COMPLEXITY PARAMETERS FOR LEARNING MULTI-LABEL CONCEPT CLASSES

A Thesis

Submitted to the Faculty of Graduate Studies and Research

In Partial Fulfillment of the Requirements

For the Degree of

Doctor of Philosophy

In

Computer Science

University of Regina

By

Rahim Samei

Regina, Saskatchewan

March, 2015

UNIVERSITY OF REGINA

FACULTY OF GRADUATE STUDIES AND RESEARCH

SUPERVISORY AND EXAMINING COMMITTEE

Rahim Samei, candidate for the degree of Doctor of Philosophy in Computer Science**,** has presented a thesis titled, ***Complexity Parameters for Learning Multi-Label Concept Classes,*** in an oral examination held on March 24, 2015. The following committee members have found the thesis acceptable in form and content, and that the candidate demonstrated satisfactory knowledge of the subject material.

External Examiner:	*Dr. Vladimir Pestov, University of Ottawa
Co-Supervisor:	Dr. Sandra Zilles, Department of Computer Science
Co-Supervisor	**Dr. Boting Yang, Department of Computer Science
Committee Member:	Dr. Howard Hamilton, Department of Computer Science
Committee Member:	Dr Malik Mouhoub, Department of Computer Science
Committee Member:	Dr. Shaun Fallat, Department of Mathematics & Statistics
Chair of Defense:	Dr. Susan Johnston, Department of English

*via SKYPE
**Not present at defense

Abstract

In Computational Learning Theory, one way to model a concept is to consider it as a member of the Cartesian products of instances (sets), where each instance may correspond to a binary or multi-valued domain. A concept class is a set of concepts, and the goal of learning algorithms is to identify the target concept in a concept class from a small number of examples, i.e., labeled instances. This thesis studies multi-label concept classes and three important learning complexity parameters for these classes.

The first parameter examined in this work is the Vapnik-Chervonenkis-dimension (VCD) and its previously studied analogues for multi-label concept classes such as the Graph-dimension, Pollard's pseudo-dimension and the Natarajan dimension. This thesis also extends the study of sample compression schemes to multi-label concept classes. A sample compression scheme (SCS) for a concept class C compresses every set S of labeled examples for some concept in C to a subset, which is decompressed to some concept that is consistent with S. The size of the SCS is the cardinality of its largest compressed set. The best possible size of an SCS is the second parameter studied below. This work formulates a sufficient condition, which we call the reduction property, for a notion of VC-dimension (VCD_Ψ) to yield labeled compression schemes for maximum classes of VCD_Ψ d in which the compression sets have size at most d. This compression scheme is in fact an extension of Floyd and Warmuth's binary

scheme to the multi-label case. The same condition also yields a so-called tight SCS, which we define to generalize Kuzmin and Warmuth's unlabeled binary scheme to the multi-label case. The Graph dimension satisfies our sufficient condition, while neither Pollard's pseudo-dimension nor the Natarajan dimension does. Moreover, this thesis shows that any class of Graph-dimension 1 has an SCS of size 1.

The third parameter studied in this thesis is the recursive teaching dimension (RTD), a complexity parameter of the recently introduced recursive teaching model, which is of interest in the study of binary concept classes because of its connection to the VCD. In the recursive teaching model, the teacher and learner agree to start the learning process with the easiest concept in the class, i.e., a concept with the minimum teaching dimension (minimum number of labeled examples needed to identify a concept among all other concepts in the class). This concept is then removed from the concept class and a concept with the minimum teaching dimension in the remaining class is chosen to teach and remove, and the process continues until no concepts are left in the class. This procedure is called a recursive teaching plan and the recursive teaching dimension of the class is the largest minimum teaching dimension encountered in the plan.

This thesis establishes a further connection between RTD and VCD in both the multi-label and binary cases, by providing an upper bound on the size of classes of a given RTD, analogous to Sauer's bound on the size of classes of a given VCD_Ψ. Maximum (largest possible) classes of a given VCD_Ψ are proven to be RTD-maximum as well. It is shown that for any VCD_Ψ-maximum class C where VCD_Ψ fulfills the reduction property, there is a teaching plan for C in which the recursive teaching sets coincide with the compression sets resulting from a tight compression scheme. Methodologically, an algebraic approach turns out to be useful, for example to prove an interesting graph-theoretic property of VCD_Ψ-maximum classes.

Acknowledgments

I am deeply indebted to my supervisors, Dr. Sandra Zilles and Dr. Boting Yang, for their constant support, encouragement and consideration throughout my Ph.D. program. I greatly appreciate their advice and patience in helping me improve my presentation and writing skills. Dr Zilles's expertise and insight is highly invaluable for guiding me through this thesis research. I am particularly grateful for the valuable suggestions and comments Dr. Zilles and Dr. Yang made on my thesis writing.

My special thanks to Dr. Pavel Semukhen, a former PIMS Postdoctoral researcher at the University of Regina, for sharing his expert knowledge of Linear Algebra with me. I highly appreciate Dr. Semukhin's contribution in the first paper that I published at the ALT conference in 2012.

I would like to thank Dr. Shaun Fallat, Dr. Howard Hamilton and Dr. Malek Mouhoub for being on my committee. I would like to express my gratitude to my committee, for the time and effort they have put into reading and examining my thesis.

I also acknowledge the financial support from my supervisors, the Faculty of Graduate Studies and Research and the Department of Computer Science, which allowed me to focus solely on my PhD program during the past four years.

This thesis work is dedicated to my beloved wife, Raheleh, who has been a constant source of support and encouragement during the challenges of graduate school and life.

Contents

Abstract **i**

Acknowledgments **iii**

Dedication **iv**

Contents **v**

List of Tables **viii**

List of Figures **xi**

1 Introduction **1**

1.1 The VC-dimension 5

1.1.1 Binary Case 5

1.1.2 Multi-label Case 6

1.2 Sample Compression Schemes 7

1.3 The Recursive Teaching Dimension 8

1.4 Contributions of This Thesis 10

1.5 Organization 13

2 Background and Notation **15**

3 Linear Algebraic Approach **22**

3.1 Sauer's Bound in the Binary Case 25

3.2 Generalized Sauer Bound 27

3.3 Algebraic Characterization of Teaching Sets 33

3.4 Shortest-path Closedness in One-inclusion Hypergraphs 36

4 The Reduction Property **40**

4.1 The Graph Dimension 44

4.2 Pollard's pseudo-dimension 52

4.3 The Natarajan Dimension 54

5 Sample Compression Schemes **56**

5.1 PAC-learnability of Multi-label Classes 57

5.2 Generalizing Floyd and Warmuth's Compression Scheme 61

5.3 Generalizing the Kuzmin-Warmuth Unlabeled Scheme 71

5.4 Connection to the One-inclusion Hypergraph 99

5.5 Sample compression for classes of VCD_{Ψ} 1 102

5.5.1 The Graph Dimension 103

5.5.2 Pollard's Pseudo-dimension 110

5.5.3 The Natarajan Dimension 111

6 Recursive Teaching Dimension **113**

6.1 RTD-maximum Classes 116

6.1.1 Teaching Plans of RTD-maximum Classes 131

6.2 RTD-maximal Classes 136

7 Conclusions **141**

7.1 Linear Algebraic Approach 141

7.2 The Reduction Property 142
7.3 Sample Compression Schemes 142
7.4 Recursive Teaching Dimension 143
7.5 Future Research Directions 143

References **146**

Appendix A Binary RTD-maximum Classes **152**

List of Tables

2.1 A multi-label concept class and its restriction and reduction. 16

2.2 A concept class C with $\mathrm{VCD}(C) = 2$, $\mathrm{TD}(C) = 2$ and $\mathrm{RTD}(C) = 1$. . 17

2.3 A multi-label concept class and label mapping 19

2.4 Two images of applying mappings from the Natarajan family on the class C from Table 2.3. 20

3.1 A concept class used for illustration 23

3.2 The VCD_{Ψ_G}-maximum class obtained from Figure 3.1 32

4.1 A VCD_{Ψ_P}-maximum class that does not fulfill the reduction property 53

4.2 Reductions of C where C is the VCD_{Ψ_P}-maximum class from Table 4.1 53

4.3 Mappings of the concept class C from Table 4.1 54

4.4 A VCD_{Ψ_N}-maximum class that does not fulfill the reduction property 55

4.5 Reductions of C where C is the VCD_{Ψ_N}-maximum class from Table 4.4 55

5.1 A VCD_{Ψ_G}-maximum class and the extension of Floyd and Warmuth's compression scheme 62

5.2 C^{X_1} and $(C^{X_1})^{X_2}$ where C is the VCD_{Ψ_G}-maximum class from Table 5.1 63

5.3 $C|_Y$, $(C|_Y)^{X_1}$ and $((C|_Y)^{X_1})^{X_2}$ where $Y = \{X_1, X_2, X_3\}$ and C is the VCD_{Ψ_G}-maximum class from Table 5.1 67

5.4 C^{X_2} and $(C^{X_2})^{X_1}$ where C is the VCD_{Ψ_G}-maximum class from Table 5.1 69
5.5 A VCD_{Ψ_G}-maximum class and representatives resulting from Algorithm 2 74
5.6 C^{X_1} and $\mathrm{tail}_{X_1}(C)$ where C is the VCD_{Ψ_G}-maximum class from Table 5.5 81
5.7 C^{X_1} and $(C^{X_1})^{X_2}$ where C is the VCD_{Ψ_G}-maximum class from Table 5.5 84
5.8 $\mathrm{tail}_{X_2}(C^{X_1})$, $\mathrm{tail}_{X_2}((C^{X_1})^{X_3})$ and $\mathrm{tail}_{X_2}(C^{X_1} - X_3)$ where C is the VCD_{Ψ_G}-maximum class from Table 5.5 84
5.9 $C - X_4$, $(C - X_4)^{X_1}$, C^{X_1} and $C^{X_1} - X_4$ where C is the VCD_{Ψ_G}-maximum class from Table 5.5 85
5.10 VCD_{Ψ_P}-maximum class from Table 4.1 with a tight compression scheme 99
5.11 A VCD_{Ψ_G}-maximal class of VCD_{Ψ_G} 1 that is not VCD_{Ψ_G}-maximum . 103
5.12 Illustration of the proof of Lemma 5.44 104
5.13 A concept class of VCD 2 with 4 concepts of teaching dimension 3 . . 105
5.14 A concept class of VCD_{Ψ_G} 1 and its compression sets of size 1 106
5.15 A concept class of VCD_{Ψ_P} 1 with 2 concepts of teaching dimension 2 111
5.16 A concept class of VCD_{Ψ_N} 1 with 3 concepts of teaching dimension 2 112

6.1 Summary of the main results of Sections 6.1 and 6.2 119
6.2 An RTD-maximum class whose restriction is not RTD-maximum . . . 131
6.3 An RTD-maximum class that is shortest-path closed, but not VCD-maximum 132
6.4 A VCD-maximal class that does not shatter some subset of the instance space 137
6.5 C is RTD-maximal but not RTD-maximum (the class was found by computer experiments) 140

A.1 C is RTD-maximum but $\overline{C}$ is not 153

A.2 C_1 and $\overline{C_1}$ are RTD-maximum but neither C_1 ($\{X_1, X_2, X_3\}$ is shattered) nor $\overline{C_1}$ ($\{X_2, X_3\}$ is shattered) is VCD-maximum 154

List of Figures

1.1 A concept class in which each concept represents an axis-aligned rectangle in a finite 2-dimensional grid. 2

3.1 A geometric example of a VCD_{Ψ_G}-maximum class 32

3.2 A concept class and its one-inclusion graph 37

Chapter 1

Introduction

Machine Learning is concerned with the study of systems, such as robots and computer programs, that can gain knowledge from data. *Computational Learning Theory*, in which learning is modeled as a computational process, proposes and analyzes formal models for machine learning problems and their corresponding algorithms.

In the literature, various models have been proposed to cover the broad spectrum of machine learning problems. The models that are considered in this thesis have the following framework in common. The learning environment is modeled as a set of attributes which are called *instances*. Each attribute can be discretized as a binary or multi-valued instance. For example, the attribute "gender" with values "male" and "female" is binary, while "eye colour" with the value set {"blue", "brown", "green", "grey"} is multi-valued. We can encode these two attributes as instances called X_1 and X_2 where the value set of X_1 is $\{0, 1\}$ and that of X_2 is $\{0, 1, 2, 3\}$. A *concept* is a set of pairs such that each pair contains an instance and a *label*, i.e., some value for that instance, and each instance must appear in exactly one pair. For example, assuming attributes "gender" and "eye colour", possible concepts are {(gender, female), (eye colour, grey)} (represented as $\{(X_1, 1), (X_2, 3)\}$) or {(gender,

male), (eye colour, blue)} (represented as $\{(X_1, 0), (X_2, 0)\}$). A *concept class* is then a set of concepts, namely a set of assignments to all instances.[1] For example, a data table listing the gender and eye colour of every person in a specific group of people would be a concept class. When all instances are binary, then a concept class over these instances is called a *binary* class; it is called a *multi-label* class otherwise. Multi-label concept classes are used to model learning environments with multi-valued attributes. An *example* is an instance-label pair and a *sample* is a set of examples. The learner is provided with a sample from the *target* concept (training examples), which is among the concepts in the concept class, and the goal of the learner is to identify the target concept in the concept class.

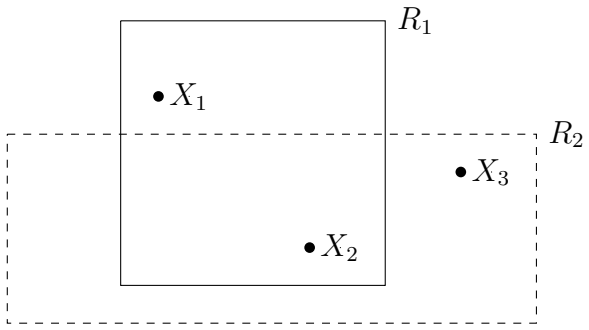

Figure 1.1: A concept class in which each concept represents an axis-aligned rectangle in a finite 2-dimensional grid.

