Sample Complexity versus Approximation Error

Peter L. Bartlett and Robert C. Williamson
Department of Systems Engineering
Research School of Information Sciences and Engineering
Australian National University, Canberra 0200 Australia

Abstract

We consider the problem of learning an unknown real-valued function from a sequence of values of the function at randomly chosen points when the estimates are constrained to some class of functions called the hypothesis class. Ideally, the hypothesis class should be able to approximate a wide variety of target functions, yet not need an excessive number of examples to ensure that a learning algorithm can choose a near-optimal approximation to the target function. We show that these two objectives are incompatible, in the sense that as the approximation error of the hypothesis class decreases, a lower bound on the sample complexity must increase. The approximation error is the largest distance between an element of the class of candidate target functions and its best approximation in the hypothesis class. The sample complexity is the number of examples necessary for a near-optimal estimate of any target function. The relationship between approximation error and sample complexity depends on a combinatorial dimension of the class of candidate target functions which are to be approximated. We give an example of this relationship for the case of Lipschitz spaces.

1 Introduction

Suppose we wish to design an algorithm to estimate a real-valued function, given a sequence of values of the function at randomly-chosen points. There are two sources of error for such an algorithm: approximation error and estimation error. We refer to the class of all functions that the algorithm might output as the hypothesis class. In general, the target function is not a member of this class; the error in approximating the target function is the distance to the closest function in the hypothesis class. We would expect that as we consider hypothesis classes that are more complex (in some sense), this approximation error would decrease.

Estimation error arises because the algorithm must choose a hypothesis on the basis of a finite data sequence, so it is unlikely to choose the best approximation to the target function. For a fixed length data sequence, we would expect the estimation error to increase as the complexity of the hypothesis class is increased. The amount of data necessary for a learning algorithm to reliably choose a near-optimal approximation is called the sample complexity.

Approximation error and sample complexity have been determined for specific hypothesis classes (see for example [3]). It appears that there is a general tradeoff between these two quantities—if we choose a powerful hypothesis class so that the approximation error is small, we will need a large data sequence to choose a near-optimal function from the hypothesis class. In this paper, we quantify this tradeoff.

The approximation error is defined as the worst case error over all functions in some class K of functions that we wish to approximate. Recent results [4] show that the sample complexity of a hypothesis class depends on a measure of its complexity known as its fat-shattering function. We relate the fat-shattering function of a hypothesis class to its approximation error in terms of properties of the class K of approximated functions.

This result is analogous to the Cramér-Rao bound in parameter estimation [5]; it shows that if we want to choose a hypothesis class to achieve a certain approximation error, there is an unavoidable statistical penalty. (The analogue is between the bias in the Cramér-Rao bound and the approximation error, and between variance and the estimation error.)

In Section 2, we give formal definitions of the learning problem, sample complexity, and approximation error. In Section 3, we present the tradeoff. We apply this result in Section 4 to specific target function classes, the homogeneous Lipschitz spaces. Section 5 contains some remarks and discussion about open problems.

2 Definitions and Notation

Throughout the paper, we use log to denote the logarithm to base two, $\mathbb{N}$ to denote the natural numbers, and let $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$.

Define $X = [0,1]^n$, let P be a probability distribution on X, and let t be a $[0,1]$-valued function defined on X. We consider the problem of estimating t from a sequence $((x_1, t(x_1)), \ldots, (x_m, t(x_m)))$, where the x_i are chosen according to P. The estimates are restricted to some class F of $[0,1]$-valued functions[1] defined on X which we call the hypothesis class. The aim is to choose a function f in F so that the expected absolute difference

[1] In this paper, all functions are assumed to be measurable, and the function class F satisfies the mild measurability condition given in [4].

between t and f is small. Define the **error** of f with respect to P and t as

$$\mathbf{er}_{P,t}(f) = \int_X |f(x) - t(x)| \, dP(x).$$

For $x \in X^m$, let $\mathrm{sam}(x,t) = ((x_1, t(x_1)), \ldots, (x_m, t(x_m)))$. A learning algorithm is a mapping from the space of samples of this form, $\cup_m (X \times [0,1])^m$, to F. The sample complexity of F is the number of examples necessary for some algorithm to choose a function from F with near-minimal error.

Definition 1 *Let m be a positive integer and $0 < \epsilon, \delta < 1$. We say that F is (ϵ, δ)-***learnable** *from m examples if there is an algorithm $A : (X \times [0,1])^m \to F$ such that, for all probability distributions P on X and all measurable functions $t : X \to [0,1]$, with probability $1 - \delta$ A chooses a function in F that has error within ϵ of the infimum over F,*

$$P^m \left\{ x \in X^m : \mathbf{er}_{P,t}(A(\mathrm{sam}(x,t))) > \inf_{f \in F} \mathbf{er}_{P,t}(f) + \epsilon \right\} < \delta.$$

The **sample complexity** *of (ϵ, δ)-learning with F is the smallest m for which F is (ϵ, δ)-learnable from m examples.*

Note that this definition of learning differs slightly from that in [4], since we make no assumptions about the target function t. Results in [4] imply that the sample complexity (as given in Definition 1) depends nearly linearly on a scale-sensitive dimension of F called the fat-shattering function.

Definition 2 (Fat-Shattering Function) *Let γ be a positive real number. A sequence $x \in X^m$ is γ-shattered by F if there is an $r = (r_1, \ldots, r_m)$ in $\mathbb{R}^m$ such that for all sign assignments $s = (s_1, \ldots, s_m)$ in $\{-1, 1\}^m$ there is a function f in F for which $s_i(f(x_i) - r_i) \geq \gamma$ for $i = 1, 2, \ldots, m$.*

The fat-shattering function fat_F *of F is defined by*

$$\mathrm{fat}_F(\gamma) = \max \{ m : \exists x \in X^m, \ x \text{ is } \gamma\text{-shattered by } F \}$$

when the maximum exists, otherwise we say $\mathrm{fat}_F(\gamma)$ *is infinite.*

The following theorem follows easily from the proof of Theorem 21 and Theorem 22 in [4].

Theorem 3 *Suppose F is a class of $[0,1]$-valued functions defined on X, and $0 < \epsilon, \delta < 1/8$, and let $m_0(\epsilon, \delta)$ be the sample complexity of (ϵ, δ)-learning with F. Then*

$$m_0(\epsilon, \delta) = \Omega\left(\frac{\mathrm{fat}_F(65\epsilon)}{\log 1/\epsilon} \right),$$

and

$$m_0(\epsilon, \delta) = O\left(\frac{\mathrm{fat}_F(\epsilon/16)}{\epsilon^2} \left(\log^2 \frac{\mathrm{fat}_F(\epsilon/16)}{\epsilon^4} + \log \frac{1}{\delta} \right) \right).$$

Approximation error

Suppose t is a $[0,1]$-valued function defined on X, P is a probability distribution on X, A is a learning algorithm, and $x \in X^m$. Then the error of the algorithm's hypothesis ($\mathbf{er}_{P,t}(A(\mathrm{sam}(x,t)))$) can be divided into two parts, the error of approximating t using a function from F ($\inf_{f\in F} \mathbf{er}_{P,t}(f)$), and the additional error that arises because the algorithm must estimate t from a finite data sample.

