Bounding the Fat Shattering Dimension of a Composition Function Class Built Using a Continuous Logic Connective

Hubert Haoyang Duan
University of Ottawa
hduan065@uottawa.ca

June 23, 2012

Abstract: The paper deals with an important combinatorial parameter of a function class, the Fat Shattering dimension. An important known result in statistical learning theory is that a function class is distribution-free Probably Approximately Correct learnable if it has finite Fat Shattering dimension on every scale.

As the main new result, we explore the construction of a new function class from a collection of existing ones, obtained by forming compositions with a continuous logic connective (a uniformly continuous function from the unit hypercube to the unit interval). Vidyasagar had proved that such a composition function class has finite Fat Shattering dimension of all scales if the classes in the original collection do; however, no estimates of the dimension were known. Using results by Mendelson-Vershynin and Talagrand, we bound the Fat Shattering dimension of scale ϵ of this new function class in terms of a sum of the Fat Shattering dimensions of the collection's classes.

1 Introduction

In the area of statistical learning theory, the Probably Approximately Correct (PAC) learning model formalizes the notion of learning by using sample data points to produce valid hypotheses through algorithms.

Our main new result provides an upper bound on the Fat Shattering dimension of a function class, which consists of functions from a domain X to the unit interval $[0, 1]$, built using a continuous logic connective. An introduction to PAC learning is included in the paper to provide all the necessary prerequisites for stating our result. Hence, we first introduce the PAC learning model applied to learning a concept class $\mathcal{C}$, a collection of subsets of X, and more generally, a function class $\mathcal{F}$. We also explain the Vapnik-Chervonenkis and the Fat Shattering dimensions and cover some known results relating learning under this model to these dimensions.

This paper involves mostly concepts from analysis and some concepts from probability theory; the reader is recommended to have a good understanding of basic notions in measure theory.

Outline of Paper

Section 2 provides a brief overview of measure theory and analysis. In Section 3, we give two definitions of PAC learning, one for a concept class $\mathcal{C}$ and the other for a function class $\mathcal{F}$. Then, in Sections 4 and 5, we explore two combinatorial parameters, the Vapnik-Chervonenkis (VC) dimension and the Fat Shattering dimension of scale ϵ, for $\mathcal{C}$ and $\mathcal{F}$, respectively. We also discuss how these dimensions relate to the PAC learnability of concept and function classes.

In Section 6, as the main original result of our research, given function classes $\mathcal{F}_1, \ldots, \mathcal{F}_k$ and a "continuous logic connective" (that is, a continuous function $u : [0,1]^k \to [0,1]$), we consider the construction of a new composition function class $u(\mathcal{F}_1, \ldots, \mathcal{F}_k)$, consisting of functions $u(f_1, \ldots, f_k)$ defined by

$$u(f_1, \ldots, f_k)(x) = u(f_1(x), \ldots, f_k(x))$$

for $f_i \in \mathcal{F}_i$. We then bound the Fat Shattering dimension of scale ϵ of this class in terms of a sum of the Fat Shattering dimensions of scale $\delta(\epsilon, k)$ of $\mathcal{F}_1, \ldots, \mathcal{F}_k$, where $\delta(\epsilon, k)$ only depends on ϵ and k. There is a previously known analogous estimate for a composition of concept classes built using a usual connective of classical logic [Vid97]. We deduce our new bound using results from Mendelson-Vershynin and Talagrand.

In this paper, any propositions or examples given with proofs, unless mentioned otherwise, are done by us and are independent of any sources.

2 Brief Overview of Measure Theory and Analysis

This section lists some definitions and results in measure theory and analysis, found in standard textbooks, such as [Doo94], [Vid97], and [AC05], which are used in this paper.

Probability Spaces

A *measurable space* $(X, \mathcal{S})$ is a set X equipped with a *σ-algebra* $\mathcal{S}$, a non-empty collection of subsets of X closed under complements and countable unions. If $(X, \mathcal{S})$ and $(Y, \mathcal{T})$ are two measurable spaces, a function $f : X \to Y$ is called *measurable* if $f^{-1}(T) \in \mathcal{S}$ for all $T \in \mathcal{T}$.

Suppose $(X, \mathcal{S})$ is a measurable space; a *measure* is a function $\mu : \mathcal{S} \to \mathbb{R}^+ = \{r \in \mathbb{R} : r \geq 0\}$ satisfying $\mu(\emptyset) = 0$ and

$$\mu\left(\bigcup_{i \in \mathbb{N}} A_i\right) = \sum_{i \in \mathbb{N}} \mu(A_i),$$

for every collection $\{A_i \in S : i \in \mathbb{N}\}$ of pairwise disjoint sets. The triple $(X, \mathcal{S}, \mu)$ is called a *measure space*. If in addition, μ satisfies $\mu(X) = 1$, then μ is a *probability measure* and

$(X, \mathcal{S}, \mu)$ is called a *probability space*.

Given a probability space $(X, \mathcal{S}, \mu)$, one can measure the difference between two subsets $A, B \in \mathcal{S}$ of X by looking at their symmetric difference $A \triangle B = (A \cup B) \setminus (A \cap B)$. More generally, given two measurable functions $f, g : X \to [0, 1]$, one can look at the expected value of their absolute difference by integrating with respect to μ:

$$\int_X |f(x) - g(x)|\, d\mu(x).$$

This paper does not go into any details involving the Lebesgue integral nor does it discuss any integrability or measurability issues; we assume that integration of measurable functions to the real numbers, which is a measure space, makes sense and is linear and order-preserving.

Validating hypotheses in the PAC learning model uses the idea of measuring the symmetric difference of two subsets of a probability space $(X, \mathcal{S}, \mu)$ and calculating the expected value of the difference of $f, g : X \to [0, 1]$. The structure of metric spaces arises naturally from these two notions.

Metric Spaces

A *metric space* (M, d) is a set M equipped with a *metric* $d : M \times M \to \mathbb{R}^+$, which is symmetric and satisfies the triangle inequality and the condition that $d(m_1, m_2) = 0$ if and only if $m_1 = m_2$. Given a metric space (M, d), a *metric sub-space* of M (which is a metric space in its own right) is a nonempty subset $M' \subseteq M$ equipped with the distance $d_{|_{M'}}$, the restriction of d to M'.

A *normed vector space* (V, ρ) is a vector space V over $\mathbb{R}$ equipped with a *norm* $\rho : V \to \mathbb{R}^+$ satisfying

1. $\rho(v_1) = 0$ if and only if $v_1 = 0$
2. $\rho(rv_1) = |r|\rho(v_1)$
3. $\rho(v_1 + v_2) \leq \rho(v_1) + \rho(v_2)$

for all $v_1, v_2 \in V$ and $r \in \mathbb{R}$. The structure of a metric space exists in every normed vector space since the function $d : V \times V \to \mathbb{R}^+$ defined by $d(u, v) = \rho(u - v)$ is always a metric on V. In this case, d is called the *metric induced by the norm* ρ on V.

The following subsection provides a few examples of metric spaces which will be encountered in this paper.

Examples of Metric Spaces

The real numbers $(\mathbb{R}, \rho)$, with the absolute value norm $\rho(r) = |r|$ for $r \in \mathbb{R}$, is a normed vector space so $\mathbb{R}$ can be equipped with the metric $d(r, r') = \rho(r - r') = |r - r'|$. The unit interval $[0, 1]$ is a subset of $\mathbb{R}$, so it is a metric sub-space of $(\mathbb{R}, d)$.

In addition, given a probability space $(X, \mathcal{S}, \mu)$, the set V of all bounded measurable functions from X to $\mathbb{R}$ is a vector space, with point-wise addition and scalar multiplication.

The function $\rho : V \to \mathbb{R}^+$ defined by

$$\rho(f) = \sqrt{\left(\int_X (f(x))^2 d\mu(x)\right)}$$

is a norm on V if any two functions $f, g : X \to \mathbb{R}$ which agree on a subset of X with full measure, $\mu(\{x \in X : f(x) = g(x)\}) = 1$, are identified via an equivalence relation. The norm ρ is called the $L_2(\mu)$ *norm* on V and we normally write $||f||_2 = \rho(f)$ for $f \in V$. As a result, V can be turned into a metric space.

