Measurability Aspects of the Compactness Theorem for Sample Compression Schemes

Damjan Kalajdzievski

Thesis submitted to the Faculty of Graduate and Postdoctoral Studies
in partial fulfillment of the requirements for the degree of Master of Science in Mathematics [1]

Department of Mathematics and Statistics
Faculty of Science
University of Ottawa

[1] The M.Sc. program is a joint program with Carleton University, administered by the Ottawa-Carleton Institute of Mathematics and Statistics

Abstract

In 1998, it was proved by Ben-David and Litman that a concept space has a sample compression scheme of size d if and only if every finite subspace has a sample compression scheme of size d. In the compactness theorem, measurability of the hypotheses of the created sample compression scheme is not guaranteed; at the same time measurability of the hypotheses is a necessary condition for learnability.

In this thesis we discuss when a sample compression scheme, created from compression schemes on finite subspaces via the compactness theorem, have measurable hypotheses. We show that if X is a standard Borel space with a d-maximum and universally separable concept class $\mathcal{C}$, then $(X, \mathcal{C})$ has a sample compression scheme of size d with universally Borel measurable hypotheses. Additionally we introduce a new variant of compression scheme called a copy sample compression scheme.

Acknowledgements

I would like to thank my supervisor Dr. Pestov for all that he has taught me and for his time and support. I am also grateful to have had the Ontario Graduate Scholarship as financial support, and the financial support of the Department of Mathematics at the University of Ottawa. I would like to express my thanks to both thesis examiners, Dr. J. Levy and Dr. S. Zilles, for their careful reading of the thesis, and suggesting numerous improvements.

Contents

Introduction **1**

1 Vapnik-Chervonenkis Dimension **5**

1.1 Vapnik-Chervonenkis Dimension 5

1.2 Maximum and Maximal Classes 8

1.3 Concepts as Relations 12

2 Sample Compression Schemes **16**

2.1 Introduction of Sample Compression Schemes 16

2.2 Compactness Theorem 19

2.3 Sample Compression Schemes and VC Dimension 21

2.4 Extended Sample Compression Schemes 22

3 Learnability **27**

3.1 Learnability 27

3.2 Sample Complexity of Compression Schemes 30

4 Measurability of Sample Compression Schemes and Learning Rules **37**

4.1 Measurability of Compression Schemes 37

Open Questions **45**

A **46**

A.1 Measure Theory Preliminaries 46

A.2 Nets and Filters Preliminaries 48

B **50**

B.1 Well Behaved Hypothesis Spaces and Other Forms of Measurability 50

B.2 Sample Complexity of Copy Sample Compression Schemes 53

Bibliography **64**

Introduction

The common context for Statistical Learning Theory is the setting of a concept space. A concept space is a set X and a family $\mathcal{C}$ of subsets of X called concepts. The goal of Statistical Learning Theory is to be able to, given a hidden concept C in $\mathcal{C}$, "learn" the concept by sampling finite subsets of X labelled according to their membership in C. Learning is usually provided by a function or algorithm which takes a labelled sample as an input, and outputs a subset of X called a hypothesis. The function or algorithm is said to be consistent if the hypothesis, for a given sample labelled according to a concept, agrees with the labelling on this sample. The distance between the hypothesis and the target concept can be quantified in varying ways depending on the model of "learning".

The PAC, or "Probably Approximately Correct", model for learning was introduced by Valiant in the 80's [18]. A concept space $(X, \mathcal{C})$ is PAC learnable with respect to a family $\mathcal{P}$ of probability measures on X if there exists a function, called a "learning rule", mapping labelled samples in X to subsets of X, where for any $0 < \varepsilon \leq 1,\ 0 < \delta \leq 1$, there is a positive integer m such that given any $C \in \mathcal{C}$, $P \in \mathcal{P}$, the probability (according to P) of sample of size greater than m being mapped to a hypothesis with error from C greater than ε, is less than δ (error of a hypothesis from a concept is meant to be the probability (according to P) of their symmetric difference). In particular if $\mathcal{P}$ is the family of all probability measures on

X, one speaks of distribution-free PAC learnability, and this is the context that will be of exclusive interest to us.

Learnability is intimately related to the concept of VC dimension, or "Vapnik-Chervonenkis Dimension", introduced by Vapnik and Chervonenkis in [19]. VC dimension is a parameter quantifying the combinatorial complexity of a concept space defined from the idea of a subset of X being "shattered" by concepts. $A \subseteq X$ is shattered by $\mathcal{C}$ if any subset of A is equal to some concept in $\mathcal{C}$ intersected with A; the VC dimension of a concept space is the supremum of the cardinalities of all finite subsets which are shattered by $\mathcal{C}$. In 1986 Blumer et al. related VC dimension to PAC learnability by proving that (given a measurability condition called "well behaved") a concept space has finite VC dimension if and only if it is PAC learnable if and only if any consistent learning rule which provides concepts as hypotheses, learns the concept space with sample sizes (sample complexity) bounded above by a formula involving the VC dimension of the concept space. The bounds mentioned are improved by Shawe-Taylor et al. in [17] who assumed stronger measurability conditions.

A natural class of consistent learning algorithms are sample compression schemes which were introduced by Littlestone and Warmuth in 1986 [12]. A sample compression scheme of size d maps subsets (labelled or unlabelled depending on the variant) of X of size at most d, to hypotheses, such that each sample labelled according to a concept has a subsample of size at most d which is mapped to a hypothesis agreeing with the concept on the initial sample. A sample compression scheme of size d can be thought to save every sample labelled according to a concept to some subsample of size at most d. Also in [12] it is shown that, given measurability conditions of the sample compression scheme and concept space, every sample compression scheme is a PAC learning rule with sample complexity bounded above by a formula involving the size of the sample compression scheme (the bounds are due to Floyd and Warmuth in [8]). For a concept space of VC dimension d, a sample compression scheme of size d learns with bounds on sample complexity better than that of the bounds for general

consistent learning rules provided by the VC dimension (illustrated in Figure 3.1 in chapter 3). In chapter 2 we define our own variant of sample compression scheme called "copy sample compression schemes", of which sample compression schemes are a special case. A copy sample compression scheme can be thought of as an algorithm which checks sample compression schemes of varying sizes, and for different concept classes, and picks a hypothesis for the concepts in any of these different concept classes. Copy sample compression schemes also add other flexibilities as will be exhibited in chapter 3, and may have better bounds in some instances for sample complexity than that of sample compression schemes.

A natural open question posed in [12] is whether or not a concept space with VC dimension d also has a sample compression scheme of size $O(d)$. The question currently remains open. The existence of an unlabelled sample compression scheme of size d does imply that the concept space has VC dimension at most d, and in some cases like that of maximum classes, VC dimension d is enough to provide a sample compression scheme of size d [10].

In [2], it was proved by Ben-David and Litman, using a proof based on the compactness theorem of predicate logic, that a concept space $(X, \mathcal{C})$ has a sample compression scheme of size d if and only if every finite subspace has a sample compression scheme of size d ($(Y, \mathcal{C}')$ is a finite subspace if $Y \subseteq X$ is finite and $\mathcal{C}' = \{C \cap Y : C \in \mathcal{C}\}$). We provide a different, and technically simpler, proof of this using an approach with ultralimits normally used in Analysis. In either case the result, named the "compactness theorem" for sample compression schemes, did not take any measurability considerations into account, and as our example in chapter 4 shows the resulting hypotheses need not be measurable; chapter 4 of this thesis explores when the compression scheme resulting from the compactness theorem has measurable hypotheses. Perhaps the most useful of these results is that, when X is a standard Borel space with a d-maximum and universally separable concept class $\mathcal{C}$, then $(X, \mathcal{C})$ has a sample compression scheme of size d with universally Borel

measurable hypotheses. In the appendix B.1 we also collect differing measurability conditions rules from varying relevant papers. These conditions are defined to be utilized in proving different bounds for sample complexity of consistent learning rules and for sample complexity of sample compression schemes in these various papers.

This Thesis starts in chapter 1 by outlining VC dimension and important concepts such as the concepts of maximality due to [20] and embeddability due to [2]. In chapter 2 sample compression schemes and their variants are presented and discussed. In Chapter 3 we introduce PAC learnability, and investigate sample complexity for spaces with finite VC dimension and sample complexity for spaces with sample compression schemes or copy sample compression schemes. Finally in chapter 4 we investigate measurability of hypotheses from sample compression schemes generated by the compactness theorem. We also include two appendices, with appendix A consisting of some preliminaries, and appendix B consisting of some excluded proofs and a list of varying technical measurability conditions which are referenced throughout the thesis.

Chapter 1

Vapnik-Chervonenkis Dimension

1.1 Vapnik-Chervonenkis Dimension

We begin with the definitions of a concept space and the VC dimension associated to a concept space.

Definition 1.1.1 *A **concept space** is a pair $(X, \mathcal{C})$ consisting of a set X equipped with a set $\mathcal{C}$ of subsets of X. X is referred to as the **domain**, and $\mathcal{C}$ is referred to as the **concept class**. For a subset A of X, denote*

$$\mathcal{C} \sqcap A = \{C \cap A : C \in \mathcal{C}\},$$

*and we say that $(Y, \mathcal{C}')$ is a **subspace** of $(X, \mathcal{C})$ if $Y \subseteq X$ and $\mathcal{C}' = \mathcal{C} \sqcap Y$.*

Definition 1.1.2 ([19]) *We say that a subset A of X is **shattered** by $\mathcal{C}$ if $\mathcal{C} \sqcap A = 2^A$.*

Definition 1.1.3 ([19]) *The **Vapnik-Chervonenkis dimension** or **VC-dimension** of $(X, \mathcal{C})$ (denoted $\mathrm{VC}(X, \mathcal{C})$, or $\mathrm{VC}(\mathcal{C})$ when X is understood) is*

$$\mathrm{VC}(\mathcal{C}) = \sup\{|A| : A \subseteq X, A \text{ is finite, } A \text{ is shattered by } \mathcal{C}\}.$$

In particular if the value is infinite, we say $\mathrm{VC}(\mathcal{C}) = \infty$.

The following are some elementary or well known examples of VC dimension which can be found in every text on statistical learning.

Example 1.1.4 Let X be any infinite set and $\mathcal{C} = 2^X$, Then clearly $\mathrm{VC}(X, \mathcal{C}) = \infty$ because every (finite) $A \subseteq X$ has $\mathcal{C} \sqcap A = 2^A$ and so A is shattered.

Example 1.1.5 Let X be any totally ordered set with at least two elements, and let

$$\mathcal{C} = \{I_x : x \in X\} \cup \{\emptyset\},$$

where $I_x = \{y \in X : y \leq x\}$ is an initial segment of $(X, <)$. For any $x, y \in X$ where $x \neq y$, without loss of generality $y < x$, we have

$$\mathcal{C} \sqcap \{x, y\} = \{\emptyset, \{y\}, \{x, y\}\},$$

hence $\{y\}$ is shattered, however $\{x\} \notin \mathcal{C} \sqcap \{x, y\}$ and so $\{x, y\}$ is not shattered. Therefore $\mathrm{VC}(X, \mathcal{C}) = 1$.

Example 1.1.6 Let $X = \mathbb{R}^2$ and

$$\mathcal{C} = \{[a, b] \times [c, d] : a, b, c, d \in \mathbb{R}\}.$$

Clearly $\mathcal{C}$ shatters $\{(-1, 0), (0, 1), (0, 1), (0, -1)\}$. Now let

$$A = \{(a_1, a_2), (b_1, b_2), (c_1, c_2), (d_1, d_2), (e_1, e_2)\}$$

be given. Without loss of generality (a_1, a_2) is the leftmost point, (b_1, b_2) is the highest point, (c_1, c_2) is the rightmost point, and (d_1, d_2) is the lowest point. Since

$$\{(a_1, a_2), (b_1, b_2), (c_1, c_2), (d_1, d_2)\} \subseteq [a, b] \times [c, d] \in \mathcal{C},$$

we have

$$(e_1, e_2) \in [a_1, c_1] \times [d_2, b_2] \subseteq [a, b] \times [c, d],$$

and so

$$\{(a_1, a_2), (b_1, b_2), (c_1, c_2), (d_1, d_2)\} \notin \mathcal{C} \sqcap A.$$

Therefore $\mathrm{VC}(X, \mathcal{C})$=4.

Unless otherwise specified, from now on we will consider $(X, \mathcal{C})$ to be our concept space, and $d, k, m, n \in \mathbb{N}$.

Definition 1.1.7 ([19]) *The **n'th shatter coefficients** of $\mathcal{C}$ are defined to be*

$$s(\mathcal{C}, n) = max\{|\mathcal{C} \sqcap A| : A \subseteq X,\ |A| = n\}.$$

Note that $\mathrm{VC}(\mathcal{C}) = \sup\{n \in \mathbb{N} : s(\mathcal{C}, n) = 2^n\}$.

Notation 1.1.8 Let $\binom{n}{\leq d}$ denote

$$\binom{n}{\leq d} = \sum_{i=0}^{d} \binom{n}{i}.$$

Theorem 1.1.9 (Sauer-Shelah Lemma [16]) *Let $\mathrm{VC}(\mathcal{C}) = d$. Then $\forall n \in \mathbb{N}$*

$$s(\mathcal{C}, n) \leq \binom{n}{\leq d} \leq \left(\frac{en}{d}\right)^d.$$

We can consider $\mathcal{C}$ as "function class"; a family of $\{0, 1\}$ valued functions on X: Let

$$\mathcal{F}_{\mathcal{C}} = \{\chi_C : C \in \mathcal{C}\}$$

where χ_C is the indicator function of C on X. Similarly, if $\mathcal{F}$ is a family of $\{0, 1\}$ valued functions on X we can get a concept class

$$\mathcal{C}_{\mathcal{F}} = \{C \in 2^X : \chi_C = f, \text{ for some } f \in \mathcal{F}\}.$$

Defining shattering for a function class $\mathcal{F}$ as: $A \subseteq X$ is shattered by $\mathcal{F}$ if $\{f_{|A} : f \in \mathcal{F}\} = 2^A$. We can see that $\mathcal{F}$ shatters A iff $\mathcal{C}_{\mathcal{F}}$ shatters A, and $\mathcal{C}$ shatters A iff $\mathcal{F}_{\mathcal{C}}$ shatters A so the two notions are equivalent.

In the future we will consider concepts as functions, but will still use set relations and operations on concepts, which will have the obvious meaning; for instance $x \in C$ will be the same as $C(x) = 1$, $C \subseteq C'$ the same as $\text{support}(C) \subseteq \text{support}(C')$, $C \cap C'$ the same as $\min\{C, C'\}$, etc.

1.2 Maximum and Maximal Classes

The following definitions are due to [20].

Definition 1.2.1 *Let $d \in \mathbb{N}$. A concept class $\mathcal{C}$ is **d-maximum** if for every $A \subseteq X$ finite,*

$$|\mathcal{C} \sqcap A| = \binom{|A|}{\leq d}.$$

Definition 1.2.2 *A concept class $\mathcal{C}$ is **d-maximal** if $\text{VC}(\mathcal{C}) = d$, and for any $D \in 2^X \setminus \mathcal{C}$ we have $\text{VC}(\mathcal{C} \cup \{D\}) > d$.*

Note that if $\mathcal{C}$ is d-maximum, then $\text{VC}(\mathcal{C}) = d$ because for $A \subseteq X$, if $|A| = d$ then

$$|\mathcal{C} \sqcap A| = \binom{d}{\leq d} = 2^d = 2^{|A|},$$

so A is shattered, and if $|A| > d$ then

$$|\mathcal{C} \sqcap A| = \binom{|A|}{\leq d} < 2^{|A|},$$

so A is not shattered.

As a consequence of Zorn's Lemma every concept class of VC dimension d is contained in a d-maximal concept class.

Maximum does not necessarily imply maximal and vice versa. Also note that if $(X, \mathcal{C})$ is d-maximum, any subspace of $(X, \mathcal{C})$ is d-maximum as well, but this is not necessarily the case for d-maximal.

Example 1.2.3 Let $X = \{1, 2, 3, 4\}$,
$\mathcal{C} = \{\{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 4\}, \{2, 4\}, \{3, 4\}, \{1, 2, 3\}\}$. It is easy to check $\mathcal{C}$ is 2-maximal but not 2-maximum since

$$|\mathcal{C}| = 10 < 11 = \binom{4}{\leq 2}.$$

Example 1.2.4 ([8]) Let $X = \{1, 2, 3, 4\}$,
$\mathcal{C} = \{\{1\}, \{2\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 4\}, \{2, 4\}, \{3, 4\}, \{1, 2, 3\}, \{1, 2, 4\}\}$. It is easy to check $\mathcal{C}$ is 2-maximal but not 2-maximum since

$$|\mathcal{C}| = 10 < 11 = \binom{4}{\leq 2}.$$

Example 1.2.5 Let $X = \mathbb{R}$ and $\mathcal{C} = \{(-\infty, a) : a \in \mathbb{Q}\}$. For any $A = \{x_1, ..., x_n\} \subseteq X$ finite, without loss of generality with $x_1 < ... < x_n$, we have that

$$|\mathcal{C} \sqcap A| = |\{\emptyset, \{x_1\}, \{x_1, x_2\}, ..., \{x_1, ..., x_n\}\}| = |A| + 1 = \binom{|A|}{\leq 1},$$

thus $\mathcal{C}$ is 1-maximum. However, $\mathcal{C}$ is not 1-maximal since $\mathrm{VC}(\mathcal{C} \cup \{X\}) = 1$. Note that any concept space where X is totally ordered with no minimal element, and where $\mathcal{C}$ is the set of all initial segments, is 1-maximum. This is also the case if X has at least two elements, where $\mathcal{C}$ is the set of all initial segments and the empty set.