Suppose a given concept class has a large number m of instances (attributes), such as, for example a binary concept class representing a set of axis-aligned rectangular areas in a finite 2-dimensional rectangular grid, where each instance corresponds to one of m points in the grid and carries the label 1 if the point is contained in the rectangle represented by the concept, and the label 0 if it is not. The target concept then represents one specific rectangle (set of grid points) in the concept class, which represents a set of potential target rectangles. Figure 1.1 shows such a concept class

[1] Concept classes can have finite or infinite instance sets, but in this thesis we mostly consider the finite case.

with two concepts R_1, R_2 over a grid containing the three instances X_1, X_2, and X_3, where $R_1 \supseteq \{(X_1, 1), (X_2, 1), (X_3, 0)\}$ and $R_2 \supseteq \{(X_1, 0), (X_2, 1), (X_3, 1)\}$. A learner would be presented with a number of instance-label pairs, i.e., a list of points together with the information whether or not they belong to the target rectangle. The goal of the learner would be to identify the target rectangle, where the term "identify" may have different interpretations, depending on the learning model, as discussed below.

Computational Learning Theory is concerned with the feasibility of learning, where learning algorithms are often evaluated with respect to

- time and space complexity,
- sample complexity, i.e., the number of training examples required by the algorithm in order to fulfill the learning task, or
- learning error, i.e., the difference between the hypothesis (the concept that the learning algorithm conjectures) and the target concept.

This thesis considers two learning models, the *PAC-learning* model and the *teaching* model, and for both it focuses on the best possible sample complexity that any learning algorithm can achieve, where sample complexity is measured by the worst case over all potential target concepts. Sample complexity has been studied extensively by the machine learning community for many years for different reasons such as, (i) the limited availability of training examples in many applications forces us to come up with input-efficient learning algorithms, and (ii) a theoretical investigation of sample complexity yields formal guarantees concerning the number of training examples that need to be obtained in practice.

In the *Probably Approximately Correct* (PAC for short) learning model, which was proposed by Valiant in 1984 [Val84], the learner intends to obtain a good approximation of the target concept in the concept class with a high probability, from *randomly*

chosen examples. The learner is thus given ϵ (accuracy parameter), δ (confidence parameter) and a subset of instances that are randomly chosen w.r.t. some probability distribution and labeled consistently with the target concept. The cardinality of this subset is upper-bounded by a polynomial in $\frac{1}{\epsilon}$ and $\frac{1}{\delta}$. The learner is then required to predict an ϵ-approximation of the target concept with probability of at least $1 - \delta$. A learning algorithm, with parameters ϵ and δ, PAC-learns the target concept in a concept class if and only if, for any probability distribution, any ϵ, and any δ the algorithm outputs a hypothesis with a probability greater than $1 - \delta$, such that the difference between the hypothesis and the target concept is less than ϵ. The difference is measured as the probability that the target concept and the hypothesis disagree on an instance drawn with respect to the underlying distribution. In the rectangle example, the learner would be required to find, with probability at least $1 - \delta$ over all samples, a rectangle R such that the set of grid points in the symmetric difference of R and the target rectangle R^* has a cumulative probability of at most ϵ.

The teaching model was introduced by Goldman and Kearns [GK91] and Shinohara and Miyano [SM91] independently. In this model, the learner is provided with *well-chosen* examples by a *teacher*, as opposed to the PAC-learning model in which the learner is given randomly chosen examples, and is required to identify the target concept precisely. In short, the teacher gives the learner a *teaching set* for the target concept, a set of examples that are labeled consistently only with the target concept and not with any other concept in the class; thus, the learner can identify the target concept immediately. The *teaching dimension* of a concept is measured by the number of examples in the smallest teaching set of that concept. The teaching dimension of a concept class is the worst-case teaching dimension of concepts in the concept class, i.e., the complexity of teaching the most difficult concept in the class. In the rectangle example, the teaching dimension of the rectangle $R_{\max}$ containing all

grid points is 2, since $R_{\max}$ is the only concept in the class containing both the topmost/leftmost and the bottommost/rightmost grid point (recall that the grid points are arranged in a rectangular shape). Together, these two points, each with the label 1, would be sufficient to identify $R_{\max}$ in the concept class. The teaching dimension of the concept class equals the number m of grid points, since the empty rectangle can only be identified uniquely after m examples with label 0 have been presented.

This thesis focuses on three learning complexity parameters: the VC-*dimension*, the size of *Sample Compression Schemes*, and the *Recursive Teaching Dimension*. The first two are combinatorial parameters that have been studied in connection with the PAC learning model, while the third one is a combinatorial parameter introduced in the study of teaching models.

1.1 The VC-dimension

The notion of VC-dimension, which is a complexity parameter for learning from random examples, has been studied in both the binary and multi-label case.

1.1.1 Binary Case

The Vapnik-Chervonenkis dimension (VCD for short), which was first introduced for binary concept classes [VC71], is an essential learning complexity parameter, as Blumer et al. [BEHW89] connected the PAC-learnability of a class to its VCD. In particular, Blumer et al. showed that if the VCD of a concept class is finite, then there is a PAC-learner for the class with a sample complexity that is upper-bounded by a polynomial in $\frac{1}{\epsilon}$, $\frac{1}{\delta}$ and the VCD [BEHW89].

Sauer [Sau72] and Shelah [She72] found an upper bound (typically called Sauer's bound) on the size of a concept class with a given VCD. Later in 1997, Gurvits [Gur97]

and Smolensky [Smo97] also established Sauer's bound using a different approach, namely a *linear algebraic* approach. Sauer's bound has proven helpful — if not essential — for a variety of studies, most notably for the definition and analysis of VCD-maximum classes.

A concept class (over a finite instance space) of a given VCD is VCD-maximum if its size meets Sauer's bound [Wel87]. The *restriction* of a concept class C to a subset of the instance space is the projection of C onto the instances in that set. In the case of an infinite instance space, a concept class C of a given VCD is VCD-maximum if the restriction of C to any finite subset of the instance space is VCD-maximum [Wel87]. VCD-maximum classes exhibit a number of interesting structural properties, e.g., their complements as well as their restrictions to subsets of the instance space are VCD-maximum [Wel87, Flo89, RBR09]. As another fascinating property, Kuzmin and Warmuth proved that when a class is VCD-maximum, then in some graph representation, called the *one-inclusion graph* of the class, the length of the shortest path between any two concepts is equal to the symmetric difference of those concepts [KW07, RBR09]; such a concept class is then called *shortest-path closed*.

1.1.2 Multi-label Case

Most prior work on multi-label classes concerns the combinatorial structure of such classes, and in particular various options for defining analogues of the VCD [Alo83, Nat89, Vap89, Pol90, Gur97, FS12] that coincide with the VCD in the binary case. The framework proposed by Gurvits [Gur97] generalizes over many of these notions. It turns out that, as in the binary case, the finiteness of most of the VCD notions studied in the literature is sufficient and necessary for the PAC-learnability of multi-label classes [BCHL95]. Moreover, Haussler and Long generalized Sauer's bound to

multi-label classes for a variety of such analogues of VCD [HL95]. Gurvits as well presented a generalization of Sauer's bound for multi-label classes by exploiting Linear Algebra [Gur97].

More recent studies show that results relating the VCD to the density of the one-inclusion graph of a concept class can be also extended to some of the multi-label analogues [RBR09, SS10], and provide sample bounds for various learning models and strategies [RBR09, DSBS11]. Daniely and Shalev-Shwartz revisited the one-inclusion learner proposed in [RBR09] and derived more efficient learners with the same sample complexity as the one-inclusion learner [DS14].

1.2 Sample Compression Schemes

To the best of our knowledge, the notion of Sample Compression Scheme (SCS) has been studied exclusively for binary concept classes. The smallest possible size of an SCS for a concept class is considered as a complexity parameter for learning from random examples.

An SCS for a concept class C compresses every set S of labeled examples for some concept in C to a subset, which is decompressed to some concept that is a superset of S [LW86]. The size of the SCS is the cardinality of its largest compression set. Littlestone and Warmuth proved that if a concept class C has an SCS of size k, then there exists a PAC-learning algorithm for C (based on that SCS) with a sample complexity that is upper-bounded by a polynomial in $\frac{1}{\epsilon}$, $\frac{1}{\delta}$ and k [LW86]. This result raises the question whether the smallest possible size of an SCS for C is bounded linearly in the VCD of C [LW86, FW95]. To date, there is no general answer to this question. Partial answers in the literature concern mostly the case of VCD-maximum concept classes and the case of concept classes of VCD 1. In the

rectangle example, the compression function keeps the leftmost, rightmost, topmost and bottommost instances with the label 1 from any given set of examples. So, the compression function saves at most 4 grid points from any sample set. The decompression function returns as a hypothesis the smallest axis-aligned rectangle that contains these points. This hypothesis is guaranteed to be consistent with the original set of examples [FW95]. Therefore, this class has an SCS of size 4. Note that this class is of VCD 4 [BEHW89].

Floyd and Warmuth showed that every VCD-maximum concept class of VCD d has an SCS of size d [FW95]. Since an SCS for a concept class C also applies to all subclasses of C, Floyd and Warmuth's result implies that every concept class of VCD 1 has an SCS of size 1. This is due to the fact that every concept class of VCD 1 is contained in a VCD-maximum class of VCD 1 over the same instance space [WW87].

An astonishing observation was made by Kuzmin and Warmuth, who proved that each VCD-maximum class of VCD d has an *unlabeled* SCS of size d, i.e., an SCS in which the compression sets have no label information [KW07]. Ben-David and Litman demonstrated a special-case of this result earlier [BL98].

Rubinstein and Rubinstein [RR08, RR12] and Rubinstein et al. [RBR09] characterized some geometric and topological properties of the one-inclusion graph of VCD-maximum classes and connected these results to SCSs for VCD-maximum classes. In addition, they found the answer to some of the conjectures in [KW07] by focusing on geometric instances of VCD-maximum classes.

1.3 The Recursive Teaching Dimension

To the best of our knowledge, the notion of Recursive Teaching Dimension (RTD) has only been studied for binary concept classes.

RTD is a complexity parameter in one of the models of *cooperative* pairs of teachers and learners, namely the *recursive teaching* model [DSZ10, ZLHZ11]. In cooperative models [Bal08, DSZ10, ZLHZ11], a teacher and a learner make some agreements on the selection of examples in order to reduce the sample complexity of teaching. In the recursive teaching model, the teacher and learner agree to start the learning process with the easiest concept in the class, i.e., a concept with the minimum teaching dimension. This concept will then be removed from the concept class. Next, a concept with the minimum teaching dimension in the remaining class is chosen to teach and remove, and the process continues until no concepts are left in the class to teach (learn). This procedure is called a recursive teaching plan and the recursive teaching dimension of the class is the largest minimum teaching dimension in the plan. In the rectangle example, there is a teaching plan in which each concept is recursively taught using at most two examples, each with the label 1, namely the two examples for two points diagonally across from each other on the border of the rectangle (details are omitted). So, the RTD of the class is 2.

Although Goldman and Kearns showed that there is no general relation between the teaching dimension and the VCD [GK91, GK95], recent work by Doliwa et al. indicates connections between the VCD and the RTD; besides maximum classes, several other types of concept classes are shown to have an RTD upper-bounded by their VCD [DSZ10, DFSZ14]. Moreover, Doliwa et al. showed that for a binary VCD-maximum class C, the compression sets resulting from Kuzmin and Warmuth's unlabeled scheme [KW07] can be used as recursive teaching sets for C; there is a teaching plan for C in which the recursive teaching sets coincide with the compression sets resulting from Kuzmin and Warmuth's unlabeled scheme [DSZ10].

1.4 Contributions of This Thesis

Since a vast number of applications in Machine Learning deal with learning environments that contain multi-valued attributes (e.g. multi-class classification), the study of multi-label concept classes on a formal level certainly deserves the attention of the learning theory community. This work mainly focuses on multi-label concept classes and studies three complexity parameters, namely the VCD, the size of SCSs and the RTD.

It carefully examines the well-known learning complexity measures for multi-label concept classes, such as the Graph-dimension ($\mathrm{VCD}_{\Psi_\mathrm{G}}$) [Nat89], Pollard's pseudo-dimension ($\mathrm{VCD}_{\Psi_\mathrm{P}}$) [Pol90] and the Natarajan dimension ($\mathrm{VCD}_{\Psi_\mathrm{N}}$) [Nat89] and identifies the *concept class reduction property* (hereafter referred to as the reduction property), which makes a novel distinction between these notions. Given a binary concept class C over an instance space X, the *reduction* of C with respect to an instance $X_t \in X$ is defined as the set of all concepts c in the restriction of C to $X \setminus \{X_t\}$ for which both the concepts $c \cup \{(X_t, 0)\}$ and $c \cup \{(X_t, 1)\}$ are contained in C. In the multi-label case, it is not at all obvious how the reduction should even be defined: should a concept c in the reduction with respect to X_t have all $|X_t|$ possible extensions contained in C (i.e., $c \cup \{(X_t, \ell)\} \in C$ for all $\ell \in X_t$) or should we only require there to be at least two different extensions of c in C? VCD_Ψ *fulfills* the reduction property if for any VCD_Ψ-maximum class C and for any instance X_t, any concept c in the restriction of C to $X \setminus \{X_t\}$ has either a unique extension to C or all $|X_t|$ possible extensions to C. Thus, if the reduction property is fulfilled, these two possible definitions of reduction coincide.