To define approximation error, we assume that t is a member of a set K of $[0,1]$-valued functions defined on X. Typically, K is a class of "well-behaved" functions that we wish to approximate — without such a restriction on t, the following worst case definition of approximation error would make no sense. Examples of function classes K include functions satisfying a Lipschitz or other smoothness constraint [10], or classes of functions whose Fourier transforms have bounded moments [3].

An obvious way to define how well F approximates K is in terms of the worst case error,

$$\sup_P \sup_{t\in K} \inf_{f\in F} \mathbf{er}_{P,t}(f),$$

where the first supremum is over all distributions P on X. Obviously, this provides an upper bound on the L_1 approximation error of F with respect to K defined by

$$D_1(K,F) = \sup_{t\in K} \inf_{f\in F} \int_X |f(x) - t(x)|\, dx.$$

More generally, we can define approximation error in L_p for $p = 1, 2, \ldots, \infty$ as follows.

Definition 4 *Let p be in $\mathbb{N} \cup \{\infty\}$. Let K and F be subsets of the space $L_p(X,[0,1])$. The* **approximation error in L_p of F with respect to K** *is defined as*

$$D_p(K,F) = \sup_{t\in K} \inf_{f\in F} \|t - f\|_p,$$

where

$$\|g\|_p = \begin{cases} \left(\int_X |g|^p\, dx\right)^{1/p} & p \in \mathbb{N} \\ \sup_{x\in X} g(x) & p = \infty \end{cases}$$

In the next section, we present a tradeoff between the approximation error $D_p(K,F)$ and the fat-shattering function of F.

3 Results

The following definition gives a measure of complexity of the approximated function class K that is useful when approximation error is measured using a p-norm and $p \in \mathbb{N}$.

Definition 5 (Setwise Fat-Shattering Function) *Let S be a bounded Lebesgue-measurable subset of $\mathbb{R}^n$, K a class of $[0,1]$-valued functions defined on X, and γ a positive real number. Let $b = (b_1, b_2, \ldots, b_m) \in X^m$, and define $b_i + S = \{b_i + a : a \in S\}$. The sequence b is* **S-setwise γ-shattered by K** *if*

1. $b_i + S \subset X$, *and*

2. *there is an* $r = (r_1, \ldots, r_m)$ *in* $\mathbb{R}^m$ *such that for all sign assignments* $s = (s_1, \ldots, s_m)$ *in* $\{-1, 1\}^m$ *there is a function* t *in* K *for which, for all* i *in* $\{1, 2, \ldots, m\}$ *and for all* x_i *in* $(b_i + S)$,

$$s_i(t(x_i) - r_i) \geq \gamma.$$

The **setwise fat-shattering function of** K **with set size** Δ *is*

$$\mathrm{fat}_{K,\Delta}(\gamma) = \sup_S \{m : \exists b \in X^m,\ b \text{ is } S\text{-setwise } \gamma\text{-shattered by } K\},$$

where the supremum is over all Lebesgue-measurable subsets S *of* $\mathbb{R}^n$ *satisfying* $\int_S dx = \Delta$.

The following theorem is the main result of the paper. It shows that there is a trade-off between the approximation error of F ($D_p(K, F)$) and fat_F, which determines the sample complexity of learning with F.

Theorem 6 *Let* K *be a set of* $[0, 1]$*-valued functions defined on* X.

a. *For all* $\gamma > D_\infty(K, F)$,

$$\mathrm{fat}_F(\gamma) \geq \mathrm{fat}_K(2\gamma).$$

b. *Let* $p \in \mathbb{N}$, $\Delta \in \mathbb{R}^+$. *For all* γ *satisfying* $D_p(K, F) \leq 2\Delta^{1/p}\gamma$,

$$\mathrm{fat}_F(\gamma) \geq \frac{\mathrm{fat}_{K,\Delta}(12\gamma) - 4}{\log^2\left(4\mathrm{fat}_{K,\Delta}(12\gamma)/\gamma^2\right)}.$$

The proof of (a) follows easily from the definitions.

Proof (a) Suppose $D_\infty(K, F) < \gamma$ and $\mathrm{fat}_K(2\gamma) = m$. Then there is an $x \in X^m$, an $r \in \mathbb{R}^m$, and a set K' of 2^m functions in K so that, for all s in $\{-1, 1\}^m$ there is a g_s in K' with $s_i(g_s(x_i) - r_i) \geq 2\gamma$ for $i = 1, 2, \ldots, m$. By hypothesis, for each g_s in K', there is a function f_s in F that has $\|f_s - g_s\|_\infty < \gamma$, so $s_i(f_s(x_i) - r_i) > \gamma$. Hence, x is γ-shattered by F, so $\mathrm{fat}_F(\gamma) \geq m$. □

The proof of (b) uses some concepts from the theory of metric spaces (see, for example, [8]).

Definition 7 *Let* (Z, ρ) *be a metric space,* $B \subset Z$, *and* $\epsilon > 0$. *The set* B *is* ϵ*-separated if for all distinct* a *and* b *in* B $\rho(a, b) \geq \epsilon$. *If* A *is a subset of* Z, *let* $\mathcal{A}_\rho(A, B, \epsilon)$ *denote the number of elements of* A *that can be approximated to within* ϵ *by* B,

$$\mathcal{A}_\rho(A, B, \epsilon) = \left|\{a \in A : \exists b \in B\, \rho(a, b) < \epsilon\}\right|.$$

The metric spaces we will consider contain real-valued functions defined on X, with distances

$$\rho_x(a, b) = \max_{1 \leq i \leq m} |a(x_i) - b(x_i)|,$$

where $x = (x_1, x_2, \ldots, x_m) \in X^m$.

To prove (b), we use the following lemma to show that if $\mathrm{fat}_F(\gamma)$ is sufficiently small in terms of $\mathrm{fat}_{K,\Delta}(12\gamma)$, then F cannot accurately approximate a certain finite subset of K at any $\mathrm{fat}_{K,\Delta}(12\gamma)$ points. We then use a probabilistic argument to show that in this case there exists a function g in K for which $\inf_{f \in F} \|f - g\|_p$ is at least $2\Delta^{1/p}\gamma$, so $D_p(K, F)$ is at least that large.

Lemma 8 *Let F be a class of $[0,1]$-valued functions defined on X, $0 < \alpha, \gamma < 1$, $m \in \mathbb{N}$, and*

$$\mathrm{fat}_F(\gamma) < \frac{m - \log \frac{2}{\alpha}}{\log^2 \frac{4m}{\gamma^2}}.$$

For all $x \in X^m$ and all $\delta \geq 12\gamma$, if A is a δ-separated set with respect to ρ_x then

$$\mathcal{A}_{\rho_x}(A, F, \delta/3) \leq \alpha 2^m.$$

The proof of Lemma 8 is in the appendix.

Proof (of Theorem 6(b)) Let $\delta = 12\gamma$ and $m = \mathrm{fat}_{K,\Delta}(\delta)$. We will show that, if

$$\mathrm{fat}_F(\gamma) < \frac{m-4}{\log^2 \frac{4m}{\gamma^2}}$$

then $D_p(K, F) > 2\Delta^{1/p}\gamma$.