Remark 1.2.6 If $(X, \mathcal{C})$ is finite, then d-maximum implies d-maximal.
If $\mathcal{C}$ is d-maximum, then any $A \in 2^X \setminus \mathcal{C}$ has

$$|\mathcal{C} \cup \{A\}| = |\mathcal{C}| + 1 = \binom{|X|}{\leq d} + 1 > \binom{|X|}{\leq d}$$

hence by Sauer's Lemma $\mathrm{VC}(\mathcal{C} \cup \{A\}) > d$, and therefore $\mathcal{C}$ is d-maximal.

Lemma 1.2.7 ([20]) *Let $(X, \mathcal{C})$ be finite with VC-dimension d. For $x \in X$, there are at most $\binom{|X|-1}{\leq d-1}$ sets $C \in \mathcal{C}$ such that $x \in C$ and $C \setminus \{x\} \in \mathcal{C}$.*

Proof: Let $x_0 \in X$, $Y = X \setminus \{x_0\}$, and

$$\mathcal{C}' = \{C \in \mathcal{C} : x_0 \in C \text{ and } C \setminus \{x_0\} \in \mathcal{C}\}.$$

Suppose

$$|\mathcal{C}'| > \binom{|X|-1}{\leq d-1}.$$

Then

$$|\mathcal{C}' \sqcap Y| = |\mathcal{C}'| > \binom{|X|-1}{\leq d-1} = \binom{|Y|}{\leq d-1},$$

thus by Sauer's Lemma $\mathrm{VC}(Y, \mathcal{C}' \sqcap Y) > d-1$. Let $\{x_1, ..., x_d\}$ be d points in Y shattered by $\mathcal{C}' \sqcap Y \subseteq \mathcal{C}$, and let $A = \{x_0, x_1, ..., x_d\}$. Now by the definition of $\mathcal{C}'$, for each $C \in \mathcal{C}' \sqcap \{x_1, ..., x_d\} = 2^{\{x_1,...,x_d\}}$ there is $C_{x_0} \in \mathcal{C}$ such that $C_{x_0} = C \cup \{x_0\}$, hence

$$\mathcal{C} \sqcap A \supset 2^{\{x_1,...,x_d\}} \cup \{B \cup \{x_0\} : B \in 2^{\{x_1,...,x_d\}}\} = 2^A,$$

contradicting $\mathrm{VC}(\mathcal{C}) = d$. ∎

Theorem 1.2.8 ([20]) *Let $(X, \mathcal{C})$ be finite with VC-dimension d. The concept space $(X, \mathcal{C})$ is d-maximum if and only if*

$$|\mathcal{C}| = \binom{|X|}{\leq d}.$$

Proof: If $(X, \mathcal{C})$ is d-maximum then by the definition

$$|\mathcal{C}| = \binom{|X|}{\leq d}.$$

For the converse, we will use induction on $|X| = n \geq d$.

If $n = d$, then $\mathcal{C} = 2^X$ is maximum and

$$|\mathcal{C}| = 2^d = \binom{d}{\leq d}.$$

Assume the statement of the theorem is true for all $(X, \mathcal{C})$ where $|X| \leq n$, and let $(X, \mathcal{C})$ have $|X| = n + 1$. Let $x_0 \in X$ and let $Y = X \setminus \{x_0\}$. By the induction hypothesis, it suffices to show that

$$|\mathcal{C} \sqcap Y| = \binom{n}{\leq d}.$$

By lemma 1.2.7,
$\mathcal{C}' = \{C \in \mathcal{C} : x_0 \in C \text{ and } C \setminus \{x_0\} \in \mathcal{C}\}$ has size at most $\binom{n}{\leq d-1}$. Define

$$\pi : \mathcal{C} \setminus \mathcal{C}' \to \mathcal{C} \sqcap Y \text{ by } \pi(C) = C \cap Y.$$

We will show π is injective. Suppose there is $C_1 \neq C_2$ in $\mathcal{C} \setminus \mathcal{C}'$ such that

$$\pi(C_1) = C_1 \cap Y = C_2 \cap Y = \pi(C_2).$$

If $x_0 \in C_1 \cap C_2$, then

$$C_1 = (C_1 \cap Y) \cup \{x_0\} = (C_2 \cap Y) \cup \{x_0\} = C_2,$$

and if $x_0 \notin C_1 \cup C_2$, then

$$C_1 = C_1 \cap Y = C_2 \cap Y = C_2,$$

so without loss of generality $x_0 \in C_1 \setminus C_2$. We get that

$$C_1 \setminus \{x_0\} = C_1 \cap Y = C_2 \cap Y = C_2 \in \mathcal{C}$$

hence $C_1 \in \mathcal{C}'$, a contradiction, therefore π is injective. Finally,

$$|\mathcal{C} \sqcap Y| \geq |\mathcal{C} \setminus \mathcal{C}'| = |\mathcal{C}| - |\mathcal{C}'| \geq \binom{n+1}{\leq d} - \binom{n}{\leq d-1} = \binom{n}{\leq d}.$$

∎

1.3 Concepts as Relations

In this section we will look at concept spaces defined as a relation on a pair of sets. This will allow us to characterize useful notions of embeddings for concept spaces as found in [2]. It will also allow us to define the dual concept space of a concept space.

We can define a concept class on a domain X via a relation $R \subseteq X \times Y$ for some set Y, by $\mathcal{C}_R = \{C_y : y \in Y\}$ where $C_y = \{x \in X : (x, y) \in R\}$. Similarly given $(X, \mathcal{C})$, the corresponding space in the form (X, Y, R) is $(X, \mathcal{C}, \in)$. A subclass of (X, Y, R) is $(A, B, R_{|_{A \times B}})$ where $A \subseteq X$, and $B \subseteq Y$. This is convenient for defining the idea of a dual to a concept space as follows:

Definition 1.3.1 *Given a concept space (X, Y, R), the **dual concept space** of (X, Y, R), denoted*

$$(X, Y, R)^*,$$

is

$$(Y, X, R^*), \text{ where } R^* = \{(y, x) : (x, y) \in R\}.$$

The dual concept space of a space represented as $(X, \mathcal{C})$, can be thought of as

$$(\mathcal{C}, \{\{C \in \mathcal{C} : x \in C\} : x \in X\}).$$

Definition 1.3.2 ([2]) *Let (X, Y, R), (X', Y', R') be concept spaces. An **embedding** from (X, Y, R) to (X', Y', R') is a function $\pi : X \times Y \to X' \times Y'$ such that for every*

$$(x, y) \in X \times Y, \ (x, y) \in R \text{ iff } \pi((x, y)) \in R'.$$

*A **generalized embedding** from (X, Y, R) to (X', Y', R') is a function $\tau \in 2^X$ and a function $\pi : X \times Y \to X' \times Y'$ such that for every $(x, y) \in X \times Y$,*

$$\text{if } \tau(x) = 0 \text{ then } (x, y) \in R \text{ iff } \pi((x, y)) \in R',$$

$$\textit{if } \tau(x) = 1 \textit{ then } (x,y) \in R \textit{ iff } \pi((x,y)) \notin R'.$$

(X,Y,R) *is* ***weakly (generalized) embeddable*** *in* (X',Y',R') *if every finite subclass* $(A,B,R_{|_{A\times B}})$ *of* (X,Y,R) *is (generalized) embeddable in* (X',Y',R').

The above notions partially order any set of concept spaces; if there exists an embedding or generalized embedding from (X,Y,R) to (X',Y',R'), we will denote that

$$(X,Y,R) \preceq_{emb} (X',Y',R')$$

or

$$(X,Y,R) \preceq_{gemb} (X',Y',R')$$

respectively.
If (X,Y,R) is weakly embeddable in (X',Y',R'), or weakly generalized embeddable in (X',Y',R'), we will denote that

$$(X,Y,R) \preceq^{w}_{emb} (X',Y',R')$$

or

$$(X,Y,R) \preceq^{w}_{gemb} (X',Y',R')$$

respectively.

Definition 1.3.3 *Let us say that* (X,Y,R) *and* (X',Y',R') *are* ***bi-embeddable*** *if* $(X,Y,R) \preceq_{emb} (X',Y',R')$ *and* $(X',Y',R') \preceq_{emb} (X,Y,R)$.

A concept space (X,Y,R) may have some redundant points in $X \times Y$ as far as R is concerned, but we can reduce it to its essential information by setting:

$$x \sim x' \text{ in } X \text{ iff } \forall y \in Y,\ (x,y) \in R \iff (x',y) \in R,$$
$$y \sim y' \text{ in } Y \text{ iff } \forall x \in X,\ (x,y) \in R \iff (x,y') \in R.$$

$$R_{\sim} = \{([x]_{\sim},[y]_{\sim}) \in X/\sim \times Y/\sim : (x,y) \in R\}$$

separates the points of $X/\sim \times Y/\sim$ and $(X/\sim, Y/\sim, R_\sim)$ is bi-embeddable to (X, Y, R) via the quotient map for $(X, Y, R) \preceq_{emb} (X/\sim, Y/\sim, R_\sim)$, and mapping each equivalence class to its (choose any) representative for
$(X/\sim, Y/\sim, R_\sim) \preceq_{emb} (X, Y, R)$.

Remark 1.3.4

(1) $(X, Y, R) \preceq_{emb} (X', Y', R') \Rightarrow (X, Y, R) \preceq_{gemb} (X', Y', R') \Rightarrow (X, Y, R) \preceq^w_{gemb} (X', Y', R')$.

(2) $(X, Y, R) \preceq^w_{emb} (X', Y', R') \Rightarrow (X, Y, R) \preceq^w_{gemb} (X', Y', R')$.

Notation 1.3.5 In the proof of the next proposition and throughout the further text we use the notation $\triangle$ for symmetric difference of a set; i.e. $V \triangle W := (V \backslash W) \cup (W \backslash V)$

Proposition 1.3.6 ([2]) *If* $(X, Y, R) \preceq^w_{gemb} (X', Y', R')$ *then* $\mathrm{VC}(X, Y, R) \leq \mathrm{VC}(X', Y', R')$.

Proof: Let A be a finite subset of X that is shattered, let

$$B = \{b_D \in Y : D \subseteq A,\ C_{b_D} \cap A = D\},$$

and let τ, $\pi = (\pi_1, \pi_2)$ be the generalized embedding from (A, B, R) into (X', Y', R'). π_2 is injective because for $b_1, b_2 \in B,\ b_1 \neq b_2$, there exists $x \in C_{b_1} \setminus C_{b_2} \cup C_{b_2} \setminus C_{b_1}$. Without loss of generality $x \in C_{b_1} \setminus C_{b_2}$. We have:

if $\tau(x) = 0,$ then $x \in C_{b_1}$ implies $\pi_1(x) \in C_{\pi_2(b_1)}$ and $x \notin C_{b_2}$ implies $\pi_1(x) \notin C_{\pi_2(b_2)}$;

if $\tau(x) = 1,$ then $x \in C_{b_1}$ implies $\pi_1(x) \notin C_{\pi_2(b_1)}$ and $x \notin C_{b_2}$ implies $\pi_1(x) \in C_{\pi_2(b_2)}$.

In either case $\pi_1(x) \in C_{\pi_2(b_1)} \triangle C_{\pi_2(b_2)}$ and so $\pi_2(b_2) \neq \pi_2(b_1)$. This also shows that $C_b \mapsto C_{\pi_2(b)} \cap \pi_1(A)$ is injective, hence

$$2^{|A|} \geq |\{C_{\pi_2(b)} \cap \pi_1(A) : b \in B\}| \geq |\{C_b : b \in B\}| = 2^{|A|}$$

and therefore $\pi_1(A)$ is shattered in (X', Y', R'). ∎

Theorem 1.3.7 ([11]) *For any class* (X, Y, R)*:*

$$\log_2(\mathrm{VC}(X, Y, R)) - 1 < \mathrm{VC}((X, Y, R)^*) < 2^{\mathrm{VC}(X,Y,R)+1}.$$

Proof: Since $((X, Y, R)^*)^* = (X, Y, R)$, it suffices to show the first inequality. Let A be a set of cardinality $\lfloor \log_2(\mathrm{VC}(X, Y, R)) \rfloor$. One has $(A, 2^A, \in) \preceq_{emb} (2^A, 2^{2^A}, \in)^*$ via $\pi(x, y) = (\{B \subseteq A : x \in B\}, y)$. Noting that $(2^A, 2^{2^A}, \in)$ is embeddable in any class of the same or greater VC-dimension, $(2^A, 2^{2^A}, \in) \preceq_{emb} (X, Y, R)$, and thus $(2^A, 2^{2^A}, \in)^* \preceq_{emb} (X, Y, R)^*$. Therefore $(A, 2^A, \in) \preceq_{emb} (X, Y, R)^*$ and so $\lfloor \log_2(\mathrm{VC}(X, Y, R)) \rfloor \leq \mathrm{VC}((X, Y, R)^*)$. ∎

Corollary 1.3.8 $\mathrm{VC}(X, Y, R) < \infty$ *if and only if* $\mathrm{VC}((X, Y, R)^*) < \infty$.

Chapter 2

Sample Compression Schemes

2.1 Introduction of Sample Compression Schemes

Sample compression schemes, introduced by Littlestone and Warmuth ([12]), are naturally arising algorithms which learn concepts by saving finite samples of concepts to subsets of size at most d.

The following notations will be used in the definitions of sample compression schemes, and throughout the text.

Notation 2.1.1 For $d \in \mathbb{N} \cup \{\infty\}$ let

$$[X]^{<d} = \{A \subseteq X : |A| < d\},$$

let

$$\mathcal{C}_{|_A} = \{C_{|_A} : C \in \mathcal{C}\},$$

where $A \subseteq X$ and $C_{|_A}$ is the function C restricted to the domain A, and let

$$\mathcal{C}_{|_{[X]^{<d}}} = \{C_{|_A} : C \in \mathcal{C},\ A \subseteq X,\ |A| < d\}.$$

We can similarly define

$$[X]^{\leq d},\ \mathcal{C}_{|_{[X]\leq d}},\ [X]^{=d},\ \text{and}\ \mathcal{C}_{|_{[X]=d}}.$$

Notation 2.1.2 For two functions f,g with $\mathrm{dom}(g) \subseteq \mathrm{dom}(f)$, let

$$g \sqsubseteq f$$

be the notation for f extending g.

Definition 2.1.3 *For $d \in \mathbb{N}$, an **unlabelled sample compression scheme of size d** on $(X, \mathcal{C})$ is a function*

$$\mathcal{H} : [X]^{\leq d} \to 2^X$$

with the property that

$$\forall f \in \mathcal{C}_{|_{[X]<\infty}},\ \exists \sigma \in [\mathrm{dom}(f)]^{\leq d},\ \textit{such that}\ f \sqsubseteq \mathcal{H}(\sigma).$$

*A **labelled sample compression scheme of size d** on $(X, \mathcal{C})$ is a function*

$$\mathcal{H} : \mathcal{C}_{|_{[X]\leq d}} \to 2^X$$

with the property that

$$\forall f \in \mathcal{C}_{|_{[X]<\infty}},\ \exists g \in \mathcal{C}_{|_{[X]\leq d}},\ \textit{such that}\ g \sqsubseteq f \sqsubseteq \mathcal{H}(g).$$

*We will call the range of a sample compression scheme the **hypothesis class** and denote it by $\mathfrak{H}$.*

Example 2.1.4 Let X be any totally ordered set, and let $\mathcal{C} = \{I_x : x \in X\}$ be the set of all initial segments of X. Defining

$$\mathcal{H} : \{x\} \mapsto I_x,\ \emptyset \mapsto \emptyset,$$
$$\text{and}\ \mathcal{H}' : \{x\} \mapsto I_x \setminus \{x\},\ \emptyset \mapsto X,$$

we will show $\mathcal{H}$ and $\mathcal{H}'$ are unlabelled sample compression schemes of size 1 on $(X, \mathcal{C})$. Given a sample $f \in \mathcal{C}_{|_{[X]<\infty}}$, if $f = 0$ on its domain then $\emptyset \in [\mathrm{dom}(f)]^{\leq 1}$ and $f \sqsubseteq \mathcal{H}(\emptyset) = \emptyset$. Otherwise $x_f = \max\{x : f(x) = 1\}$ exists, and so

$$\{x_f\} \in [\mathrm{dom}(f)]^{\leq 1},\ f \sqsubseteq \mathcal{H}(\{x_f\}) = I_{x_f}.$$

Thus $\mathcal{H}$ is a sample compression scheme of size 1 on $(X, \mathcal{C})$.
Similarly for $\mathcal{H}'$, if $f = 1$ on its domain then $\emptyset \in [\mathrm{dom}(f)]^{\leq 1}$ and $f \sqsubseteq \mathcal{H}(\emptyset) = X$. Otherwise $x_f = \min\{x : f(x) = 0\}$ exists, and so

$$\{x_f\} \in [\mathrm{dom}(f)]^{\leq 1},\ f \sqsubseteq \mathcal{H}(\{x_f\}) = I_{x_f} \setminus \{x_f\}.$$

Therefore $\mathcal{H}'$ is also a sample compression scheme of size 1 on $(X, \mathcal{C})$.