This thesis shows that any analogue of the VCD that fulfills the reduction property connects the study of machine learning problems in the binary case to the ones in the

multi-label case. More precisely, it shows that if VCD_Ψ fulfills the reduction property, then for any VCD_Ψ-maximum class of VCD_Ψ d, the reduction of C with respect to any instance is also VCD_Ψ-maximum of VCD_Ψ $d-1$ — analogous to the result by Welzl for VCD-maximum classes [Wel87]. We will soon see that this result has a remarkable effect in extending the statements for VCD-maximum classes to VCD_Ψ-maximum classes. It is proven that while the Graph-dimension has the reduction property, neither Pollard's pseudo-dimension nor the Natarajan dimension fulfill it.

This work also extends the study of sample compression schemes to multi-label concept classes. It is mainly motivated by the connection between PAC-learnability and SCSs, that is, the size of an SCS for a multi-label concept class C yields sample bounds for a PAC-learner for C. Although this connection in the multi-label case is easily established analogously to that in the binary case, SCSs have never been studied for multi-label concept classes. Obviously, the study of SCSs in the multi-label case is at least as hard as in the binary case. However, it is not trivial at all to determine whether the existence of compression schemes for binary classes leads to the existence of compression schemes for their analogues in the multi-label case, even for VCD-maximum classes. This shows a gap between two research avenues in computational learning theory, namely learning binary classes and learning multi-label classes; this gap is bridged in this thesis.

We generalize some prominent existing SCSs for binary concept classes to the multi-label case. Although unlabeled SCSs cannot exist for multi-label classes, we characterize the *tightness* property for SCSs, which was exploited in Kuzmin and Warmuth's unlabeled scheme, and generalize unlabeled schemes to tight labeled schemes for multi-label concept classes. In short, (i) we extend Floyd and Warmuth's compression scheme to VCD_Ψ-maximum classes where VCD_Ψ fulfills the reduction property, (ii) we prove that if VCD_Ψ fulfills the reduction property, then any VCD_Ψ-maximum

class has a tight SCS of size VCD_Ψ of the class, (iii) we show that classes of Graph-dimension 1 have a compression scheme of size 1. The proof of the latter is not a straightforward translation of the proof of the similar result for VCD 1 binary classes [WW87]. In fact, we show that, as opposed to the binary case, classes of Graph-dimension 1 may not be contained in maximum size classes of Graph-dimension 1.

Finally, this thesis studies the RTD of both binary and multi-label concept classes. RTD turned out to be an interesting and significant parameter for binary concept classes because of its connection to the VCD [DSZ10, DFSZ14]. An open question is whether or not the RTD has an upper bound linear in the VCD. To the best of our knowledge, recursive teaching is the only model known so far that could potentially establish a close connection between the complexity of learning from a teacher (RTD) and the complexity of learning from randomly chosen examples (VCD). Motivated by this statement, this thesis establishes a further connection between RTD and VCD in both the multi-label and binary cases and reinforces the relation between VCD and RTD.

We first use an algebraic approach to show that a Sauer-type function upper-bounds the size of classes of a given RTD. We then connect RTD-maximum classes, which are classes of a given RTD whose size meets the Sauer-type bound, to VCD_Ψ-maximum classes. In fact, we generalize some of the most interesting properties of VCD-maximum classes to VCD_Ψ-maximum classes along with RTD-maximum classes. In particular, we consider a well-known graph representation of binary concept classes, namely the *one-inclusion graph*. In this graph, the vertices are concepts and there is an edge between two concepts if they disagree on exactly one instance. The definition of the one-inclusion graph is easily extended to the one-inclusion hypergraph for multi-label concept classes. We apply an algebraic approach and show

that the one-inclusion hypergraph of VCD_Ψ-maximum classes is shortest-path closed, i.e., the length of the shortest path between any two concepts is equal to the symmetric difference of those concepts. Our proof is an alternative proof for the same result on the one-inclusion graph of VCD-maximum classes in the binary case [KW07]. Moreover, we establish a connection between tight compression schemes and recursive teaching plans for VCD_Ψ-maximum classes in the multi-label case, which is a generalization of the similar result in the binary case by Doliwa et al. [DSZ10]. In fact, we show that for any VCD_Ψ-maximum class C where VCD_Ψ fulfills the reduction property, there is a teaching plan for C in which the recursive teaching sets coincide with the compression sets resulting from a tight compression scheme.

1.5 Organization

In Chapter 2, we introduce notation and basic definitions. The notion of label mapping, which is used for defining analogues of VCD for multi-label concept classes, is also discussed.

In Chapter 3, we discuss the application of Linear Algebra for establishing results on combinatorial parameters in Computational Learning Theory. We review the linear algebraic approach used by Gurvits [Gur97] and Smolensky [Smo97] in proving Sauer's lemma for binary concept classes. We also describe Gurvits's idea of extending Sauer's lemma and upper-bounding the size of multi-label classes [Gur97]. This upper bound then allows us to define VCD_Ψ-maximum classes. We finally introduce our idea of using Linear Algebra to characterize teaching sets and use this characterization in proving the shortest-path closedness in one-inclusion hypergraphs.

Chapter 4 discusses the reduction property. We introduce the reduction property and show that when VCD_Ψ fulfills the reduction property, the reduction of any VCD_Ψ-

maximum class is also VCD_Ψ-maximum. We then prove that the Graph-dimension fulfills the reduction property. Counterexamples are provided for Pollard's pseudo-dimension and the Natarajan dimension.

In Chapter 5, we study sample compression schemes for multi-label concept classes. To motivate this study, we show that, as in the binary case, the smallest possible size of a sample compression scheme yields sample bounds for PAC-learning in the multi-label case. We basically prove that the reduction property is a sufficient condition for VCD_Ψ that yields a labeled SCS for VCD_Ψ-maximum classes. We extend Floyd and Warmuth's compression scheme [FW95] to VCD_Ψ-maximum classes and introduce and study the notion of a *tight* compression scheme for VCD_Ψ-maximum classes which is in fact a generalization of Kuzmin and Warmuth's unlabeled compression scheme [KW07]. We finally show that any concept class of Graph-dimension 1 has a sample compression scheme of size 1.

Chapter 6 discusses the recursive teaching dimension (RTD) for multi-label concept classes. As our first observation, we prove that classes of Graph-dimension 1 have RTD 1. We then provide the main result of the chapter by establishing a Sauer-type upper bound on the size of classes of a given RTD, which leads to the definition of RTD-maximum classes. Some interesting similarities and dissimilarities between RTD-maximum and VCD_Ψ-maximum classes are discussed, and then some nontrivial results on the teaching plans of RTD-maximum classes are proved. At last, we discuss RTD-maximal classes, which are classes whose RTD increases when adding any new concept.

We conclude our work in Chapter 7, discussing open problems and conjectures.

Chapter 2

Background and Notation

Let $\mathbb{N}^+$ be the set of all positive integers. For $m \in \mathbb{N}^+$, let $[m] = \{1, \ldots, m\}$. For $m \in \mathbb{N}^+$, the set $X = \{X_1, \ldots, X_m\}$ is called an *instance space*, where each instance X_i is associated with the value set $X_i = \{0, \ldots, N_i\}$, $N_i \in \mathbb{N}^+$, for all $i \in [m]$. We call $c \in \prod_{i=1}^{m} X_i$ a *(multi-label) concept* on X, and a *(multi-label) concept class* C is a set of concepts on X, i.e., $C \subseteq \prod_{i=1}^{m} X_i$. For $c \in C$, let $c(X_i)$ denote the ith coordinate of c. We will always implicitly assume that a given concept class C is a subset of $\prod_{i=1}^{m} X_i$ for some $m \in \mathbb{N}^+$, where $X_i = \{0, \ldots, N_i\}$, $N_i \in \mathbb{N}^+$. When $N_i = 1$ for all $i \in [m]$, C is a *binary* concept class. Table 2.1 shows a multi-label concept class $C \subseteq \{0, 1, 2, 3\} \times \{0, 1, 2\}^2$ (top, left).

A *sample* is a set of *labeled examples*, i.e., of pairs $(X_t, \ell) \in X \times \mathbb{N}$. For a sample S, we define $X(S) = \{X_i \in X \mid (X_i, \ell) \in S \text{ for some } \ell\}$. A sample S is called C-*realizable* when S is consistent with some concept in the concept class C, that is, there is a concept $c \in C$ such that $c|_{X(S)} = S$. For $t \in [m]$ and $C' \subseteq \prod_{i=1,\ i \neq t}^{m} X_i$, a concept $c \in C$ is an *extension* of a concept $c' \in C'$ iff $c = c' \cup \{(X_t, l)\}$, for some $l \in X_t$. Then c' is *extended* to c with (X_t, l).

For $Y = \{X_{i_1}, \ldots, X_{i_k}\} \subseteq X$ with $i_1 < \cdots < i_k$, we denote the *restriction*

of a concept c to Y by $c|_Y$ and define it as $c|_Y = (c(X_{i_1}), \ldots, c(X_{i_k}))$. Similarly, $C|_Y = \{c|_Y \mid c \in C\}$ denotes the restriction of C to Y. We use size$(C|_Y)$ instead of $|C|_Y|$ to avoid confusion. We also denote $c|_{X\setminus\{X_t\}}$ and $C|_{X\setminus\{X_t\}}$ by $c - X_t$ and $C - X_t$, respectively. See Table 2.1 for illustration.

$c \in C$	X_1	X_2	X_3
c_1	0	0	0
c_2	0	0	1
c_3	0	0	2
c_4	1	2	0
c_5	1	2	2
c_6	3	0	0
c_7	3	0	1
c_8	2	2	2

$c' \in C - X_3$	X_1	X_2
c'_1	0	0
c'_2	1	2
c'_3	3	0
c'_4	2	2

$c'' \in [C]^{X_3}_{\geq 2}$	X_1	X_2
c''_1	0	0
c_2	1	2
c_3	3	0

$c''' \in C^{X_3}$	X_1	X_2
c'''_1	0	0

Table 2.1: A concept class $C \subseteq \{0,1,2,3\} \times \{0,1,2\}^2$, its restriction to $\{X_1, X_2\}$ (i.e., $C - X_3$), and the two possible reductions of C w.r.t. X_3 (i.e., $[C]^{X_3}_{\geq 2}$ and C^{X_3}).

In the binary case, the *reduction* C^{X_t} of C w.r.t. $X_t \in X$ consists of all concepts in $C - X_t$ that have both possible extensions to concepts in C, i.e., $C^{X_t} = \{c \in C - X_t \mid c \times \{0,1\} \subseteq C\}$. It is not obvious how the definition of reduction should be extended to the multi-valued case. One could consider the class of concepts in $C - X_t$ that have at least two distinct extensions, or of those that have all $N_t + 1$ extensions to concepts in C. We denote the former with $[C]^{X_t}_{\geq 2}$ and the latter with C^{X_t}. See Table 2.1 for illustration.

In the binary case, $Y \subseteq X$ is *shattered* by C iff $C|_Y = \prod_{X_i \in Y} X_i = \{0,1\}^{|Y|}$. The size of the largest set shattered by C is the *VC-dimension* of C, denoted VCD(C).

Example 2.1. *Consider the class C in Table 2.2. C shatters $\{X_1, X_2\}$ and C cannot shatter $\{X_1, X_2, X_3\}$, because $C|_{\{X_1,X_2,X_3\}} = C \neq \{0,1\}^3$. Therefore,* VCD$(C) = 2$.

$c \in C$	X_1	X_2	X_3
c_1	1	0	$\underline{1}$
c_2	$\underline{1}$	1	0
c_3	0	$\underline{1}$	0
c_4	0	0	0

Table 2.2: A concept class C with $\mathrm{VCD}(C) = 2$ ($\{X_1, X_2\}$ is shattered), $\mathrm{TD}(C) = 2$ and $\mathrm{RTD}(C) = 1$. Recursive teaching sets are underlined

The literature offers a variety of VCD notions for the non-binary case [Alo83, Nat89, Vap89, Pol90, Gur97]. Gurvits' framework [Gur97] generalizes over many of these notions. We first need to introduce *label mappings* and the required related notation.

Let Ψ_i be a family of mappings $\psi_i : X_i \to \{0,1\}$ and let $\Psi = \Psi_1 \times \cdots \times \Psi_m$. For a concept $c \in \prod_{i=1}^{m} X_i$ and $\overline{\psi} = (\psi_1, \ldots, \psi_m) \in \Psi$, we denote the vector $(\psi_1(c(X_1)), \ldots, \psi_m(c(X_m)))$ by $\overline{\psi}(c)$. For a concept class $C \subseteq \prod_{i=1}^{m} X_i$, define $\overline{\psi}(C) = \{\overline{\psi}(c) \mid c \in C\}$. So, $\overline{\psi}(C)$ is a subset of the boolean cube $\{0,1\}^m$. The worst-case learning complexity of a multi-label class C is reflected in the following definition by using maximum over the VCD of all possible binary images of C.

Definition 2.2. [Gur97] *Let Ψ_i, $1 \le i \le m$, be a family of mappings $\psi_i : X_i \to \{0,1\}$. Let $\Psi = \Psi_1 \times \cdots \times \Psi_m$. We denote the* VC*-dimension of C w.r.t. Ψ by* $\mathrm{VCD}_\Psi(C)$ *and define it by* $\mathrm{VCD}_\Psi(C) = \max_{\overline{\psi} \in \Psi} \mathrm{VCD}(\overline{\psi}(C))$.

Specific families of mappings yield specific notions of dimension. The most general case is the family Ψ^* of *all* m-tuples $(\psi_1, \ldots, \psi_m)$ with $\psi_i : X_i \to \{0,1\}$.