From the definition of the setwise fat-shattering function, there is a set S with $\int_S dx = \Delta$, a $b \in X^m$, an $r \in \mathbb{R}^m$, and a set $K' \subset K$ with $|K'| = 2^m$ so that

1. $b_i + S \subset X$ for all i, and
2. for all s in $\{-1,1\}^m$, there is a function t in K' such that, for all i in $\{1,\ldots,m\}$ and all x in $b_i + S$, we have $s_i(t(x_i) - r_i) \geq \delta$.

Let q be a function that maps from K' to F, and define an average approximation error on K' as

$$\mathcal{I}_p(K', q) = \frac{1}{|K'|} \sum_{g \in K'} \|q(g) - g\|_p.$$

It is shown in the appendix that

$$\frac{1}{|K'|} \sum_{g \in K'} \inf_{f \in F} \|f - g\|_p \geq \inf_q \mathcal{I}_p(K', q), \tag{1}$$

where the infimum is over all functions $q : K' \to F$. We will show that

$$\inf_q \mathcal{I}_p(K', q) \geq \delta\Delta^{1/p}/6. \tag{2}$$

This and Equation (1) implies there is a function $g \in K'$ with

$$\inf_{f \in F} \|f - g\|_p \geq \delta\Delta^{1/p}/6.$$

But $D_p(K, F)$ is the supremum over all $g \in K$ of this infimum, so the result follows.

To see that (2) is true, fix any $q : K' \to F$. Then

$$\begin{aligned} \mathcal{I}_p(K', q) &= \frac{1}{|K'|} \sum_{g \in K'} \left(\int_X |q(g)(\theta) - g(\theta)|^p \, d\theta \right)^{1/p} \\ &\geq \frac{1}{|K'|} \sum_{g \in K'} \left(\int_A |q(g)(\theta) - g(\theta)|^p \, d\theta \right)^{1/p} \end{aligned}$$

where

$$A = \bigcup_{i=1}^{m} (b_i + S).$$

Now, the sets $b_i + S$ are pairwise disjoint (since b is S-setwise δ-shattered and $\delta > 0$), so

$$\begin{aligned}
\mathcal{I}_p(K', q) &\geq \frac{1}{|K'|} \sum_{g \in K'} \left(\int_S \sum_{i=1}^{m} |q(g)(\theta + b_i) - g(\theta + b_i)|^p \, d\theta \right)^{1/p} \\
&\geq \frac{1}{|K'|} \sum_{g \in K'} \left(\int_S \left(\max_{i \in \{1,\ldots,m\}} |q(g)(\theta + b_i) - g(\theta + b_i)| \right)^p \, d\theta \right)^{1/p}.
\end{aligned}$$

Hölder's inequality implies that

$$\int_S |f(t)| \, dt \leq \left(\int_S |f(t)|^p \, dt \right)^{1/p} \left(\int_S dt \right)^{(1-1/p)}$$

for $p \in \mathbb{N}$ and measurable f (see [8]), so

$$\begin{aligned}
\mathcal{I}_p(K', q) &\geq \frac{1}{|K'|} \sum_{g \in K'} \frac{1}{\Delta^{(1-1/p)}} \int_S \max_{i \in \{1,\ldots,m\}} |q(g)(\theta + b_i) - g(\theta + b_i)| \, d\theta \\
&= \frac{1}{|K'| \Delta^{(1-1/p)}} \int_S \left(\sum_{g \in K'} \max_{i \in \{1,\ldots,m\}} |q(g)(\theta + b_i) - g(\theta + b_i)| \right) d\theta.
\end{aligned}$$

Now, for all $\theta \in S$, $x(\theta) = (\theta + b_1, \ldots, \theta + b_m)$ is in X^m and K' is 2δ-separated with respect to $\rho_{x(\theta)}$. So Lemma 8 (with $\alpha = 1/2$) implies $\mathcal{A}_{\rho_{x(\theta)}}(K', F, \delta/3) \leq 2^{m-1}$, and

$$\begin{aligned}
\mathcal{I}_p(K', q) &\geq \frac{1}{|K'| \Delta^{(1-1/p)}} \int_S \left(\sum_{g \in K'} \rho_{x(\theta)}(q(g), g) \right) d\theta \\
&\geq \frac{1}{2^m \Delta^{(1-1/p)}} \Delta 2^{m-1} \frac{\delta}{3} \\
&\geq 2\Delta^{1/p} \gamma.
\end{aligned}$$

□

4 Example

In this section we give an example of the application of Theorem 6 to a certain class of functions K, the homogeneous Lipschitz spaces. We choose these spaces because they are simple to define, they enable us to give a constructive proof of lower bounds on $\mathrm{fat}_{K,\Delta}$, and they show that the bounds of Theorem 6 are tight in a certain sense.

Definition 9 (Homogeneous Lipschitz Spaces [10]) *Let $\| \cdot \|_2$ be the Euclidean norm on $\mathbb{R}^n$. Let $\Delta_h g(x) = g(x + h) - g(x)$. Let $j = (j_1, \ldots, j_n) \in \mathbb{N}_0^n$ be a multi-index with*

$|j| = j_1 + \cdots + j_n$, *and let* $\mathrm{D}^j f$ *represent the corresponding partial derivative of a function* $f: X \to \mathbb{R}$. *The class of functions* $\mathrm{Lip}_k(X)$ *is defined by*

$$\mathrm{Lip}_k(X) := \{f \in C^{k-1}(X,[0,1]):\ \textit{there exists an } M \in \mathbb{R} \textit{ such that } |\Delta_h \mathrm{D}^j f(x)| \le M\|h\|_2 \textit{ for } |j| = k-1,\ h \in \mathbb{R}^n,\ x, x+h \in X\}.$$

The norm $\|f\|_{\mathrm{Lip}_k(X)}$ *is defined as the infimum of all* M *satisfying* $|\Delta_h \mathrm{D}^j f(x)| \le M\|h\|_2$ *for all* $h \in \mathbb{R}^n$ *and all* j *and* x *satisfying* $|j| = k-1$ *and* $x, x+h \in X$. *We write*

$$\mathrm{Lip}_k(X, M) := \{f \in \mathrm{Lip}_k(X): \|f\|_{\mathrm{Lip}_k(X)} \le M\}.$$

Theorem 10 *Let* $n, p, k \in \mathbb{N}$ *and* $M \in \mathbb{R}^+$. *There are constants* $c_1, c_2, c_3 \in \mathbb{R}^+$ *(that depend only on* n, p, k, M*) such that*
(a) *If* $\gamma > D_\infty\left(\mathrm{Lip}_k(X, M), F\right)$ *then*

$$\mathrm{fat}_F(\gamma) \ge c_1 \gamma^{-n/k}.$$

(b) *If* $\gamma > c_2\left[D_p\left(\mathrm{Lip}_k(X, M), F\right)\right]^{kp/(kp+n)}$ *then*

$$\mathrm{fat}_F(\gamma) \ge c_3 \left(\frac{\gamma^{-n/k}}{\log^2(1/\gamma)}\right).$$

To prove Theorem 10, we give lower bounds on $\mathrm{fat}_{\mathrm{Lip}_k(X,M)}$ and $\mathrm{fat}_{\mathrm{Lip}_k(X,M),\Delta}$ by exhibiting functions in $\mathrm{Lip}_k(X, M)$ that γ-shatter and setwise γ-shatter points in X. We start with the following one-dimensional bump functions. (The proof of Lemma 11 is in the appendix.)