Proposition 2.1.5 *If $(X, \mathcal{C})$ has an unlabelled compression scheme of size d, then $(X, \mathcal{C})$ has a labelled compression scheme of size d.*

Proof: Let $(X, \mathcal{C})$ have an unlabelled compression scheme $\mathcal{H}$ of size d. For every $f \in \mathcal{C}_{|_{[X]<\infty}}$ there is $\sigma_f \in [\mathrm{dom}(f)]^{\leq d}$ such that $f \sqsubseteq \mathcal{H}(\sigma_f)$, and so any function $\mathcal{H}' : \mathcal{C}_{|_{[X]\leq d}} \to 2^X$ where $\mathcal{H}'(f_{|_{\sigma_f}}) = \mathcal{H}(\sigma_f)$ will be a labelled compression scheme of size d. ∎

From now on we will only be dealing with unlabelled sample compression schemes unless otherwise mentioned.

Proposition 2.1.6 ([2]) *If $(X, Y, R) \preceq^{w}_{gemb} (X', Y', R')$ and (X', Y', R') has a (labelled or unlabelled) sample compression scheme of size d, then (X, Y, R) also has a sample compression scheme of size d and of the same type.*

Corollary 2.1.7 *If $(X, \mathcal{C})$ has a sample compression scheme of size d, then every subspace has a sample compression scheme of size d.*

2.2 Compactness Theorem

Theorem 2.2.1 (Compactness Theorem, Ben-David and Litman [2]) *A concept space $(X, \mathcal{C})$ has a sample compression scheme of size d if and only if every finite subspace of $(X, \mathcal{C})$ has a sample compression scheme of size d.*

The compactness theorem is true for both types of sample compression schemes and similarly for all forms of extended sample compression schemes given in a following section. We will provide the proof of the theorem for unlabelled sample compression schemes. The proof we provide is simpler and more direct than the proof in [2] which is based on the Compactness Theorem of Predicate Logic. We use an approach with ultralimits, normally used in Analysis. (For preliminary information on filters and ultrafilters, see appendix A.2)

Proof: *Necessity*: By corollary 2.1.7 if $(X, \mathcal{C})$ has a sample compression scheme of size d every (finite) subspace of $(X, \mathcal{C})$ has a sample compression scheme of size d. *Sufficiency*: For all $A \in [X]^{<\infty}$ denote the sample compression scheme of size d for $(A, \mathcal{C} \sqcap A)$ as $\mathcal{H}_A$. Let $\mathfrak{U}$ be an ultrafilter on $[X]^{<\infty}$ containing the filter base

$$\{\{B \in [X]^{<\infty} : F \subseteq B\} : F \in [X]^{<\infty}\}.$$

Define $\mathcal{H} : [X]^{\leq d} \to 2^X$ as

$$\mathcal{H}(\sigma)(x) = 1 \iff \{B \in [X]^{<\infty} : \sigma \cup \{x\} \subseteq B,\ \mathcal{H}_B(\sigma)(x) = 1\} \in \mathfrak{U}.$$

Note for given $\sigma \in [X]^{\leq d},\ x \in X$, $\mathcal{H}(\sigma)(x)$ is defined as the ultralimit of the net of zeros and ones $\{\mathcal{H}_A(\sigma)(x)\}_{\sigma \subseteq A \in [X]^{<\infty}}$ along $\mathfrak{U}$.

We will show $\mathcal{H}$ is a sample compression scheme of size d on $(X, \mathcal{C})$. Let $f \in \mathcal{C}_{|_{[X]^{<\infty}}}$, and denote $\mathrm{dom}(f) = D$. Note that

$$\forall B \in [X]^{<\infty},\ D \subseteq B,\ \text{we have } f \in (\mathcal{C} \sqcap B)_{|_{[X]^{<\infty}}},\ \text{and so } \exists \sigma_B \in [D]^{\leq d} \text{ such that } f \sqsubseteq \mathcal{H}_B(\sigma_B). \tag{1}$$

We have that $[D]^{\leq d}$ is finite so let $\{\sigma_1, ..., \sigma_m\} = [D]^{\leq d}$. For $i \in \{1, ..., m\}$ letting

$$\mathcal{S}_i = \{B \in [X]^{<\infty} : D \subseteq B,\ f \sqsubseteq \mathcal{H}_B(\sigma_i)\},$$

by (1) we see that

$$\bigcup_{i=1}^{m} \mathcal{S}_i = \{B \in [X]^{<\infty} : D \subseteq B\} \in \mathfrak{U}$$

thus, by a property of ultrafilters, $\exists i_0$ such that $\mathcal{S}_{i_0} \in \mathfrak{U}$. Let $x \in D$ and let

$$\mathcal{S}^t_{i_0} = \{B \in [X]^{<\infty} : D \subseteq B,\ \mathcal{H}_B(\sigma_{i_0})(x) = t\} \text{ (where } t \in \{0, 1\}).$$

We have

$$\begin{aligned}
f(x) = 1 &\Rightarrow \forall B \in \mathcal{S}_{i_0},\ \mathcal{H}_B(\sigma_{i_0})(x) = 1 \\
&\Rightarrow \mathcal{S}_{i_0} \subseteq \mathcal{S}^1_{i_0} \subseteq \{B \in [X]^{<\infty} : \sigma_{i_0} \subseteq B,\ \mathcal{H}_B(\sigma_{i_0})(x) = 1\} \in \mathfrak{U} \\
&\Rightarrow \mathcal{H}(\sigma_{i_0})(x) = 1; \\
f(x) = 0 &\Rightarrow \forall B \in \mathcal{S}_{i_0},\ \mathcal{H}_B(\sigma_{i_0})(x) = 0 \\
&\Rightarrow \mathcal{S}_{i_0} \subseteq \mathcal{S}^0_{i_0} \subseteq \{B \in [X]^{<\infty} : \sigma_{i_0} \subseteq B,\ \mathcal{H}_B(\sigma_{i_0})(x) = 0\} \in \mathfrak{U} \\
&\Rightarrow \{B \in [X]^{<\infty} : \sigma_{i_0} \subseteq B,\ \mathcal{H}_B(\sigma_{i_0})(x) = 0\}^c \notin \mathfrak{U} \\
&\Rightarrow \{B \in [X]^{<\infty} : \sigma_{i_0} \subseteq B,\ \mathcal{H}_B(\sigma_{i_0})(x) = 1\} \notin \mathfrak{U} \\
&\Rightarrow \mathcal{H}_B(\sigma_{i_0})(x) = 0.
\end{aligned}$$

Therefore $f \sqsubseteq \mathcal{H}(\sigma_{i_0})$ and so $\mathcal{H}$ is a sample compression scheme of size d on $(X, \mathcal{C})$. ∎

We would like to point out that even though the above proof is essentially the same as the original proof in [2], it is reformulated using ultralimits, as usually done in analysis, and does not use logic. As such, it may be easier to understand.

A point of concern with the compactness theorem is that the compression scheme resulting from the finite domains need not have measurable hypothesis spaces, and we construct an example to prove this point in chapter 4. This problem is the main focus of chapter 4.

2.3 Sample Compression Schemes and VC Dimension

The following remark is a simple observation.

Proposition 2.3.1 *If $(X, \mathcal{C})$ has an unlabelled sample compression scheme of size d, then* $\mathrm{VC}(\mathcal{C}) \leq d$.

Proof: Suppose $\mathrm{VC}(\mathcal{C}) > d$ and let $A \subseteq X$ be a set of size $d+1$ which is shattered. We have that $|\mathcal{C} \sqcap A| = 2^{d+1}$, but there are

$$\binom{d+1}{\leq d} < \binom{d+1}{\leq d+1} = 2^{d+1}$$

subsets of A of size at most d. Therefore $(A, \mathcal{C} \sqcap A)$ cannot have a sample compression scheme of size d, which is a contradiction. ∎

Note that the proof only applies to unlabelled sample compression schemes and the same is not true for labelled sample sample compression schemes. However if a labelled sample compression scheme of size d exists on $(X, \mathcal{C})$ then $\mathrm{VC}(\mathcal{C}) \leq 5d$ [8].

It is a major open question posed by Littlestone and Warmuth, in paper [12], whether or not a concept space $(X, \mathcal{C})$ has an (unlabelled or labelled) sample compression scheme of size $\mathrm{O}(\mathrm{VC}(\mathcal{C}))$. In the case of d-maximum spaces, a sample compression scheme of size d exists.

Theorem 2.3.2 ([10]) *If $(X, \mathcal{C})$ is d-maximum, then it has an unlabelled compression scheme of size d.*

This thesis' motivation is to try and generalize this result to obtain measurable hypotheses. We succeed in chapter 4 under some additional assumptions.

Using remark 1.2.6 we have the following corollary.

Corollary 2.3.3 *If $(X, \mathcal{C})$ is finite with VC-dimension d and*

$$|\mathcal{C}| = \binom{|X|}{\leq d},$$

then it has an unlabelled compression scheme of size d.

2.4 Extended Sample Compression Schemes

In this section we will introduce a new variant of compression scheme called copy sample compression schemes. All other discussed compression schemes are special cases of copy sample compression schemes. The initial motivation for copy sample compression schemes was the ability to collect sample compression schemes of varying sizes, and for different concept classes, into one function; a copy sample compression scheme can be thought of as an algorithm which checks sample compression schemes of varying sizes, and for different concept classes, and picks a hypothesis for the concepts in any of these different concept classes. This is formalized in proposition 2.4.7. Copy sample compression schemes also add other flexibilities, and may in some instances have better bounds for sample complexity than regular sample compression schemes.

Definition 2.4.1 ([2]) *Let $\mathbf{b}$ be a symbol not in X.*
*An **array sample compression scheme of size k** for $(X, \mathcal{C})$ is a function*

$$\mathcal{H} : (X \cup \{\mathbf{b}\})^d \to 2^X$$

with the property that

$$\forall f \in \mathcal{C}_{|_{[X]<\infty}},\ \exists \sigma \in (X \cup \{\mathbf{b}\})^d,\ \textit{such that}\ \operatorname{range}(\sigma) \subseteq \operatorname{dom}(f) \cup \{\mathbf{b}\}\ \textit{and}\ f \sqsubseteq \mathcal{H}(\sigma)$$

(where, for a sequence $\sigma = (a_1, ..., a_k)$, $\operatorname{range}(\sigma)$ is the set $\{a_1, ..., a_k\}$).

Definition 2.4.2 ([12]) *An **extended sample compression scheme of size k using b bits** for $(X, \mathcal{C})$ is a function*

$$\mathcal{H} : \bigcup_{i=0}^{k} ([X]^{=i} \times 2^b) \to 2^X$$

with the property that

$$\forall f \in \mathcal{C}_{|_{[X]<\infty}},\ \exists \sigma \in [\mathrm{dom}(f)]^{\leq k} \text{ and } \tau \in 2^b,\ \text{such that } f \sqsubseteq \mathcal{H}(\sigma \times \tau).$$

The preceding definitions are special cases of the following new variant of an extended sample compression scheme.

Definition 2.4.3 *Let $k \in \mathbb{N}$, and let $\{n_i\}_{i=0}^{k}$ be a finite sequence in $\mathbb{N}$.*
*A $\{\mathbf{n_i}\}_{\mathbf{i=0}}^{\mathbf{k}}$**-copy unlabelled sample compression scheme of size k** on $(X, \mathcal{C})$ is a function*

$$\mathcal{H} : \bigcup_{i \in \{j \in \mathbb{N} : 0 \leq j \leq k,\ n_j \neq 0\}} ([X]^{=i} \times \{1, ..., n_i\}) \to 2^X$$

with the property that

$$\forall f \in \mathcal{C}_{|_{[X]<\infty}},\ \exists \sigma \in [\mathrm{dom}(f)]^{\leq k} \text{ and } i \in \{1, ..., n_{|\sigma|}\},\ \text{such that } f \sqsubseteq \mathcal{H}(\sigma \times i).$$

*If $\{n_i\}_{i=0}^{k}$ is just a constant sequence $\{n\}$ for some $n \in \mathbb{N}$, we will call it an **n-copy unlabelled sample compression scheme of size k**. We can define $\{n_i\}_{i=0}^{k}$-copy labelled sample compression schemes of size k similarly.*

Note that a sample compression scheme of size d defines a 1-copy sample compression scheme of size d, and a 1-copy sample compression scheme of size d defines a sample compression scheme of size d. A compactness theorem can also be proven for copy sample compression schemes (and the other versions of extended sample compression schemes), namely, $(X, \mathcal{C})$ has a $\{n_i\}_{i=0}^{k}$-copy sample compression scheme of size k if and only if every finite subspace $(Y, \mathcal{C} \sqcap Y)$ has a $\{n_i\}_{i=0}^{k}$-copy sample compression scheme of size k.

Proposition 2.4.4 *Let $|X| = m$, let $(X, \mathcal{C})$ have a sample compression scheme of size d, and let $k \leq d$. Whenever*

$$n\binom{m}{\leq k} \geq \binom{m}{\leq d},$$

$(X, \mathcal{C})$ has an n-copy sample compression scheme of size k.

Having an n-copy sample compression scheme of size k for $(X, \mathcal{C})$ does not imply there is a sample compression scheme of size d when

$$n\binom{m}{\leq k} \leq \binom{m}{\leq d}.$$

Example 2.4.5 Let $X = \{1, 2, 3, 4, 5\}$, $\mathcal{C} = 2^{\{1,2\}}$. Enumerate $\mathcal{C} = \{C_l : l \in \{1, 2, 3, 4\}\}$ in any way and define $\mathcal{H} : [X]^{=0} \times \{1, 2, 3, 4\} \to 2^X$ by $\mathcal{H}(\emptyset \times l) = C_l$, $l \in \{1, 2, 3, 4\}$. $\mathcal{H}$ is a 4-copy sample compression scheme of size 0, but since $\mathrm{VC}(X, \mathcal{C}) = 2$, $(X, \mathcal{C})$ has no sample compression scheme of size 1 although

$$4\binom{5}{0} = 4 < 5 = \binom{5}{1}$$

Example 2.4.6 Let $X = \{1, 2, 3, 4\}$,
$\mathcal{C} = \{\{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 4\}, \{2, 4\}, \{3, 4\}, \{1, 2, 3\}\}$. Recall that $\mathcal{C}$ is 2-maximal but not 2-maximum since

$$|\mathcal{C}| = 10 < 11 = \binom{4}{\leq 2}.$$

We can define a 2-copy sample compression scheme of size 1 by

$$\begin{aligned}
&\mathcal{H}(\emptyset \times 1) = \{1, 2\},\ \mathcal{H}(\emptyset \times 2) = \{3, 4\},\\
&\mathcal{H}(\{1\} \times 1) = \{3\},\ \mathcal{H}(\{1\} \times 2) = \{1, 3\},\\
&\mathcal{H}(\{2\} \times 1) = \{1\},\ \mathcal{H}(\{2\} \times 2) = \{2, 4\},\\
&\mathcal{H}(\{3\} \times 1) = \{2\},\ \mathcal{H}(\{3\} \times 2) = \{1, 2, 3\},
\end{aligned}$$

$$\mathcal{H}(\{4\} \times 1) = \{2, 3\},\ \mathcal{H}(\{4\} \times 2) = \{1, 4\}.$$

Note that

$$2\binom{4}{\leq 1} = 10,$$

however $(X, \mathcal{C})$ has no sample compression scheme of size less than 2.

Copy sample compression schemes for a concept space $(X, \mathcal{C})$ can be defined from multiple sample compression schemes for concept classes which cover $\mathcal{C}$. This can allow us to split up $\mathcal{C}$ into spaces which are known to have sample compression schemes, and then form a copy sample compression scheme.

Proposition 2.4.7 *Let $|X| = m$, and $\mathcal{C} \subseteq \bigcup_{j=1}^{n} \mathcal{C}_j$ where each $(X, \mathcal{C}_j)$ has a sample compression scheme $\mathcal{H}_j$ of size d_j. Define*

$$k = \max(\{d_j\}_{j=1}^{n}),$$

and

$$n_i = |\{j : d_j \geq i\}|$$

for $0 \leq i \leq k$. Then $(X, \mathcal{C}_j)$ has a $\{n_i\}_{i=0}^{k}$-copy sample compression scheme of size k.