Example 2.1. Following the notations in the paragraph above, V is a metric space with distance d defined by

$$d(f, g) = ||f - g||_2 = \sqrt{\left(\int_X (f(x) - g(x))^2 d\mu(x)\right)}.$$

Write $[0,1]^X$ for the set of all measurable functions from a probability space $(X, \mathcal{S}, \mu)$ to $[0,1]$. Then, it is a metric sub-space of V with distance induced by the $L_2(\mu)$ norm on V, restricted of course to $[0,1]^X$.

Given metric spaces $(M_1, d_1), \dots, (M_k, d_k)$, their product $M_1 \times \dots \times M_k$ always has a natural metric structure, defined as follows.

Example 2.2. If $(M_1, d_1), \dots, (M_k, d_k)$ are metric spaces, then their product $M_1 \times \dots \times M_k$ is a metric space with distance d^2 defined by

$$d^2((m_1, \dots, m_k), (m'_1, \dots, m'_k)) = \sqrt{((d_1(m_1, m'_1))^2 + \dots + (d_k(m_k, m'_k))^2)}.$$

The distance d^2 is normally referred to as the L_2 *product distance* on $M_1 \times \dots \times M_k$.

Consequently, the set $[0,1]^k$, which denotes the set-theoretic product $[0,1] \times \dots \times [0,1]$, is then a metric space with the L_2 product distance. Also, following Examples 2.1 and 2.2, if $\mathcal{F}_1, \dots, \mathcal{F}_k$ are sets of measurable functions from a probability space $(X, \mathcal{S}, \mu)$ to the unit interval, then $\mathcal{F}_i \subseteq [0,1]^X$ for each $i = 1, \dots, k$. Therefore, the product $\mathcal{F}_1 \times \dots \times \mathcal{F}_k$ is a metric space with the L_2 distance as well.

3 The Probably Approximately Correct Model

Let $(X, \mathcal{S})$ be a measurable space. A *concept class* $\mathcal{C}$ on X is a subset of $\mathcal{S}$, and an element $A \in \mathcal{C}$, which is a measurable subset of X, is called a *concept.* A *function class* $\mathcal{F}$ is a collection of measurable functions from X to the unit interval $[0,1]$. Unless stated otherwise, from this section onwards, the following notations will be used:

1. $X = (X, \mathcal{S})$: a *measurable space*
2. μ: a *probability measure* $\mathcal{S} \to \mathbb{R}^+$

3. $\mathcal{C}$: a *concept class* and $\mathcal{F}$: a *function class*

This section provides the definitions of learning $\mathcal{C}$ and $\mathcal{F}$ in the Probably Approximately Correct (PAC) learning model, introduced in 1984 by Valiant.

Concept class PAC learning involves producing a valid hypothesis for every concept $A \in \mathcal{C}$ by first drawing random points, forming a training sample, from X labeled with whether these points are contained in A. In other words, a labeled sample of m points $x_1, \ldots, x_m \in X$ for A consists of these points and the evaluations $\chi_A(x_1), \ldots, \chi_A(x_m)$ of the indicator function $\chi_A : X \to \{0,1\}$, where

$$\chi_A(x) = 1 \text{ if and only if } x \in A.$$

The set of all labeled samples of m points can then be identified with $(X \times \{0,1\})^m$, and producing a hypothesis for A with a labeled sample is exactly the process of associating the sample to a concept $H \in \mathcal{C}$ (i.e. this process is a function from the set of all labeled samples to the concept class).

Here is the precise definition of a concept class being learnable.

Definition 3.1 ([Val84]). A concept class $\mathcal{C}$ is *distribution-free Probably Approximately Correct learnable* if there exists a function (a learning algorithm) $\mathcal{L} : \cup_{m \in \mathbb{N}} (X \times \{0,1\})^m \to \mathcal{C}$ with the following property: for every $\epsilon > 0$, for every $\delta > 0$, there exists a $M \in \mathbb{N}$ such that for every $A \in \mathcal{C}$, for every probability measure μ, for every $m \geq M$, for any $x_1, \ldots, x_m \in X$, we have $\mu(H_m \bigtriangleup A) < \epsilon$ with confidence at least $1 - \delta$, where $H_m = \mathcal{L}((x_1, \chi_A(x_1)), \ldots, (x_m, \chi_A(x_m)))$.

Confidence of at least $1 - \delta$ in the definition above, keeping to the same notations, simply means that the (product) measure of the set of all m-tuples $(x_1, \ldots, x_m) \in X^m$, where $\mu(H_m \bigtriangleup A) < \epsilon$ for $H_m = \mathcal{L}((x_1, \chi_A(x_1)), \ldots, (x_m, \chi_A(x_m)))$, is at least $1 - \delta$. An equivalent statement to $\mathcal{C}$ being distribution-free PAC learnable is that for every $\epsilon, \delta > 0$, there exists $M \in \mathbb{N}$ such that for every $A \in \mathcal{C}$, probability measure μ, and $m \geq M$,

$$\mu^m(\{(x_1, \ldots, x_m) \in X^m : \mu(H_m \bigtriangleup A) \geq \epsilon\}) \leq \delta,$$

for $H_m = \mathcal{L}((x_1, \chi_A(x_1)), \ldots, (x_m, \chi_A(x_m)))$. (The symbol μ^m denotes the product measure on X^m; the reader can refer to [Doo94] for the details.)

A concept class $\mathcal{C}$ is distribution-free learnable in the PAC learning model if a hypothesis H can always be constructed from an algorithm $\mathcal{L}$ for every concept $A \in \mathcal{C}$, using any labeled sample for A, such that the measure of their symmetric difference $H \bigtriangleup A$ is arbitrarily small with respect to every probability measure and with arbitrarily high confidence, as long as the sample size is large enough.

Every concept $A \in \mathcal{C}$ is a subset of X and can be associated to its indicator function $\chi_A : X \to \{0,1\}$. Even more generally, χ_A is a function from X to $[0,1]$; in other words, every concept class $\mathcal{C}$ can be identified as a function class $\mathcal{F}_{\mathcal{C}} = \{\chi_A : X \to [0,1] : A \in \mathcal{C}\}$, so it is natural to generalize Definition 3.1 for any function class $\mathcal{F}$.

Definition 3.1 involves the symmetric difference of two concepts and its generalization to measurable functions $f, g : X \to [0,1]$ is the expected value of their absolute difference

$\mathbb{E}_\mu(f,g)$, as seen in the previous section:

$$\mathbb{E}_\mu(f,g) = \int_X |f(x) - g(x)| \, d\mu(x).$$

A simple exercise can show that if $f, g \in [0,1]^X$ are indicator functions of two concepts $A, B \subseteq X$, then $\mathbb{E}_\mu(f,g)$ coincides with the measure of their symmetric difference: $\mathbb{E}_\mu(f,g) = \mu(A \triangle B)$, where $f = \chi_A$ and $g = \chi_B$.

With this generalization of the symmetric difference, distribution-free PAC learning for any function class can be defined. In the context of function class learning, a labeled sample of m points $x_1, \ldots, x_m \in X$ for a function $f \in \mathcal{F}$ consists of these points and the evaluations $f(x_1), \ldots, f(x_m)$. Then, the set of all labeled samples of m points can be identified with $(X \times [0,1])^m$, and producing a hypothesis is the process of associating a labeled sample to a function $H \in \mathcal{F}$ (just as in concept class learning).

Definition 3.2 ([Vid97]). A function class $\mathcal{F}$ is *distribution-free Probably Approximately Correct learnable* if there exists a function (a learning algorithm) $\mathcal{L} : \cup_{m \in \mathbb{N}} (X \times [0,1])^m \to \mathcal{F}$ with the following property: for every $\epsilon > 0$, for every $\delta > 0$, there exists a $M \in \mathbb{N}$ such that for every $f \in \mathcal{F}$, for every probability measure μ, for every $m \geq M$, for any $x_1, \ldots, x_m \in X$, we have $\mathbb{E}_\mu(H_m, f) < \epsilon$ with confidence at least $1 - \delta$, where $H_m = \mathcal{L}((x_1, f(x_1)), \ldots, (x_m, f(x_m)))$.