Lemma 11 *For all* $k \in \mathbb{N}$ *there is a constant* $c \in \mathbb{R}^+$ *such that for all* $M \in \mathbb{R}+$ *there is a function* $f_k : \mathbb{R} \to \mathbb{R}$ *that satisfies*

1. $f \in \mathrm{Lip}_k(\mathbb{R}, M)$.
2. f_k *vanishes outside* $(0, 1)$.
3. $f_k^{(j)}(0) = f_k^{(j)}(1/2) = f_k^{(j)}(1) = 0$ *for all* $j \in \{1, \ldots, k-1\}$.
4. f_k *increases monotonically on* $(0, 1/2)$.
5. $f_k(1/4) = cM$.

By rotating these functions, we can obtain bump functions in $\mathbb{R}^n$. We can construct a set of functions that setwise γ-shatters a set of points in X that are arranged in a uniform rectangular lattice by summing a number of translated, dilated, and possibly inverted copies of these functions; this gives the following lower bound on the setwise fat-shattering function (the proof is in the appendix).

Lemma 12 *For* $n, k \in \mathbb{N}$, *there are constants* c_1 *and* c_2 *such that, for any* $M \in \mathbb{R}^+$, *if* $\Delta = c_1(\gamma/M)^{n/k}$, *then*

$$\mathrm{fat}_{\mathrm{Lip}_k(X,M),\Delta}(\gamma) \ge c_2(\gamma/M)^{-n/k}.$$

(It is easy to check that the bound in Lemma 12 is within a constant factor of the best possible for $k = 1$.) Theorem 10 follows easily from this lemma and Theorem 6, and from the inequality

$$\mathrm{fat}_{\mathrm{Lip}_k(X,M)}(\gamma) \geq \mathrm{fat}_{\mathrm{Lip}_k(X,M),\Delta}(\gamma).$$

We can compare the bounds in Theorem 10 with bounds on $D_p(\mathrm{Lip}_k(X,M),F)$ and fat_F for specific function classes F. For $n, \nu \in \mathbb{N}$, let $\mathcal{P}_{n,\nu}$ be the class of polynomials defined on the n variables $x_1, \ldots, x_n$, with degree at most ν in each variable and range $[0,1]$ on $X = [0,1]^n$. Members of this class can be specified with $d = (\nu+1)^n$ parameters. Lorentz [9, Theorem 8, p90] shows that, for any $n, \nu, k \in \mathbb{N}$ and $M \in \mathbb{R}^+$, there is a constant c_1 such that

$$D_\infty\left(\mathrm{Lip}_k(X,M), \mathcal{P}_{n,\nu}\right) \leq c_1 d^{-k/n},$$

which implies

$$D_p\left(\mathrm{Lip}_k(X,M), \mathcal{P}_{n,\nu}\right) \leq c_1 d^{-k/n}$$

for any $p \in \mathbb{N}$. Theorem 10(a) shows that there is a constant c_2 such that, if $\gamma > c_1 d^{-k/n}$ then $\mathrm{fat}_F(\gamma) \geq c_2 d$. Theorem 10(b) shows that there are constants c_3 and c_4 such that, if $\gamma > c_4 d^{-k^2 p/(nkp+n^2)}$ then $\mathrm{fat}_F(\gamma) \geq c_5 d^{kp/(kp+n)} / \log^2 d$. It is easy to show that $\mathrm{fat}_{\mathcal{P}_{n,\nu}}(\gamma) = d$ for $\gamma \leq 1/2$ (see for example [6, Theorem 4, p109]). It follows that Theorem 10(a) is tight within a constant factor, and Theorem 10(b) is tight within a log factor as kp/n becomes large. Notice that we have used a degree of approximation result for $p = \infty$ to infer a result for finite p; it seems likely that this bound on $D_p\left(\mathrm{Lip}_k(X,M), \mathcal{P}_{n,\nu}\right)$ is already loose for small p.

Finally, since fat_F is a non-increasing function, Theorem 10 has the following corollary.

Corollary 13 *Let $n, k \in \mathbb{N}$, $M \in \mathbb{R}^+$, $X = [0,1]^n$. There is a constant $c \in \mathbb{R}^+$ such that if $F \subset [0,1]^X$ has approximation error $D_1\left(\mathrm{Lip}_k(X,M), F\right) < \epsilon$ for some $0 < \epsilon < 1$, then*

$$\mathrm{fat}_F(\epsilon) \geq \frac{c\epsilon^{-n/(k+n)}}{\log^2 1/\epsilon}.$$

5 Conclusions

We have shown that there is a tradeoff between sample complexity and approximation error in learning $[0,1]$-valued functions defined on $[0,1]^n$. The same results give a similar tradeoff for the problem of learning probabilistic concepts, introduced by Kearns and Schapire [7]. A probabilistic concept can be viewed as a distribution on $X \times \{0,1\}$; the aim of the learning algorithm is to estimate the conditional probability distribution $\Pr(y = 1|x)$ from a random sequence of (x,y) pairs. Kearns and Schapire show that the fat-shattering function provides a lower bound for the sample complexity of learning probabilistic concepts, so our results can be applied directly.

Several natural open problems remain. Our definition of sample complexity is rather arbitrary. First, can similar results be obtained when the error $\mathbf{er}_{P,t}$ is defined using a p-norm instead of a 1-norm? Second, we require that the learning algorithm finds a near-optimal function for any target function $t : X \to [0,1]$. It seems natural to restrict t to the class K of functions that we wish to approximate. Unfortunately, the general lower bounds

of [4] do not apply in this case; indeed an example in [4] shows that no such lower bounds exist in terms of the fat-shattering function without further restrictions on K. The reason is that a single real value can convey an arbitrary amount of information to the learning algorithm. This unusual behaviour does not occur when the function values are corrupted with random [4] or deterministic [2] observation noise, nor does it occur in the problem of learning probabilistic concepts [7]; in these cases, the amount of information conveyed to the learning algorithm is limited.

Acknowledgements

This research was supported by the Australian Telecommunications and Electronics Research Board and the Australian Research Council.

References

[1] N. Alon, S. Ben-David, N. Cesa-Bianchi and D. Haussler, *Scale-sensitive dimensions, uniform convergence, and learnability*, Symposium on Foundations of Computer Science, 1993.

[2] M. Anthony and P. L. Bartlett, *Function learning from interpolation*, Department of Systems Engineering, Australian National University, Technical Report, Canberra, Australia, 1994.

[3] A. R. Barron, *Approximation and Estimation Bounds for Artificial Neural Networks*, in Proceedings of the Fourth Annual Workshop on Computational Learning Theory, L. G. Valiant and M. K. Warmuth, eds., Morgan Kaufmann, San Mateo, 1991, pp. 243–249.

[4] P. L. Bartlett, P. M. Long and R. C. Williamson, *Fat-shattering and the learnability of real-valued functions (extended abstract)*, Proceedings of the Seventh Annual ACM Conference on Computational Learning Theory, 1994.

[5] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley and Sons, New York, 1991.

[6] D. Haussler, *Decision theoretic generalizations of the PAC model for neural net and other learning applications*, Information and Computation, 100 (1992), pp. 78–150.

[7] M. J. Kearns and R. E. Schapire, *Efficient distribution-free learning of probabilistic concepts (extended abstract)*, Proceedings of the 31st Annual Symposium on the Foundations of Computer Science, 1990.

