k}{\alpha} \ln(3C) + \frac{k}{2} \ln k + \frac{k}{\alpha} \ln \frac{1}{\epsilon} + \ln \frac{1}{\delta})$$

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 $s = (\frac{\epsilon}{3C})^{1/\alpha} \cdot \frac{1}{\sqrt{k}}$; 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 $R < (\epsilon/3C)^{1/\alpha}$ and, by Lemma 4, piecewise linear interpolation over Delaunay simplices will guarantee approximation error $|f - \tilde{f}| \leq 3CR^\alpha < 3C((\frac{\epsilon}{3C})^{1/\alpha})^\alpha = \epsilon$.

Denote by E the event that, after $m = n \ln n + nb$ samples, at least one cube is empty. By the same analysis of Theorem 1, we have $P[E] < e^{-b}$.

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

$$\begin{aligned} m = n \ln n + nb = n(\ln n + b) &\geq (\frac{1}{s})^k (\ln(\frac{1}{s})^k + \ln \frac{1}{\delta}) \\ &= ((\frac{3C}{\epsilon})^{1/\alpha} \sqrt{k})^k (\ln((\frac{3C}{\epsilon})^{1/\alpha} \sqrt{k})^k + \ln \frac{1}{\delta}) \\ &= (3C)^{k/\alpha} \cdot k^{k/2} \cdot (\frac{1}{\epsilon})^{k/\alpha} \cdot (\frac{k}{\alpha} \ln(3C) + \frac{k}{2} \ln k + \frac{k}{\alpha} \ln \frac{1}{\epsilon} + \ln \frac{1}{\delta}). \quad \square \end{aligned}$$

A similar result exists for functions over the more general cube $[0, \rho]^k$ as follows.

Corollary 6. *Let f be Hölder with constant C and order α, $f : [0,1]^k \to \mathbb{R}$. If*

$$m \geq (3\rho C)^{k/\alpha} \cdot k^{k/2} \cdot (\frac{1}{\epsilon})^{k/\alpha} \cdot (\frac{k}{\alpha} \ln(3\rho C) + \frac{k}{2} \ln k + \frac{k}{\alpha} \ln \frac{1}{\epsilon} + \ln \frac{1}{\delta})$$

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.

In this case,

$$m = O((\frac{1}{\epsilon})^{k/\alpha} \cdot (\ln \frac{1}{\epsilon} + \ln \frac{1}{\delta})),$$

so learning is still polynomial in $1/\epsilon$ and $1/\delta$. However, the number of samples becomes troublesome when the order of the Hölder function is near zero.

4. Necessary Samples for Reliable C^2 Function Learning

Theorem 1 establishes an upper bound on the number of random samples with which we can reliably learn any C^2 function $f : [0,1]^k \to \mathbb{R}$ by piecewise linear interpolation. We now seek to establish a lower bound on the number of random samples required to reliably learn any such function.

Lemma 7. *Consider a set of sample points from* $[0,1]^k$. *Suppose* $\exists \mathbf{x_0} \in [0,1]^k$ *such that, for every sample* $\mathbf{s}$, $|\mathbf{x_0} - \mathbf{s}| \geq \sqrt{\epsilon/C}$ $(\epsilon, C > 0)$. *Then, for the* C^2 *function defined by* $f(\mathbf{x}) = C \cdot |\mathbf{x} - \mathbf{x_0}|^2$, *whose graph is a paraboloid, we have* $|f(\mathbf{x_0}) - \tilde{f}(\mathbf{x_0})| \geq \epsilon$.

Proof. Denote by $\mathbf{s_0}$ the sample point nearest to $\mathbf{x_0}$. Because $\tilde{f}$ is piecewise linear and f increases with distance from $\mathbf{x_0}$, $\tilde{f}(\mathbf{x_0}) \geq f(\mathbf{s_0})$, so

$$|f(\mathbf{x_0}) - \tilde{f}(\mathbf{x_0})| = |0 - \tilde{f}(\mathbf{x_0})| = \tilde{f}(\mathbf{x_0}) \geq f(\mathbf{s_0})$$
$$= C \cdot |\mathbf{s_0} - \mathbf{x_0}|^2 \geq C \cdot (\sqrt{\frac{\epsilon}{C}})^2 = \epsilon. \quad \square$$

It remains to show that with too few random samples from $[0,1]^k$ such a point $\mathbf{x_0}$ will exist with probability δ or greater.

Theorem 8. *Consider a set of* m *sample points taken from the uniform distribution on* $[0,1]^k$. *Given* $C > 0$ *and given* ϵ *and* δ $(0 < \epsilon < C/4,\ 0 < \delta < 1)$, *if*

$$m < (\frac{Ck}{4})^{k/2} \cdot (\frac{1}{\epsilon})^{k/2} \cdot \ln \frac{1}{\delta},$$

then there exists $f \in C^2[0,1]^k$ *with a bound of* $2C$ *on its partial second derivatives which cannot be reliably learned in* $[\sqrt{\frac{\epsilon}{C}}, 1 - \sqrt{\frac{\epsilon}{C}}]^k$ *by any piecewise linear approximation* $\tilde{f}$ *over simplices on the sample points because, for some* $\mathbf{x_0} \in [\sqrt{\frac{\epsilon}{C}}, 1 - \sqrt{\frac{\epsilon}{C}}]^k$, $P[|f(\mathbf{x_0}) - \tilde{f}(\mathbf{x_0})| \geq \epsilon] \geq \delta$.