Both definitions of PAC learning contain the ϵ and δ parameters. The accuracy error ϵ is used because the hypothesis cannot be, in general, expected to have zero error - only an arbitrarily small error. The risk parameter δ exists because there is no guarantee that any collection of sufficiently large training points leads to a valid hypothesis; the learning algorithm is only expected to produce a valid hypothesis with the sample points with confidence at least $1 - \delta$. Hence, the name “Probably (δ) Approximately (ϵ) Correct” is used [KV94].

The following example illustrates that the set of all axis-aligned rectangles in $\mathbb{R}^2$ is distribution-free PAC learnable. Both the statement and its proof can be found in Chapter 3 of [Vid97] and Chapter 1 of [KV94].

Example 3.1. In $X = \mathbb{R}^2$, the concept class $\mathcal{C} = \{[a,b] \times [c,d] : a,b,c,d \in \mathbb{R}\}$ is distribution-free PAC learnable.

Proof. Let $\epsilon, \delta > 0$. Given a concept A and any sample of m training points $x_1, \ldots, x_m \in X$, define the hypothesis concept H_m to be the intersection of all rectangles containing only training points x_i such that $\chi_A(x_i) = 1$. In other words, H_m is the smallest rectangle that contains only the sample points *in* A.

Let μ be any probability measure, and in fact, $H_m \triangle A = A \setminus H_m$, which can be broken down into four sections $T_1, \ldots, T_4$. If we can conclude that

$$\mu\left(\bigcup_{i=1}^{4} T_i\right) < \epsilon,$$

with confidence at least $1 - \delta$, then the proof is complete.

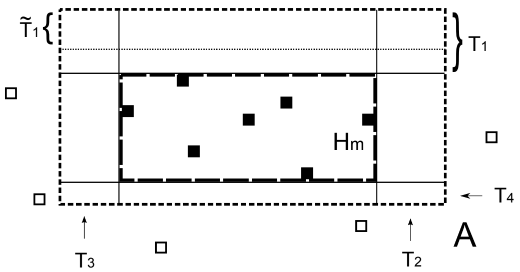

Figure 3.1: Learning an axis-aligned rectangle.

Consider the top section T_1 and define $\tilde{T}_1$ to be the rectangle along the top parts of A whose measure is exactly $\epsilon/4$. The event $\tilde{T}_1 \subseteq T_1$, which is equivalent to $\mu(T_1) \geq \epsilon/4$, holds exactly when no points in the sample $x_1, \ldots, x_m$ fall in $\tilde{T}_1$, and the probability of this event (which is the measure of all such m-tuples of $(x_1, \ldots, x_m) \in X^m$ where $x_i \notin \tilde{T}_1$ for all $i = 1, \ldots, m$) is $(1 - \epsilon/4)^m$. Similarly, the same holds for the other three sections $T_2, \ldots, T_4$. Therefore, the probability that there exists at least one T_i such that $\mu(T_i) \geq \epsilon/4$, where $i \in \{1, \ldots, 4\}$, is at most $4(1 - \epsilon/4)^m$. Hence, as long as we pick m large enough that $4(1 - \epsilon/4)^m \leq \delta$, with confidence (probability) at least $1 - \delta$, $\mu(T_i) < \epsilon/4$ for every $i = 1, \ldots, 4$ and thus,

$$\mu(H_m \triangle A) = \mu\left(\bigcup_{i=1}^{4} T_i\right) \leq \mu(T_1) + \ldots + \mu(T_4) < 4\left(\frac{\epsilon}{4}\right) = \epsilon.$$

Please note that this argument, though very intuitive, actually requires the classical Glivenko-Cantelli theorem, see e.g. [Bil95]. Figure 3.1 provides a visual illustration of the rectangles.

In summary, as long as $m \geq (4/\epsilon)\ln(4/\delta)$, with confidence at least $1 - \delta$, $\mu(H_m \triangle A) < \epsilon$. We note that this estimate of the sample size only depends on ϵ and δ, so $\mathcal{C}$ is indeed distribution-free PAC learnable. □

In the next section, a fundamental theorem which characterizes concept class distribution-free PAC learning will be stated. However, in order to state this theorem, the notion of shattering, which is essential in learning theory, must be introduced.

4 The Vapnik-Chervonenkis Dimension

The Vapnik-Chervonenkis dimension is a combinatorial parameter which is defined using the notion of shattering, developed first in 1971 by Vapnik and Chervonenkis.

Definition 4.1 ([VC71]). Given any set X and a collection $\mathcal{A}$ of subsets of X, the collection $\mathcal{A}$ *shatters* a finite subset $S \subseteq X$ if for every $B \subseteq S$, there exists $A \in \mathcal{A}$ such that $A \cap S = B$.

There is an equivalent condition, which is sometimes easier to work with, to shattering, expressed in terms of characteristic functions of subsets of X.

Proposition 4.1. The collection $\mathcal{A}$ shatters a subset $S = \{x_1, \ldots, x_n\} \subseteq X$ if and only if for every $e = (e_1, \ldots, e_n) \in \{0, 1\}^n$, there exists $A \in \mathcal{A}$ such that $\chi_A(x_i) = e_i$, for all $i = 1, \ldots, n$.

Definition 4.2 ([VC71]). The *Vapnik-Chervonenkis (VC) dimension* of the collection $\mathcal{A}$, denoted by $\mathrm{VC}(\mathcal{A})$, is defined to be the cardinality of the largest finite subset $S \subseteq X$ shattered by $\mathcal{A}$. If $\mathcal{A}$ shatters arbitrarily large finite subsets of X, then the VC dimension of $\mathcal{A}$ is defined to be ∞.

The VC dimension is defined for every collection $\mathcal{A}$ of subsets of any set X, so in particular, $X = (X, \mathcal{S})$ can be a measurable space and $\mathcal{A} = \mathcal{C}$ can be a concept class.

The following is an example, which we believe to be original, illustrating the calculation of the VC dimension for a concept class in the context of $X = \mathbb{R}^n$. In order to prove the VC dimension of a concept class $\mathcal{C}$ is d, we must provide a subset $S \subseteq X$ with cardinality d which is shattered by $\mathcal{C}$ and prove that no subset with cardinality $d+1$ can be shattered by $\mathcal{C}$. The reader can refer to [KV94] and [Pes10b] for more examples on calculating VC dimensions.

Example 4.1. Consider the space $X = \mathbb{R}^n$. A hyperplane $H_{\vec{a},b}$ is defined by a nonzero vector $\vec{a} = (a_1, \ldots, a_n) \in \mathbb{R}^n$ and a scalar $b \in \mathbb{R}$:

$$\begin{aligned} H_{\vec{a},b} &= \{\vec{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n : \vec{x} \cdot \vec{a} = b\} \\ &= \{\vec{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n : x_1 a_1 + \ldots + x_n a_n = b\}. \end{aligned}$$

Write $\mathcal{C}$ as the set of all hyperplanes: $\mathcal{C} = \{H_{\vec{a},b} : \vec{a} \in \mathbb{R}^n \setminus \{\vec{0}\}, b \in \mathbb{R}\}$. Then $\mathrm{VC}(\mathcal{C}) = n$.

Proof. Consider the subset $S = \{\vec{e}_1, \ldots, \vec{e}_n\} \subseteq \mathbb{R}^n$, where $\vec{e}_i$ is the vector with 1 on the i-th component and 0 everywhere else. Suppose $B \subseteq S$ and there are two cases to consider:

1. If $B = \emptyset$, then let $\vec{a} = (1, 1, \ldots, 1) \in \mathbb{R}^n$ and the hyperplane $H_{\vec{a},-1} = \{\vec{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n : x_1 + \ldots + x_n = -1\}$ is disjoint from S.

2. If $B \neq \emptyset$, then set $\vec{a} = (a_1, \ldots, a_n) \in \mathbb{R}^n \setminus \{\vec{0}\}$, where $a_i = \chi_B(\vec{e}_i)$. Then the hyperplane $H_{\vec{a},1} = \{\vec{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n : x_1 a_1 + \ldots + x_n a_n = 1\}$ satisfies

$$H_{\vec{a},1} \cap S = B.$$

Moreover, no subset $S = \{\vec{x}_1, \ldots, \vec{x}_n, \vec{x}_{n+1}\} \subseteq \mathbb{R}^n$ with cardinality $n+1$ can be shattered by $\mathcal{C}$. At best, there exists a unique hyperplane $H_{\vec{a},b}$ containing n of these points, say $\{\vec{x}_1, \ldots, \vec{x}_n\}$, so if $\vec{x}_{n+1} \in H_{\vec{a},b}$, then there are no hyperplanes that include $\vec{x}_1, \ldots, \vec{x}_n$, but not $\vec{x}_{n+1}$. Otherwise, if $\vec{x}_{n+1} \notin H_{\vec{a},b}$, then there are no hyperplanes that include $\vec{x}_1, \ldots, \vec{x}_n, \vec{x}_{n+1}$. □

The VC dimension is central to the PAC learning model for concept classes. In fact, the PAC learnability of a concept class is completely determined by its VC dimension.