[8] A. N. Kolmogorov and S. V. Fomin, *Introductory Real Analysis*, Dover, New York, 1970.

[9] G. G. Lorentz, *Approximation of Functions*, Holt, Rinehart and Winston, New York, 1966.

[10] H. Wallin, *New and Old Function Spaces*, in Function Spaces and Applications (Lecture Notes in Mathematics 1302), Springer Verlag, Berlin, 1988, pp. 99–114.

Appendix

Proof of Lemma 8

If (Z, ρ) is a metric space, $B \subset Z$, and $\epsilon > 0$, the ϵ-packing number $\mathcal{M}_\rho(\epsilon, B)$ of B is the size of the largest ϵ-separated subset of B. The following lemma relates a packing number of F to $\mathrm{fat}_F(\gamma)$. It combines Lemma 14 and 15 in [1].

Lemma 14 *If F is a class of $[0,1]$-valued functions defined on X with $\mathrm{fat}_F(\gamma) \leq d$, $d \geq 1$, $\epsilon \geq 4\gamma$, $x \in X^m$, and $m \geq \log c + 1$, then*

$$\mathcal{M}_{\rho_x}(\epsilon, F) \leq 2(mb^2)^{\log c}$$

where $b = \lfloor 1/(2\gamma) \rfloor + 1$ and

$$c = \sum_{i=1}^{d} \binom{m}{i} b^i.$$

It follows that if we choose m sufficiently large under these conditions, we can make $\mathcal{M}_{\rho_x}(\epsilon, F)$ significantly smaller than 2^m for any $x \in X^m$.

Lemma 15 *If $0 < \alpha, \gamma < 1$ and $m \in \mathbb{N}$ satisfies*

$$\mathrm{fat}_F(\gamma) < \frac{m - \log \frac{2}{\alpha}}{\log^2 \frac{4m}{\gamma^2}} \tag{3}$$

then for all $x \in X^m$ and all $\epsilon \geq 4\gamma$

$$\mathcal{M}_{\rho_x}(\epsilon, F) < \alpha 2^m.$$

Proof Let

$$d = \left\lfloor \frac{m - \log \frac{2}{\alpha}}{\log^2 \frac{4m}{\gamma^2}} \right\rfloor.$$

By hypothesis, $\mathrm{fat}_F(\gamma) \leq d$, so we must have $d \geq 0$. If $d = 0$, for all $x \in X$ and any $f_1, f_2 \in F$ we have $|f_1(x) - f_2(x)| < 2\gamma$, so $\mathcal{M}_{\rho_x}(\epsilon, F) \leq 1$ for any $\epsilon \geq 2\gamma$. Since $m \geq \log(2/\alpha)$ (because $d \geq 0$), $\alpha 2^m \geq 2$ so $\mathcal{M}_{\rho_x}(\epsilon, F) < \alpha 2^m$ in this case.

Assume then that $d \geq 1$. Let b and c be defined as in Lemma 14. Then $b < 2/\gamma$ and

$$\begin{aligned} \log c &< \log \sum_{i=1}^{d} \binom{m}{i} \left(\frac{2}{\gamma}\right)^i \\ &< \log \left(d \binom{m}{d} \left(\frac{2}{\gamma}\right)^d \right) \\ &< d \log(2m/\gamma) + \log d, \end{aligned}$$

so if

$$m \geq d \log(2m/\gamma) + \log d, \tag{4}$$

then Lemma 14 implies

$$\mathcal{M}_{\rho_x}(\epsilon, F) < 2\left(\frac{4m}{\gamma^2}\right)^{(d\log(2m/\gamma)+\log d)}.$$

To ensure that this quantity is no bigger than $\alpha 2^m$, we must show that

$$m \geq \log\frac{2}{\alpha} + \log\frac{4m}{\gamma^2}\left(d\log\frac{2m}{\gamma} + \log d\right),$$

so

$$\begin{aligned} m &\geq \log\frac{2}{\alpha} + d\log^2\frac{4m}{\gamma^2} \\ \Leftrightarrow d &\leq \frac{m - \log\frac{2}{\alpha}}{\log^2\frac{4m}{\gamma^2}} \end{aligned}$$

will suffice. In addition, this will ensure that (4) holds. □

Lemma 8 follows from Lemma 15 and the following elementary result which relates $\mathcal{A}_\rho(A, F, \epsilon)$ to $\mathcal{M}_\rho(\epsilon, F)$, for ϵ-separated A.

Lemma 16 *If A and F are in the metric space (Z, ρ), $\epsilon > 0$, and A is a finite ϵ-separated set, then*

$$\mathcal{A}_\rho(A, F, \epsilon/3) \leq \mathcal{M}_\rho(\epsilon/3, F).$$

Proof Since A is finite, $\mathcal{A}_\rho(A, F, \epsilon/3)$ is finite. Let $n = \mathcal{A}_\rho(A, F, \epsilon/3)$. From the definition of $\mathcal{A}_\rho$, there is a set $A' \subset A$ of size n that can be approximated to accuracy $\epsilon/3$ by F. Construct a set $F' \subset F$ by choosing, for each a in A', a single f in F that has $\rho(f, a) < \epsilon/3$. Clearly $|F'| = n$, because if a and b are distinct elements of A', $f \in F'$, and $\rho(f, a) < \epsilon/3$, then $\rho(f, b) \geq \rho(a, b) - \rho(f, a) > 2\epsilon/3$ by the triangle inequality for ρ.

Now, for any distinct f_1 and f_2 in F', choose the a_1 and a_2 in A' that satisfy $\rho(f_1, a_1) < \epsilon/3$ and $\rho(f_2, a_2) < \epsilon/3$. Then the triangle inequality implies

$$\begin{aligned} \rho(f_1, f_2) &\geq \rho(f_1, a_2) - \rho(f_2, a_2) \\ &> \rho(a_1, a_2) - \rho(f_1, a_1) - \epsilon/3 \\ &> \epsilon/3. \end{aligned}$$

That is, F' is $\epsilon/3$-separated, so $\mathcal{M}_\rho(\epsilon/3, F)$ is at least n. □

Proof of Inequality 1

Let K' and F be sets of $[0, 1]$-valued functions where K' is finite. Let $p \in \mathbb{N}$. In this section, we prove the inequality

$$\frac{1}{|K'|}\sum_{g \in K'} \inf_{f \in F} \|f - g\|_p \geq \inf_q \mathcal{I}_p(K', q),$$

where the infimum is over all functions q from K' to F, and

$$\mathcal{I}_p(K', q) = \frac{1}{|K'|} \sum_{g \in K'} \|q(g) - g\|_p.$$

For $\beta > 0$, we can choose a $q_\beta : K' \to F$ such that

$$\inf_{f \in F} \|f - g\|_p \leq \|q_\beta(g) - g\|_p < \inf_{f \in F} \|f - g\|_p + \beta.$$

Then

$$\begin{aligned} \inf_q \mathcal{I}(q, K', p) &\leq \mathcal{I}(q_\beta, K', p) \\ &= \frac{1}{|K'|} \sum_{g \in K'} \|q_\beta(g) - g\|_p \\ &< \beta + \frac{1}{|K'|} \sum_{g \in K'} \inf_{f \in F} \|f - g\|_p. \end{aligned}$$

That is, for any $\beta > 0$

$$\inf_q \mathcal{I}(q, K'.p) < \beta + \frac{1}{|K'|} \sum_{g \in K'} \inf_{f \in F} \|f - g\|_p,$$

so $\inf_q \mathcal{I}(q, K', p) \leq 1/|K'| \sum_{g \in K'} \inf_{f \in F} \|f - g\|_p$, which is the desired result.