Proof: Without loss of generality assume $j < l$ implies $d_j \geq d_l$. We will show that $\mathcal{H}$ defined by $\mathcal{H}(\sigma \times l) = \mathcal{H}_l(\sigma)$ for $l \in \{1, ..., n_{|\sigma|}\}$ is a $\{n_i\}_{i=0}^{k}$-copy sample compression scheme of size k for $(X, \mathcal{C})$. Note that $\mathcal{H}_l$ is defined at σ because

$$|\{1, ..., n_{|\sigma|}\}| = n_{|\sigma|} = |\{j : d_j \geq |\sigma|\}|$$

implies

$$d_1 \geq |\sigma|,\ d_2 \geq |\sigma|, ..., d_{n_{|\sigma|}} \geq |\sigma|$$

so in particular $d_l \geq |\sigma|$.

Let $f \in \mathcal{C}_{|_{[X]<\infty}}$ and let $C_f \in \mathcal{C} \subseteq \bigcup_{j=1}^{n} \mathcal{C}_j$ be such that C_f extends f to X. Since $C_f \in \mathcal{C}_l$ for some $1 \leq l \leq n$, there exists

$$\sigma \in [\mathrm{dom}(f)]^{\leq d_l} \subseteq [\mathrm{dom}(f)]^{\leq k}$$

such that

$$f \sqsubseteq \mathcal{H}_l(\sigma) = \mathcal{H}(\sigma \times l).$$

∎

Corollary 2.4.8 *If there exists a family* $\{\mathcal{C}_j\}_{j=1}^n$ *of concept classes such that* $\mathcal{C} \subseteq \bigcup_{j=1}^n \mathcal{C}_j$*, and for each* $\mathcal{C}_j$ *there is a* d*-maximum* $\mathcal{C}'_j$ *containing* $\mathcal{C}_j$*, then there is an* n*-copy sample compression scheme of size* d *for* $(X, \mathcal{C})$*.*

Chapter 3

Learnability

3.1 Learnability

The PAC, or "Probably Approximately Correct", model for learning was introduced by Valiant [18]. In this chapter we introduce PAC learnability, and investigate sample complexity for spaces with finite VC dimension, and for spaces with sample compression schemes or copy sample compression schemes.

For this section, let $(X, \mathfrak{A})$ be a measurable space with a concept class $\mathcal{C} \subseteq \mathfrak{A}$ and a family of probability measures $\mathcal{P}$ on $(X, \mathfrak{A})$. We will also reference measurability conditions (M1),(M2),...,(M5) defined and discussed in appendix B.1.

Definition 3.1.1 ([18]) *A **learning rule** for $(X, \mathcal{C})$ is a function*

$$\mathcal{L} : \bigcup_{n=1}^{\infty} (X^n \times 2^n) \to \mathfrak{A}$$

which satisfies the following measurability condition we will label "(M1)": for every $C \in \mathcal{C}$, every $n \geq 1$ and every $P \in \mathcal{P}$, the function

$$A \mapsto P(\mathcal{L}(A, C_{|A}) \triangle C)$$

from $(X, \mathfrak{A})^n$ *to* $\mathbb{R}$ *is measurable.*

We will call $\mathfrak{H} = \text{range}(\mathcal{L}) \subseteq \mathfrak{A}$ *the* ***hypothesis space****.*

A learning rule is ***consistent*** *with* $\mathcal{C}$ *if for every* $C \in \mathcal{C}$ *and* $A \in X^n$*,*

$$\mathcal{L}(A, C_{|_A})_{|_A} = C_{|_A}.$$

The domain $\bigcup_{n=1}^{\infty}(X^n \times 2^n)$ for learning rules represents all finite labelled samples where (A, τ), for $A = (a_1, ..., a_n) \in X^n$ and $\tau = (l_1, ..., l_n) \in 2^n$, represents the function $\{(a_1, l_1), ..., (a_n, l_n)\}$, and $(A, C_{|_A})$ in the definition is shorthand for $(A, (\chi_{C\cap A}(a_1), ..., \chi_{C\cap A}(a_n)))$.

Any sample compression scheme $\mathcal{H}$ of size d on $(X, \mathcal{C})$ defines a (in general, non-unique) function $\mathcal{L}_{\mathcal{H}}$, with $\mathcal{L}_{\mathcal{H}}$ being a consistent learning rule for $(X, \mathcal{C})$ if it satisfies (M1):

For each $A \in [X]^{<\infty}$, $C \in \mathcal{C} \sqcap A$ pick $\sigma_{C,A} \in [A]^{\leq d}$ such that $C = \mathcal{H}(\sigma_{C,A})_{|_A}$. Define

$$\mathcal{L}_{\mathcal{H}}(A, \tau) = \begin{cases} \mathcal{H}(\sigma_{C,A}), & \text{if } \tau = C \text{ for some } C \in \mathcal{C} \sqcap A \\ \emptyset, & \text{otherwise} \end{cases}.$$

Similarly a $\{n_i\}_{i=0}^{k}$-copy sample compression scheme $\mathcal{H}'$ of size k defines a (in general, non-unique) function $\mathcal{L}_{\mathcal{H}'}$, with $\mathcal{L}_{\mathcal{H}'}$ being a a consistent learning rule for $(X, \mathcal{C})$ if it satisfies (M1):

For each $A \in [X]^{<\infty}$, $C \in \mathcal{C} \sqcap A$ pick $\sigma_{C,A} \in [A]^{\leq d}$, $l \in \{1, ..., n_{|\sigma|}\}$ such that $C = \mathcal{H}'(\sigma_{C,A} \times l)_{|_A}$. Define

$$\mathcal{L}_{\mathcal{H}'}(A, \tau) = \begin{cases} \mathcal{H}'(\sigma_{C,A} \times l), & \text{if } \tau = C \text{ for some } C \in \mathcal{C} \sqcap A \\ \emptyset, & \text{otherwise} \end{cases}.$$

Definition 3.1.2 ([18]) *A learning rule* $\mathcal{L}$ *for a space* $(X, \mathcal{C})$ *is* ***probably approximately correct (PAC)*** *under* $\mathcal{P}$ *if for every* $\varepsilon > 0$

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}} \sup_{C\in\mathcal{C}} P^n(\{A \in X^n : P(\mathcal{L}(A, C_{|_A}) \triangle C) > \varepsilon\}) = 0.$$

We will say a concept space $(X, \mathcal{C})$ *is* ***probably approximately correct (PAC) learnable*** *(under* $\mathcal{P}$*) if there exists a PAC learning rule (under* $\mathcal{P}$*) for* $(X, \mathcal{C})$*. For a given PAC learning rule* $\mathcal{L}$ *and given* $0 < \varepsilon \leq 1,\ 0 < \delta \leq 1$ *we will define the* ***sample complexity*** *of* $\mathcal{L}$*,*

$$m_{\mathcal{L}}(\varepsilon, \delta),$$

to be the least integer such that for all $n \geq m_{\mathcal{L}}(\varepsilon, \delta)$*,*

$$\sup_{P \in \mathcal{P}} \sup_{C \in \mathcal{C}} P^n(\{A \in X^n : P(\mathcal{L}(A, C_{|_A}) \triangle C) > \varepsilon\}) < \delta.$$

We call ε *the* ***accuracy*** *and* δ *the* ***risk****.*

Under the measurability condition that $\mathcal{C}$ is well behaved ((M4) in appendix B.1), the following is a result due to Blumer et al in [4]:

Theorem 3.1.3 ([4]) *The following conditions are equivalent:*

(1) $\mathrm{VC}(X, \mathcal{C}) < \infty$*.*

(2) $(X, \mathcal{C})$ *is PAC learnable.*

(3) Every consistent learning rule $\mathcal{L} : \bigcup_{n=1}^{\infty}(X^n \times 2^n) \to \mathcal{C}$ *is PAC for* $(X, \mathcal{C})$ *with*

$$m_{\mathcal{L}}(\varepsilon, \delta) \leq \max\left(\frac{4}{\varepsilon}\log_2(\frac{2}{\delta}), \frac{8d}{\varepsilon}\log_2(\frac{13}{\varepsilon})\right)$$

for $d = \mathrm{VC}(X, \mathcal{C})$*.*

Given a consistent learning rule with $\mathfrak{H} \subseteq \mathcal{C}$ as in (3) in the above theorem, if we assume a measurability condition (M3) (appendix B.1) which is stronger than the condition of being well behaved, it is shown by Shawe-Taylor et al. in [17] that we can improve the bounds: Every consistent learning rule with $\mathfrak{H} \subseteq \mathcal{C}$, and satisfying (M3), has sample complexity

$$m_{\mathcal{L}}(\varepsilon, \delta) \leq \frac{1}{1-\beta}\left(\frac{1}{\varepsilon}\ln(\frac{2}{\delta}) + \frac{2d\ln 2}{\varepsilon} + \frac{d}{\varepsilon}\ln(\frac{1}{\varepsilon\beta^2})\right)$$

for any $0 < \beta < 1$, where d is the VC dimension of $(X, \mathcal{C})$.

3.2 Sample Complexity of Compression Schemes

Theorem 3.2.1 ([12]) *Let P be any probability measure on a measurable space $(X, \mathfrak{A})$, C a concept in $\mathcal{C} \subseteq \mathfrak{A}$, and $\mathcal{H}$ any function from $[X]^{\leq d}$ to 2^X, satisfying measurability condition (M5). Then the probability that $A \subseteq X$, $|A| = m \geq d$, contains a subset σ of size at most d such that $P(\mathcal{H}(\sigma) \triangle C) > \varepsilon > 0$ and $\mathcal{H}(\sigma)_{|A} = C_{|A}$, is at most*

$$\sum_{n=0}^{d} \binom{m}{n} (1-\varepsilon)^{m-n}.$$

Proof: Let $C \in \mathcal{C}$ and ε be given. First we consider the probability that a set of size m has a subset of size exactly $n \leq d$ with the property $P(\mathcal{H}(\sigma) \triangle C) > \varepsilon$ and $\mathcal{H}(\sigma)_{|A} = C_{|A}$. For $A = (a_1, ..., a_m) \in X^m$ and $J = \{j_1, ..., j_n\} \subseteq \{1, ..., m\}$, let $A_{|J}$ denote $\{a_{j_1}, ..., a_{j_n}\}$.

There are $\binom{m}{n}$ many subsets of A of size n, hence fixing J a subset of $\{1, ..., m\}$ of size n, the probability we wish to bound from above is at most

$$P^m(\{A \in X^m : \exists I \subseteq \{1, ..., m\} \text{ of size } n \text{ where } P(\mathcal{H}(A_{|I}) \triangle C) > \varepsilon \text{ and } \mathcal{H}(A_{|I})_{|A} = C_{|A}\})$$
$$= \binom{m}{n} P^m(\{A \in X^m : P(\mathcal{H}(A_{|J}) \triangle C) > \varepsilon \text{ and } \mathcal{H}(A_{|J})_{|A} = C_{|A}\}).$$

Since permuting J to some other subset of size n in $\{1, ..., m\}$ does not affect the above probability, we can assume $J = \{1, ..., n\}$.

We will prove at this point that $\{A \in X^m : P(\mathcal{H}(A_{|J}) \triangle C) > \varepsilon \text{ and } \mathcal{H}(A_{|J})_{|A} = C_{|A}\}$ is measurable due to the hypothesis that $\mathcal{H}$ satisfies (M5): Let $1 \leq p < q \leq m$ and let $\pi_{p,q}$ be the (measurable) function from X^m to X^{p+1} mapping $(x_1, ..., x_m) \mapsto (x_1, ..., x_p, x_q)$. By (M5) and the measurability of C we have

$$\{A \in X^m : \mathcal{H}(A_{|J})_{|A} = C_{|A}\} = \{A \in X^m : A \in ((\mathcal{H}(A_{|J}) \triangle C)^c)^m\}$$
$$= \bigcap_{q=1}^{m} \{A \in X^m : (\mathcal{H}(A_{|J}) \triangle C)^c)(A_{|\{q\}}) = 1\}$$

$$= \bigcap_{q=1}^{m} \pi_{n,q}^{-1}(\{(x_1, ..., x_{n+1}) \in X^{n+1} : (\mathcal{H}(\{x_1, ..., x_n\})\triangle C)^c)(x_{n+1}) = 1\}) \in \mathfrak{A}^m.$$

Also $\{A \in X^m : P(\mathcal{H}(A_{|_J})\triangle C) > \varepsilon\}$ is measurable since

$$B = \{(x_1, ..., x_{m+1}) \in X^{m+1} : (\mathcal{H}(\{x_1, ..., x_n\})\triangle C)^c(x_{m+1}) = 1\}$$
$$= \pi_{n,m+1}^{-1}(\{(x_1, ..., x_{n+1}) \in X^{n+1} : (\mathcal{H}(\{x_1, ..., x_n\})\triangle C)^c(x_{n+1}) = 1\})$$

is measurable by (M5) and the measurability of C, and a straightforward application of Fubini's theorem gives us that the map

$$(x_1, ..., x_m) \mapsto \int_X \chi_B(x_1, ..., x_{m+1}) dP(x_{m+1}) =$$
$$= P(\{y : (x_1, ..., x_m, y) \in B\})$$
$$= P(\{y : y \in \mathcal{H}(\{x_1, ..., x_n\})\triangle C)^c\})$$
$$= P(\mathcal{H}(\{x_1, ..., x_n\})\triangle C)^c)$$

is measurable.

Now let

$$E_C := \{A \in X^m : \mathcal{H}(A_{|_J})_{|_{A_{|_{\{n+1,...,m\}}}}} = C_{|_{A_{|_{\{n+1,...,m\}}}}}\}$$
$$= \{A \in X^m : A_{|_{\{n+1,...,m\}}} \in ((\mathcal{H}(A_{|_J})\triangle C)^c)^{m-n}\}$$
$$= \bigcap_{q=n+1}^{m} \{A \in X^m : (\mathcal{H}(A_{|_J})\triangle C)^c)(A_{|_{\{q\}}}) = 1\}$$
$$= \bigcap_{q=n+1}^{m} \pi_{n,q}^{-1}(\{(x_1, ..., x_{n+1}) \in X^{n+1} : (\mathcal{H}(\{x_1, ..., x_n\})\triangle C)^c)(x_{n+1}) = 1\}) \in \mathfrak{A}^m.$$

and

$$E_\varepsilon := \{A \in X^n : P(\mathcal{H}(A)\triangle C) > \varepsilon\}$$
$$= \{A \in X^n : P((\mathcal{H}(A)\triangle C)^c) \leq (1 - \varepsilon)\}.$$

E_ε is measurable since $B = \{(x_1, ..., x_{n+1}) \in X^{n+1} : (\mathcal{H}(\{x_1, ..., x_n\})\triangle C)^c(x_{n+1}) = 1\}$ is measurable by (M5) and the measurability of C, and a straightforward application

of Fubini's theorem gives us that the map

$$(x_1, ..., x_n) \mapsto \int_X \chi_B(x_1, ..., x_{n+1}) dP(x_{n+1}) =$$
$$= P(\{y : (x_1, ..., x_n, y) \in B\})$$
$$= P(\{y : y \in \mathcal{H}(\{x_1, ..., x_n\}) \triangle C)^c\})$$
$$= P(\mathcal{H}(\{x_1, ..., x_n\}) \triangle C)^c)$$

is measurable.