Example 2.3. *Consider the concept class C on the left of Table 2.3. Let $\overline{\psi} = (\psi_1, \psi_2, \psi_3)$ where*

$$\psi_1(x) = \begin{cases} 1 & \text{if } x = 2 \\ 0 & \text{otherwise,} \end{cases} \quad \text{and} \quad \psi_2(x) = \psi_3(x) = \begin{cases} 1 & \text{if } x = 0 \\ 0 & \text{otherwise.} \end{cases}$$

As shown in the middle part of Table 2.3, $\{X_2, X_3\}$ *is shattered by* $\overline{\psi}(C)$*. No binary class resulting from* C *can shatter* X*, since* C *has only* 5 *concepts. Thus* $\mathrm{VCD}_{\Psi^*}(C) = 2$*.*

The term *Graph-dimension* [Nat89] refers to VCD_{Ψ_G}, where $\Psi_G = \Psi_{G_1} \times \cdots \times \Psi_{G_m}$, and for all $i \in [m]$, $\Psi_{G_i} = \{\psi_{G,k} \mid 0 \leq k \leq N_i\}$ and

$$\psi_{G,k}(x) = \begin{cases} 1 & \text{if } x = k \\ 0 & \text{otherwise.} \end{cases}$$

That means, one considers all ways of mapping the values in each column to 1, if they equal some value k and to 0, if they differ from k. For each column, a different value of k may be used. The largest possible VC-dimension over the resulting binary classes is the Graph-dimension.

Example 2.4. *The class* C *on the left of Table 2.3 has Graph-dimension* 2*, as witnessed by the tuple of mappings that uses* 2 *as the value of* k *for* X_1*, and* 0 *as the value of* k *for* X_2 *and* X_3*, i.e., the tuple* $(\psi_{G,2}, \psi_{G,0}, \psi_{G,0})$ *where*

$$\psi_{G,2}(x) = \begin{cases} 1 & \textit{if } x = 2 \\ 0 & \textit{otherwise,} \end{cases} \quad \textit{and} \quad \psi_{G,0}(x) = \begin{cases} 1 & \textit{if } x = 0 \\ 0 & \textit{otherwise.} \end{cases}$$

This tuple transforms C *to the binary class* C' *shown in the middle part of the table. Here the set* $\{X_1, X_3\}$ *is shattered by* C'*. Note that not every tuple of mappings yields a VC-dimension of* 2*, as shown in the right part of the table: the class* C'' *is obtained using the tuple* $(\psi_{G,2}, \psi_{G,0}, \psi_{G,2})$*, that is, when the value of* k *is set to* 2 *for both* X_1 *and* X_3*, while it is* 0 *for* X_2*.*

By *Pollard's pseudo-dimension* [Pol90] we refer to VCD_{Ψ_P}, where $\Psi_P = \Psi_{P_1} \times$

$c \in C$	X_1	X_2	X_3
c_1	2	0	1
c_2	1	1	1
c_3	1	2	2
c_4	0	2	0
c_5	2	0	0

$c' \in C'$	X_1	X_2	X_3
c'_1	1	1	0
c'_2	0	0	0
c'_3 *	0	0	0
c'_4	0	0	1
c'_5	1	1	1

$c'' \in C''$	X_1	X_2	X_3
c''_1	1	1	0
c''_2	0	0	0
c''_3	0	0	1
c''_4 *	0	0	0
c''_5 *	1	1	0

Table 2.3: A concept class C (left) and two binary classes obtained by applying column-wise label mappings to C. Duplicate concepts introduced by the mappings are marked with *.

$\cdots \times \Psi_{P_m}$ and for all $i \in [m]$, $\Psi_{P_i} = \{\psi_{P,k} \mid 0 \leq k \leq N_i\}$ and

$$\psi_{P,k}(x) = \begin{cases} 1 & \text{if } x \geq k \\ 0 & \text{otherwise.} \end{cases}$$

Example 2.5. *Consider the multi-label class C on the left of Table 2.3. By the choice of $\overline{\psi} = (\psi_{P,2}, \psi_{P,2}, \psi_{P,1})$, $\overline{\psi}(C)$ shatters $\{X_2, X_3\}$ and thus $\mathrm{VCD}_{\Psi_\mathrm{P}}(C) = 2$.*

The term *Natarajan-dimension* [Nat89] refers to $\mathrm{VCD}_{\Psi_\mathrm{N}}$, where $\Psi_N = \Psi_{N_1} \times \cdots \times \Psi_{N_m}$ and for all $i \in [m]$, $\Psi_{N_i} = \{\psi_{N,k,k'} : k, k' \in X_i, k \neq k'\}$ and

$$\psi_{N,k,k'}(x) = \begin{cases} 1 & \text{if } x = k \\ 0 & \text{if } x = k' \\ * & \text{otherwise.} \end{cases}$$

Here technically, ψ_i maps to $\{0, 1, *\}$, where $*$ is a null element to be ignored when computing the VC-dimension.

Example 2.6. *One can verify that the multi-label class C on the left of Table 2.3 is of $\mathrm{VCD}_{\Psi_\mathrm{N}}$ 1. For example, consider the tuples $\overline{\psi} = (\psi_{N,0,1}, \psi_{N,2,0}, \psi_{N,0,1})$ and $\overline{\psi'} = (\psi_{N,1,2}, \psi_{N,1,0}, \psi_{N,0,1})$. Table 2.4 shows the image of applying $\overline{\psi}$ and $\overline{\psi'}$ on C. Obviously, none of $\overline{\psi}(C)$ and $\overline{\psi'}(C)$ shatters a set of size 2.*

$c \in \overline{\psi}(C)$	X_1	X_2	X_3
c_1	$*$	0	0
c_2	0	$*$	0
c_3	0	1	$*$
c_4	1	1	1
c_5	$*$	0	1

$c' \in \overline{\psi'}(C)$	X_1	X_2	X_3
c'_1	0	0	0
c'_2	1	1	0
c'_3	1	$*$	$*$
c'_4	$*$	$*$	1
c'_5	0	0	1

Table 2.4: $\overline{\psi}(C)$ and $\overline{\psi'}(C)$, where C is the class from Table 2.3, $\overline{\psi} = (\psi_{N,0,1}, \psi_{N,2,0}, \psi_{N,0,1})$ and $\overline{\psi'} = (\psi_{N,1,2}, \psi_{N,1,0}, \psi_{N,0,1})$.

Clearly, VCD_{Ψ^*} upper-bounds all VCD notions. Also, $\mathrm{VCD}_{\Psi_\mathrm{P}} \geq \mathrm{VCD}_{\Psi_\mathrm{N}}$ and $\mathrm{VCD}_{\Psi_\mathrm{G}} \geq \mathrm{VCD}_{\Psi_\mathrm{N}}$ [HL95]. However, $\mathrm{VCD}_{\Psi_\mathrm{P}}$ and $\mathrm{VCD}_{\Psi_\mathrm{G}}$ are incomparable [BCHL95].

As in the binary case [FW95], a *forbidden labeling* of C with $\mathrm{VCD}_\Psi(C) = d < |X|$, is a set of $d+1$ examples that is inconsistent with all concepts in C. For $Y = \{X_{i_1}, \ldots, X_{i_{d+1}}\} \subseteq X$, $\mathrm{Forb}(C, Y) = X_{i_1} \times \cdots \times X_{i_{d+1}} \setminus C|_Y$ is the set of forbidden labelings on Y and $\mathrm{Forb}(C) = \bigcup_{Y \subseteq X,\ |Y|=d+1} \mathrm{Forb}(C, Y)$ is the set of forbidden labelings of size $d+1$. For $d = |X|$, we define $\mathrm{Forb}(C, Y) = \mathrm{Forb}(C) = \emptyset$.

Example 2.7. *Consider the multi-label class C on the left of Table 2.3. As discussed in Example 2.4,* $\mathrm{VCD}_{\Psi_\mathrm{G}}(C) = 2$. *For* $Y = \{X_1, X_2, X_3\}$, $\mathrm{Forb}(C, Y) = \mathrm{Forb}(C) = X_1 \times X_2 \times X_3 \setminus C$. *In particular,* $(0,0,0), (2,2,1) \in \mathrm{Forb}(C, Y)$.

A sample S is a *teaching set* for a concept c in a class C, if c is the only concept from C that is consistent with S. The collection of all teaching sets for c in C is denoted $\mathrm{TS}(c, C)$. For simplicity, if S is a teaching set for c with respect to C, we also call $X(S)$ a teaching set for c with respect to C, since the labels of examples from S are uniquely determined by $X(S)$ and c. The *teaching dimension* of c in C is $\mathrm{TD}(c, C) = \min\{|S|\colon S \in \mathrm{TS}(c, C)\}$. The teaching dimension of C is $\mathrm{TD}(C) = \max_{c \in C} \mathrm{TD}(c, C)$ [GK95, SM91].

Example 2.8. *Consider the class C in Table 2.2. The sample $S = \{(X_1, 1), (X_2, 1)\}$ only is consistent with c_2 in C, so S is a teaching set for c_2. In particular,* $\mathrm{TS}(c_2, C) =$

$\{\{(X_1,1),(X_2,1)\},\{(X_1,1),(X_3,0)\},\{(X_1,1),(X_2,1),(X_3,0)\}\}$ *and thus* $\mathrm{TD}(c_2,C) = \min\{2,2,3\} = 2$. *One can see that for each concept* $c \in C$, $1 \leq \mathrm{TD}(c,C) \leq 2$, *so* $\mathrm{TD}(C) = 2$.

The following definitions are based on previous literature on recursive teaching [DSZ10, ZLHZ11]. A *teaching plan* for a concept class C is a sequence $P = ((c_1,S_1),\ldots,(c_n,S_n))$, where $C = \{c_1,\ldots,c_n\}$ and $S_i \in \mathrm{TS}(c_i,\{c_i,\ldots,c_n\})$ for all $i = 1,\ldots,n$. The *order* of the teaching plan P is $\mathrm{ord}(P) = \max_{i=1,\ldots,n}|S_i|$. The *recursive teaching dimension* of C is

$$\mathrm{RTD}(C) = \min\{\mathrm{ord}(P) : \ P \text{ is a teaching plan for } C\}.$$

A teaching plan of C whose order equals $\mathrm{RTD}(C)$ is called an *optimal teaching plan* for C. For an optimal teaching plan $P = ((c_1,S_1),\ldots,(c_n,S_n))$ for C, the set S_i is called a *recursive teaching set* for c_i in C with respect to the plan P, and $|S_i|$ is called the *recursive teaching dimension* of c_i in C with respect to the plan P, denoted $\mathrm{RTD}(c_i,C)$.

Example 2.9. *Consider the concept class* C *in Table 2.2. There is teaching plan* $P = \{(c_1,S_1),\ldots,(c_4,S_4)\}$ *for* C *where* S_1,S_2,S_3 *are of size* 1 *(the underlined entries in the table) and* $S_4 = \emptyset$. *So,* $\mathrm{ord}(P) = 1$, *and since there does not exist a teaching plan for* C *of a smaller order,* $\mathrm{RTD}(C) = 1$.

Chapter 3

Linear Algebraic Approach

This chapter discusses the application of Linear Algebra for establishing results on combinatorial parameters in Computational Learning Theory.

Gurvits [Gur97] and Smolensky [Smo97] took advantage of the linear algebraic method for the first time in proving Sauer's lemma (Theorem 3.3) [Sau72] in which an upper bound is established on the size of a binary concept class with a given VC-dimension. This approach was also exploited by Gurvits in generalizing Sauer's bound to the multi-label case [Gur97]. Section 3.1 reviews the results by Gurvits and Smolensky in the binary case. In Section 3.2, we describe Gurvits's idea for extending Sauer's lemma and upper-bounding the size of multi-label classes.

In Section 3.3, we introduce our idea of using Linear Algebra to characterize teaching sets which, in Section 6.1, will turn out to be essential in proving a Sauer-type upper bound on the size of concept classes of a given RTD.

We also use this algebraic characterization to extend one of the significant results for binary concept classes to the multi-label case. In short, Kuzmin and Warmuth proved an interesting property about the graph representation of binary classes of a given VCD whose size meet Sauer's bound [KW07]. In Section 3.4 we extend that

result to the hypergraph representation of multi-label concept classes whose size meet the generalized Sauer bound.

We first go through the basics of the linear algebraic method. Let $C \subseteq \prod_{i=1}^{m} X_i$ and $C = \{c_1, c_2, \ldots, c_n\}$. Clearly, any function F on C is a multi-variable function over variables corresponding to $X_1, \ldots, X_m$. In fact, by defining a real-valued function F on C, we restrict the domain of $F : X_1 \times \cdots \times X_m \to \mathbb{R}$ to the concepts (tuples) in C. In particular, $F(C)$ is equivalent with the vector $\mathbf{F} = (f_1, \ldots, f_n)$ where

$$f_i = F(c_i) := F(c_i(X_1), \ldots, c_i(X_m)), \quad \text{for all } i \in \{1, \ldots, n\}.$$

$c \in C$	X_1	X_2	X_3
c_1	0	0	1
c_2	1	0	1
c_3	1	1	1
c_4	1	0	0

Table 3.1: A concept class used for illustration.