4.1 Characterization of Concept Class PAC Learning

The following is one of the main theorems concerning PAC learning, whose proof results from Vapnik and Chervonenkis' paper [VC71] in 1971 and the 1989 paper [BEHW89] by Blumer et al.

Theorem 4.2 ([VC71] and [BEHW89]). Let $\mathcal{C}$ be a concept class of a measurable space $(X, \mathcal{S})$. The following are equivalent:

1. $\mathcal{C}$ is distribution-free Probably Approximately Correct learnable.
2. $\mathrm{VC}(\mathcal{C}) < \infty$.

Both directions of the proof for this result require expressing the number of sample training points required for learning in terms of the VC dimension of $\mathcal{C}$; a crucial lemma used in the proof is Sauer's Lemma, seen in [Sau72]. Given a concept class $\mathcal{C}$ with finite VC dimension, the lemma states that the growth of $|\{A \cap C : C \in \mathcal{C}\}|$ for any finite set A, with $|A| = n$, is bounded above by a polynomial function in n as n grows to infinity.

Using Theorem 4.2, one can more easily determine whether a given concept class is distribution-free PAC learnable.

Example 4.2. The set of all hyperplanes $\mathcal{C} = \{H_{\vec{a},b} : \vec{a} \in \mathbb{R}^n \setminus \{\vec{0}\}, b \in \mathbb{R}\}$, as defined in Example 4.1, is distribution-free PAC learnable.

Every concept class $\mathcal{C}$ can be viewed as a function class $\mathcal{F}_{\mathcal{C}} = \{\chi_A : X \to [0,1] : A \in \mathcal{C}\}$, as seen in Section 3, so a natural question is whether the notion of shattering can be generalized. Indeed, the next section introduces the Fat Shattering dimension of scale ϵ, which is a generalization of the VC dimension.

5 The Fat Shattering Dimension

Let $\epsilon > 0$ from this section onwards. A combinatorial parameter which generalizes the Vapnik-Chervonenkis dimension is the Fat Shattering dimension of scale ϵ, defined first by Kearns and Schapire in 1994.

This dimension, assigned to function classes, involves the notion of *ϵ-shattering*, but similar to the notion of (regular) shattering, it can be defined for any collection of functions $f : X \to [0,1]$, where X is any set. For the sake of this paper, the following sections (still) assume $X = (X, \mathcal{S})$ is a measurable space and the collection of functions is a function class $\mathcal{F}$.

Definition 5.1 ([KS94]). Let $\mathcal{F}$ be a function class. Given a subset $S = \{x_1, \ldots, x_n\} \subseteq X$, the class $\mathcal{F}$ *ϵ-shatters S*, with *witness* $c = (c_1, \ldots, c_n) \in [0,1]^n$, if for every $e \in \{0,1\}^n$, there exists $f \in \mathcal{F}$ such that

$$f(x_i) \geq c_i + \epsilon \text{ for } e_i = 1, \text{ and } f(x_i) \leq c_i - \epsilon \text{ for } e_i = 0.$$

Figure 5.1 illustrates the notion of ϵ-shattering for the subset $S = \{x_1, \ldots, x_6\}$, with witness $c = (c_1, \ldots, c_6)$. Given the binary vector $e = (101011)$, there is a function $f \in \mathcal{F}$ that passes above $c_1 + \epsilon, c_3 + \epsilon, c_5 + \epsilon, c_6 + \epsilon$ at the points x_1, x_3, x_5, x_6, respectively, but passes below $c_2 - \epsilon, c_4 - \epsilon$ at x_2, x_4.

Definition 5.2 ([KS94]). The *Fat Shattering dimension of scale* $\epsilon > 0$ of $\mathcal{F}$, denoted by $\mathrm{fat}_\epsilon(\mathcal{F})$, is defined to be the cardinality of the largest finite subset of X that can be ϵ-shattered by $\mathcal{F}$. If $\mathcal{F}$ can ϵ-shatter arbitrarily large finite subsets, then the Fat Shattering dimension of scale ϵ of $\mathcal{F}$ is defined to be ∞.

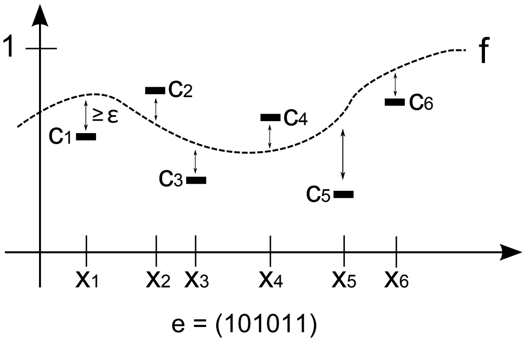

Figure 5.1: Diagram of ϵ-shattering.

When the function class $\mathcal{F}$ consists of only functions taking values in $\{0,1\}$, then the Fat Shattering dimension of any scale $\epsilon \leq 1/2$ of $\mathcal{F}$ agrees with the VC dimension of the corresponding collection of subsets of X, induced by the (indicator) functions in $\mathcal{F}$.

With the generalization from a concept class to a function class, a natural question is whether the finiteness of the Fat Shattering dimension of all scales ϵ for a function class $\mathcal{F}$ is equivalent to $\mathcal{F}$ being distribution-free PAC learnable. This question is addressed in the following subsection.

5.1 Sufficient Condition for Function Class PAC Learning

One direction of Theorem 4.2 can be generalized and stated in terms of the Fat Shattering dimension of scale ϵ of a function class.

Theorem 5.1 ([ABDCBH97] and [Vid97]). Let $\mathcal{F}$ be a function class. If $\mathrm{fat}_\epsilon(\mathcal{F}) < \infty$ for all $\epsilon > 0$, then $\mathcal{F}$ is distribution-free PAC learnable.

However, the converse to Theorem 5.1 is false. There exists a distribution-free PAC learnable function class with infinite Fat Shattering dimension of some scale ϵ.

In fact, for every concept class $\mathcal{C}$ with cardinality $\aleph_0$ or $2^{\aleph_0}$, there is an associated function class $\mathcal{F}_\mathcal{C}$ defined as follows. Set up a bijection $b : \mathcal{C} \to [0,1/3]$ or to $[0,1/3] \cap \mathbb{Q}$, depending on the cardinality of $\mathcal{C}$, and for every $A \in \mathcal{C}$, define a function $f_A : X \to [0,1]$ by

$$f_A(x) = \chi_A(x) + (-1)^{\chi_A(x)} b(A).$$

Now, write $\mathcal{F}_\mathcal{C} = \{f_A : A \in \mathcal{C}\}$. Note that $\mathcal{F}_\mathcal{C}$ can be thought of the collection of all indicator functions of $A \in \mathcal{C}$, except that each "indicator" function f_A has two unique identifying points $b(A)$ and $1 - b(A)$, instead of simply 0 and 1. The following proposition provides many counterexamples to the converse of Theorem 5.1, which are much simpler than the one found in [Vid97].

The construction of the function class $\mathcal{F}_\mathcal{C}$ and the proposition below are developed from an idea of Example 2.10 in [Pes10a].

Proposition 5.2. Let $\mathcal{C}$ be a concept class. The associated function class $\mathcal{F}_{\mathcal{C}} = \{f_A : A \in \mathcal{C}\}$, defined in the previous paragraph, is always distribution-free PAC learnable; this class has infinite Fat Shattering dimension of all scales $\epsilon < 1/6$ if $\mathcal{C}$ has infinite VC dimension.

Proof. The function class $\mathcal{F}_{\mathcal{C}}$ is distribution-free PAC learnable because every function $f_A \in \mathcal{F}_{\mathcal{C}}$ can be uniquely identified with just one point $x_0 \in X$ in any labeled sample: $f_A(x_0) \in \{b(A), 1 - b(A)\}$ uniquely determines A and thus, f_A.