We have that

$$P^m(\{A \in X^m : P(\mathcal{H}(A_{|_J}) \triangle C) > \varepsilon \text{ and } \mathcal{H}(A_{|_J})_{|_A} = C_{|_A}\})$$
$$\leq P^m(\{A \in X^m : P(\mathcal{H}(A_{|_J}) \triangle C) > \varepsilon \text{ and } \mathcal{H}(A_{|_J})_{|_{A_{|_{\{n+1,...,m\}}}}} = C_{|_{A_{|_{\{n+1,...,m\}}}}}\})$$
$$= P^m(E_C \cap (E_\varepsilon \times X^{m-n})).$$

By Fubini's theorem

$$P^m(E_C \cap (E_\varepsilon \times X^{m-n})) = \int_{E_\varepsilon \times X^{m-n}} \chi_{E_C}(x_1, ..., x_m) dP^m$$
$$= \int_{E_\varepsilon} \left(\int_{X^{m-n}} \chi_{E_C}(x_1, ..., x_m) dP^{m-n} \right) dP^n.$$

Now

$$(x_1, ..., x_n) \times X^{m-n} \cap E_C = (x_1, ..., x_n) \times \{A \in X^{m-n} : A_{|_{\{n+1,...,m\}}} \in ((\mathcal{H}(A_{|_J}) \triangle C)^c)^{m-n}\}$$

and since $(x_1, ..., x_n) \in E_\varepsilon$, the inner integral is at most $(1 - \varepsilon)^{m-n}$ and so

$$P^m(E_C \cap (E_\varepsilon \times X^{m-n})) \leq (1 - \varepsilon)^{m-n}.$$

Therefore summing over all subsets J of $\{1, ..., m\}$ of size at most d, the probability that $A \subseteq X$, $|A| = m \geq d$, contains a subset σ of size at most d such that $P(\mathcal{H}(\sigma) \triangle C) > \varepsilon > 0$ and $\mathcal{H}(\sigma)_{|_A} = C_{|_A}$, is at most

$$\sum_{n=0}^{d} \binom{m}{n} (1 - \varepsilon)^{m-n}.$$

∎

Lemma 3.2.2 ([8]) *Let $0 < \varepsilon \leq 1,\ 0 < \delta \leq 1$ and m, d positive integers. If there is $0 < \beta < 1$ where*

$$m \geq \frac{1}{1-\beta}\left(\frac{1}{\varepsilon}\ln(\frac{1}{\delta}) + d + \frac{d}{\varepsilon}\ln(\frac{1}{\beta\varepsilon})\right),$$

then

$$\sum_{n=0}^{d}\binom{m}{n}(1-\varepsilon)^{m-n} \leq \delta.$$

Proof: Let $\varepsilon, \delta, \beta, m, d$ be as in the statement of the lemma. Then

$$\frac{1}{\varepsilon}\ln(\frac{1}{\delta}) + d + \frac{d}{\varepsilon}\left(-\ln(d) + \ln(\frac{d}{\beta\varepsilon}) - 1 + \frac{\beta\varepsilon}{d}m + 1\right)$$

$$= \beta m + \frac{1}{\varepsilon}\ln(\frac{1}{\delta}) + d + \frac{d}{\varepsilon}\ln(\frac{1}{\beta\varepsilon}) \leq m.$$

We will use the fact that $\ln(m) \leq -\ln(\alpha) - 1 + \alpha m$ for all $\alpha > 0$. With $\alpha = \frac{\beta\varepsilon}{d}$ we get

$$\ln(m) \leq \ln(\frac{d}{\beta\varepsilon}) - 1 + \frac{\beta\varepsilon}{d}m$$

thus

$$\frac{1}{\varepsilon}\ln(\frac{1}{\delta}) + d + \frac{d}{\varepsilon}\left(-\ln(d) + \ln(m) + 1\right)$$

$$\leq \frac{1}{\varepsilon}\ln(\frac{1}{\delta}) + d + \frac{d}{\varepsilon}\left(-\ln(d) + \ln(\frac{d}{\beta\varepsilon}) - 1 + \frac{\beta\varepsilon}{d}m + 1\right) \leq m.$$

Hence we have

$$\ln(\frac{1}{\delta}) + d(-\ln(d) + \ln(m) + 1) \leq \varepsilon(m-d)$$

and so

$$\left(\frac{em}{d}\right)^{d} \leq e^{\varepsilon(m-d)}\delta.$$

Therefore since $m \geq d$,

$$\sum_{i=0}^{d}\binom{m}{i}(1-\varepsilon)^{m-i} \leq \binom{m}{\leq d}(1-\varepsilon)^{m-d} \leq \left(\frac{em}{d}\right)^{d}(1-\varepsilon)^{m-d} \leq \left(\frac{em}{d}\right)^{d}e^{-\varepsilon(m-d)} \leq \delta.$$

Theorem 3.2.1 and the last lemma lead to bounds for the sample complexity of sample compression schemes. We illustrate in Figure 3.1 on the next page, that these bounds for sample compression schemes of size d are better than the ones for general consistent learning rules on a space with VC dimension d. Note that 0.05 is one of the standard choices for risk in statistics.

Theorem 3.2.3 ([8]) *If $(X, \mathcal{C})$ has a sample compression scheme $\mathcal{H}$ of size d satisfying (M5), then for $0 < \varepsilon \leq 1,\ 0 < \delta \leq 1$, if $\mathcal{L}_{\mathcal{H}}$ is a learning rule (if $\mathcal{L}_{\mathcal{H}}$ satisfies (M2)) it has sample complexity at most*

$$m_{\mathcal{L}_{\mathcal{H}}}(\varepsilon, \delta) \leq \frac{1}{1-\beta}\left(\frac{1}{\varepsilon}\ln(\frac{1}{\delta}) + d + \frac{d}{\varepsilon}\ln(\frac{1}{\beta\varepsilon})\right),$$

for any $0 < \beta < 1$.

Proof: Let $\mathcal{H}$ be as in the statement of the theorem, fix $0 < \varepsilon \leq 1,\ 0 < \delta \leq 1,\ 0 < \beta < 1,\ C \in \mathcal{C},\ P \in \mathcal{P}$, and let

$$m \geq \frac{1}{1-\beta}\left(\frac{1}{\varepsilon}\ln(\frac{1}{\delta}) + d + \frac{d}{\varepsilon}\ln(\frac{1}{\beta\varepsilon})\right).$$

Since $\mathcal{H}$ is consistent and satisfies (M5), by theorem 3.2.1

$$\begin{aligned}
&P^m(\{A \in X^m : P(\mathcal{L}_{\mathcal{H}}(A, C_{|_A}) \triangle C) > \varepsilon\}) \\
&\leq P^m(\{A \in X^m : \exists \sigma \in [A]^{\leq d} \text{ where } P(\mathcal{H}(\sigma) \triangle C) > \varepsilon\}) \\
&= P^m(\{A \in X^m : \exists \sigma \in [A]^{\leq d} \text{ where } P(\mathcal{H}(\sigma) \triangle C) > \varepsilon \text{ and } \mathcal{H}(\sigma)_{|_A} = C_{|_A}\}) \\
&\leq \sum_{n=0}^{d} \binom{m}{n} (1-\varepsilon)^{m-n},
\end{aligned}$$

and by lemma 3.2.2

$$\sum_{n=0}^{d} \binom{m}{n} (1-\varepsilon)^{m-n} \leq \delta.$$

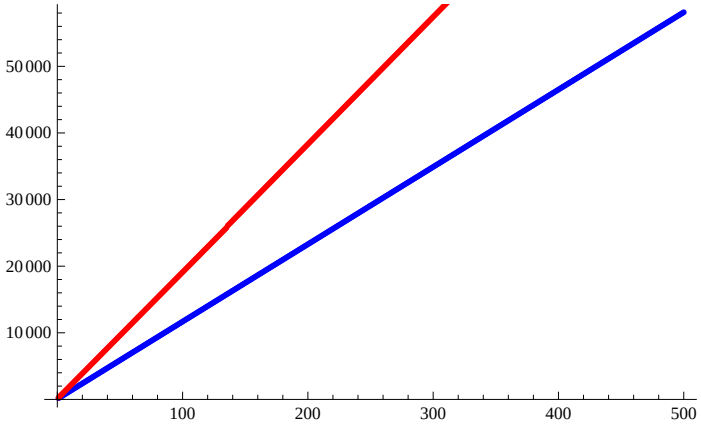

Figure 3.1: Given $\varepsilon = 0.05,\ \delta = 0.05$, we plot a graph with d on the x-axis and:
$f(d) = \frac{1}{1-\beta}\left(\frac{1}{\varepsilon}\ln(\frac{1}{\delta}) + d + \frac{d}{\varepsilon}\ln(\frac{1}{\beta\varepsilon})\right)$ in blue, and
$g(d) = \frac{1}{1-\beta}\left(\frac{1}{\varepsilon}\ln(\frac{2}{\delta}) + \frac{2d\ln 2}{\varepsilon} + \frac{d}{\varepsilon}\ln(\frac{1}{\varepsilon\beta^2})\right)$ in red,
where we optimize β in each function for each value of d.

The sample complexity for our copy sample compression schemes can be bounded in similar fashion to that for sample compression schemes. The proofs are very similar to those for sample compression schemes and can be found in the appendix B.2.

Theorem 3.2.4 *Let P be any probability measure on a measurable space $(X, \mathfrak{A})$, C a concept in $\mathcal{C} \subseteq \mathfrak{A}$, $\{n_i\}_{i=0}^{k} \subseteq \mathbb{N}$, and $\mathcal{H}$ any function from $\bigcup_{i\in\{j\in\mathbb{N}:0\leq j\leq k,\ n_j\neq 0\}}([X]^{=i} \times \{1, ..., n_i\})$ to 2^X satisfying measurability condition (M5). Then the probability that $A \subseteq X$, $|A| = m \geq k$, contains a subset σ of size at most k, and $l \in \{1, ..., n_{|\sigma|}\}$ such that $P(\mathcal{H}(\sigma \times l)\triangle C) > \varepsilon > 0$ and $\mathcal{H}(\sigma \times l)_{|A} = C_{|A}$, is at most*

$$\sum_{i=0}^{k} n_i \binom{m}{i} (1-\varepsilon)^{m-i}.$$

Lemma 3.2.5 *Let* $0 < \varepsilon \leq 1,\ 0 < \delta \leq 1$, m, k *positive integers, and* $n = \max(\{n_i\}_{i=0}^{d})$. *if there is* $0 < \beta < 1$ *where*

$$m \geq \frac{1}{1-\beta}\left(\frac{1}{\varepsilon}\ln(\frac{n}{\delta}) + k + \frac{k}{\varepsilon}\ln(\frac{1}{\beta\varepsilon})\right),$$

then

$$\sum_{i=0}^{k} n_i \binom{m}{i} (1-\varepsilon)^{m-i} \leq \delta.$$

Theorem 3.2.4 and the last lemma lead to bounds for the sample complexity of copy sample compression schemes.

Theorem 3.2.6 *If* $(X, \mathcal{C})$ *has a* $\{n_i\}_{i=0}^{k}$*-copy sample compression scheme* $\mathcal{H}$ *of size* k *satisfying (M5), and* $n = \max(\{n_i\}_{i=0}^{d})$, *then for* $0 < \varepsilon \leq 1,\ 0 < \delta \leq 1$, *if* $\mathcal{L}_{\mathcal{H}}$ *is a learning rule (if* $\mathcal{L}_{\mathcal{H}}$ *satisfies (M2)) it has sample complexity at most*

$$m_{\mathcal{L}_{\mathcal{H}}}(\varepsilon, \delta) \leq \frac{1}{1-\beta}\left(\frac{1}{\varepsilon}\ln(\frac{n}{\delta}) + k + \frac{k}{\varepsilon}\ln(\frac{1}{\beta\varepsilon})\right),$$

for any $0 < \beta < 1$.

Compared to the bounds for sample complexity of sample compression schemes, copy sample compression schemes may be better in some instances. For example, in any concept space with $|X| = 884$ and a sample compression scheme of size 7 there exists (by proposition 2.4.4) a 18418-copy sample compression scheme of size 5. In this case, if we wish to learn with accuracy $\varepsilon = 0.05$ and risk $\delta = 0.05$, the previous theorem guarantees the 18418-copy sample compression scheme of size 5 will achieve this with sample size 879, but the bounds for sample complexity of sample compression schemes (for all $\beta \in (0, 1)$) exceed 884.

Chapter 4

Measurability of Sample Compression Schemes and Learning Rules

4.1 Measurability of Compression Schemes

Given a sample compression scheme $\mathcal{H}$ of size d and $\sigma \in [X]^{\leq d}$, the corresponding hypothesis $H(\sigma)$ is not necessarily measurable with respect to a given sigma algebra on X. If we are creating our compression scheme via the compactness theorem we would like to be able to see when the resulting compression scheme will have $H(\sigma)$ measurable $\forall \sigma \in [X]^{\leq d}$. This condition is necessary for sample compression schemes to be considered as a learning rule, but it is not sufficient and in the appendix B.1 we will discuss a sufficient condition "(M5)" which allowed for the sample complexity bounds in theorem 3.2.3.

Notation 4.1.1 Let $(X, \mathfrak{A})$ be a measurable space, $Pr(X)$ the set of all probability measures on $(X, \mathfrak{A})$, and let $\mathfrak{A}(X)$ be the set of all real valued bounded functions on

X which are measurable with respect to $\mathfrak{A}$.

In the future we will fix $(X, \mathfrak{A})$ to be our measurable space unless otherwise stated.

The following remark is obvious.

Remark 4.1.2 Note that if a net of $\{0,1\}$-valued functions $\{f_\alpha\}_{\alpha\in I}$ converges pointwise to some $f \in \mathbb{R}^X$, then

$$\forall A \in [X]^{<\infty}\ \exists \alpha_0 \in I\ \forall \alpha \in I,\ \alpha \geq \alpha_0\ \Rightarrow\ f_{\alpha|_A} = f_{|_A}$$

and in particular $f \in 2^X$.

Lemma 4.1.3 ([6]) *Let $\mathcal{F} \subseteq 2^X$ be d-maximal for some d. Then $\mathcal{F}$ is closed under the topology of pointwise convergence of nets on $\mathbb{R}^X$, the product topology on $\mathbb{R}^X$. That is, if $\{f_\alpha\}_{\alpha\in I}$ is a net of functions in $\mathcal{F}$ converging pointwise to some real valued function f, then $f \in \mathcal{F}$ (for the remainder of this section when we write $f_\alpha \to f$, we mean that the convergence is pointwise unless mentioned otherwise).*

Proof: Let $\{f_\alpha\}_{\alpha\in I}$ be a net in $\mathcal{F}$ such that $f_\alpha \to f \in \mathbb{R}^X$. For $A \in [X]^{<\infty}$, by remark 4.1.2 $f \in 2^X$. Also $\exists \alpha \in I$ such that $f_{\alpha|_A} = f_{|_A}$ and so

$$f_{|_A} \in \{f_{\alpha|_A}\}_{\alpha\in I} \subseteq \mathcal{F}_{|_A}.$$

Now suppose $f \notin \mathcal{F}$. Since $\mathcal{F}$ is d-maximal, $\mathrm{VC}(\mathcal{F} \cup \{f\}) > d$ and so $\exists A \in [X]^{<\infty}$ of cardinality $d+1$ such that $2^A = \mathcal{F}_{|_A} \cup \{f_{|_A}\} = \mathcal{F}_{|_A}$, contradicting $\mathrm{VC}(\mathcal{F}) = d$. ∎

One would wonder whether in a measurable space $(X, \mathfrak{A})$ with $\mathcal{C} \subseteq \mathfrak{A}$ and $\mathrm{VC}(\mathcal{C}) < \infty$, the closure of $\mathcal{C}$ in $\mathbb{R}^X$ lies within $\mathfrak{A}$. The answer is negative in general as shown by our following example (the definition of universally measurable is found in appendix A.1):

Example 4.1.4 Let $(X, \mathcal{B})$ be an uncountable standard Borel space and $\mathfrak{A}$ the sigma algebra of universally Borel measurable sets. Let κ be the cardinal

$$\kappa = \min\{\kappa' : \exists A \subseteq X \text{ non } \mathfrak{A}\text{-measurable and } |A| = \kappa'\}.$$

Of course, $\aleph_0 < \kappa \leq \mathfrak{c}$, but the value of κ cannot be specified in ZFC and would require additional set theoretic axioms; for example under Martin's Axiom $\kappa = \mathfrak{c}$. Fix $Y \subseteq X$ non $\mathfrak{A}$-measurable with $|Y| = \kappa$. Well order Y with $\prec$ such that each initial segment $I_x^{\prec} = \{y : y \preceq x\}$ of Y has cardinality strictly less than Y (in other words fix a minimal well ordering on Y). By the definition of κ, each $I_x^{\prec}$ is universally measurable. Now set for $x \in Y$,

$$f_x = \chi_{I_x^{\prec}}.$$

One has $\{f_x\}_{x \in Y} \subseteq \mathfrak{A}$ and $\mathrm{VC}(\{f_x\}_{x \in Y}) = 1$ ($(Y, \{f_x\}_{x \in Y})$ is 1-maximum). We have that $f_x \to \chi_Y$ as a net with Y directed by $\prec$ because for every $A \in [X]^{<\infty}$, letting $x \succeq a = \max_{\prec} A \cap Y$ (or pick any $a \in Y$ if $A \cap Y = \emptyset$), we have

$$f_{x|_A} = \chi_{I_x^{\prec} \cap Y|_A} = \chi_{I_x^{\prec} \cap A \cap Y|_A} = \chi_{A \cap Y|_A} = \chi_{Y|_A}.$$

We can use the previous example to show that the compactness theorem, using sample compression schemes on finite subdomains, can create a sample compression scheme returning a nonmeasurable hypothesis.

Example 4.1.5 In the previous example let

$$\mathcal{C} = \{I_x^{\prec} : I_x^{\prec} \text{ is an initial segment of} Y\}.$$

Note that $(Y, \mathcal{C})$ is 1-maximum and by example 2.1.4,

$$\mathcal{H}(\{x\}) = I_x^{\prec} \setminus \{x\}, \ \mathcal{H}(\emptyset) = Y$$

is a sample compression scheme of size 1 for $(Y, \mathcal{C})$. We have that $\mathcal{H}(\emptyset) = Y$ is not measurable even though $\mathcal{C}$ consists only of measurable sets.