Example 3.1. *Consider the class C in Table 3.1. For the function $H(X_1, X_2, X_3) = X_1X_2 + X_3$, $H(c_1) = 0 + 1 = 1$, $H(c_2) = 0 + 1 = 1$, $H(c_3) = 1 + 1 = 2$ and $H(c_4) = 0 + 0 = 0$. Hence, $\mathbf{H} = H(C) = (H(c_1), H(c_2), H(c_3), H(c_4)) = (1, 1, 2, 0)$.*

Now, consider a family of real-valued functions $F_1, \ldots, F_k$ on C, for some positive integer k. Each function $F_j : C \to \mathbb{R}$ is an n-dimensional vector $\mathbf{F_j} \in \mathbb{R}^n$ and the family constitutes a vector $\overline{\mathbf{F}} = (\mathbf{F_1}, \ldots, \mathbf{F_k})$. That is,

$$\begin{aligned} \mathbf{F_1} &= (F_1(c_1), F_1(c_2), \ldots, F_1(c_n)) \in \mathbb{R}^n \\ \mathbf{F_2} &= (F_2(c_1), F_2(c_2), \ldots, F_2(c_n)) \in \mathbb{R}^n \\ &\vdots \\ \mathbf{F_k} &= (F_k(c_1), F_k(c_2), \ldots, F_k(c_n)) \in \mathbb{R}^n. \end{aligned} \quad \Rightarrow \quad \overline{\mathbf{F}}^T = \begin{pmatrix} \mathbf{F_1} \\ \mathbf{F_2} \\ \vdots \\ \mathbf{F_k} \end{pmatrix}$$

The functions $F_1, \ldots, F_k$ span the n-dimensional vector space $\mathbb{R}^n$, if and only if any vector in $\mathbb{R}^n$ can be represented as a linear combination of vectors $\mathbf{F_1}, \ldots, \mathbf{F_k}$. Moreover, any vector $\mathbf{v} = (v_1, \ldots, v_n) \in \mathbb{R}^n$ corresponds to a function $F_{\mathrm{v}} : C \to \mathbb{R}$, such that $F_{\mathrm{v}}(c_i) = v_i$, for all $i \in \{1, \ldots, n\}$. Hence, a family of real-valued functions $F_1, \ldots, F_k$ on C is spanning on C if and only if any real-valued function on C can be expressed as a linear combination of functions $F_1, \ldots, F_k$. Note that we need at least n functions (vectors) to span $\mathbb{R}^n$, i.e., $k \geq n$.

Example 3.2. *Consider the class C in Table 3.1 and functions $F_1(X_1, X_2, X_3) = X_1^2$, $F_2(X_1, X_2, X_3) = X_1 + X_2$, $F_3(X_1, X_2, X_3) = X_1 X_3^3$ and $F_4(X_1, X_2, X_3) = 1 - X_3$. Then,*

$$\begin{aligned} \mathbf{F_1} &= (F_1(c_1), F_1(c_2), F_1(c_3), F_1(c_4)) = (0,1,1,1) \\ \mathbf{F_2} &= (F_2(c_1), F_2(c_2), F_2(c_3), F_2(c_4)) = (0,1,2,1) \\ \mathbf{F_3} &= (F_3(c_1), F_3(c_2), F_3(c_3), F_3(c_4)) = (0,1,1,0) \\ \mathbf{F_4} &= (F_4(c_1), F_4(c_2), F_4(c_3), F_4(c_4)) = (0,0,0,1) \end{aligned} \quad \Rightarrow \quad \overline{\mathbf{F}}^T = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 0 & 1 & 2 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

One can see that this family is not spanning, because the vector $\mathbf{H} = (1,1,2,0)$, which corresponds to the function H in Example 3.1, cannot be expressed as a linear combination of $\mathbf{F_1}$, $\mathbf{F_2}$, $\mathbf{F_3}$ and $\mathbf{F_4}$.

We define $P^d(N_1, \ldots, N_m)$, $0 \leq d \leq m$, to be the following collection of monomials with variables in $X = \{X_1, \ldots, X_m\}$:

$$P^d(N_1, \ldots, N_m) = \{\ X_{i_1}^{n_{i_1}} \cdots X_{i_k}^{n_{i_k}} \mid\ 1 \leq i_1 < \cdots < i_k \leq m, 0 \leq k \leq d, \text{ and } 0 \leq n_{i_t} \leq N_{i_t}, \text{ for all } t \in \{1, \ldots, k\}\}.$$

For $k = 0$, we define $X_{i_1}^{n_{i_1}} \cdots X_{i_k}^{n_{i_k}}$ to be the constant polynomial 1.

Let $\Phi_d(N_1, \ldots, N_m) = |P^d(N_1, \ldots, N_m)|$. It is easy to verify that

$$\Phi_d(N_1, \ldots, N_m) = 1 + \sum_{1 \le i \le m} N_i + \sum_{1 \le i_1 < i_2 \le m} N_{i_1} N_{i_2} + \cdots + \sum_{1 \le i_1 < i_2 < \cdots < i_d \le m} N_{i_1} N_{i_2} \cdots N_{i_d}.$$

When for all $i \in [m]$, X_i has a binary domain, we replace $P^d(1, \ldots, 1)$ and $\Phi_d(1, \ldots, 1)$ with $P^d(m)$ and $\Phi_d(m)$, respectively. That is,

$$P^d(m) = \{X_{i_1} \cdots X_{i_k} \mid 1 \le i_1 < \cdots < i_k \le m, 0 \le k \le d\} \text{ and } \Phi_d(m) = \sum_{i=0}^{d} \binom{m}{i}.$$

Note that $\Phi_d(m)$ is different from $\Phi_d(N_1) = |P^d(N_1)| = |\{1, X_1, \ldots, X_1^{N_1}\}|$. Although they represent the same values when $N_1 = m$ and $d \in \{0, 1\}$, this should not cause confusion as the exact meaning of such notation will be clear from the context.

3.1 Sauer's Bound in the Binary Case

In this section, we briefly review an important combinatorial result proven by Sauer [Sau72] and independently by Shelah [She72] and its consequences in Learning Theory. In this section, we only consider binary concept classes. In particular, let $m \in \mathbb{N}^+$, $X_i = \{0, 1\}$, for all $i \in [m]$ and $C \subseteq \prod_{i=1}^{m} X_i$.

Theorem 3.3 (**Sauer's bound**). [Sau72, She72] *Let* $\mathrm{VCD}(C) = d$. *Then* $|C| \le \Phi_d(m)$.

Sauer's bound resulted in the following definition that identifies maximum size classes with respect to VC-dimension.

Definition 3.4 (VCD-**maximum class**). [Wel87] *Let* $\mathrm{VCD}(C) = d$. C *is called* VCD-*maximum if* $|C| = \Phi_d(m)$.

Remark 3.5. *For the case that X is infinite, Welzl called a class of* VCD d VCD-*maximum if, for any finite subset $Y \subseteq X$,* $\text{size}(C|_Y) = \Phi_d(|Y|)$. *For finite X, these two definitions are equivalent* [Wel87].

Gurvits [Gur97] and Smolensky [Smo97] both used Linear Algebra in proving Sauer's bound. Although they applied the same approach in their proofs, they did it independently with different formulations.

Theorem 3.6. [Gur97, Smo97] *Let* $\text{VCD}(C) = d$. *The set of monomials $P^d(m)$ spans $\mathbb{R}^{|C|}$.*

Proof. As explained at the beginning of this chapter, the set of monomials $P^d(m) = \{X_{i_1} \cdots X_{i_k} : 1 \leq i_1 < \cdots < i_k \leq m, 0 \leq k \leq d\}$ spans $\mathbb{R}^{|C|}$, when any real-valued function on C can be expressed as a linear combination of monomials in $P^d(m)$.

Consider a function $F : X_1 \times \cdots \times X_m \to \mathbb{R}$. Note that F can be interpolated by a polynomial $p(X_1, \ldots, X_m)$ over monomials from $P^m(N_1, \ldots, N_m)$. Since X_i is a binary variable, for all $i \in [m]$, p is in fact a polynomial over monomials in $P^m(1, \ldots, 1) = P^m(m)$. We now show that any monomial of size greater than d in $P^m(m)$ can be expressed as a linear combination of monomials in $P^d(m)$, and consequently p is a linear combination of monomials $P^d(m)$.

W.l.o.g., consider the monomial $X_1 \cdots X_{d+1}$. Let $\mathbf{v} = (v_1, \ldots, v_{d+1})$ be a boolean vector such that, for any $c \in C|_{\{X_1,\ldots,X_{d+1}\}}$, $c \neq (v_1, \ldots, v_{d+1})$. There must exist such a tuple because $\text{VCD}(C) < d + 1$ and C does not shatter $\{X_1, \ldots, X_{d+1}\}$. We define $p'(X_1, \ldots, X_m) = Z_1 \cdots Z_{d+1}$ such that

$$Z_i = \begin{cases} X_i & \text{if } v_i = 1 \\ 1 - X_i & \text{if } v_i = 0 \end{cases}$$

for all $i \in \{1, \ldots, d+1\}$. On the one hand,

$$p'(X_1, \ldots, X_m) = X_1 \cdots X_{d+1} + L(X_1, \ldots, X_m),$$

where $L(X_1, \ldots, X_m)$ is a linear combination of monomials in $P^d(m)$.

On the other hand, for any $(u_1, \ldots, u_m) \in \{0,1\}^m$, $p'(u_1, \cdots, u_m) = 1$ if and only if $u_i = v_i$, for all $i \in \{1, \ldots, d+1\}$, and $p'(u_1, \cdots, u_m) = 0$ otherwise. In particular, for any $c \in C$, since $c|_{\{X_1,\ldots,X_{d+1}\}} \neq (v_1, \ldots, v_{d+1})$, $p'(c) = p'(c(X_1), \ldots, c(X_m)) = 0$.

So,

$$X_1 \cdots X_{d+1} + L(X_1, \ldots, X_m) = 0,$$

when we restrict the domain of p' to C. Consequently, $X_1 \cdots X_{d+1} = L(X_1, \ldots, X_m)$, that is, $X_1 \cdots X_{d+1}$ can be expressed as a linear combination of monomials in $P^d(m)$. This argument is true for any monomial of size $d+1$ over variables in X.

By using induction, we can replace any monomial of size greater than d in $P^m(m)$ with a linear combination of monomials in $P^d(m)$. □

The immediate consequence of Theorem 3.6 is another justification of Sauer's bound. In fact, since the size of a spanning set cannot be smaller than the dimension of the vector space, we conclude that $|P^d(m)| \geq |C|$, or equivalently, $|C| \leq \Phi_d(m)$.

3.2 Generalized Sauer Bound

In this section, we explain Gurvits' idea of applying Linear Algebra to generalize Sauer's bound to multi-label concept classes [Gur97]. Since Gurvits' results works for VCD_Ψ, where Ψ is the direct product of spanning families of mappings, we need to explain when a family of mappings is called spanning.

Let $k \geq 1$ and Ψ_i, $i \in [m]$, be a family of mappings $\psi_j : X_i \to \{0,1\}$, for $j \in \{1,\ldots,k\}$. Each $\psi_j \in \Psi_i$ corresponds to a vector $\boldsymbol{\psi}_j = (\psi_j(0),\ldots,\psi_j(N_i))$ and Ψ_i is equivalent with the vector $\boldsymbol{\Psi}_i = (\boldsymbol{\psi}_1,\ldots,\boldsymbol{\psi}_k)$. That is,

$$\begin{array}{l} \boldsymbol{\psi}_1 = (\psi_1(0),\ldots,\psi_1(N_i)) \in \mathbb{R}^{N_i+1} \\ \boldsymbol{\psi}_2 = (\psi_2(0),\ldots,\psi_2(N_i)) \in \mathbb{R}^{N_i+1} \\ \qquad\qquad \vdots \\ \boldsymbol{\psi}_k = (\psi_k(0),\ldots,\psi_k(N_i)) \in \mathbb{R}^{N_i+1} \end{array} \quad \Rightarrow \quad \boldsymbol{\Psi}_i^T = \begin{pmatrix} \boldsymbol{\psi}_1 \\ \boldsymbol{\psi}_2 \\ \vdots \\ \boldsymbol{\psi}_k \end{pmatrix}$$

The statement "Ψ_i spans $\mathbb{R}^{N_i+1}$" or "Ψ_i is spanning on X_i" means that any vector in $\mathbb{R}^{N_i+1}$ can be represented as a linear combination of vectors $\boldsymbol{\psi}_1,\ldots,\boldsymbol{\psi}_k$. Note that each real-valued function f on X_i corresponds to a vector $(f(0),f(1),\ldots,f(N_i)) \in \mathbb{R}^{N_i+1}$. So, $\Psi_i = \{\psi_1,\ldots,\psi_k\}$ is spanning on X_i iff any real-valued function on X_i can be expressed as a linear combination of mappings from Ψ_i.

Example 3.7. *Let* $X_1 = \{0,1,2\}$ *and* $\Psi_1 = \{\psi_1,\psi_2,\psi_3\}$ *where,*

$$\psi_1(x) = \begin{cases} 1 & \textit{if } x = 0 \\ 0 & \textit{otherwise,} \end{cases} \qquad \psi_2(x) = \begin{cases} 1 & \textit{if } x = 0 \\ 1 & \textit{if } x = 1 \\ 0 & \textit{otherwise,} \end{cases} \qquad \psi_2(x) = \begin{cases} 1 & \textit{if } x = 2 \\ 0 & \textit{otherwise.} \end{cases}$$

Then

$$\begin{array}{l} \boldsymbol{\psi}_1 = (\psi_1(0),\psi_1(1),\psi_1(2)) = (1,0,0) \\ \boldsymbol{\psi}_2 = (\psi_2(0),\psi_2(1),\psi_2(2)) = (1,1,0) \\ \boldsymbol{\psi}_3 = (\psi_3(0),\psi_3(1),\psi_3(2)) = (0,0,1) \end{array} \quad \Rightarrow \quad \boldsymbol{\Psi}_i^T = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

One can verify that any vector in $\mathbb{R}^3$ *can be represented as a linear combination of* $\boldsymbol{\psi}_1$, $\boldsymbol{\psi}_2$ *and* $\boldsymbol{\psi}_3$. *In other words, any function* $f : \{0,1,2\} \to \mathbb{R}$ *can be expressed as a linear combination of* ψ_1, ψ_2 *and* ψ_3. *For example, let* $f(x) = x^2$. *Then* $f(x) = -\psi_1(x) + \psi_2(x) + 4\psi_3(x)$.