Furthermore, suppose $\mathcal{C}$ has infinite VC dimension. Let $n \in \mathbb{N}$ be arbitrary and because $\mathrm{VC}(\mathcal{C}) = \infty$, there exists $S = \{x_1, \ldots, x_n\}$ such that $\mathcal{C}$ shatters S. Suppose $\epsilon < 1/6$ and we claim that $\mathcal{F}_{\mathcal{C}}$ ϵ-shatters S with witness $c = (0.5, \ldots, 0.5) \in [0,1]^n$. Indeed, let $e \in \{0,1\}^n$ and there exists $A \in \mathcal{C}$ such that

$$\chi_A(x_i) = e_i,$$

for all $i = 1, \ldots, n$, by Proposition 4.1. As a result,

$$f_A(x_i) = 1 - b(A) \geq 0.5 + \epsilon \text{ for } e_i = 1$$

and

$$f_A(x_i) = b(A) \leq 0.5 - \epsilon \text{ for } e_i = 0.$$

Consequently, $\mathcal{F}_{\mathcal{C}}$ has infinite Fat Shattering dimension of all scales $\epsilon < 1/6$. □

One research topic we would like to consider in the future is to come up with a new combinatorial parameter for a function class, related to the notion of shattering, which would characterize PAC distribution-free learning. This new parameter would have to solve the problem of unique identifications of functions, a problem that does not occur with concept classes.

The next section explains the main result of our research: bounding the Fat Shattering dimension of scale ϵ of a composition function class which is built with a continuous logic connective.

6 The Fat Shattering Dimension of a Composition Function Class

The goals of this section are to construct a new function class from old ones by means of a continuous logic connective and to bound the Fat Shattering dimension of scale ϵ of the new function class in terms of the dimensions of the old ones. The following subsection provides this construction, which can be found in Chapter 4 of [Vid97], in the context of concept classes using a connective of classical logic.

6.1 A Review of the Construction in the Context of Concept Classes

Let $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k$ be concept classes, where $k \geq 2$, and let $u : \{0,1\}^k \to \{0,1\}$ be any function, commonly known as a connective of classical logic. A new collection of subsets of X arises from $\mathcal{C}_1, \ldots, \mathcal{C}_k$ as follows.

As mentioned earlier in this paper, every element $A \in \mathcal{C}_i$ can be identified as a binary function $f : X \to \{0,1\}$, namely its characteristic function $f = \chi_A$, and vice versa. Now, for any k functions $f_1, \ldots, f_k : X \to \{0,1\}$, where $f_i \in \mathcal{C}_i$ with $i = 1, \ldots, k$, consider a new function $u(f_1, \ldots, f_k) : X \to \{0,1\}$ defined by

$$u(f_1, \ldots, f_k)(x) = u(f_1(x), \ldots, f_k(x)).$$

The set of all possible $u(f_1, \ldots, f_k)$, denoted by $u(\mathcal{C}_1, \ldots, \mathcal{C}_k)$, is given by

$$u(\mathcal{C}_1, \ldots, \mathcal{C}_k) = \{u(f_1, \ldots, f_k) : f_i \in \mathcal{C}_i\}.$$

For instance, when $k = 2$, we can consider the "Exclusive Or" connective $\oplus : \{0,1\}^2 \to \{0,1\}$ defined by

$$p \oplus q = (p \wedge \neg q) \vee (\neg p \wedge q),$$

which corresponds to the symmetric difference operation. Then, our new concept class constructed from $\mathcal{C}_1$ and $\mathcal{C}_2$ is

$$\{A_1 \triangle A_2 : A_1 \in \mathcal{C}_1, A_2 \in \mathcal{C}_2\}.$$

The next known theorem states that if $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k$ all have finite VC dimension to start with, then regardless of u, the new collection $u(\mathcal{C}_1, \ldots, \mathcal{C}_k)$ always has finite VC dimension.

Theorem 6.1 ([Vid97]). Let $k \geq 2$. Suppose $\mathcal{C}_1, \ldots, \mathcal{C}_k$ are concept classes, each viewed as a collection of binary functions, and $u : \{0,1\}^k \to \{0,1\}$ is any function. If the VC dimension of $\mathcal{C}_i$ is finite for all $i = 1, \ldots, k$. Then there exists a constant $\alpha = \alpha_k$, which depends only on k, such that

$$\mathrm{VC}(u(\mathcal{C}_1, \ldots, \mathcal{C}_k)) < d\alpha_k,$$

where $d = \max\limits_{i=1}^{k} \mathrm{VC}(\mathcal{C}_i)$.

The proof of this theorem can be found in [Vid97] and uses Sauer's Lemma to bound the VC dimension of $u(\mathcal{C}_1, \ldots, \mathcal{C}_k)$. The main objective of our research is to generalize this theorem for function classes, in terms of the Fat Shattering dimension of scale ϵ, but the connective of classical logic u would have to be replaced by a *continuous logic connective*, which is simply a continuous function $u : [0,1]^k \to [0,1]$.

6.2 Construction of New Function Class with Continuous Logic Connective

In first-order logic, there are only two truth-values 0 or 1, so a connective is a function $\{0,1\}^k \to \{0,1\}$ in the classical sense. However, in continuous logic, truth-values can be found anywhere in the unit interval $[0,1]$. Therefore, we should consider a function $u : [0,1]^k \to [0,1]$, which will transform function classes, and require that u be a continuous logic connective.

In other words, u should be continuous from the (product) metric space $[0,1]^k$ to the unit interval [YBHU08]; in fact, because u is continuous from a compact metric space to a metric space, it is automatically uniformly continuous.

The following provides the definition of a uniformly continuous function u from any metric space to another, but we must first qualify u with a modulus of uniform continuity.

Definition 6.1 (See e.g. [YBHU08]). A *modulus of uniform continuity* is any function $\delta : (0,1] \to (0,1]$.

Definition 6.2 (See e.g. [YBHU08]). Let (M_1, d_1) and (M_2, d_2) be two metric spaces. A function $u : M_1 \to M_2$ is *uniformly continuous* if there exists (a modulus of uniform continuity) $\delta : (0,1] \to (0,1]$ such that for all $\epsilon \in (0,1]$ and $m_1, m_2 \in M_1$, if $d_1(m_1, m_2) < \delta(\epsilon)$, then $d_2(u(m_1), u(m_2)) < \epsilon$.

Such a δ is called a *modulus of uniform continuity for u*.

Given function classes $\mathcal{F}_1, \ldots, \mathcal{F}_k$ and a uniformly continuous function $u : [0,1]^k \to [0,1]$, consider the new function class $u(\mathcal{F}_1, \ldots, \mathcal{F}_k)$ defined by

$$u(\mathcal{F}_1, \ldots, \mathcal{F}_k) = \{u(f_1, \ldots, f_k) : f_i \in \mathcal{F}_i\},$$

where $u(f_1, \ldots, f_k)(x) = u(f_1(x), \ldots, f_k(x))$ for all $x \in X$, just as in Section 6.1 for concept classes, with $f_i \in \mathcal{F}_i$ and $i = 1, \ldots, k$. Our main result states that the Fat Shattering dimension of scale ϵ of $u(\mathcal{F}_1, \ldots, \mathcal{F}_k)$ is bounded by a sum of the Fat Shattering dimensions of scale $\delta(\epsilon, k)$ of $\mathcal{F}_1, \ldots, \mathcal{F}_k$, where $\delta(\epsilon, k)$ is a function of the modulus of uniform continuity $\delta(\epsilon)$ for u and k. It is a known result, seen in Chapter 5 of [Vid97], that this new class $u(\mathcal{F}_1, \ldots, \mathcal{F}_k)$ has finite Fat Shattering dimension of all scales $\epsilon > 0$ (and thus, it is distribution-free PAC learnable) if each of $\mathcal{F}_1, \ldots, \mathcal{F}_k$ has finite Fat Shattering dimension of all scales, but no bounds were previously known.

6.3 Main Result

Fix $k \geq 2$ and the following theorem is our main new result.