Note however by example 2.1.4

$$\mathcal{H}(\{x\}) = I_x^{\prec},\ \mathcal{H}(\emptyset) = \emptyset$$

is also a sample compression scheme of size 1 for $(Y, \mathcal{C})$, and has measurable hypotheses. However, in this thesis we are unable to find a concept space with VC dimension d and measurable concepts, in which any sample compression scheme of size d must have a nonmeasurable hypothesis.

Lemma 4.1.6 *Let $(X, \mathcal{C})$ be a concept space. For every finite subspace $(A, \mathcal{C} \sqcap A)$, let $\mathcal{H}_A$ be a compression scheme of size d on $(A, \mathcal{C} \sqcap A)$, let $\mathcal{H}$ be the compression scheme of size d on X defined by $\{\mathcal{H}_A\}_{A \in [X]^{<\infty}}$ via the compactness theorem (theorem 2.2.1), and let $\sigma \in [X]^{\leq d}$. If for all $A \in [X]^{<\infty}$ with $\sigma \subseteq A$, a function $f_{A,\sigma} \in 2^X$ is any extension of $\mathcal{H}_A(\sigma)$ to X, we have that $\mathcal{H}(\sigma)$ is a cluster point of the net $\{f_{A,\sigma}\}_{\sigma \subseteq A \in [X]^{<\infty}}$ indexed by A (where $\{A \in [X]^{<\infty} : \sigma \subseteq A\}$ is directed by inclusion) in the topology of pointwise convergence; the product topology on 2^X.*

Proof: Let $\sigma \in [X]^{\leq d}$, let $\{f_{A,\sigma}\}_{\sigma \subseteq A \in [X]^{<\infty}}$, $\{\mathcal{H}_A\}_{A \in [X]^{<\infty}}$, $\mathcal{H}$ be as in the statement of the theorem, and let $\mathfrak{U}$ be the ultrafilter on $[X]^{<\infty}$ defined as in our proof of the compactness theorem. Let $x \in X$,

$$\mathcal{A}_x = \{A \in [X]^{<\infty} : \sigma \cup \{x\} \subseteq A,\ \mathcal{H}_A(\sigma)(x) = \mathcal{H}(\sigma)(x)\}.$$

$\mathcal{A}_x \in \mathfrak{U}$ because

$$\mathcal{H}(\sigma)(x) = 1 \Rightarrow \mathcal{A}_x = \{A \in [X]^{<\infty} : \sigma \cup \{x\} \subseteq A,\ \mathcal{H}_A(\sigma)(x) = 1 = \mathcal{H}(\sigma)(x)\} \in \mathfrak{U}$$

and

$$\mathcal{H}(\sigma)(x) = 0 \Rightarrow \mathcal{A}_x^c = \{A \in [X]^{<\infty} : \sigma \cup \{x\} \subseteq A,\ \mathcal{H}_A(\sigma)(x) = 1 \neq \mathcal{H}(\sigma)(x)\} \notin \mathfrak{U} \Rightarrow \mathcal{A}_x \in \mathfrak{U}.$$

Now fix a basic open neighbourhood

$$U = \{\mathcal{H}(\sigma)(x_1)\}_{x_1} \times ... \times \{\mathcal{H}(\sigma)(x_n)\}_{x_n} \times \prod_{x \notin \{x_1,...,x_n\}} \{0,1\}_x$$

of $\mathcal{H}(\sigma)$ in 2^X, and let $B \in [X]^{<\infty}$. There is

$$B' \in \bigcap_{i=1}^{i=n} \mathcal{A}_{x_i} \cap \bigcap_{b \in B} \mathcal{A}_b =$$

$$= \{A \in [X]^{<\infty} : \sigma \cup \{x_1, ..., x_n\} \cup B \subseteq A,\ \mathcal{H}_A(\sigma)_{|_{\{x_1,...,x_n\}}} = \mathcal{H}(\sigma)_{|_{\{x_1,...,x_n\}}}\} \in \mathfrak{U},$$

hence $\sigma \cup \{x_1, ..., x_n\} \cup B \subseteq B'$ and $\mathcal{H}_{B'}(\sigma)_{|_{\{x_1,...,x_n\}}} = \mathcal{H}(\sigma)_{|_{\{x_1,...,x_n\}}}$. There is $f_{B',\sigma}$ such that $f_{B',\sigma|_{B'}} = \mathcal{H}_{B'}(\sigma)$ thus $f_{B',\sigma|_{\{x_1,...,x_n\}}} = \mathcal{H}(\sigma)_{|_{\{x_1,...,x_n\}}}$ and therefore $f_{B',\sigma} \in U$. ∎

Lemma 4.1.7 (Measurable Compactness Lemma) *Let $(X, \mathfrak{A})$ be a measurable space. $(X, \mathcal{C})$ has a sample compression scheme of size d with measurable hypotheses if and only if every finite subspace $(A, \mathcal{C} \sqcap A)$ of $(X, \mathcal{C})$ has a sample compression scheme of size d, $\mathcal{H}_A$, where for all $\sigma \in [X]^{\leq d}$, there is $d_\sigma \in \mathbb{N}$ and $\mathcal{M}_\sigma \subseteq \mathfrak{A}$ d_σ-maximal, such that*

$$\{\mathcal{H}_A(\sigma)\}_{\sigma \subseteq A \in [X]^{<\infty}} \subseteq \{f_{|_A} : f \in \mathcal{M}_\sigma,\ \sigma \subseteq A \in [X]^{<\infty}\}.$$

Proof: *Necessity*: Let $\mathcal{H}$ be a sample compression scheme of size d with measurable hypotheses on $(X, \mathcal{C})$, and let $\sigma \in [X]^{\leq d}$. Now set $\mathcal{M}_\sigma = \{\mathcal{H}(\sigma)\}$, noting that $\mathcal{M}_\sigma$ is 0-maximal. Let $A \subseteq X$ be a finite subset, we have that

$$\mathcal{H}_A : [A]^{\leq d} \to 2^A,\ \sigma \mapsto \mathcal{H}(\sigma)_{|_A} \in \mathcal{M}_{\sigma|_A}$$

is a sample compression scheme of size d on A.

Sufficiency: Let $\{\mathcal{H}_A\}_{A \in [X]^{<\infty}}$, $\{d_\sigma\}_{\sigma \in [X]^{\leq d}}$, $\{\mathcal{M}_\sigma\}_{\sigma \in [X]^{\leq d}}$ be as in the statement of the theorem, and let $\mathcal{H}$ be the sample compression scheme of size d on $(X, \mathcal{C})$ defined by $\{\mathcal{H}_A\}_{A \in [X]^{<\infty}}$ via the compactness theorem (theorem 2.2.1). Let $\sigma \in [X]^{\leq d}$ be given. By lemma 4.1.3, $\mathcal{M}_\sigma$ is closed in 2^X since it is maximal. Since $\mathcal{H}_A(\sigma) \in \mathcal{M}_{\sigma|_A}$ for all $\sigma \subseteq A \in [X]^{<\infty}$, by lemma 4.1.6,

$$\mathcal{H}(\sigma) \in \overline{\mathcal{M}_\sigma} = \mathcal{M}_\sigma \subseteq \mathfrak{A}.$$

∎

Theorem 4.1.8 *If $(X, \mathfrak{A})$ is a measurable space and $\mathcal{C} \subseteq \mathfrak{A}$ is a d-maximum and d-maximal class, then $(X, \mathcal{C})$ has a sample compression scheme of size d with measurable hypotheses. In this case, the sample compression scheme maps to concepts in $\mathcal{C}$.*

Proof: $(X, \mathcal{C})$ is d-maximum implies that: $(X, \mathcal{C})$ has a sample compression scheme $\mathcal{H}$ of size d, and every finite subspace $(A, \mathcal{C} \sqcap A)$ is d-maximum. For every finite subspace $(A, \mathcal{C} \sqcap A)$, we have $\mathcal{H}_A : [A]^{\leq d} \to 2^A$, $\sigma \mapsto \mathcal{H}(\sigma)_{|A}$ is a sample compression scheme of size d on A. Since $(A, \mathcal{C} \sqcap A)$ is d-maximum, for all $\sigma \in [A]^{\leq d}$, $\mathcal{H}_A(\sigma) \in \mathcal{C} \sqcap A$. This shows that the hypothesis of lemma 4.1.7 in the converse direction holds for $\mathcal{M}_\sigma = \mathcal{C}$ and so $(X, \mathcal{C})$ has a sample compression scheme of size d with measurable hypotheses. In particular, from the proof of the theorem, we see that

$$\{\mathcal{H}(\sigma) : \sigma \in [X]^{\leq d}\} \subseteq \bigcup_{\sigma \in [X]^{\leq d}} \mathcal{M}_\sigma = \mathcal{C}.$$

∎

Notation 4.1.9 $\mathbf{C}(X)$ is notation for the set of continuous functions from X to $\mathbb{R}$, where $\mathbb{R}$ has the usual topology and the topology on X will be clear in the context.

The following theorem is an important result from [5] (theorem 2F p. 854), which will allow us to drop the maximality condition of the corollary in some circumstances.

Theorem 4.1.10 *Let X be any Hausdorff space and $\mathcal{C} \subseteq \mathbf{C}(X)$ a pointwise bounded set. Then $\mathcal{C}$ is relatively compact in the collection of universally Borel measurable real valued functions on X under the topology of pointwise convergence if and only if for*

every $K \subseteq X$ compact, every $\{C_n\}_{n\in\mathbb{N}} \subseteq \mathcal{C}$, and every $\alpha < \beta$ in $\mathbb{R}$, there exists $I \subseteq \mathbb{N}$ such that

$$\{x \in K : C_n(x) \leq \alpha \ \forall n \in I,\ C_n(x) \geq \beta \ \forall n \in \mathbb{N} \setminus I\} = \emptyset.$$

Lemma 4.1.11 *Let X be any Hausdorff space with a d-maximum concept class $\mathcal{C}$ consisting of clopen sets. Then $(X, \mathcal{C})$ has a sample compression scheme of size d with universally Borel measurable hypotheses.*

Proof: Clearly $\mathcal{C} \subseteq \mathbf{C}(X)$, and $\mathcal{C}$ is a pointwise bounded set of functions. Suppose there is $\{C_n\}_{n\in\mathbb{N}} \subseteq \mathcal{C}$ such that $\forall I \subseteq \mathbb{N}$, there is $x \in X$ where $x \notin C_n \ \forall n \in I$ and $x \in C_n \ \forall n \in \mathbb{N} \setminus I$. This implies that $\mathrm{VC}((X, \mathcal{C})^*) = \infty$, which contradicts $\mathrm{VC}(X, \mathcal{C}) < \infty$. Thus for every $\{C_n\}_{n\in\mathbb{N}} \subseteq \mathcal{C}$, there exists $I \subseteq \mathbb{N}$ such that

$$\{x \in X : C_n(x) = 0 \ \forall n \in I,\ C_n(x) = 1 \ \forall n \in \mathbb{N} \setminus I\} = \emptyset,$$

and so by the previous theorem $\mathcal{C}$ is relatively compact in the collection of universally Borel measurable real valued functions on X. Since $\mathcal{C}$ is d-maximum, for all $\sigma \in [X]^{\leq d}$, $\sigma \subseteq A \in [X]^{<\infty}$, we have $\mathcal{H}_A(\sigma) \in \mathcal{C}_{|A}$, and so using lemma 4.1.6 similarly to lemma 4.1.7 we have that $(X, \mathcal{C})$ has a sample compression scheme $\mathcal{H}$ of size d such that $\{\mathcal{H}(\sigma) : \sigma \in [X]^{\leq d}\} \subseteq \overline{\mathcal{C}}$. Therefore $\{\mathcal{H}(\sigma) : \sigma \in [X]^{\leq d}\}$ consists of universally Borel measurable sets. ∎

The following definitions can be found in reference [15] pp. 38-39.

Definition 4.1.12 *A family $\mathcal{C}$ of functions has a subset $\mathcal{D}$ which is **universally dense** in $\mathcal{C}$ if every $C \in \mathcal{C}$ is the pointwise limit of a sequence in $\mathcal{D}$.*
*A concept space $(X, \mathcal{C})$ is **universally separable** if there exists a countable universally dense subset $\mathcal{D}$ of $\mathcal{C}$.*

Remark 4.1.13 $\mathcal{D}'$ is universally dense in $\mathcal{D}$, is equivalent to $\mathcal{D}'$ is dense in $\mathcal{D}$ in the topology generated by the $L^1(P)$ seminorm for every $P \in Pr(X)$.

Remark 4.1.14 Note that for every $A \in [X]^{<\infty}$, if $\mathcal{D}$ is universally dense in $\mathcal{C}$, then $\mathcal{D} \sqcap A = \mathcal{C} \sqcap A$. This implies that $\mathcal{H}$ is a sample compression scheme of size d on $(X, \mathcal{D})$ if and only if it is a sample compression scheme of size d on $(X, \mathcal{C})$. Therefore, to prove something about the sample compression schemes for a universally separable concept class, we only need to consider sample compression schemes for countable concept classes.

We will use the following classical result in descriptive set theory (which can be found in [9] pp. 83) about standard Borel spaces to get that all d-maximum universally separable concept spaces, in which the countable universally dense subset consists of Borel sets, have a sample compression scheme of size d with universally Borel measurable hypotheses:

Lemma 4.1.15 *Let X be a standard Borel space, and let $\mathcal{C}$ be a countable family of Borel sets. There is a refinement of the topology on X to a Polish topology generating the same Borel sigma algebra in which $\mathcal{C}$ consists of clopen sets.*

By combining the last two lemmas (lemmas 4.1.11 and 4.1.15) and noting remark 4.1.14, we get the following potentially useful result:

Theorem 4.1.16 *Let X be a standard Borel space with a d-maximum and universally separable concept class $\mathcal{C}$. Then $(X, \mathcal{C})$ has a sample compression scheme of size d with universally Borel measurable hypotheses.*

Open Questions

We have raised the following open question: even assuming that a sample compression scheme exists for every finite concept subspace of a given space, can we conclude that such a scheme consisting of measurable hypotheses will exist for a concept class consisting of Borel sets on a standard Borel domain? A stronger version of this question can also be asked for the measurability condition (M5), that one needs for the proof of the main sample complexity bound for sample compression schemes: Given that a sample compression scheme exists for every finite concept subspace of a given space, can we conclude that such a scheme also satisfying (M5) will exist for a concept class consisting of Borel sets on a standard Borel domain? Finally, one can ask if there is validity of either question under a stronger assumption on the concept class: e.g., image admissible Suslin, or universally separable.

It is an open question of whether or not there exists a sample compression scheme of size $O(d)$ for a concept space of VC dimension d. We may ask similar but weaker questions of copy sample compression schemes that may help in clarifying methods to prove, or disprove, the original question. For example: Given a concept space of VC dimension d is there an $O(d)$-copy sample compression scheme of size $O(d)$?

Appendix A

A.1 Measure Theory Preliminaries

The following are some preliminary measure theoretic definitions, and can be found in [3] and [9].

Definition A.1.1 *A **sigma algebra** $\mathfrak{A}$ on a set X, is a set of subsets of X satisfying:*

(1) $X \in \mathfrak{A}$

(2) $A \in \mathfrak{A}$ *implies* $A^c \in \mathfrak{A}$

(3) $\{A_i\}_{i=0}^{\infty} \subseteq \mathfrak{A}$ *implies* $\bigcup_{i=0}^{\infty} A_i \in \mathfrak{A}$.

*A **measurable space** $(X, \mathfrak{A})$ is a set X equipped with a sigma algebra $\mathfrak{A}$.*

It is easy to see that an intersection of sigma algebras is also a sigma algebra.

Definition A.1.2 *A **sigma algebra generated** by a family $\mathfrak{B} \subseteq 2^X$, denoted $\sigma(\mathfrak{B})$, is the intersection of all sigma algebras on X containing $\mathfrak{B}$.*

Definition A.1.3 *A **measure** on measurable space $(X, \mathfrak{A})$, is a function $\mu : \mathfrak{A} \to [0, \infty]$ satisfying:*

(1) $\mu(\emptyset) = 0$

(2) if $\{A_i\}_{i\in I} \subseteq \mathfrak{A}$ *is a countable pairwise disjoint family, then*

$$\mu(\bigcup_{i\in I} A_i) = \sum_{i\in I} \mu(A_i).$$

A ***probability measure*** *on measurable space* $(X, \mathfrak{A})$*, is a measure* P *also satisfying* $P(X) = 1$*. We denote the set of all probability measures on* $(X, \mathfrak{A})$ *by* $Pr(X)$*. A* ***measure space*** $(X, \mathfrak{A}, \mu)$ *is a measurable space* $(X, \mathfrak{A})$ *equipped with a measure* μ*.*

Definition A.1.4 *The* ***completion*** *of a measure space* $(X, \mathfrak{A}, \mu)$ *is the measure space* $(X, \mathfrak{A}_\mu, \widehat{\mu})$ *where*

$$\mathfrak{A}_\mu = \sigma(\mathfrak{A} \cup \{B \in 2^X : \exists A \in \mathfrak{A},\ B \subseteq A \text{ and } \mu(A) = 0\}),$$

and

$$\widehat{\mu} : \mathfrak{A}_\mu \to [0, \infty]; B \mapsto \inf\{\mu(A) : A \in \mathfrak{A},\ B \subseteq A\}.$$

It is true that $\widehat{\mu}$ is the unique measure extending μ to $\mathfrak{A}_\mu$, and any $B \in \mathfrak{A}_\mu$ is of the form $A \cup C$ where $A \in \mathfrak{A},\ C \in \{Y \in 2^X : \exists Y' \in \mathfrak{A},\ Y \subseteq Y' \text{ and } \mu(Y') = 0\}$.