Remark 3.8. *Let Ψ_i be a spanning family of mappings on X_i. Then for each $p, q \in X_i$ with $p \neq q$, there must exist a mapping $\psi_{p\neq q} \in \Psi_i$ such that $\psi_{p\neq q}(p) \neq \psi_{p\neq q}(q)$. W.l.o.g., we always assume that*

$$\psi_{p\neq q}(x) = \begin{cases} 0 & \text{if } x = p \\ 1 & \text{if } x = q \\ 0 \text{ or } 1 & \text{otherwise.} \end{cases}$$

To prove the generalized Sauer bound (Theorem 3.10), Gurvits first observed the following.

Lemma 3.9. [Gur97] *Suppose Ψ_i, for all $i \in [m]$, is a spanning family of mappings on X_i and $\Psi = \Psi_1 \times \cdots \times \Psi_m$. Then the family of mappings $\Pi_\Psi = \{\overline{\psi} : X_1 \times \cdots \times X_m \to \{0,1\} \mid \overline{\psi} \in \Psi\}$ is spanning on $X_1 \times \cdots \times X_m$, where for $\overline{\psi} = (\psi_1, \psi_2, \ldots, \psi_m) \in \Psi$ and $(x_1, x_2, \ldots, x_m) \in X_1 \times \cdots \times X_m$ we define*

$$(\psi_1, \psi_2, \ldots, \psi_m)(x_1, x_2, \ldots, x_m) = \psi_1(x_1) \cdot \psi_2(x_2) \cdot \ \ldots \ \cdot \psi_m(x_m).$$

Theorem 3.10. [Gur97] *Let Ψ_i, $1 \leq i \leq m$, be a spanning family of mappings $\psi_i : X_i \to \{0,1\}$, and $\Psi = \Psi_1 \times \cdots \times \Psi_m$. If $\mathrm{VCD}_\Psi(C) = d$ then the monomials from $P^d(N_1, \ldots, N_m)$ span the vector space $\mathbb{R}^{|C|}$.*

Proof. We show that any function on C can be expressed as a linear combination of monomials from $P^d(N_1, \ldots, N_m)$.

By Lemma 3.9, we know that if Ψ_i is spanning on X_i, for all $i \in [m]$, then Π_Ψ is spanning on $X_1 \times \cdots \times X_m$. In particular, any function on C can be expressed as a linear combination of products $\psi_1(X_1) \cdot \ldots \cdot \psi_m(X_m)$, $\psi_i \in \Psi_i$.

Consider any of these products $\psi_1(X_1) \cdot \ldots \cdot \psi_m(X_m)$. Let $\overline{\psi} = (\psi_1, \ldots, \psi_m)$, $X_i' = \psi_i(X_i)$, for all $i \in [m]$, and $C' = \overline{\psi}(C)$. C' is a binary class over m binary

instances and, by Definition 2.2, $\mathrm{VCD}(C') \leq \mathrm{VCD}_\Psi(C) = d$. By Theorem 3.6, the monomial $X'_1 \cdot \ldots \cdot X'_m$ can be expressed as a linear combination of short products in

$$\{X'_{i_1} \cdot \ldots \cdot X'_{i_k} \mid 1 \leq i_1 < \cdots < i_k \leq m \text{ and } k \leq d\}.$$

It follows that $\psi_1(X_1) \cdot \ldots \cdot \psi_m(X_m)$ can be expressed as a linear combination of short products in $\{\psi_{i_1}(X_{i_1}) \cdot \ldots \cdot \psi_{i_k}(X_{i_k}) \mid 1 \leq i_1 < \cdots < i_k \leq m \text{ and } k \leq d\}$.

Moreover, we can use interpolation to represent any mapping $\psi_i(X_i)$ by a polynomial of degree at most N_i, such that $\psi_i(X_i) = a_{N_i} X_i^{N_i} + a_{N_i-1} X_i^{N_i-1} + \cdots + a_0$. By replacing each ψ_{i_j}, $j \in \{1, \ldots, k\}$, in a short product $\psi_{i_1}(X_{i_1}) \cdot \ldots \cdot \psi_{i_k}(X_{i_k})$ with the interpolating polynomial, we can express it as a linear combination of monomials in

$$\{X_{i_1}^{n_{i_1}} \cdots X_{i_k}^{n_{i_k}} : k \leq d, \text{ and } 0 \leq n_{i_t} \leq N_{i_t} \text{ for all } t,\ 1 \leq t \leq k\}.$$

So, any function on C (any vector from $\mathbb{R}^{|C|}$) can be expressed as a linear combination of monomials in $P^d(N_1, \ldots, N_m)$ and hence $P^d(N_1, \ldots, N_m)$ spans the vector space $\mathbb{R}^{|C|}$. □

One immediately obtains the following generalization of Sauer's bound.

Corollary 3.11 (**generalized Sauer bound**). *Let Ψ_i, $1 \leq i \leq m$, be a spanning family of mappings $\psi_i : X_i \to \{0, 1\}$, and $\Psi = \Psi_1 \times \cdots \times \Psi_m$. If $\mathrm{VCD}_\Psi(C) = d$ then $|C| \leq \Phi_d(N_1, \ldots, N_m)$.*

Since Ψ_{G}, Ψ_{P} and Ψ^* are products of spanning families, this bound and the following general definition of maximum classes applies to them.

Definition 3.12 (VCD_Ψ**-maximum class**). *Let Ψ_i, $1 \leq i \leq m$, be a spanning family of mappings $\psi_i : X_i \to \{0, 1\}$. Let $\Psi = \Psi_1 \times \cdots \times \Psi_m$. C is called VCD_Ψ-maximum*

if, for some $d \in \mathbb{N}$, $\mathrm{VCD}_\Psi(C) = d$ *and* $|C| = \Phi_d(N_1, \ldots, N_m)$.

The class of all sets of size up to $\mathrm{VCD}(C)$, which is the standard example of a VCD-maximum class in the binary case, has a straightforward extension to a $\mathrm{VCD}_{\Psi_\mathrm{G}}$-maximum ($\mathrm{VCD}_{\Psi^*}$-maximum) multi-label class, namely the class of concepts that have at most $\mathrm{VCD}_{\Psi_\mathrm{G}}(C)$ ($\mathrm{VCD}_{\Psi^*}(C)$) many non-zero elements. As another intuitive example of a maximum multi-label class, consider the following geometric example of a class that is maximum of VCD_{Ψ^*} 2 and $\mathrm{VCD}_{\Psi_\mathrm{G}}$ 2.

Example 3.13. *X corresponds to m lines in general position on the plane, i.e., no two lines are parallel and no three lines share a common point. Then (i) the number of regions is $1 + m + m(m-1)/2$; (ii) the number of segments and rays is m^2; (iii) the number of intersection points is $m(m-1)/2$. Summing these numbers yields $1 + 2m^2 = \Phi_2(2, \ldots, 2)$. All regions, segments, rays and intersection points form a natural multi-label class concept class that is VCD_{Ψ^*}-maximum and $\mathrm{VCD}_{\Psi_\mathrm{G}}$-maximum of dimension 2. Each instance takes values in $\{-1, 0, +1\}$, depending on which side of the line the concept is on (and 0 if the concept is contained within the line itself). Each region is a concept with instance values -1 or $+1$. Each segment/ray is a concept with value 0 in one particular instance and values -1 or $+1$ in all the other instances. Each intersection point is a concept with value 0 on exactly two instances. One can verify that no set of three instances is shattered using any label mapping to a binary class. Figure 3.1 illustrates such a class for $m = 3$, and Table 3.2 shows the corresponding concepts.*

In the binary case, restrictions and reductions of maximum classes are again maximum [Wel87]. For the multi-label case, the corresponding result is known for restrictions.

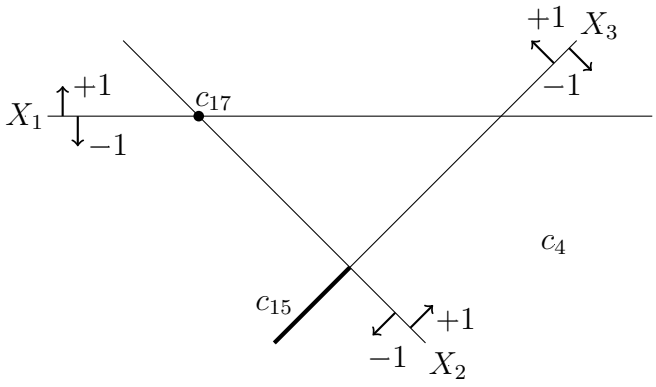

Figure 3.1: The geometric class described in Example 3.13 for $m = 3$.

$c \in C$	X_1	X_2	X_3
c_1	$+1$	-1	$+1$
c_2	$+1$	$+1$	$+1$
c_3	$+1$	$+1$	-1
c_4	-1	$+1$	-1
c_5	-1	-1	-1
c_6	-1	-1	$+1$
c_7	-1	$+1$	$+1$
c_8	0	-1	$+1$
c_9	$+1$	0	$+1$
c_{10}	0	$+1$	$+1$
c_{11}	$+1$	$+1$	0
c_{12}	0	$+1$	-1
c_{13}	-1	$+1$	0
c_{14}	-1	0	-1
c_{15}	-1	-1	0
c_{16}	-1	0	$+1$
c_{17}	0	0	$+1$
c_{18}	0	$+1$	0
c_{19}	-1	0	0

Table 3.2: The VCD_{Ψ_G}-maximum class obtained from Figure 3.1.

Theorem 3.14. [Gur97] *Let Ψ_i, $1 \leq i \leq m$, be a spanning family of mappings $\psi_i : X_i \to \{0, 1\}$, and $\Psi = \Psi_1 \times \cdots \times \Psi_m$. Let C be VCD_Ψ-maximum with $\mathrm{VCD}_\Psi(C) = d$, and $Y \subseteq X$ with $|Y| \geq d$. Then $C|_Y$ is VCD_Ψ-maximum with $\mathrm{VCD}_\Psi(C|_Y) = d$.*

Proof. Let $Y = \{X_{i_1}, \ldots, X_{i_k}\}$ and assume there is a linear dependency between some monomials in $P^d(N_{i_1}, \ldots, N_{i_k})$ on $C|_Y$. Since $P^d(N_{i_1}, \ldots, N_{i_k}) \subseteq P^d(N_1, \ldots, N_m)$, there is a linear dependency between some monomials from $P^d(N_1, \ldots, N_m)$ on $C|_Y$. By the definition of restriction, a linear dependency on $C|_Y$ results in a linear dependency on C. So, the monomials from $P^d(N_1, \ldots, N_m)$ are linearly dependent on C. This contradicts the fact that C is VCD_Ψ-maximum, and so the monomials from $P^d(N_{i_1}, \ldots, N_{i_k})$ are independent on $C|_Y$. Therefore,

$$\text{size}(C|_Y) \geq |P^d(N_{i_1}, \ldots, N_{i_k})| = \Phi_d(N_{i_1}, \ldots, N_{i_k}).$$

Furthermore, $\text{VCD}_\Psi(C|_Y) \leq d$, and by Theorem 3.10, the monomials from $P^d(N_{i_1}, \ldots, N_{i_k})$ span the vector space $\mathbb{R}^{\text{size}(C|_Y)}$. So, $\text{size}(C|_Y) \leq |P^d(N_{i_1}, \ldots, N_{i_k})| = \Phi_d(N_{i_1}, \ldots, N_{i_k})$. Hence, $\text{size}(C|_Y) = \Phi_d(N_{i_1}, \ldots, N_{i_k})$. Considering the size of $C|_Y$, $\text{VCD}_\Psi(C|_Y)$ cannot be smaller than d. Hence $C|_Y$ is a VCD_Ψ-maximum class of dimension $d = \text{VCD}_\Psi(C)$. □

3.3 Algebraic Characterization of Teaching Sets

Here, we identify an algebraic characterization of teaching sets, a result which is contained in our publication in the Journal of Theoretical Computer Science [SSYZ14a]. In Section 3.4 and later in Section 6.1, we will see the usefulness of this characterization.

We first need to introduce some notation. For each $i \in \{1, \ldots, m\}$ and $k \in \{0, \ldots, N_i\}$, let $p_{i,k} : \mathbb{R} \to \{0, 1\}$ be a polynomial of degree N_i that satisfies the

following conditions:

$$p_{i,k}(X_i) = \begin{cases} 1 & \text{if } X_i = k \\ 0 & \text{otherwise.} \end{cases} \tag{3.1}$$

We can find such a polynomial using interpolation. In the binary case, there is no need for interpolation, since

$$p_{i,0}(X_i) = 1 - X_i \text{ and } p_{i,1}(X_i) = X_i. \tag{3.2}$$

Definition 3.15. *Let $C = \{c_1, \ldots, c_n\}$, $|C| = n$. We associate each concept $c_i \in C$, $i \in \{1, \ldots, n\}$, with the ith standard basis vector $\mathbf{c_i} = (0, \ldots, 1, \ldots, 0)$ of $\mathbb{R}^n$.*

Let $p(X_1, \ldots, X_m) \in \mathbb{R}[X_1, \ldots, X_m]$ be a polynomial. As discussed before, p corresponds to a vector $\mathbf{p} = (p_1, \ldots, p_n) \in \mathbb{R}^n$, where $p_i = p(c_i(X_1), \ldots, c_i(X_m))$, for all $i \in \{1, \ldots, n\}$. The phrase "$p(X_1, \ldots, X_m) = 0$ on C" means that p corresponds to the zero vector in $\mathbb{R}^{|C|}$.

Let $\mathcal{P}$ be a collection of polynomials from $\mathbb{R}[X_1, \ldots, X_m]$. We say that $c_i \in C$ lies in the span of $\mathcal{P}$, or c_i can be expressed as a linear combination of elements in $\mathcal{P}$, if the vector $\mathbf{c_i}$ can be expressed as a linear combination of vectors that correspond to polynomials from $\mathcal{P}$. In other words, $c_i \in C$ lies in the span of $\mathcal{P}$ (c_i can be expressed as a linear combination of elements in $\mathcal{P}$) if $\mathbf{c_i} = \mathbf{p}$, where $\mathbf{p}$ corresponds to a polynomial $p(X_{i_1}, \ldots, X_{i_k})$ and $p(X_{i_1}, \ldots, X_{i_k})$ is a linear combination of elements in $\mathcal{P}$.