Theorem 6.2. Let $\epsilon > 0$, $\mathcal{F}_1, \ldots, \mathcal{F}_k$ be function classes of X, and $u : [0,1]^k \to [0,1]$ be a uniformly continuous function with modulus of continuity $\delta(\epsilon)$. Then

$$\mathrm{fat}_\epsilon(u(\mathcal{F}_1, \ldots, \mathcal{F}_k)) \leq \left(\frac{K \log(4c'k\sqrt{k}/(\delta(\epsilon/(2c'))\epsilon))}{K' \log(2)} \right) \sum_{i=1}^{n} \mathrm{fat}_{c \frac{\delta(\epsilon/(2c'))\epsilon}{k\sqrt{k}}}(\mathcal{F}_i),$$

where c, c', K, K' are some absolute constants.

Extracting the actual values of these absolute constants is not easy, and we hope to find them in future research. For this reason, comparing the bound in Theorem 6.2 with the existing estimate for the VC dimension of a composition concept class is difficult; however, in statistical learning theory, estimates for function class learning are generally much worse than estimates for concept class learning.

In order to prove Theorem 6.2, for clarity, we will introduce an auxiliary function $\phi : \mathcal{F}_1 \times \ldots \times \mathcal{F}_k \to [0,1]^X$ and prove the following.

Lemma 6.1. Let $\epsilon > 0$. If $u : [0,1]^k \to [0,1]$ is uniformly continuous with modulus of continuity $\delta(\epsilon)$, then the function $\phi : \mathcal{F}_1 \times \ldots \times \mathcal{F}_k \to [0,1]^X$ defined by

$$\phi(f_1, \ldots, f_k)(x) = u(f_1(x), \ldots, f_k(x))$$

is also uniformly continuous with modulus of continuity $\frac{\delta(\epsilon/2)\epsilon}{2k}$, from the metric space $\mathcal{F}_1 \times \ldots \times \mathcal{F}_k$ with distance $\tilde{d}^2$ to $[0,1]^X$. Also, $\phi(\mathcal{F}_1 \times \ldots \times \mathcal{F}_k) = u(\mathcal{F}_1, \ldots, \mathcal{F}_k)$, where the symbol $\phi(\mathcal{F}_1 \times \ldots \times \mathcal{F}_k)$ simply represents the image of ϕ.

Then, we will prove the next lemma, and our main result will follow directly.

Lemma 6.2. Let $\epsilon > 0$, $\mathcal{F}_1, \ldots, \mathcal{F}_k$ be function classes of X, and $\phi : \mathcal{F}_1 \times \ldots \times \mathcal{F}_k \to [0,1]^X$ be uniformly continuous with some modulus of continuity $\delta(\epsilon, k)$, a function of ϵ and k. Then

$$\mathrm{fat}_{c'\epsilon}(\phi(\mathcal{F}_1 \times \ldots \times \mathcal{F}_k)) \leq \left(\frac{K \log(2\sqrt{k}/\delta(\epsilon, k))}{K' \log(2)} \right) \sum_{i=1}^{k} \mathrm{fat}_{c \frac{\delta(\epsilon,k)}{\sqrt{k}}}(\mathcal{F}_i),$$

where c, c', K, K' are some absolute constants.

6.4 Proofs

This subsection provides all the proofs for our main theorem.

Proof of Lemma 6.1. Suppose $u : [0,1]^k \to [0,1]$ is uniformly continuous with a modulus of continuity $\delta(\epsilon)$, where $[0,1]^k$ is a metric space with the L_2 product distance d^2. We claim that the function $\phi : \mathcal{F}_1 \times \ldots \times \mathcal{F}_k \to [0,1]^X$ defined by

$$\phi(f_1, \ldots, f_k)(x) = u(f_1(x), \ldots, f_k(x))$$

is uniformly continuous with modulus of continuity $\frac{\delta(\epsilon/2)\epsilon}{2k}$. Let $\epsilon > 0$ and

$$(f_1, \ldots, f_k), (f'_1, \ldots, f'_k) \in \mathcal{F}_1 \times \ldots \times \mathcal{F}_k.$$

Suppose

$$\begin{aligned}\tilde{d}^2((f_1, \ldots, f_k), (f'_1, \ldots, f'_k)) &= \sqrt{((||f_1 - f'_1||_2)^2 + \ldots + (||f_k - f'_k||_2)^2)} \\ &< \frac{\delta(\epsilon/2)\epsilon}{2k} = \sqrt{\frac{\delta(\epsilon/2)^2(\epsilon/2)^2}{k^2}}.\end{aligned}$$

Hence, for each $i = 1, \ldots, k$,

$$||f_i - f'_i||_2 = \sqrt{\left(\int_X (f_i(x) - f'_i(x))^2 \, d\mu(x) \right)} < \sqrt{\frac{\delta(\epsilon/2)^2(\epsilon/2)^2}{k^2}}.$$

Write $A_i = \{x \in X : |f_i(x) - f_i'(x)| \geq \sqrt{\frac{\delta(\epsilon/2)^2}{k}}\}$ and we must have that $\mu(A_i) < \frac{(\epsilon/2)^2}{k}$, for each $i = 1, \ldots, k$. Otherwise,

$$\begin{aligned}
\int_X (f_i(x) - f_i'(x))^2 \, d\mu(x) &= \int_{A_i} (f_i(x) - f_i'(x))^2 \, d\mu(x) + \int_{X \setminus A_i} (f_i(x) - f_i'(x))^2 \, d\mu(x) \\
&\geq \int_{A_i} \left(\sqrt{\frac{\delta(\epsilon/2)^2}{k}}\right)^2 d\mu(x) + \int_{X \setminus A_i} (f_i(x) - f_i'(x))^2 \, d\mu(x) \\
&= \mu(A_i) \left(\sqrt{\frac{\delta(\epsilon/2)^2}{k}}\right)^2 + \int_{X \setminus A_i} (f_i(x) - f_i'(x))^2 \, d\mu(x) \\
&\geq \frac{(\epsilon/2)^2}{k} \frac{\delta(\epsilon/2)^2}{k} + \int_{X \setminus A_i} (f_i(x) - f_i'(x))^2 \, d\mu(x) \\
&\geq \frac{\delta(\epsilon/2)^2 (\epsilon/2)^2}{k^2},
\end{aligned}$$

which is a contradiction. Now, write $A = A_1 \cup \ldots \cup A_k$ and we have that $X \setminus A = \{x \in X : |f_i(x) - f_i'(x)| < \sqrt{\frac{\delta(\epsilon/2)^2}{k}},$ for all $i = 1, \ldots, k\}$. Suppose $x \in X \setminus A$ and then

$$\begin{aligned}
d^2((f_1(x), \ldots, f_k(x)), (f_1'(x), \ldots, f_k'(x))) &= \sqrt{|f_1(x) - f_1'(x)|^2 + \ldots + |f_k(x) - f_k'(x)|^2} \\
&< \sqrt{\left(\frac{\delta(\epsilon/2)^2}{k} + \ldots + \frac{\delta(\epsilon/2)^2}{k}\right)} < \delta(\epsilon/2).
\end{aligned}$$

Consequently, by the uniform continuity of u, for all $x \in X \setminus A$,

$$|u(f_1(x), \ldots, f_k(x)) - u(f_1'(x), \ldots, f_k'(x))| < \epsilon/2.$$

Finally,

$$\begin{aligned}
||\phi(f_1, \ldots, f_k) - \phi(f_1', \ldots, f_k')||_2 &= \sqrt{\left(\int_X (u(f_1(x), \ldots, f_k(x)) - u(f_1'(x), \ldots, f_k'(x)))^2 \, d\mu(x)\right)} \\
&\leq \sqrt{\left(\int_{X \setminus A} (u(f_1(x), \ldots, f_k(x)) - u(f_1'(x), \ldots, f_k'(x)))^2 \, d\mu(x)\right)} \\
&+ \sqrt{\left(\int_A (u(f_1(x), \ldots, f_k(x)) - u(f_1'(x), \ldots, f_k'(x)))^2 \, d\mu(x)\right)} \\
&< \sqrt{\left(\int_{X \setminus A} (\epsilon/2)^2 \, d\mu(x)\right)} + \sqrt{\left(\int_A 1 \, d\mu(x)\right)} \\
&\leq (\epsilon/2) + (\epsilon/2) = \epsilon,
\end{aligned}$$

as $\mu(A) \leq \sum_{i=1}^{k} \mu(A_i) \leq k\left(\frac{(\epsilon/2)^2}{k}\right) = (\epsilon/2)^2$. □

Now, in order to prove Lemma 6.2, we first introduce the concept of an ϵ-covering number for any metric space, based on [MV03], and relate this number for a function class to its Fat Shattering dimension of scale ϵ by using results from Mendelson and Vershynin [MV03] and Talagrand [Tal03].