Proof. Let $\mathcal{R} = [\sqrt{\frac{\epsilon}{C}}, 1 - \sqrt{\frac{\epsilon}{C}}]^k$. For an arbitrary $\mathbf{x} \in \mathcal{R}$, the probability that none of the samples lie within $\sqrt{\epsilon/C}$ of $\mathbf{x}$ is $(1 - V_k(S_{\mathbf{x}, \sqrt{\frac{\epsilon}{C}}}))^m$. When this probability is δ or more, we can show (by Lemma 7) the existence of a C^2 function with partial second derivatives at most $2C$ for which $P[|f - \tilde{f}| \geq \epsilon] \geq \delta$. Solving for m establishes a lower bound:

$$(1 - V_k(S_{\mathbf{x}, \sqrt{\frac{\epsilon}{C}}}))^m \geq \delta \ln(1 - V_k(S_{\mathbf{x}, \sqrt{\frac{\epsilon}{C}}}))^m \geq \ln \delta \quad \text{[okay, since } \epsilon < C/4\text{]}$$
$$m \cdot \ln(1 - V_k(S_{\mathbf{x}, \sqrt{\frac{\epsilon}{C}}})) \geq \ln \delta m \leq \frac{\ln \delta}{\ln(1 - V_k(S_{\mathbf{x}, \sqrt{\frac{\epsilon}{C}}}))} m$$
$$< \frac{\ln \delta}{\ln(1 - (2\sqrt{\frac{\epsilon}{Ck}})^k)} m < \frac{\ln \delta}{\ln(1 - (\frac{4\epsilon}{Ck})^{k/2})}, \quad (1)$$

where $(2\sqrt{\frac{\epsilon}{Ck}})^k$ in inequality (1) is the volume of a cube inscribed in $S_{\mathbf{x}, \sqrt{\frac{\epsilon}{C}}}$.

Now,

$$\frac{\ln \delta}{\ln(1-(\frac{4\epsilon}{Ck})^{k/2})} > \frac{\ln \delta}{-(\frac{4\epsilon}{Ck})^{k/2}} = \frac{\ln \frac{1}{\delta}}{(\frac{4\epsilon}{Ck})^{k/2}} = (\frac{Ck}{4})^{k/2} \cdot (\frac{1}{\epsilon})^{k/2} \cdot \ln \frac{1}{\delta},$$

establishing $(\frac{Ck}{4})^{k/2} \cdot (\frac{1}{\epsilon})^{k/2} \cdot \ln \frac{1}{\delta}$ as a lower bound on the number of samples necessary to learn arbitrary C^2 functions over our domain. □

As for Lipschitz functions [2], the number of samples which are necessary for reliable C^2 function learning is polynomial in $1/\epsilon$ and $1/\delta$:

$$m = \Omega((\frac{1}{\epsilon})^{k/2} \cdot \ln \frac{1}{\delta}).$$

5. Necessary Samples for Reliable Hölder Function Learning

Theorem 5 establishes an upper bound on the number of random samples with which we can reliably learn any Hölder function $f : [0,1]^k \to \mathbb{R}$ by piecewise linear interpolation. We now seek to establish a lower bound on the number of random samples required to reliably learn any such function.

Lemma 9. *Consider a set of sample points from $[0,1]^k$. Suppose $\exists \mathbf{x_0} \in [0,1]^k$ such that, for every sample $\mathbf{s}$, $|\mathbf{x_0} - \mathbf{s}| \geq (\epsilon/C)^{1/\alpha}$ ($\epsilon, C > 0$, $0 < \alpha \leq 1$). Then, for the Hölder function defined by $f(\mathbf{x}) = C \cdot |\mathbf{x} - \mathbf{x_0}|^\alpha$, we have $|f(\mathbf{x_0}) - \tilde{f}(\mathbf{x_0})| \geq \epsilon$.*

Proof. Denote by $\mathbf{s_0}$ the sample point nearest to $\mathbf{x_0}$. Because $\tilde{f}$ is piecewise linear and f increases with distance from $\mathbf{x_0}$, $\tilde{f}(\mathbf{x_0}) \geq f(\mathbf{s_0})$, so

$$|f(\mathbf{x_0}) - \tilde{f}(\mathbf{x_0})| = |0 - \tilde{f}(\mathbf{x_0})| = \tilde{f}(\mathbf{x_0}) \geq f(\mathbf{s_0}) = C \cdot |\mathbf{s_0} - \mathbf{x_0}|^\alpha$$
$$\geq C \cdot ((\frac{\epsilon}{C})^{1/\alpha})^\alpha = \epsilon. \quad \square$$

It remains to show that with too few random samples from $[0,1]^k$ such a point $\mathbf{x_0}$ will exist with probability δ or greater.