When every concept in C lies in the span of $\mathcal{P}$, $\mathcal{P}$ spans $\mathbb{R}^{|C|}$ and vice versa.

Lemma 3.16. *If a set of instances $\{X_{i_1}, \ldots, X_{i_k}\} \subseteq X$ is a teaching set for a concept $c \in C$, then c lies in the span of $P^k(N_{i_1}, \ldots, N_{i_k})$.*

Proof. Let $n = |C|$. Fix some ordering on the concepts in C and assume that c

is the tth concept in that ordering. We show that $\mathbf{c_t}$ can be expressed as a linear combination of vectors that correspond to monomials from $P^k(N_{i_1}, \ldots, N_{i_k})$.

Let $\{(X_{i_1}, n_{i_1}), \ldots, (X_{i_k}, n_{i_k})\}$ be a teaching set for c_t in C, and $p(X_{i_1}, \ldots, X_{i_k}) = p_{i_1,n_{i_1}}(X_{i_1}) \times \cdots \times p_{i_k,n_{i_k}}(X_{i_k})$. Since $\{(X_{i_1}, n_{i_1}), \ldots, (X_{i_k}, n_{i_k})\}$ is not consistent with any concept in $C \setminus \{c_t\}$, we conclude that $\mathbf{p} \in \mathbb{R}^n$, which corresponds to $p(X_{i_1}, \ldots, X_{i_k})$, and $\mathbf{c_t} \in \mathbb{R}^n$ are equivalent, i.e., $\mathbf{p} = \mathbf{c_t}$.

Since each $p_{i_t,n_{i_t}}(X_{i_t})$ is a polynomial of degree N_{i_t}, we can write $p(X_{i_1}, \ldots, X_{i_k})$ as a linear combination of monomials from $P^k(N_{i_1}, \ldots, N_{i_k})$. So, $\mathbf{p}$, and consequently $\mathbf{c_t}$, can be expressed as a linear combination of vectors that correspond to monomials from $P^k(N_{i_1}, \ldots, N_{i_k})$. □

The next lemma is a stronger result where the VCD_Ψ of the class comes into play.

Lemma 3.17. *Let* $\text{VCD}_\Psi(C) = d$. *A set of instances* $\{X_{i_1}, \ldots, X_{i_k}\} \subseteq X$ *is a teaching set for a concept* $c \in C$, *if and only if* c *lies in the span of* $P^d(N_{i_1}, \ldots, N_{i_k})$.

Proof. Let $\{(X_{i_1}, n_{i_1}), \ldots, (X_{i_k}, n_{i_k})\}$ be a teaching set for c. By Lemma 3.16, c corresponds to a polynomial p that is a linear combination of monomials in $P^k(N_{i_1}, \ldots, N_{i_k})$. So, for the case when $k \leq d$, p is a linear combination of monomials in $P^d(N_{i_1}, \ldots, N_{i_k})$, and therefore, c lies in the span of $P^d(N_{i_1}, \ldots, N_{i_k})$.

We now consider the case $k > d$. By Theorem 3.14, $C|_{\{X_{i_1},\ldots,X_{i_k}\}}$ is also VCD_Ψ-maximum of dimension d and thus by Theorem 3.10, any real-valued function on $C|_{\{X_{i_1},\ldots,X_{i_k}\}}$ can be expressed as a linear combination of monomials in $P^d(N_{i_1}, \ldots, N_{i_k})$. In particular, we can express p as a linear combination of monomials in $P^d(N_{i_1}, \ldots, N_{i_k})$. Hence, c is in the span of $P^d(N_{i_1}, \ldots, N_{i_k})$.

To prove the other implication, fix some ordering on the concepts in C and assume that c is the uth concept in that ordering, that is, c corresponds to the uth standard basis vector of $\mathbb{R}^{|C|}$. Assume that c_u lies in the span of $P^d(N_{i_1}, \ldots, N_{i_k})$. In

particular, assume that $\mathbf{c_u} = \mathbf{p}$, where $\mathbf{p}$ corresponds to a polynomial $p(X_{i_1}, \ldots, X_{i_k})$ over monomials in $P^d(N_{i_1}, \ldots, N_{i_k})$. So, $\mathbf{p}$ is the uth standard basis vector of $\mathbb{R}^{|C|}$ and has only one non-zero coordinate.

For the purpose of contradiction, assume that $\{X_{i_1}, \ldots, X_{i_k}\}$ is not a teaching set for c. Hence there is another concept $c_v \neq c_u$ from C which coincides with c_u on $\{X_{i_1}, \ldots, X_{i_k}\}$, that is, $c_u(X_{i_j}) = c_v(X_{i_j})$ for all $j = 1, \ldots, k$. Thus the following equalities hold

$$p_u = p(c_u(X_{i_1}), \ldots, c_u(X_{i_k})) = p(c_v(X_{i_1}), \ldots, c_v(X_{i_k})) = p_v.$$

So, the uth and vth coordinates of the vector $\mathbf{p} = (p_1, \ldots, p_n)$ are equal. In particular, $\mathbf{p}$ must have at least two coordinates equal to 1, namely, the uth and vth coordinates — a contradiction. □

3.4 Shortest-path Closedness in One-inclusion Hypergraphs

In this section, we illustrate the power of linear algebraic methods by solving a complex problem in computational learning theory. Kuzmin and Warmuth proved that when a class is VCD-maximum, then in some graph representation, namely, the one-inclusion graph of the class, the length of the shortest path between any two concepts is equal to the symmetric difference of those concepts [KW07]. Such a concept class is then called shortest-path closed.

For $c, c' \in C$, $c \triangle c'$ denotes the set of instances on which c and c' differ, i.e.,

$$c \triangle c' = \{X_i \in X \mid c(X_i) \neq c'(X_i)\}.$$

Definition 3.18. [AHW87] *The* one-inclusion graph $G(C)$ *of a concept class* C *is the labeled graph* G *with* $V(G) = C$ *and* $E(G) = \{\{c, c'\} : |c \triangle c'| = 1\}$. *Every edge* $\{c, c'\} \in E(G)$ *is labeled by the instance from* $c \triangle c'$.

A class C is *shortest-path closed* if for any two concepts $c, c' \in C$, there is a path in $G(C)$ between c and c', which is labeled by instances of $c \triangle c'$ and has a length of $|c \triangle c'|$.

Example 3.19. *Consider the concept class* C *and its one-inclusion graph in Figure 3.2. One can see that* C *is shortest-path closed.*

$c \in C$	X_1	X_2	X_3	X_4
c_1	0	0	0	0
c_2	0	0	0	1
c_3	0	1	0	0
c_4	0	0	1	1
c_5	1	0	0	1
c_6	1	0	1	1

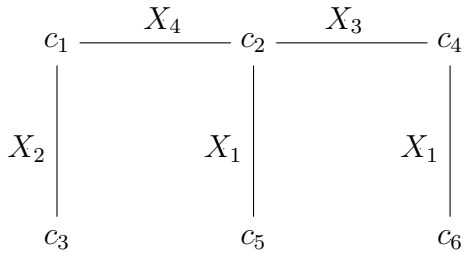

Figure 3.2: A concept class and its one-inclusion graph.

Here, we show that the one-inclusion hypergraph for a VCD_Ψ-maximum class is also shortest-path closed. The one-inclusion hypergraph is a natural extension of the one-inclusion graph to the multi-label case and was defined by Rubinstein et al. [RBR09] as follows.

Definition 3.20. [RBR09] *The* one-inclusion hypergraph $G(C)$ *of a multi-label concept class* C *is the labeled undirected graph* $G(C) = (V, E)$ *with the vertex set* $V(G) = C$ *and the set of hyperedges* $E(G) = \{\{c_{i_1}, \ldots, c_{i_t}\} : |c_{i_j} \triangle c_{i_k}| = 1,$ *for all* $j, k \in \{1, \ldots, t\}, j \neq k, t \geq 2\}$. *The label of a hyperedge* $\{c_{i_1}, \ldots, c_{i_t}\}$ *is the instance* X_p

where $c_{i_j} \triangle c_{i_k} = \{X_p\}$*, for all* $j, k \in \{1, \ldots, t\}$*,* $j \neq k$*. For a concept* $c \in C$*,* $I_C(c)$ *denotes the set of instances labeling hyperedges containing* c*, that is,*

$$I_C(c) = \{X_t \in X \mid \textit{there exists a concept } c' \in C \setminus \{c\} \textit{ such that } c - X_t = c' - X_t\}.$$

Definition 3.21. *Let* $G(C) = (V, E)$ *be the one-inclusion hypergragh of* C*.* $c, c' \in C$ *are called* Hamming-connected *when*

$$|c \triangle c'| = \textit{ the length of the shortest path between } c \textit{ and } c'.$$

That is, there is a path in $G(C)$ *between* c *and* c'*, which is labeled by instances of* $c \triangle c'$ *and has a length of* $|c \triangle c'|$*.* C *is called* shortest-path closed *iff any two concepts* $c, c' \in C$ *are Hamming-connected.*

Kuzmin and Warmuth proved that the one-inclusion graph of a binary maximum class is shortest-path closed [KW07]. Here, we extend this result to VCD_Ψ-maximum classes. We exploit an algebraic approach to show that the one-inclusion hypergraph of any VCD_Ψ-maximum class is shortest-path closed.

We first present a lemma that is a generalization of Lemma 17 in [KW07] in which Kuzmin and Warmuth proved that when C is VCD-maximum, then for any $c \in C$, the set of instances corresponding to incident edges for c in the one-inclusion graph of C, is a teaching set for c. While Kuzmin and Warmuth used a combinatorial argument in their proof, we apply Linear Algebra here. Note that our proof is an alternative proof for the analogue result in the binary case.

Lemma 3.22. [KW07] *Let* C *be a* VCD_Ψ*-maximum class. Then for every* $c \in C$*,* $I_C(c)$ *is a teaching set for* c*.*

Proof. By Theorem 3.10, the monomials from $P^d(N_1, \ldots, N_m)$ span the vector space

$\mathbb{R}^{|C|}$, and since $|P^d(N_1,\ldots,N_m)|= \Phi_d(N_1,\ldots,N_m) = |C|$, the set $P^d(N_1,\ldots,N_m)$ is a basis for $\mathbb{R}^{|C|}$. We use this fact to prove our claim.

Let $c \in C$ and let $S = \{(X_{i_1}, n_{i_1}),\ldots,(X_{i_k}, n_{i_k})\}$ be a minimal teaching set for c in the sense that no proper subset of S is a teaching set for c. For purposes of contradiction, suppose $I_C(c) \neq X(S)$. In particular, assume that $X(S) \nsubseteq I_C(c)$ and let $X_t \in X(S) \setminus I_C(c)$. By Lemma 3.17, there is a linear combination p of monomials from $P^d(N_{i_1},\ldots,N_{i_k})$ such that $\mathbf{c} = \mathbf{p}$. Note that $X \setminus \{X_t\}$ is also a teaching set for c, since otherwise $X_t \in I_C(c)$. Thus, there is a linear combination p' of monomials from $P^d(N_1,\ldots,N_{t-1},N_{t+1},\ldots,N_m)$ with $\mathbf{c} = \mathbf{p'}$ and thus $\mathbf{p} = \mathbf{p'}$. In particular, the polynomials p and p' are equivalent on C. Since $P^d(N_1,\ldots,N_m)$ is a basis for $\mathbb{R}^{|C|}$, there is no linear dependency between the monomials in $P^d(N_1,\ldots,N_m)$. As p' does not depend on X_t, p does not depend on X_t either. Thus p depends only on variables from $S \setminus \{X_t\}$. By Lemma 3.17, $S \setminus \{X_t\}$ is a teaching set for c, which contradicts the minimality of S. Therefore $S \subseteq I_C(c)$ and $I_C(c)$ is a teaching set for c. □

The proof of the following theorem is completely different from the proof of Lemma 14 in [KW07], which states the same result for the binary case.

Theorem 3.23. *If C is a VCD_Ψ-maximum class, then C is shortest-path closed.*

Proof. We prove that any two concepts c_1 and c_2 in C are Hamming-connected, by induction on $|c_1 \triangle c_2|$. For $|c_1 \triangle c_2|= 1$ the proof is obvious. Suppose $|c_1 \triangle c_2|= n$ and any two concepts c, c' with $|c \triangle c'| < n$ are Hamming-connected. By Lemma 3.22, $I_C(c_1)$ is a teaching set for c_1, and therefore, it cannot be disjoint from $c_1 \triangle c_2$. Hence, there is an $X_i \in I_C(c_1) \cap (c_1 \triangle c_2)$, $i \in [m]$. Let c' be the concept from C such that $c_1 \triangle c' = \{X_i\}$. Then $|c' \triangle c_2|= n - 1$ and by the induction hypothesis c' and c_2 are Hamming-connected. Therefore, c_1 and c_2 are Hamming-connected. □

Chapter 4

The Reduction Property

In this chapter, we identify the *reduction property* for VCD_Ψ which turns out to be an essential structural feature of VCD_Ψ-maximum classes. All our results in this chapter are published in the proceedings of the 25th International Conference on Algorithmic Learning Theory [SYZ14].

As shown by Floyd and Warmuth [FW95], every binary *maximum* class C has a compression scheme of size $\mathrm{VCD}(C)$. This result was strengthened by showing the existence of unlabeled schemes (in which the compression sets are subsets of X without label information) of size $\mathrm{VCD}(C)$ [KW07]. Both results rely on the fact that, for $\mathrm{VCD}(C) = d < m$, restrictions and reductions of binary maximum classes w.r.t. a single instance are maximum of VCD d and $d-1$, respectively [Wel87].