Definition 6.3. Let $\epsilon > 0$ and suppose (M, d) is a metric space. The *ϵ-covering number*, denoted by $N(M, \epsilon, d)$, of M is the minimal number N such that there exists elements $m_1, m_2, \ldots, m_N \in M$ with the property that for all $m \in M$, there exists $i \in \{1, 2, \ldots, N\}$ for which

$$d(m, m_i) < \epsilon.$$

The set $\{m_1, m_2, \ldots, m_N\}$ is called a *(minimal) ϵ-net* of M.

The following proposition relates the ϵ-covering number of a product of metric spaces, with the L_2 product distance d^2, $M_1 \times \ldots \times M_k$ to the $\frac{\epsilon}{\sqrt{k}}$-covering number of each space M_i.

Proposition 6.3. Let $\epsilon > 0$ and suppose $(M_1, d_1), \ldots, (M_k, d_k)$ are metric spaces, each with finite $\frac{\epsilon}{\sqrt{k}}$-covering numbers, $N_i = N(M_i, \frac{\epsilon}{\sqrt{k}}, d_i)$ for $i = 1, \ldots, k$. Then

$$N(M_1 \times \ldots \times M_k, \epsilon, d^2) \leq \prod_{i=1}^{k} N_i.$$

Proof. Let $C_i = \{a_1^i, \ldots, a_{N_i}^i\}$ be a minimal $\frac{\epsilon}{\sqrt{k}}$-net for M_i with respect to distance d_i, where $i = 1, \ldots, k$ and suppose $(a^1, \ldots, a^k) \in M_1 \times \ldots \times M_k$. Then, for each $i = 1, \ldots, k$, there exists $a_{j_i}^i \in C_i$, where $1 \leq j_i \leq N_i$ such that $d_i(a^i, a_{j_i}^i) < \frac{\epsilon}{\sqrt{k}}$. Hence,

$$\begin{aligned} d^2((a^1, \ldots, a^k), (a_{j_1}^1, \ldots, a_{j_k}^k)) &= \sqrt{\left((d_1(a^1, a_{j_1}^1))^2 + \ldots + (d_k(a^k, a_{j_k}^k))^2\right)} \\ &< \sqrt{\left(\left(\frac{\epsilon}{\sqrt{k}}\right)^2 + \ldots + \left(\frac{\epsilon}{\sqrt{k}}\right)^2\right)} = \epsilon, \end{aligned}$$

where each $(a_{j_1}^1, \ldots, a_{j_k}^k) \in C_1 \times \ldots \times C_k$, which has cardinality $\Pi_{i=1}^k N_i$. Therefore, $N(M_1 \times \ldots \times M_k, \epsilon, d^2) \leq \Pi_{i=1}^k N_i$. □

Also, if $u : M_1 \to M_2$ is any uniformly continuous function with a modulus of uniform continuity $\delta(\epsilon)$ from any metric space to another, then the image of a minimal $\delta(\epsilon)$-net of M_1 under u becomes an ϵ-net for $u(M_1)$.

Proposition 6.4. Let $\epsilon > 0$ and suppose (M_1, d_1) and (M_2, d_2) are two metric spaces. If a function $u : M_1 \to M_2$ is uniformly continuous with a modulus of continuity $\delta(\epsilon)$, then $N(u(M_1), \epsilon, d_2) \leq N(M_1, \delta(\epsilon), d_1)$, where $u(M_1)$ denotes the image of u.

Proof. Suppose $N = N(M_1, \delta(\epsilon), d_1)$ is the $\delta(\epsilon)$-covering number for M_1 and let $\{m_1, \ldots, m_N\}$ be a $\delta(\epsilon)$-net for M_1. Hence for every $u(m) \in u(M_1)$, where $m \in M_1$, there exists $i \in \{1, \ldots, N\}$ such that

$$d_1(m, m_i) < \delta(\epsilon),$$

which implies $d_2(u(m), u(m_i)) < \epsilon$ as u is uniformly continuous. As a result, the set

$$\{u(m_1), \ldots, u(m_N)\}$$

is an ϵ-net for $u(M_1)$, so $N(u(M_1), \epsilon, d_2) \leq N(M_1, \delta(\epsilon), d_1)$. □

In particular, we can view $\mathcal{F}_1, \ldots, \mathcal{F}_k$ as metric spaces, all with distances induced by the $L_2(\mu)$ norm and suppose $\phi : \mathcal{F}_1 \times \ldots \times \mathcal{F}_k \to [0,1]^X$ is uniformly continuous with modulus of continuity $\delta(\epsilon, k)$. Then, by Proposition 6.3, if $\mathcal{F}_1, \ldots, \mathcal{F}_k$ all have finite $\frac{\delta(\epsilon,k)}{\sqrt{k}}$-covering numbers, the metric space $\mathcal{F}_1 \times \ldots \times \mathcal{F}_k$, with the L_2 product metric $\tilde{d}^2$, also has a finite $\delta(\epsilon, k)$-covering number: if we write $N(\mathcal{F}_i, \frac{\delta(\epsilon,k)}{\sqrt{k}}, L_2(\mu))$ as the $\frac{\delta(\epsilon,k)}{\sqrt{k}}$-covering number for $\mathcal{F}_i$, then,

$$N(\mathcal{F}_1 \times \ldots \times \mathcal{F}_k, \delta(\epsilon, k), \tilde{d}^2) \leq \prod_{i=1}^{k} N(\mathcal{F}_i, \frac{\delta(\epsilon, k)}{\sqrt{k}}, L_2(\mu)).$$

Now, by Proposition 6.4,

$$\begin{aligned} N(\phi(\mathcal{F}_1 \times \ldots \times \mathcal{F}_k), \epsilon, L_2(\mu)) &\leq N(\mathcal{F}_1 \times \ldots \times \mathcal{F}_k, \delta(\epsilon, k), \tilde{d}^2) \\ &\leq \prod_{i=1}^{k} N(\mathcal{F}_i, \frac{\delta(\epsilon, k)}{\sqrt{k}}, L_2(\mu)). \end{aligned}$$

In other words, the ϵ-covering number for $\phi(\mathcal{F}_1 \times \ldots \times \mathcal{F}_k)$ is bounded by a product of the $\frac{\delta(\epsilon,k)}{\sqrt{k}}$-covering numbers of each $\mathcal{F}_i$. To prove Lemma 6.2, we now state the main theorem of a paper written by Mendelson and Vershynin, which relates the ϵ-covering number of a function class to its Fat Shattering dimension of scale ϵ.

Theorem 6.5 ([MV03]). Let $\epsilon > 0$ and let $\mathcal{F}$ be a function class. Then for every probability measure μ,

$$N(\mathcal{F}, \epsilon, L_2(\mu)) \leq \left(\frac{2}{\epsilon}\right)^{K\text{fat}_{c\epsilon}(\mathcal{F})}$$

for absolute constants c, K.

And Talagrand provides the converse.

Theorem 6.6 ([Tal03]). Following the notations of Theorem 6.5, there exists a probability measure μ such that

$$N(\mathcal{F}, \epsilon, L_2(\mu)) \geq 2^{K'\text{fat}_{c'\epsilon}(\mathcal{F})},$$

for absolute constants c', K'.