Proof of Lemma 11

The functions $\{f_i : i = 0, 1, \ldots\}$ are constructed as follows. Let $f_0 : \mathbb{R} \to \mathbb{R}$ be defined by

$$f_0(x) = \begin{cases} M & \text{if } 0 < x < 1 \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

For $k \in \mathbb{N}$, let $f_k : \mathbb{R} \to \mathbb{R}$ be the function obtained by integrating $f_k^{(k)}$ k times, where

$$f_k^{(k)}(x) = \begin{cases} f_{k-1}^{(k-1)}(2x) & 0 < x < 1/2 \\ -f_{k-1}^{(k-1)}(2x-1) & 1/2 \leq x < 1 \\ 0 & \text{otherwise,} \end{cases} \tag{6}$$

and $f_k^{(j)}(0) = 0$ for $j = 0, 1, \ldots, k-1$ ($g^{(0)}(x)$ denotes $g(x)$). Notice that f_k is in C^{k-1} and its k-th derivative is defined and equal to $f_k^{(k)}$ almost everywhere, so

$$\|f_k\|_{\mathrm{Lip}_k(\mathbb{R})} = \max_{x \in [0,1]} \left| f_k^{(k)}(x) \right|.$$

Figure 1 illustrates f_0, f_1, f_2, and their derivatives.

The following proposition is Part 1 of Lemma 11.

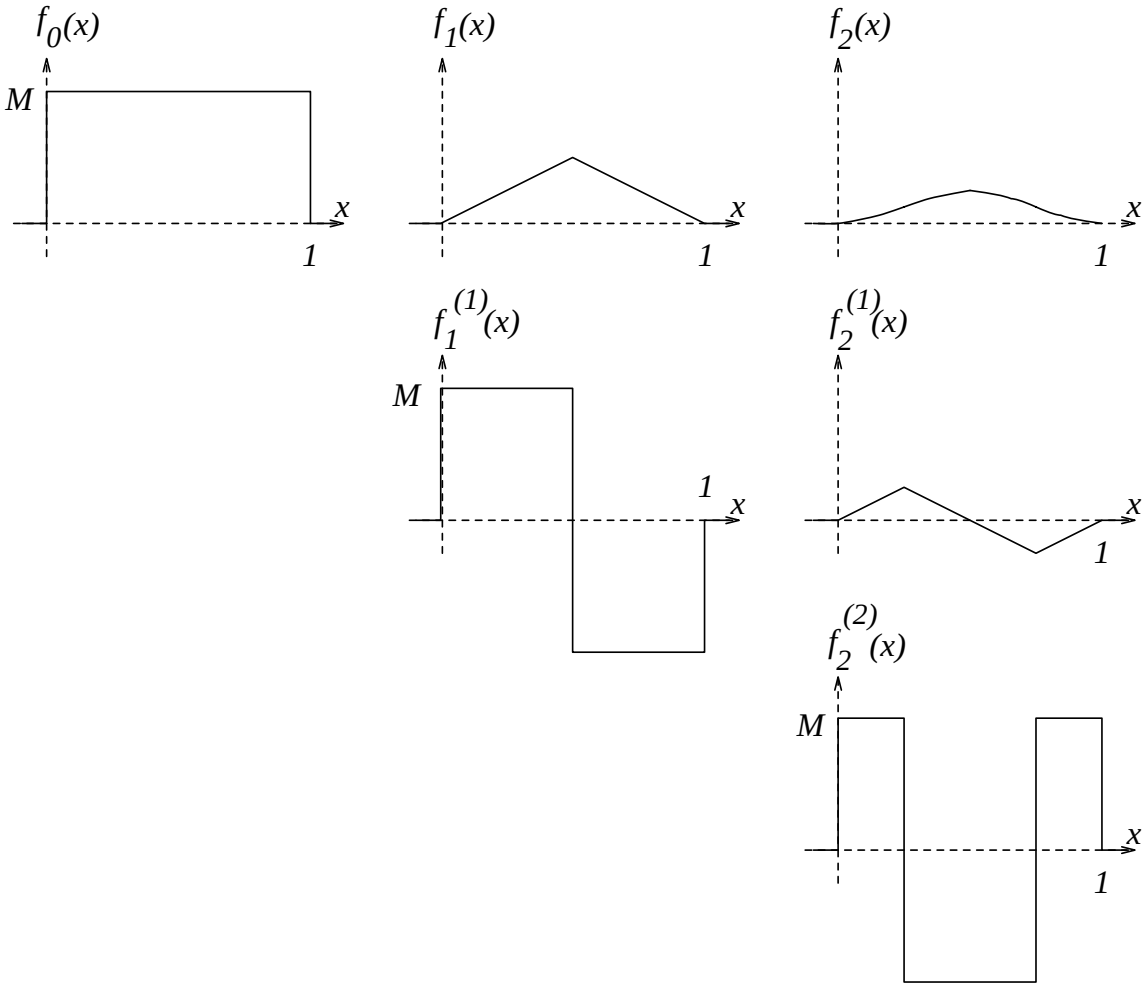

Figure 1: The functions f_0, f_1, f_2, and their derivatives.

Proposition 17 *For $k \in \mathbb{N}$, f_k is in $\mathrm{Lip}_k(\mathbb{R}, M)$.*

Proof By definition, $f_k \in C^{k-1}(X, \mathbb{R})$. We will show by induction that $\left|f_k^{(k)}(x)\right| = M$ for x in $(0,1)$. This implies the proposition. By definition, $|f_1^{(1)}(x)| = M$ for x in $(0,1)$. Suppose $|f_k^{(k)}(x)| = M$ for all x in $(0,1)$ and some k in $\mathbb{N}$. Then the definition of $f_{k+1}^{(k+1)}$ implies that $|f_{k+1}^{(k+1)}(x)| = M$ for all x in $(0,1)$. $\square$

The following two propositions give useful properties of these functions. Recall that the indicator function $1_A : \mathbb{R} \to \{0,1\}$ of a set $A \subset \mathbb{R}$ satisfies $1_A(x) = 1$ if and only if $x \in A$.

Proposition 18 *For all i in $\mathbb{N}$ and all j in $\{1, 2, \ldots, i\}$,*

$$f_i^{(j)}(x) = 2^{j-i}\left(1_{(0,1/2)}(x) f_{i-1}^{(j-1)}(2x) - 1_{[1/2,1)}(x) f_{i-1}^{(j-1)}(2x-1)\right) \tag{7}$$

for all x in $(0,1)$, and

$$f_i^{(j-1)}(1) = 0. \tag{8}$$

Notice that this proposition proves Parts 2 and 3 of Lemma 11.

Proof Let $S(i,j)$ be the proposition that (7) holds for all x in $(0,1)$ and (8) holds. We will use an inductive argument to prove the proposition.