Theorem 10. *Consider a set of m sample points taken from the uniform distribution on $[0,1]^k$. Given $C > 0$ and α ($0 < \alpha \leq 1$), and given ϵ and δ ($0 < \epsilon < C/2^\alpha$, $0 < \delta < 1$), if*

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

then there exists a Hölder function f with Hölder constant C and order α which cannot be reliably learned in $[(\frac{\epsilon}{C})^{1/\alpha}, 1-(\frac{\epsilon}{C})^{1/\alpha}]^k$ by any piecewise linear

approximation $\tilde{f}$ over simplices on the sample points because, for some $\mathbf{x_0} \in [(\frac{\epsilon}{C})^{1/\alpha}, 1-(\frac{\epsilon}{C})^{1/\alpha}]^k$, $P[|f(\mathbf{x_0}) - \tilde{f}(\mathbf{x_0})| \geq \epsilon] \geq \delta$.

Proof. Let $\mathcal{R} = [(\frac{\epsilon}{C})^{1/\alpha}, 1-(\frac{\epsilon}{C})^{1/\alpha}]^k$. For an arbitrary $\mathbf{x} \in \mathcal{R}$, the probability that none of the samples lie within $(\epsilon/C)^{1/\alpha}$ of $\mathbf{x}$ is $(1 - V_k(S_{\mathbf{x},(\frac{\epsilon}{C})^{1/\alpha}}))^m$. When this probability is δ or more, we can show (by Lemma 9) the existence of a Hölder function for which $P[|f - \tilde{f}| \geq \epsilon] \geq \delta$. Solving for m establishes a lower bound:

$$
\begin{aligned}
&(1 - V_k(S_{\mathbf{x},(\frac{\epsilon}{C})^{1/\alpha}}))^m \\
&\qquad \geq \delta \ln(1 - V_k(S_{\mathbf{x},(\frac{\epsilon}{C})^{1/\alpha}}))^m \geq \ln \delta \quad [\text{ okay, since } \epsilon < C/2^\alpha] \\
&\qquad m \cdot \ln(1 - V_k(S_{\mathbf{x},(\frac{\epsilon}{C})^{1/\alpha}})) \geq \ln \delta m \leq \frac{\ln \delta}{\ln(1 - V_k(S_{\mathbf{x},(\frac{\epsilon}{C})^{1/\alpha}}))} m \\
&\qquad\qquad < \frac{\ln \delta}{\ln(1 - (\frac{2(\epsilon/C)^{1/\alpha}}{\sqrt{k}})^k)} m < \frac{\ln \delta}{\ln(1 - (\frac{2}{\sqrt{k}})^k (\frac{\epsilon}{C})^{k/\alpha})}, \quad (2)
\end{aligned}
$$

where $(\frac{2(\epsilon/C)^{1/\alpha}}{\sqrt{k}})^k$ in inequality 2 is the volume of a cube inscribed in $S_{\mathbf{x},(\frac{\epsilon}{C})^{1/\alpha}}$.

Now,

$$
\begin{aligned}
\frac{\ln \delta}{\ln(1 - (\frac{2}{\sqrt{k}})^k (\frac{\epsilon}{C})^{k/\alpha})} > \frac{\ln \delta}{-(\frac{2}{\sqrt{k}})^k (\frac{\epsilon}{C})^{k/\alpha}} &= \frac{\ln \frac{1}{\delta}}{(\frac{2}{\sqrt{k}})^k (\frac{\epsilon}{C})^{k/\alpha}} \\
&= (\frac{C^{1/\alpha}\sqrt{k}}{2})^k \cdot (\frac{1}{\epsilon})^{k/\alpha} \cdot \ln \frac{1}{\delta},
\end{aligned}
$$

establishing $(\frac{C^{1/\alpha}\sqrt{k}}{2})^k \cdot (\frac{1}{\epsilon})^{k/\alpha} \cdot \ln \frac{1}{\delta}$ as a lower bound on the number of samples necessary to learn arbitrary Hölder functions over our domain. □

The number of samples which are necessary for reliable Hölder function learning is polynomial in $1/\epsilon$ and $1/\delta$:

$$m = \Omega((\frac{1}{\epsilon})^{k/\alpha} \cdot \ln \frac{1}{\delta}).$$

This bound becomes problematic when the order of the Hölder function is near zero.

References

[1] M. Anthony, N. Biggs, *Computational Learning Theory: An Introduction*, Cambridge University Press, Cambridge (1992).

[2] D.A. Cooper, An improved bound for learning Lipschitz functions, *Communications in Applied Analysis*, **10** (2006), 19-27.

[3] R.M. Dudley, *Real Analysis and Probability*, Wadsworth and Brooks/Cole, Pacific Grove, Calif. (1989).

[4] B.W. Mel, Connectionist robot motion planning: A neurally-inspired approach to guided reasoning, In: *Perspectives in Artificial Intelligence*, Academic Press, San Diego (1990).

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

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

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

[8] Z.B. Zabinsky, R.L. Smith, B.P. Kristinsdottir, Optimal estimation of univariate black-box Lipschitz functions with upper and lower error bounds, *Computers and Operations Research and their Application to Problems of World Concern*, **30** (2003), 1539-1553.