Definition A.1.5 *For a measure space* $(X, \mathfrak{A})$ *the sigma algebra of* ***universally measurable*** *subsets of* X *(with respect to* $\mathfrak{A}$*) is*

$$\mathfrak{A}^* = \bigcap_{P\in Pr(X)} \mathfrak{A}_P.$$

We say a set is universally measurable (with respect to $\mathfrak{A}$*), if it is an element of* $\mathfrak{A}^*$*.*

Definition A.1.6 *The* ***Borel sigma algebra*** *of a topological space* $(X, \mathcal{T})$ *is the sigma algebra* $\mathcal{B} = \sigma(\mathcal{T})$*. A* ***Borel space*** *is a measurable space* $(X, \mathcal{B})$ *on a topological space, where the sigma algebra* $\mathcal{B}$ *is the Borel sigma algebra on* X*.*

Definition A.1.7 *A* ***Polish space*** *is a topological space* $(X, \mathcal{T})$ *that is metrizable by a complete metric, and separable.*

Definition A.1.8 *A **standard Borel space** is a Borel space associated to a Polish topological space.*

Every uncountable Borel space $(X, \mathcal{B})$ is isomorphic as a measure space to $[0, 1]$ with the Borel sigma algebra (ie. there is a bijection $f : X \to [0, 1]$ such that for every $A \subseteq X$, A is measurable in X iff $f(A)$ is measurable in $[0, 1]$) [9].

A.2 Nets and Filters Preliminaries

The following are some preliminary definition that can be found in [13] and [1]

Definition A.2.1 *A **directed set** I is a set equipped with a partial order such that*

$$\forall \alpha, \beta \in I,\ \exists \gamma \in I,\ \gamma \geq \beta \text{ and } \gamma \geq \alpha.$$

*A subset J of a directed set I is **cofinal** in I if*

$$\forall \alpha \in I,\ \exists \beta \in J,\ \beta \geq \alpha.$$

Definition A.2.2 *A **net** in X is a function mapping a directed set to X. A **subnet** of a net $f : I \to X$ is the function $f \circ g$, where g is a nondecreasing function from a directed set J to I with $g(J)$ cofinal in I.*

Definition A.2.3 *A point x of is a **cluster point** of a net $\{x_\alpha\}_{\alpha \in I}$ in a topological space, if for every neighbourhood U of x the set $\{\alpha \in I : x_\alpha \in U\}$ is cofinal in I.*
*A net $\{x_\alpha\}_{\alpha \in I}$ in a topological space **converges** to a point x if for every neighbourhood U of x*

$$\exists \alpha \in I,\ \forall \beta \geq \alpha,\ x_\beta \in U.$$

A point x is a cluster point of a net if and only if there is a subnet converging to x.

Definition A.2.4 *A **filter** $\mathcal{F}$ on a set X is a set of subsets of X satisfying:*

(1) $\emptyset \notin \mathcal{F}$

(2) $A \in \mathcal{F}$ *and* $A \subseteq B$ *imply* $B \in \mathcal{F}$

(3) $A, B \in \mathcal{F}$ *implies* $A \cap B \in \mathcal{F}$.

*A **filter base** $\mathcal{F}$ on a set X is a set of subsets of X satisfying:*

(1) $\emptyset \notin \mathcal{F}$

(2) $A, B \in \mathcal{F}$ *implies* $A \cap B \in \mathcal{F}$.

*An **ultrafilter** $\mathfrak{U}$ on a set X is a filter on X where every $A \subseteq X$ has either $A \in \mathfrak{U}$ or $A^c \in \mathfrak{U}$.*

Every filter base $\mathcal{F}$ generates a unique filter that is equal to the intersection of all filters containing $\mathcal{F}$, and every filter is contained in an ultrafilter as a consequence of Zorn's Lemma.

There are two types of ultrafilters, principal and free.

Definition A.2.5 *An ultrafilter $\mathfrak{U}$ on X is **principal** if it has a least element under set inclusion, and is **free** if it is not principal.*

Principal ultrafilters are of the form $\mathfrak{U}_A = \{B \in 2^X : A \subseteq B\}$ for some $A \subseteq X$.

Appendix B

B.1 Well Behaved Hypothesis Spaces and Other Forms of Measurability

The subject of this section will be additional conditions of measurability for learning rules.

In the definition of a learning rule we had required that the hypothesis class is measurable, and that the function

$$A \mapsto P(\mathcal{L}(A, C_{|A}) \triangle C)$$

is measurable. We will call these conditions (M2) and (M1) respectively. Note that (M1) implies (M2) and the consideration of chapter 4 was the (M2) condition for sample compression schemes.

Fix a measurable space $(X, \mathfrak{A})$.

Notation B.1.1 Fix a set C, and an $\varepsilon > 0$. Define

$$\mathcal{D}_C = \{H \triangle C : H \in \mathfrak{H}\}.$$

When $\{C\} \cup \mathfrak{H} \subseteq \mathfrak{A}$, for a measure $P \in Pr(X)$ define

$$\mathcal{D}_C^{\varepsilon} = \{D \in \mathcal{D}_C : P(D) > \varepsilon\}.$$

Note that for $m \in \mathbb{N}$, $\mathfrak{A}^m$ is notation for the product sigma algebra on X^m.

Definition B.1.2 ([17], [4]) *Let $\{C\} \cup \mathfrak{H} \subseteq \mathfrak{A}$ and $\varepsilon > 0$. Define*

$$Q_\varepsilon^m = \{A \in X^m : \exists D \in \mathcal{D}_C^\varepsilon, \text{ such that } A \in (D^c)^m\}$$

and define

$$J_{\varepsilon,r}^{m+k} =$$

$$= \{A \in X^{m+k} : \exists D \in \mathcal{D}_C^\varepsilon, A_{|_{\{1,..,m\}}} \in (D^c)^m \text{ and } \exists I \subseteq \{m+1, ..., m+k\},$$
$$\text{such that } |I| \geq kr\varepsilon \text{ and } A_{|_I} \in D^{|I|}\}.$$

*We will say a hypothesis space satisfies **(M3)** if:*

$$\text{for all } m \geq 1,\ \varepsilon > 0,\ C \in \mathfrak{A},\ P \in Pr(X, \mathfrak{A}),$$

we have

$$Q_\varepsilon^m \in \mathfrak{A}^m,$$

and

$$\text{for all } m \geq \frac{4d}{\varepsilon},\ \varepsilon > 0,\ k = m(\frac{\varepsilon r m}{d} - 1),\ r = 1 - \sqrt{\frac{2}{\varepsilon k}},\ C \in \mathfrak{A},\ P \in Pr(X, \mathfrak{A})$$

such that k is in $\mathbb{N}$, we have

$$J_{\varepsilon,r}^{m+k} \in \mathfrak{A}^{m+k}.$$

(Note that $r = 1 - \sqrt{\frac{2}{\varepsilon k}}$, $k = m(\frac{\varepsilon r m}{d} - 1)$ have a solutions in $\mathbb{R}^+$ for $m \geq \frac{4d}{\varepsilon}$.)
*We will say a hypothesis space satisfies **(M4)**, or is **well behaved** if:*

$$\text{for all } m \geq 1,\ \varepsilon > 0,\ C \in \mathfrak{A},\ P \in Pr(X, \mathfrak{A}),$$

we have

$$Q_\varepsilon^m \in \mathfrak{A}^m,$$

and

$$\textit{for all } m \geq 1,\ \varepsilon > 0,\ k = m,\ r = \frac{m}{2},\ C \in \mathfrak{A},\ P \in Pr(X, \mathfrak{A}),$$

we have

$$J_{\varepsilon,r}^{m+k} \in \mathfrak{A}^{m+k}.$$

Proposition B.1.3 ([4]) *If $\mathfrak{H}$ is universally separable and (M2) then $\mathfrak{H}$ satisfies (M3) and (M4).*

Proof: Fix $m \geq 1,\ \varepsilon > 0,\ C \in \mathfrak{A},\ P \in Pr(X, \mathfrak{A})$. Since $\mathfrak{H}$ is universally separable, $\mathcal{D}_C$ is universally separable. Let $\mathcal{D}'$ be a countable universally dense subset of $\mathcal{D}_C$, let $\{\delta_i\}_{i=1}^{\infty} \subseteq \mathbb{R}$ be a decreasing sequence converging to zero, let $\{\varepsilon_i\}_{i=1}^{\infty} \subseteq \mathbb{R}$ be a decreasing sequence converging to ε, and let

$$\mathcal{D}'_{i,j} = \{D' \in \mathcal{D}' : \exists D \in \mathcal{D}_C \text{ with } P(D) \geq \varepsilon_i \text{ and } P(D' \triangle D) \leq \delta_j\}$$

We will show

$$Q_{\varepsilon}^{m} = \bigcup_{i=0}^{\infty} \bigcap_{j=0}^{\infty} \bigcup_{D' \in \mathcal{D}'_{i,j}} (D'^{c})^{m},$$

and hence $Q_{\varepsilon}^{m} \in \mathfrak{A}^{m}$.

Let $A \in \bigcup_{i=0}^{\infty} \bigcap_{j=0}^{\infty} \bigcup_{D' \in \mathcal{D}'_{i,j}} (D'^{c})^{m}$. Then $A \in (D'^{c})^{m}$ for some $D' \in \mathcal{D}'_{i,j}$ where $\varepsilon_i - \delta_i > \varepsilon$. This implies $D' \in \mathcal{D}_C^{\varepsilon}$ and so $A \in Q_{\varepsilon}^{m}$.

Now let $A \in Q_{\varepsilon}^{m}$. Then there is $D \in \mathcal{D}_C^{\varepsilon}$ and $i \in \mathbb{N}$ such that $A \in (D^{c})^{m}$ and $P(D) \geq \varepsilon_i$. Picking any sequence of sets $\{D'_i\}_{i=0}^{\infty}$ in $\mathcal{D}'$ converging pointwise to D, we have also have $\{(D'^{c}_i)^{m}\}_{i=0}^{\infty}$ converges pointwise to $(D^{c})^{m}$, and so for every $j \in \mathbb{N}$ there is D'_i where $P(D'_i \triangle D) \leq \delta_j$ and $A \in (D'^{c}_i)^{m}$. Thus $A \in \bigcup_{i=0}^{\infty} \bigcap_{j=0}^{\infty} \bigcup_{D' \in \mathcal{D}'_{i,j}} (D'^{c})^{m}$.

Using similar arguments we can show, given k and r such that $J_{\varepsilon,r}^{m+k}$ is defined, we have

$$J_{\varepsilon,r}^{m+k} = \bigcup_{i=0}^{\infty} \bigcap_{j=0}^{\infty} \bigcup_{D' \in \mathcal{D}'_{i,j}} ((D'^{c})^{m} \times X^{k} \cap \bigcup_{l=\lceil kr\varepsilon \rceil}^{k} \{A \in X^{k} : \text{ exactly } l \text{ coordinates of } A \text{ are in } D'^{c}\}).$$

∎

Definition B.1.4 *Let $\{n_i\}_{i=0}^k$ be a finite sequence in $\mathbb{N}$. We will say that a function*

$$\mathcal{H} : \bigcup_{i \in \{j \in \mathbb{N} : 0 \leq j \leq k,\ n_j \neq 0\}} ([X]^{=i} \times \{1, ..., n_i\}) \to 2^X$$

*satisfies **(M5)** if for every $1 \leq i \leq k$ and $l \in \{1, ..., n_i\}$,*

$$\{(x_1, ..., x_{i+1}) \in X^{i+1} : \mathcal{H}(\{x_1, ..., x_i\} \times \min(l, n_{|\{x_1, ..., x_i\}|}))(x_{i+1}) = 1\} \in \mathfrak{A}^{i+1}.$$

In particular, a function

$$\mathcal{H} : [X]^{\leq d} \to 2^X$$

satisfies (M5) if for every $1 \leq i \leq d$

$$\{(x_1, ..., x_{i+1}) \in X^{i+1} : \mathcal{H}(\{x_1, ..., x_i\})(x_{i+1}) = 1\} \in \mathfrak{A}^{i+1}.$$

B.2 Sample Complexity of Copy Sample Compression Schemes

Here we include the omitted proofs from section 3.2 regarding sample complexity of copy sample compression schemes.

Theorem B.2.1 *Let P be any probability measure on a measurable space $(X, \mathfrak{A})$, C a concept in $\mathcal{C} \subseteq \mathfrak{A}$, $\{n_i\}_{i=0}^k \subseteq \mathbb{N}$, and $\mathcal{H}$ any function from $\bigcup_{i \in \{j \in \mathbb{N} : 0 \leq j \leq k,\ n_j \neq 0\}} ([X]^{=i} \times \{1, ..., n_i\})$ to 2^X, satisfying measurability condition (M5). Then the probability that $A \subseteq X$, $|A| = m \geq k$, contains a subset σ of size at most k, and $l \in \{1, ..., n_{|\sigma|}\}$ such that $P(\mathcal{H}(\sigma \times l) \triangle C) > \varepsilon > 0$ and $\mathcal{H}(\sigma \times l)_{|_A} = C_{|_A}$, is at most*

$$\sum_{i=0}^{k} n_i \binom{m}{i} (1 - \varepsilon)^{m-i}.$$

Proof: Let $C \in \mathcal{C}$ and ε be given. First we consider the probability that a set of size m has a subset of size exactly $i \leq d$ with the property $P(\mathcal{H}(\sigma \times l) \triangle C) > \varepsilon$ and $\mathcal{H}(\sigma \times l)_{|_A} = C_{|_A}$ for some $l \in \{1, ..., n_{|\sigma|}\}$. For $A = (a_1, ..., a_m) \in X^m$ and $J = \{j_1, ..., j_i\} \subseteq \{1, ..., m\}$, let $A_{|_J}$ denote $\{a_{j_1}, ..., a_{j_i}\}$.

There are $\binom{m}{i}$ many subsets of A of size i, hence fixing J a subset of $\{1, ..., m\}$ of size i, the probability we wish to bound above is at most

$$P^m(\{A \in X^m : \ \exists I \subseteq \{1, ..., m\} \text{ of size } i, \ \exists t \in \{1, ..., n_i\},$$
$$\text{such that } P(\mathcal{H}(A_{|_I} \times t) \triangle C) > \varepsilon \text{ and } \mathcal{H}(A_{|_I} \times \min(t, n_{|A_{|_I}|}))_{|_A} = C_{|_A}\})$$
$$= \sum_{l=1}^{n_i} \binom{m}{i} P^m(\{A \in X^m : \ P(\mathcal{H}(A_{|_J} \times \min(l, n_{|A_{|_J}|})) \triangle C) > \varepsilon$$
$$\text{and } \mathcal{H}(A_{|_J} \times \min(l, n_{|A_{|_J}|}))_{|_A} = C_{|_A}\}).$$

Since permuting J to some other subset of size i in $\{1, ..., m\}$ does not affect the above probability, we can assume $J = \{1, ..., i\}$. Fix $l \in \{1, ..., n_i\}$.