Proof of Lemma 6.2. By Propositions 6.3 and 6.4,

$$N(\phi(\mathcal{F}_1 \times \ldots \times \mathcal{F}_k), \epsilon, L_2(\mu)) \leq \prod_{i=1}^{k} N(\mathcal{F}_i, \frac{\delta(\epsilon, k)}{\sqrt{k}}, L_2(\mu)),$$

so

$$\log(N(\phi(\mathcal{F}_1 \times \ldots \times \mathcal{F}_k), \epsilon, L_2(\mu))) \leq \sum_{i=1}^{k} \log(N(\mathcal{F}_i, \frac{\delta(\epsilon, k)}{\sqrt{k}}, L_2(\mu))).$$

By Theorem 6.5,

$$\log N(\mathcal{F}_i, \frac{\delta(\epsilon, k)}{\sqrt{k}}, L_2(\mu)) \leq K \text{fat}_{c\frac{\delta(\epsilon,k)}{\sqrt{k}}}(\mathcal{F}_i) \log(2\sqrt{k}/\delta(\epsilon, k)),$$

for any probability measure μ where c, K are absolute constants. Moreover, by Theorem 6.6 for some probability measure μ and absolute constants c', K',

$$\log(N(\phi(\mathcal{F}_1 \times \ldots \times \mathcal{F}_k), \epsilon, L_2(\mu))) \geq K' \text{fat}_{c'\epsilon}(\phi(\mathcal{F}_1 \times \ldots \times \mathcal{F}_k)) \log(2)$$

and altogether,

$$\begin{aligned} \text{fat}_{c'\epsilon}(\phi(\mathcal{F}_1 \times \ldots \times \mathcal{F}_k)) &\leq \frac{\sum_{i=1}^{k} K \text{fat}_{c\frac{\delta(\epsilon,k)}{\sqrt{k}}}(\mathcal{F}_i) \log(2\sqrt{k}/\delta(\epsilon, k))}{K' \log(2)} \\ &= \left(\frac{K \log(2\sqrt{k}/\delta(\epsilon, k))}{K' \log(2)} \right) \sum_{i=1}^{k} \text{fat}_{c\frac{\delta(\epsilon,k)}{\sqrt{k}}}(\mathcal{F}_i). \end{aligned}$$

□

Finally, we will prove our main theorem.

Proof of Theorem 6.2. By Lemma 6.1, if $u : [0,1]^k \to [0,1]$ is uniformly continuous with modulus of continuity $\delta(\epsilon)$, then $\phi : \mathcal{F}_1 \times \ldots \times \mathcal{F}_k \to [0,1]^X$ defined by

$$\phi(f_1, \ldots, f_k)(x) = u(f_1(x), \ldots, f_k(x))$$

is also uniformly continuous with modulus of continuity $\frac{\delta(\epsilon/2)\epsilon}{2k}$. Then, apply Lemma 6.2 with $\delta(\epsilon, k) = \frac{\delta(\epsilon/2)\epsilon}{2k}$ and with a simple change of variables $c'\epsilon' \to \epsilon$, Theorem 6.2 follows directly. □

Altogether, we can summarize the maps in this section in the following two diagrams (where i is the diagonal map):

$$\begin{tikzcd} X \arrow[r, "i"] & X^k \arrow[r, "f_1 \times \ldots \times f_k"] & {[0,1]^k} \arrow[r, "u"] & {[0,1]} \end{tikzcd},$$

while

$$\begin{tikzcd} \mathcal{F}_1 \times \ldots \times \mathcal{F}_k \arrow[r, "\phi"] & {[0,1]^X} \end{tikzcd}.$$

This result is potentially useful because it allows us to construct new function classes using common continuous logic connectives and bound their Fat Shattering dimensions of scale ϵ. For instance, the function $u : [0,1]^2 \to [0,1]$ defined by $u(r_1, r_2) = r_1 \cdot r_2$ (multiplication) is uniformly continuous with a modulus of continuity $\delta(\epsilon) = \frac{\epsilon}{2}$. Indeed, let $\epsilon > 0$ and consider $(r_1, r_2), (r_1', r_2') \in [0,1]^2$. Suppose $d^2((r_1, r_2), (r_1', r_2')) < \delta(\epsilon) = \frac{\epsilon}{2}$, so

$$|r_1 - r_1'| < \sqrt{|r_1 - r_1'|^2 + |r_2 - r_2'|^2} < \frac{\epsilon}{2}$$

and similarly, $|r_2 - r_2'| < \frac{\epsilon}{2}$. Then,

$$\begin{aligned} |u(r_1, r_2) - u(r_1', r_2')| &= |r_1 r_2 - r_1' r_2'| \\ &= |r_1 r_2 - r_1 r_2' + r_1 r_2' - r_1' r_2'| \\ &\le |r_1(r_2 - r_2')| + |r_2'(r_1 - r_1')| \\ &\le |r_2 - r_2'| + |r_1 - r_1'| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon. \end{aligned}$$

As a result, if $\mathcal{F}_1$ and $\mathcal{F}_2$ are two function classes with finite Fat Shattering dimensions of some scale ϵ, then the function class $u(\mathcal{F}_1, \mathcal{F}_2) = \mathcal{F}_1\mathcal{F}_2 = \{f_1 \cdot f_2 : f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2\}$, defined by point-wise multiplication, also has finite Fat Shattering dimension of scale ϵ, up to some constant factor, and Theorem 6.2 provides an upper bound.

We have made an interesting connection, which has not been explored much in the past, between continuous logic and PAC learning, and we plan to investigate this connection even further. For instance, the relationship of compositions of function classes and continuous logic may be interesting to study because compositions of uniformly continuous functions are again uniformly continuous. Furthermore, we can try to add some topological structures to concept or function classes to see how PAC learning can be affected.

7 Acknowledgements

I would like to thank Professor Vladimir Pestov from the University of Ottawa for all his help and support and for introducing me to learning theory. The work in this paper was partially funded by the NSERC Undergraduate Student Research Award in the fall of 2010. I would like to thank the reviewers of *The Waterloo Mathematics Review* for their thorough reading of the paper and their feedback.

This paper is based on the author's undergraduate honours project, supervised by Professor Vladimir Pestov and completed in April 2011 at the University of Ottawa.

References

[ABDCBH97] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler, *Scale-sensitive dimensions, uniform convergence, and learnability*, Journal of the ACM **44** (1997), no. 4, 615–631.

[AC05] G. Auliac and J. Y. Caby, *Mathématiques: Topologie et analyse*, 3rd ed., EdiScience, Belgium, 2005.

[BEHW89] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. Warmuth, *Learnability and the Vapnik-Chervonenkis Dimension*, Journal of the ACM **36** (1989), no. 4, 929 – 965.

[Bil95] P. Billingsley, *Probability and Measure*, 3rd ed., Wiley-Interscience, New York, 1995.

[Doo94] J. L. Doob, *Measure Theory*, Springer-Verlag, New York, 1994.

[KS94] M. J. Kearns and R. Schapire, *Efficient Distribution-free Learning of Probabilistic Concepts*, Journal of Computer System Sciences **48** (1994), no. 3, 464–497.

[KV94] M. J. Kearns and U. V. Vazirani, *An Introduction to Computational Learning Theory*, The MIT Press, Cambridge, Massachusetts, 1994.

[MV03] S. Mendelson and R. Vershynin, *Entropy and the Combinatorial Dimension*, Inventiones Mathematicae **152** (2003), 37 – 55.

[Pes10a] V. Pestov, *A Note on Sample Complexity of Learning Binary Output Neural Networks Under Fixed Input Distributions*, Proc. 2010 Eleventh Brazilian Symposium on Neural Networks, IEEE Computer Society, Los Alamitos-Washington-Tokyo (2010), 7 – 12.

[Pes10b] ______, *Indexability, Concentration, and VC Theory*, Proc. of the 3rd International Conf. on Similarity Search and Applications (SISAP 2010) (2010), 3 – 12.

[Sau72] N. Sauer, *On the Densities of Families of Sets*, J. Combinatorial Theory **13** (1972), 145 – 147.

[Tal03] M. Talagrand, *Vapnik-Chervonenkis Type Conditions and Uniform Donsker Classes of Functions*, Annals of Probability **31** (2003), no. 3, 1565 – 1582.

[Val84] L. G. Valiant, *A Theory of the Learnable*, Communications of the ACM **27** (1984), no. 11, 1134 – 1142.

[VC71] V. N. Vapnik and A. Y. Chervonenkis, *On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities*, Theory of Prob. and its Appl. **16** (1971), no. 2, 264 – 280.

[Vid97] M. Vidyasagar, *A Theory of Learning and Generalization: With Applications to Neural Networks and Control Systems*, Springer-Verlag London Limited, London, 1997.

[YBHU08] I. B. Yaacov, A. Berenstein, C. W. Henson, and A. Usvyatsov, *Model Theory for Metric Structures*, London Math Society Lecture Note Series **350** (2008), 315 – 427.