Clearly, $S(i,i)$ is true for all i in $\mathbb{N}$, both from the definition of $f_i^{(i)}(x)$, and because

$$\begin{aligned} f_i^{(i-1)}(1) &= \int_0^1 f_i^{(i)}(x)\,dx \\ &= \int_0^{1/2} f_{i-1}^{(i-1)}(2x)\,dx - \int_{1/2}^1 f_{i-1}^{(i-1)}(2x-1)\,dx \\ &= 0. \end{aligned}$$

Now, let $l \in \mathbb{N}_0$ and suppose $S(i,j)$ is true for all i in $\mathbb{N}$ and all j in $\{i-l, i-l+1, \ldots, i\} \cap \mathbb{N}$. Then we have, for any $i \geq l+2$,

$$\begin{aligned} f_i^{(i-l-1)}(x) &= \int_0^x f_i^{(i-l)}(\gamma)\,d\gamma \\ &= 2^{-l-1}\left(1_{(0,1/2)}(x)\int_0^{2x} f_{i-1}^{(i-l-1)}(\gamma)\,d\gamma + 1_{[1/2,1)}(x)\int_0^1 f_{i-1}^{(i-l-1)}(\gamma)\,d\gamma - \right. \\ &\qquad \left. 1_{[1/2,1)}(x)\int_0^{2x-1} f_{i-1}^{(i-l-1)}(\gamma)\,d\gamma\right) \\ &= 2^{-l-1}\left(1_{(0,1/2)}(x) f_{i-1}^{(i-l-2)}(2x) - 1_{[1/2,1)}(x) f_{i-1}^{(i-l-2)}(2x-1) + \right. \\ &\qquad \left. 1_{[1/2,1)}(x) f_{i-1}^{(i-l-2)}(1)\right) \\ &= 2^{-l-1}\left(1_{(0,1/2)}(x) f_{i-1}^{(i-l-2)}(2x) - 1_{[1/2,1)}(x) f_{i-1}^{(i-l-2)}(2x-1)\right), \end{aligned}$$

where the last line follows because $S(i-1, i-l-1)$ is true, so $f_{i-1}^{(i-l-2)}(1) = 0$. Furthermore,

$$\begin{aligned} f_i^{(i-l-2)}(1) &= \int_0^1 f_i^{(i-l-1)}(x)\,dx \\ &= 2^{-l-1}\left(\int_0^{1/2} f_{i-1}^{(i-l-2)}(2x)\,dx - \int_{1/2}^1 f_{i-1}^{(i-l-2)}(2x-1)\,dx\right) \\ &= 2^{-l-2}\left(\int_0^1 f_{i-1}^{(i-l-2)}(x)\,dx - \int_0^1 f_{i-1}^{(i-l-2)}(x)\,dx\right) \\ &= 0, \end{aligned}$$

so $S(i, i-l-1)$ is true. $\square$

Proposition 19 *For all i in $\mathbb{N}$ and all x in $(0,1)$,*

$$f_i(x) = f_i(1-x).$$

Proof It suffices to show that, for each i, $f_i(x) = f_i(1-x)$ for all x in $(0,1/2)$. The proof is by induction. By definition, $f_0(x) = f_0(1-x) = M$ for all x in $(0,1/2)$. Suppose $f_k(x) = f_k(1-x)$ for all $x \in (0,1/2)$ and some k in $\mathbb{N}_0$. Then

$$\begin{aligned} f_{k+1}(x) &= \int_0^x f_{k+1}^{(1)}(\gamma)\, d\gamma \\ &= 2^{-k} \int_0^x f_k(2\gamma)\, d\gamma, \end{aligned}$$

and

$$\begin{aligned} f_{k+1}(1-x) &= 2^{-k} \int_0^{1-x} \left(f_k(2\gamma) 1_{(0,1/2)}(\gamma) - f_k(2\gamma - 1) 1_{[1/2,1)}(\gamma) \right) d\gamma \\ &= 2^{-k-1} \int_0^{2x} f_k(\gamma)\, d\gamma \\ &= f_{k+1}(x). \end{aligned}$$

$\square$

The following proposition implies Part 4 of Lemma 11.

Proposition 20 *For $k \in \mathbb{N}$, $f_k(x) \geq 0$ for $0 \leq x \leq 1$ and $f_k^{(1)}(x) \geq 0$ for $0 \leq x \leq 1/2$.*

Proof Clearly, $f_1(x) \geq 0$ for $0 \leq x \leq 1$, and $f_1^{(1)}(x) = M \geq 0$ for $0 \leq x \leq 1/2$. Suppose $f_k(x) \geq 0$ for $0 \leq x \leq 1$ and $f_k^{(1)}(x) \geq 0$ for $0 \leq x \leq 1/2$. Then $f_{k+1}^{(1)}(x) = 2^{-k} f_k(2x) \geq 0$ for $0 \leq x \leq 1/2$. and so $f_{k+1}(x) \geq 0$ for $0 \leq x \leq 1$ (applying Proposition 19 if $x \geq 1/2$). $\square$

It remains to prove Part 5 of Lemma 11.

Proposition 21 *For $k \in \mathbb{N}_0$,*

$$f_k(1/4) = M 2^{-(k^2+3k)/2}. \tag{9}$$

Proof We first prove the following symmetry property: for all $k \in \mathbb{N}$ and all x in $[0,1/2]$,

$$f_k(x) - f_k(1/4) = f_k(1/4) - f_k(1/2 - x). \tag{10}$$

Indeed,

$$\begin{aligned} f_k(x) - f_k(1/4) &= \int_0^x f_k^{(1)}(\gamma)\, d\gamma - \int_0^{1/4} f_k^{(1)}(\gamma)\, d\gamma \\ &= 2^{-k} \left(\int_0^{2x} f_{k-1}(\gamma)\, d\gamma - \int_0^{1/2} f_{k-1}(\gamma)\, d\gamma \right) \\ &= 2^{-k} \int_{1/2}^{2x} f_{k-1}(\gamma)\, d\gamma, \end{aligned}$$

and

$$\begin{aligned} f_k(1/4) - f_k(1/2 - x) &= 2^{-k} \left(\int_0^{1/2} f_{k-1}(\gamma)\, d\gamma - \int_0^{1-2x} f_{k-1}(\gamma)\, d\gamma \right) \\ &= 2^{-k} \int_{1/2}^{2x} f_{k-1}(\gamma)\, d\gamma. \end{aligned}$$

Now, Equation (9) is trivially true for $k = 0$. To prove that it is true for k in $\mathbb{N}$, we show that

$$f_k(1/4) = M \prod_{i=2}^{k+1} 2^{-i}, \tag{11}$$

which implies Equation (9). The proof of (11) is by induction. Clearly,

$$f_1(1/4) = \int_0^{1/4} f_1^{(1)}(x)\, dx = 1/2 \int_0^{1/2} f_0(x)\, dx = M/4.$$

Suppose that Equation (11) is true for $k = 1, 2, \ldots, j$. Then

$$\begin{aligned} f_{j+1}(1/4) &= 2^{-j} \int_0^{1/4} f_{j+1}^{(1)}(x)\, dx \\ &= 2^{-j-1} \int_0^{1/2} f_j(x)\, dx \\ &= 2^{-j-1} \left(\int_0^{1/4} f_j(x)\, dx + f_j(1/4)/4 + \int_{1/4}^{1/2} (f_j(x) - f_j(1/4))\, dx \right) \\ &= 2^{-j-1} \left(\int_0^{1/4} f_j(x)\, dx + f_j(1/4)/4 + \int_{1/4}^{1/2} (f_j(1/4) - f_j(1/2 - x))\, dx \right) \\ &= 2^{-j-2} f_j(1/4) \\ &= M \prod_{i=2}^{j+2} 2^{-i}. \end{aligned}$$

□

Proof of Lemma 12

To prove the lemma, we will construct bump functions in $\mathrm{Lip}_k(\mathbb{R}^n)$ and show that these setwise shatter a sequence of points in $\mathbb{R}^n$. The following proposition will be useful to prove bounds on the norm of these bump functions in $\mathrm{Lip}_k(\mathbb{R}^n)$.