We will prove at this point that
$\{A \in X^m : \ P(\mathcal{H}(A_{|_J} \times \min(l, n_{|A_{|_J}|})) \triangle C) > \varepsilon \text{ and } \mathcal{H}(A_{|_J} \times \min(l, n_{|A_{|_J}|}))_{|_A} = C_{|_A}\}$ is measurable due to the hypothesis that $\mathcal{H}$ satisfies (M5): Let $1 \leq p < q \leq m$ and let $\pi_{p,q}$ be the (measurable) function from X^m to X^{p+1} mapping $(x_1, ..., x_m) \mapsto (x_1, ..., x_p, x_q)$. By (M5) and the measurability of C we have

$$\{A \in X^m : \mathcal{H}(A_{|_J} \times \min(l, n_{|A_{|_J}|}))_{|_A} = C_{|_A}\}$$
$$= \{A \in X^m : A \in ((\mathcal{H}(A_{|_J} \times \min(l, n_{|A_{|_J}|})) \triangle C)^c)^m\}$$
$$= \bigcap_{q=1}^{m} \{A \in X^m : (\mathcal{H}(A_{|_J} \times \min(l, n_{|A_{|_J}|}))) \triangle C)^c)(A_{|_{\{q\}}}) = 1\}$$
$$= \bigcap_{q=1}^{m} \pi_{i,q}^{-1}(\{(x_1, ..., x_{i+1}) \in X^{i+1} : (\mathcal{H}(\{x_1, ..., x_i\} \times \min(l, n_{|\{x_1, ..., x_i\}|})) \triangle C)^c)(x_{i+1}) = 1\})$$
$$\in \mathfrak{A}^m.$$

Also $\{A \in X^m : P(\mathcal{H}(A_{|_J} \times \min(l, n_{|A_{|_J}|}))\triangle C) > \varepsilon\}$ is measurable since

$$B = \{(x_1, ..., x_{m+1}) \in X^{m+1} : (\mathcal{H}(\{x_1, ..., x_i\} \times \min(l, n_{|\{x_1,...,x_i\}|}))\triangle C)^c(x_{m+1}) = 1\}$$
$$= \pi^{-1}_{i,m+1}(\{(x_1, ..., x_{i+1}) \in X^{i+1} : (\mathcal{H}(\{x_1, ..., x_i\} \times \min(l, n_{|\{x_1,...,x_i\}|}))\triangle C)^c(x_{i+1}) = 1\})$$

is measurable by (M5) and the measurability of C, and a straightforward application of Fubini's theorem gives us that the map

$$(x_1, ..., x_m) \mapsto \int_X \chi_B(x_1, ..., x_{m+1}) dP(x_{m+1}) =$$
$$= P(\{y : (x_1, ..., x_m, y) \in B\})$$
$$= P(\{y : y \in \mathcal{H}(\{x_1, ..., x_i\} \times \min(l, n_{|\{x_1,...,x_i\}|}))\triangle C)^c\})$$
$$= P(\mathcal{H}(\{x_1, ..., x_i\} \times \min(l, n_{|\{x_1,...,x_i\}|}))\triangle C)^c)$$

is measurable.

Now let

$$E_C = \{A \in X^m : \mathcal{H}(A_{|_J} \times \min(l, n_{|A_{|_J}|}))_{|_{A_{|_{\{i+1,...,m\}}}}} = C_{|_{A_{|_{\{i+1,...,m\}}}}}\}$$
$$= \{A \in X^m : A_{|_{\{i+1,...,m\}}} \in ((\mathcal{H}(A_{|_J} \times \min(l, n_{|A_{|_J}|}))\triangle C)^c)^{m-i}\}$$
$$= \bigcap_{q=i+1}^{m} \{A \in X^m : (\mathcal{H}(A_{|_J} \times \min(l, n_{|A_{|_J}|})))\triangle C)^c)(A_{|_{\{q\}}}) = 1\}$$
$$= \bigcap_{q=i+1}^{m} \pi^{-1}_{i,q}(\{(x_1, ..., x_{i+1}) \in X^{i+1} : (\mathcal{H}(\{x_1, ..., x_i\} \times \min(l, n_{|\{x_1,...,x_i\}|}))\triangle C)^c)(x_{i+1}) = 1\})$$
$$\in \mathfrak{A}^m.$$

and

$$E_\varepsilon = \{A \in X^i : P(\mathcal{H}(A \times \min(l, n_{|A_{|_J}|}))\triangle C) > \varepsilon\}$$
$$= \{A \in X^i : P((\mathcal{H}(A \times \min(l, n_{|A_{|_J}|}))\triangle C)^c) \leq (1 - \varepsilon)\}.$$

E_ε is measurable since
$B = \{(x_1, ..., x_{i+1}) \in X^{i+1} : (\mathcal{H}(\{x_1, ..., x_i\} \times \min(l, n_{|\{x_1,...,x_i\}|}))\triangle C)^c(x_{i+1}) = 1\}$ is

measurable by (M5) and the measurability of C, and a straightforward application of Fubini's theorem gives us that the map

$$\begin{aligned}(x_1,...,x_i) &\mapsto \int_X \chi_B(x_1,...,x_{i+1})dP(x_{i+1}) = \\ &= P(\{y : (x_1,...,x_i,y) \in B\}) \\ &= P(\{y : y \in \mathcal{H}(\{x_1,...,x_i\} \times \min(l, n_{|\{x_1,...,x_i\}|}))\triangle C)^c\}) \\ &= P(\mathcal{H}(\{x_1,...,x_i\} \times \min(l, n_{|\{x_1,...,x_i\}|}))\triangle C)^c)\end{aligned}$$

is measurable.

We have that

$$\begin{aligned}&P^m(\{A \in X^m : P(\mathcal{H}(A_{|_J} \times \min(l, n_{|A_{|_J}|}))\triangle C) > \varepsilon \\ &\qquad\qquad \text{and } \mathcal{H}(A_{|_J} \times \min(l, n_{|A_{|_J}|}))_{|_A} = C_{|_A}\}) \\ &\leq P^m(\{A \in X^m : P(\mathcal{H}(A_{|_J} \times \min(l, n_{|A_{|_J}|}))\triangle C) > \varepsilon \\ &\qquad\qquad \text{and } \mathcal{H}(A_{|_J} \times \min(l, n_{|A_{|_J}|}))_{|_{A_{|_{\{i+1,...,m\}}}}} = C_{|_{A_{|_{\{i+1,...,m\}}}}}\}) \\ &= P^m(E_C \cap (E_\varepsilon \times X^{m-i})).\end{aligned}$$

By Fubini's theorem

$$\begin{aligned}P^m(E_C \cap (E_\varepsilon \times X^{m-i})) &= \int_{E_\varepsilon \times X^{m-i}} \chi_{E_C}(x_1,...,x_m)dP^m \\ &= \int_{E_\varepsilon} \left(\int_{X^{m-i}} \chi_{E_C}(x_1,...,x_m)dP^{m-i} \right) dP^i.\end{aligned}$$

Now

$$\begin{aligned}&(x_1,...,x_i) \times X^{m-i} \cap E_C \\ &= (x_1,...,x_i) \times \{A \in X^{m-i} : A_{|_{\{i+1,...,m\}}} \in ((\mathcal{H}(A_{|_J} \times \min(l, n_{|A_{|_J}|}))\triangle C)^c)^{m-i}\}\end{aligned}$$

and since $(x_1,...,x_i) \in E_\varepsilon$, the inner integral is at most $(1-\varepsilon)^{m-i}$ and so

$$P^m(E_C \cap (E_\varepsilon \times X^{m-i})) \leq (1-\varepsilon)^{m-i}.$$

Therefore summing over all subsets J of $\{1,...,m\}$ of size at most k, the probability that $A \subseteq X$, $|A| = m \geq k$, contains a subset σ of size at most k, and $l \in \{1,...,n_{|\sigma|}\}$, such that $P(\mathcal{H}(\sigma \times l) \triangle C) > \varepsilon > 0$ and $\mathcal{H}(\sigma \times l)_{|_A} = C_{|_A}$, is at most

$$\sum_{i=0}^{k} \sum_{l=1}^{n_i} \binom{m}{i} (1-\varepsilon)^{m-i} = \sum_{i=0}^{k} n_i \binom{m}{i} (1-\varepsilon)^{m-i}.$$

∎

Lemma B.2.2 *Let* $0 < \varepsilon \leq 1,\ 0 < \delta \leq 1$, m, k *positive integers, and* $n = \max(\{n_i\}_{i=0}^{k})$. *If there is* $0 < \beta < 1$ *where*

$$m \geq \frac{1}{1-\beta} \left(\frac{1}{\varepsilon} \ln(\frac{n}{\delta}) + k + \frac{k}{\varepsilon} \ln(\frac{1}{\beta\varepsilon}) \right),$$

then

$$\sum_{i=0}^{k} n_i \binom{m}{i} (1-\varepsilon)^{m-i} \leq \delta.$$

Proof: Let $\varepsilon, \delta, \beta, m, k, n, \{n_i\}_{i=0}^{k}$ be as in the statement of the lemma. Then

$$\frac{1}{\varepsilon} \ln(\frac{n}{\delta}) + k + \frac{k}{\varepsilon} \left(-\ln(k) + \ln(\frac{k}{\beta\varepsilon}) - 1 + \frac{\beta\varepsilon}{k} m + 1 \right)$$

$$= \beta m + \frac{1}{\varepsilon} \ln(\frac{n}{\delta}) + k + \frac{k}{\varepsilon} \ln(\frac{1}{\beta\varepsilon}) \leq m.$$

We will use the fact that $\ln(m) \leq -\ln(\alpha) - 1 + \alpha m$ for all $\alpha > 0$. With $\alpha = \frac{\beta\varepsilon}{k}$ we get

$$\ln(m) \leq \ln(\frac{k}{\beta\varepsilon}) - 1 + \frac{\beta\varepsilon}{k} m$$

thus

$$\frac{1}{\varepsilon} \ln(\frac{n}{\delta}) + k + \frac{k}{\varepsilon} (-\ln(k) + \ln(m) + 1)$$

$$\leq \frac{1}{\varepsilon} \ln(\frac{n}{\delta}) + k + \frac{k}{\varepsilon} \left(-\ln(k) + \ln(\frac{k}{\beta\varepsilon}) - 1 + \frac{\beta\varepsilon}{k} m + 1 \right) \leq m.$$

Hence we have

$$\ln(\frac{n}{\delta}) + k(-\ln(k) + \ln(m) + 1) \leq \varepsilon(m-k)$$

and so

$$n\left(\frac{em}{k}\right)^k \leq e^{\varepsilon(m-k)}\delta.$$

Therefore since $m \geq k$,

$$\sum_{i=0}^{k} n_i \binom{m}{i}(1-\varepsilon)^{m-i} \leq n\binom{m}{\leq k}(1-\varepsilon)^{m-k} \leq n\left(\frac{em}{k}\right)^k (1-\varepsilon)^{m-k} \leq n\left(\frac{em}{k}\right)^k e^{-\varepsilon(m-k)}$$

$$\leq \delta.$$

∎

Theorem B.2.3 *If $(X, \mathcal{C})$ has a $\{n_i\}_{i=0}^k$-copy sample compression scheme $\mathcal{H}$ of size k satisfying (M5), and $n = \max(\{n_i\}_{i=0}^d)$, then for $0 < \varepsilon \leq 1,\ 0 < \delta \leq 1$, if $\mathcal{L}_{\mathcal{H}}$ is a learning rule (if $\mathcal{L}_{\mathcal{H}}$ satisfies (M2)) it has sample complexity at most*

$$m_{\mathcal{L}_{\mathcal{H}}}(\varepsilon, \delta) \leq \frac{1}{1-\beta}\left(\frac{1}{\varepsilon}\ln(\frac{n}{\delta}) + k + \frac{k}{\varepsilon}\ln(\frac{1}{\beta\varepsilon})\right),$$

for any $0 < \beta < 1$.

Proof: Let $\mathcal{H}$ be as in the statement of the theorem, fix $0 < \varepsilon \leq 1,\ 0 < \delta \leq 1,\ 0 < \beta < 1,\ C \in \mathcal{C},\ P \in \mathcal{P}$, and let

$$m \geq \frac{1}{1-\beta}\left(\frac{1}{\varepsilon}\ln(\frac{n}{\delta}) + k + \frac{k}{\varepsilon}\ln(\frac{1}{\beta\varepsilon})\right).$$

Since $\mathcal{H}$ is consistent and satisfies (M5), by the previous theorem

$$P^m(\{A \in X^m : P(\mathcal{L}_{\mathcal{H}}(A, C_{|_A}) \triangle C) > \varepsilon\})$$
$$\leq P^m(\{A \in X^m : \exists \sigma \in [A]^{\leq d},\ \exists l \in \{1, ..., n_{|\sigma|}\} \text{ where } P(\mathcal{H}(\sigma \times \min(l, n_{|\sigma|})) \triangle C) > \varepsilon\})$$
$$= P^m(\{A \in X^m : \exists \sigma \in [A]^{\leq d},\ \exists l \in \{1, ..., n_{|\sigma|}\} \text{ where } P(\mathcal{H}(\sigma \times \min(l, n_{|\sigma|})) \triangle C) > \varepsilon$$

and $\mathcal{H}(\sigma \times \min(l, n_{|\sigma|}))_{|_A} = C_{|_A}\})$

$$\leq \sum_{i=0}^{k} n_i \binom{m}{i} (1-\varepsilon)^{m-i},$$

and by the previous lemma

$$\sum_{i=0}^{k} n_i \binom{m}{i} (1-\varepsilon)^{m-i} \leq \delta.$$

∎

Index

Accuracy, 29

Array sample compression scheme, 22

Bi-embeddable, 13

Borel sigma algebra, 47

Borel space, 47

Cluster point, 48

Cofinal, 48

Compactness theorem, 19

Completion of a measure space, 47

Concept class, 5

Concept space, 5

Convergence of nets, 48

Copy sample compression scheme, 23

Direct set, 48

Domain, 5

Dual concept space, 12

Embeddable, 12

Extended sample compression scheme, 23

Filter, 49

Filter base, 49

Free ultrafilter, 49

Generalized embeddable, 12

Hypothesis class, 17

Labelled sample compression scheme, 17

Learning rule, 27

M1, 50

M2, 50

M3, 51

M4, 51

M5, 53

Maximal, 8

Maximum, 8

Measurable space, 46

Measure, 46

Measure space, 47

Net, 48

Polish space, 47

Principal ultrafilter, 49

Probability measure, 47

Probably approximately correct, 28

Risk, 29

Sample complexity, 29

Sauer-Shelah lemma, 7

Shatter coefficients, 7

Shattered, 5

Sigma algebra, 46

Sigma algebra generated, 46

Standard Borel space, 48

Subnet, 48

Subspace, 5

Ultrafilter, 49

Universally dense, 43

Universally measurable, 47

Universally separable, 43

Unlabelled sample compression scheme, 17

VC dimension, 5

Weakly embeddable, 13

Weakly generalized embeddable, 13

Bibliography

[1] J.L. Bell and A.B. Slomson. *Models and Ultraproducts: An Introduction*. Dover Books on Mathematics Series. Dover Publications, 2006.

[2] Shai Ben-David and Ami Litman. Combinatorial Variability of Vapnik-Chervonenkis Classes with Applications to Sample Compression Schemes. *Discrete Applied Mathematics*, 86:3–25, 1998.

[3] P. Billingsley. *Probability and Measure*. Wiley Series in Probability and Statistics. John Wiley & Sons, 2012.

[4] Anselm Blumer, A. Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the Vapnik-Chervonenkis Dimension. *J. ACM*, 36(4):929–965, October 1989.

[5] J Bourgain, D H Fremlin, and M Talagrand. Pointwise Compact Sets of Baire-Measurable Functions. *American Journal of Mathematics*, 100(4):845–886, 1978.

[6] R.M. Dudley. *Uniform Central Limit Theorems*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 1999.

[7] Sally Floyd. Space-Bounded Learning and the Vapnik-Chervonenkis Dimension. In *Proceedings of the second annual workshop on Computational learning theory*, COLT '89, pages 349–364, San Francisco, CA, USA, 1989. Morgan Kaufmann Publishers Inc.

[8] Sally Floyd and Manfred Warmuth. Sample Compression, Learnability, and the Vapnik-Chervonenkis Dimension. In *Machine Learning*, pages 269–304, 1995.

[9] A.S. Kechris. *Classical Descriptive Set Theory*. Graduate Texts in Mathematics. Springer-Verlag, 1995.

[10] Dima Kuzmin and Manfred K. Warmuth. Unlabeled Compression Schemes for Maximum Classes. *Journal of Machine Learning Research*, 2005:591–605, 2006.

[11] Michael C. Laskowski. Vapnik-Chervonenkis Classes of Definable Sets. *J. London Math. Soc*, 45:377–384, 1992.

[12] Nick Littlestone and Manfred K. Warmuth. Relating Data Compression and Learnability. Technical report, 1986.

[13] J.R. Munkres. *Topology*. Prentice Hall, Incorporated, 2000.

[14] Vladimir Pestov. PAC Learnability of a Concept Class under Non-Atomic Measures: A Problem by Vidyasagar. In *Proceedings of the 21st international conference on Algorithmic learning theory*, ALT'10, pages 134–147, Berlin, Heidelberg, 2010. Springer-Verlag.

[15] D. Pollard. *Convergence of Stochastic Processes*. Springer Series in Statistics. Springer-Verlag, 1984.

[16] N. Sauer. On the Density of Families of Sets. *J. Combinatorial Theory Ser. A*, 13:145–147, 1972.

[17] John Shawe-taylor, Martin Anthony, and N. L. Biggs. Bounding Sample Size with the Vapnik-Chervonenkis Dimension. *Discrete Applied Mathematics*, 42:65–73, 1993.

[18] L. G. Valiant. A Theory of the Learnable. In *Proceedings of the sixteenth annual ACM symposium on Theory of computing*, STOC '84, pages 436–445, New York, NY, USA, 1984. ACM.

[19] V. N. Vapnik and A. Ya. Chervonenkis. On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.

[20] Emo Welzl. Complete Range Spaces. Unpublished manuscript, 1987.