Proposition 22 *Suppose $\phi \in C^k(\mathbb{R})$ and $\psi \in C^k(\mathbb{R}^n)$ satisfy*

$$|D^\mu \psi(x)| \leq B_{\psi,j}$$

for $|\mu| = j$ and $x \in \mathbb{R}$, for $j = 0, \ldots, k$, and

$$\left| \phi^{(j)}(x) \right| \leq B_{\phi,j}$$

for $x \in \mathbb{R}$, for $j = 0, \ldots, k$. Then for all $l = 0, \ldots, k$ and all multi-indices μ with $|\mu| = l$,

$$|D^\mu \phi(\psi(x))| \leq 2B_{\phi,l}B_{\psi,1}^l + \sum_{m=2}^{l} B_{\phi,l-m+1}B_{\psi,k}B_{\psi,1}^{l-k}.$$

Proof Let

$$A(j,i) = \sup\left\{\left|D^\mu \phi^{(i)}(\psi(x))\right| : x \in \mathbb{R}^n, |\mu| = j\right\}.$$

Then $A(0,i) = B_{\phi,i}$. Suppose μ and ν are multi-indices with $|\mu| = j$ and $|\nu| = 1$. Then

$$\begin{aligned}
\left|D^{\mu+\nu}\phi^{(i)}(\psi(x))\right| &= \left|D^\mu\left(\phi^{(i+1)}(\psi(x))D^\nu\psi(x)\right)\right| \\
&= \left|D^\mu\left(\phi^{(i+1)}(\psi(x))\right)D^\nu\phi(x) + \phi^{(i+1)}(\psi(x))D^{\mu+\nu}\psi(x)\right| \\
&\leq A(j,i+1)B_{\psi,1} + A(0,i+1)A_{\psi,j+1}.
\end{aligned}$$

That is, $A(j+1,i) \leq A(j,i+1)B_{\psi,1} + A(0,i+1)A_{\psi,j+1}$.

We will prove by induction that

$$A(j,i) \leq 2B_{\phi,i+j}B_{\psi,1}^j + \sum_{m=2}^{j} B_{\phi,i-m+j+1}B_{\psi,m}B_{\psi,1}^{j-m}. \tag{12}$$

Indeed, $A(0,i) = B_{\phi,i}$, so (12) is true for $j = 0$ and $i \in \mathbb{N}_0$. Suppose (12) is true for $j = 0, \ldots, l$ and $i \in \mathbb{N}_0$. Then it is true for all j and i, since

$$\begin{aligned}
C(l+1,i) &\leq C(l,i+1)B_{\psi,1} + C(0,i+1)B_{\psi,l+1} \\
&\leq 2B_{\phi,i+l+1}B_{\psi,1}^{l+1} + \sum_{m=2}^{l} B_{\phi,i-m+l+2}B_{\psi,m}B_{\psi,1}^{j-k+1} + B_{\phi,i+1}B_{\psi,j+1} \\
&= 2B_{\phi,i+l+1}B_{\psi,1}^{l+1} + \sum_{m=2}^{l+1} B_{\phi,i-m+l+2}B_{\psi,m}B_{\psi,1}^{j-k+1}.
\end{aligned}$$

In particular, we have that

$$C(l,0) \leq 2B_{\phi,l}B_{\psi,1}^l + \sum_{m=2}^{l} B_{\phi,l-m+1}B_{\psi,k}B_{\psi,1}^{l-k}.$$

□

Consider the functions $p, q \in C^\infty$ defined by $p : x \mapsto \|x-a\|^2$ and $q : \alpha \mapsto (1-\alpha/r^2)/2$, where $a \in \mathbb{R}^n$ and $r \in \mathbb{R}$. Clearly, $f_k \circ p \circ q$ is k times differentiable almost everywhere, and we can apply Proposition 22 to give bounds on $\|f_k \circ p \circ q\|_{\mathrm{Lip}_k(\mathbb{R}^n)}$. Since $f_k(p(q(x))) = 0$ when $\|x-a\| > r$, we can define equivalent C^k functions p and q such that

$$B_{p,i} \leq \begin{cases} r^2 & i = 0 \\ 2r & i = 1 \\ 2 & i = 2 \\ 0 & \text{otherwise} \end{cases}$$

and

$$B_{q,i} \leq \begin{cases} 1/2 & i = 0 \\ 1/(2r^2) & i = 1 \\ 0 & \text{otherwise} \end{cases}$$

Defining $f_k^{(k)}$ (and hence f_k) as in Equations (5) and (6) with $M = L$, we have $B_{f_k,i} \leq L2^{i-k}$ for $i = 0, \ldots, k$. Applying Proposition 22 gives

$$B_{p \circ q,i} \leq \begin{cases} r^2 & i = 0 \\ 2/r & i = 1 \\ 0 & \text{otherwise,} \end{cases}$$

and hence

$$B_{f_k \circ p \circ q,i} \leq L2^{i-k+1} r^{-i}.$$

In particular, $B_{f_k \circ p \circ q,k} \leq 4L/r^k$.

Now, we prove Lemma 12 by constructing (from the bump functions $f_k \circ p \circ q$) functions in $\mathrm{Lip}_k(X, M)$ that setwise γ-shatter the sequence $(a_1, \ldots, a_m) \in X^m$. Suppose these points are $2r$-separated (that is, $\|a_i - a_j\| \geq 2r$ for $i \neq j$). For $s \in \{-1, 1\}^m$, let

$$g_s(x) = 1/2 + \sum_{i=1}^{m} s_i f_k \left(1/2 - \|x - a_i\|^2/(2r^2)\right).$$

Since the a_i's are $2r$-separated, only one term in this sum is ever positive, so $\|g_s\|_{\mathrm{Lip}_k(X)} \leq 4L/r^k$. Inside the region

$$S = \left\{x \in X : \|x - a_i\|^2 \leq r^2/2\right\},$$

we have $s_i g_s(x) \geq f_k(1/4)$. So for $\gamma = f_k(1/4)$, $\mathrm{Lip}_k(X, 4L/r^k)$ can S-setwise γ-shatter the sequence of m points. By arranging the points $\{a_1, \ldots, a_m\}$ in a regular rectangular lattice, we can ensure that $m \geq c_1(n)/r^n$, where $c_1(n)$ is a positive constant that depends only on n. Setting $L = Mr^k/4$ and $r = c_2(k)(\gamma/M)^{1/k}$ gives, for $\Delta \leq c_3(n,k)(\gamma/M)^{n/k}$, that

$$\mathrm{fat}_{\mathrm{Lip}_k(X,M),\Delta}(\gamma) \geq c_4(n,k)(\gamma/M)^{-n/k}.$$