As in the binary case, restrictions of VCD_Ψ-maximum classes are still maximum in the multi-label case. However, the definition of reduction for multi-label concept classes is not as straightforward as in the binary case. In fact, there are two possible definitions of reduction of a VCD_Ψ-maximum class C w.r.t. a single instance $X_t \in X$: one could consider the class of concepts in $C - X_t$ that have at least two distinct extensions denoted by $[C]^{X_t}_{\geq 2}$, or of those that have all $N_t + 1$ extensions to concepts

in C denoted by C^{X_t}.

We thus define a core notion of our work, namely the reduction property, and in Section 4.1, we prove that the Graph-dimension fulfills the reduction property. However, in Section 4.2 and Section 4.3, we illustrate that neither Pollard's pseudo-dimension nor the Natarajan-dimension fulfill this property. As we will show in Chapter 5, the reduction property provides a sufficient condition for maximum classes of VCD_Ψ d to have a sample compression scheme of size d, provided that Ψ is based on spanning families.

Definition 4.1 (reduction property)**.** *Let $m > 1$ and Ψ_i, $1 \leq i \leq m$, be a family of mappings. Let $\Psi = \Psi_1 \times \cdots \times \Psi_m$.* VCD_Ψ *fulfills the* reduction property *iff for any* VCD_Ψ*-maximum class $C \subseteq \prod_{i=1}^{m} X_i$, for any $t \in [m]$ and for any concept $\overline{c} \in C - X_t$, $|\{c \in C \mid c - X_t = \overline{c}\}| \in \{1, N_t + 1\}$ (i.e., $[C]^{X_t}_{\geq 2} = C^{X_t}$).*

When VCD_Ψ fulfills the reduction property, by reduction of a class C (w.r.t. an instance X_t, $t \in [m]$) we always refer to both C^{X_t} and $[C]^{X_t}_{\geq 2}$, which are equal in this case. The following theorem states the key consequence of the reduction property for VCD_Ψ-maximum classes.

Theorem 4.2. *Let Ψ_i, $1 \leq i \leq m$, be a spanning family of mappings on X_i and $\Psi = \Psi_1 \times \cdots \times \Psi_m$. Let C be a* VCD_Ψ*-maximum class with* $\mathrm{VCD}_\Psi(C) = d$*. If* VCD_Ψ *fulfills the reduction property, then C^{X_t} is* VCD_Ψ*-maximum with* $\mathrm{VCD}_\Psi(C^{X_t}) = d-1$*, for any $t \in [m]$.*

Proof. For $m = d$, the claim is obviously true. So suppose $m > d$. It suffices to prove the statement for $t = m$. We first show that $\mathrm{VCD}_\Psi(C^{X_m}) \leq d - 1$. Assume $\mathrm{VCD}_\Psi(C^{X_m}) = d$, and, w.l.o.g., C^{X_m} shatters $\{X_1, \ldots, X_d\}$. Let $\overline{\psi^{1,m-1}} =$

$(\psi_1, \ldots, \psi_{m-1})$ be a tuple of non-constant mappings $\psi_i : X_i \to \{0,1\}$ where

$$\overline{\psi^{1,m-1}}(C^{X_m})|_{\{X_1,\ldots,X_d\}} = \{0,1\}^d.$$

Let $\psi_m : X_m \to \{0,1\}$ be $\psi_{0\neq 1}$ as discussed in Remark 3.8, i.e.,

$$\psi_m(x) = \begin{cases} 0 & \text{if } x = 0 \\ 1 & \text{if } x = 1 \\ 0 \text{ or } 1 & \text{otherwise,} \end{cases}$$

and $\overline{\psi^{1,m}} = (\psi_1, \ldots, \psi_{m-1}, \psi_m)$. Since VCD_Ψ fulfills the reduction property, any concept $c \in C^{X_m}$ has all $N_m + 1$ extensions to concepts in C. In particular,

$$c|_{\{X_1,\ldots,X_d\}} \cup \{(X_m, 0)\} \in C|_{\{X_1,\ldots,X_d,X_m\}}$$

and

$$c|_{\{X_1,\ldots,X_d\}} \cup \{(X_m, 1)\} \in C|_{\{X_1,\ldots,X_d,X_m\}}.$$

So, $\overline{\psi^{1,m}}(C)|_{\{X_1,\ldots,X_d,X_m\}} = \{0,1\}^{d+1}$, which contradicts the fact that $\mathrm{VCD}_\Psi(C) = d$. Hence, $\mathrm{VCD}_\Psi(C^{X_m}) \leq d - 1$.

By the reduction property, each concept $c \in C - X_m$ either has a unique extension to concepts in C or has all $N_m + 1$ extensions to concepts in C. So,

$$|C| = |C - X_m| + N_m |C^{X_m}|.$$

Also, by Theorem 3.14, $C - X_m$ is VCD_Ψ-maximum of dimension d. So,

$$
\begin{aligned}
|C^{X_m}| &= \frac{1}{N_m}(|C| - |C - X_m|) \\
&= \frac{1}{N_m}(\Phi_d(N_1, \ldots, N_m) - \Phi_d(N_1, \ldots, N_{m-1})) \\
&= \frac{1}{N_m}(N_m + \sum_{1 \le i \le m-1} N_i N_m + \cdots + \sum_{1 \le i_1 < i_2 < \cdots < i_{d-1} \le m-1} N_{i_1} N_{i_2} \cdots N_{i_{d-1}} N_m) \\
&= \frac{1}{N_m}(N_m \Phi_{d-1}(N_1, \ldots, N_{m-1})) \\
&= \Phi_{d-1}(N_1, \ldots, N_{m-1}).
\end{aligned}
$$

Since $\mathrm{VCD}_\Psi(C^{X_m}) \le d-1$ and $|C^{X_m}| = \Phi_{d-1}(N_1, \ldots, N_{m-1})$, the reduction class C^{X_m} is VCD_Ψ-maximum with $\mathrm{VCD}_\Psi(C^{X_m}) = d-1$. □

For any set $Y \subseteq X$, we extend the definition of C^Y from the binary case to the multi-label case in the obvious way. It should be noted that C^Y is well-defined, because $(C^{X_i})^{X_j} = (C^{X_j})^{X_i}$ for all $i, j \in [m]$, as in the binary case. The proof is similar to the one in the binary case [Wel87], and we include it here for the sake of completeness.

Proposition 4.3. *For any X_i, X_j with $i \neq j$, $(C^{X_i})^{X_j} = (C^{X_j})^{X_i}$.*

Proof.

$$c \in (C^{X_i})^{X_j} \Leftrightarrow c \cup \{(X_j, l)\} \in C^{X_i}, \text{ for all } l \in \{0, \ldots, N_j\} \Leftrightarrow$$

$$\text{for each } c \cup \{(X_j, l)\} \in C^{X_i},\ \{c \cup \{(X_j, l)\}\} \cup \{(X_i, t)\} \in C, \text{ for all } t \in \{0, \ldots, N_i\} \Leftrightarrow$$

$$c \cup \{(X_j, l), (X_i, t)\} \in C \text{ for all } l \in \{0, \ldots, N_j\} \text{ and for all } t \in \{0, \ldots, N_i\} \Leftrightarrow$$

$$c \cup \{(X_i, t)\} \in C^{X_j}, \text{ for all } t \in \{0, \ldots, N_i\} \Leftrightarrow c \in (C^{X_j})^{X_i}$$

□

In Chapter 5, we present an extension of Floyd and Warmuth's compression scheme to VCD_Ψ-maximum classes when VCD_Ψ fulfills the reduction property. We also show that fulfilling the reduction property for VCD_Ψ is sufficient to generalize Kuzmin and Warmuth's unlabeled compression scheme to VCD_Ψ-maximum classes. Since an unlabeled compression scheme cannot exist in the multi-label case, we reformulate the Kuzmin-Warmuth compression scheme as a labeled scheme with some nice properties.

The rest of this chapter deals with checking the reduction property for the most well-known VCD notions for multi-label concept classes in the literature for the reduction property. In particular, we show that, while the Graph-dimension has the reduction property, Pollard's pseudo-dimension and the Natarajan-dimension do not fulfill it.

4.1 The Graph Dimension

Our main objective here is to justify the following theorem.

Theorem 4.4. $\mathrm{VCD}_{\Psi_\mathrm{G}}$ *fulfills the reduction property.*

To prove Theorem 4.4 we need a sequence of lemmas and theorems. Let id_i denote the identity mapping on X_i. We now show that for a VCD_Ψ-maximum class over a spanning family Ψ, if we only map one column to binary values and keep the other columns unchanged, the resulting class is still maximum of the same dimension.

Lemma 4.5. *Let $\Psi = \Psi_1 \times \cdots \times \Psi_m$, where each Ψ_i, for $i \in [m]$, is a spanning family of mappings on X_i, and let C be* VCD_Ψ*-maximum. Let $\varphi_t \in \Psi_t$ be a non-constant*

mapping and $\overline{\varphi_t} = (\mathrm{id}_1, \ldots, \mathrm{id}_{t-1}, \varphi_t, \mathrm{id}_{t+1}, \ldots, \mathrm{id}_m)$. *Then* $\overline{\varphi_t}(C)$ *is* VCD_Ψ*-maximum of dimension* $\mathrm{VCD}_\Psi(C)$.

Proof. Let $d = \mathrm{VCD}_\Psi(C)$. If $d = 0$, then $|C| = 1$ and the claim is trivial.

W.l.o.g., let $t = 1$, i.e, $\varphi_1 : X_1 \to \{0, 1\}$ and $\overline{\varphi_1} = (\varphi_1, \mathrm{id}_2, \ldots, \mathrm{id}_m)$. Let $X_1' = \varphi_1(X_1) = \{0, 1\}$ and $C' = \overline{\varphi_1}(C)$. Then, $\mathrm{VCD}_\Psi(C') = \max_{\overline{\psi} \in \Psi} \mathrm{VCD}(\overline{\psi}(C')) \leq \max_{\overline{\psi} \in \Psi} \mathrm{VCD}(\overline{\psi}(C)) = d$. Since $\mathrm{VCD}_\Psi(C') \leq d$, the monomials in $P^d(1, N_2, \ldots, N_m)$ with variables in $\{X_1', X_2, \ldots, X_m\}$ span $\mathbb{R}^{|C'|}$, by Theorem 3.10. If C' is not VCD_Ψ-maximum of dimension d, then the monomials in $P^d(1, N_2 \ldots, N_m)$ must be linearly dependent. We will show that a linear dependency between the monomials in $P^d(1, N_2, \ldots, N_m)$ with variables in $\{X_1', X_2, \ldots, X_m\}$ implies a linear dependency between the monomials in $P^d(N_1, \ldots, N_m)$ with variables in $\{X_1, \ldots, X_m\}$. This will contradict the assumption that C is VCD_Ψ-maximum because if $|C| = \Phi_d(N_1, \ldots, N_m)$ then the monomials from $P^d(N_1, \ldots, N_m)$ must be linearly independent.

Assume there is a linear dependency between the monomials in $P^d(1, N_2, \ldots, N_m)$, i.e., there is a non-trivial polynomial $Q(X_1', X_2, \ldots, X_m)$ that is equal to a non-trivial linear combination of the monomials from $P^d(1, N_2, \ldots, N_m)$ and $Q(X_1', X_2, \ldots, X_m) = 0$ on C'. There are two possible cases to consider:

Case 1 : X_1' does not occur in Q. So, there is a linear dependency between the monomials in $P^d(N_2, \ldots, N_m)$ with variables in $\{X_2, \ldots, X_m\}$. Hence, there is a linear dependency between the monomials in $P^d(N_1, \ldots, N_m)$ with variables in $\{X_1, \ldots, X_m\}$ and C is not VCD_Ψ-maximum.

Case 2 : X_1' occurs in $Q(X_1', X_2, \ldots, X_m)$. We convert Q to Q' as follows: for each monomial $X_1' X_{i_1}^{n_{i_1}} \cdots X_{i_t}^{n_{i_t}}$ in $Q(X_1', X_2, \ldots, X_m)$ with $t < d$, replace X_1' with a polynomial of degree n_1 that interpolates φ_1 on X_1. Note that $0 < n_1 \leq N_1$, because by our assumption φ_1 is non-constant. The result of this conversion is a polynomial

$Q'(X_1, \ldots, X_m)$ that can be expressed as a linear combination of the monomials in $P^d(N_1, \ldots, N_m)$ and furthermore $Q'(X_1, \ldots, X_m) = 0$ on C.

Now, we show that $Q'(X_1, \ldots, X_m)$ is a non-trivial polynomial. Consider one of the longest monomials $X'_1 X_{i_1}^{n_{i_1}} \cdots X_{i_t}^{n_{i_t}}$ that appear in Q. Since Q is non-trivial, there is at least one such monomial. Let $R(X_1) = a_{n_1} X_1^{n_1} + a_{n_1-1} X_1^{n_1-1} + \cdots + a_0$, where $a_i \in \mathbb{R}$ for $i \leq n_1$ and $a_{n_1} \neq 0$, be an interpolating polynomial for φ_1, that is, $R(x) = \varphi_1(x)$ for all $0 \leq x \leq N_1$. Replacing X'_1 in $X'_1 X_{i_1}^{n_{i_1}} \cdots X_{i_t}^{n_{i_t}}$ with $R(X_1)$ results in the following polynomial

$$\begin{aligned} R(X_1) X_{i_1}^{n_{i_1}} \cdots X_{i_t}^{n_{i_t}} &= (a_{n_1} X_1^{n_1} + a_{n_1-1} X_1^{n_1-1} + \cdots + a_0) X_{i_1}^{n_{i_1}} \cdots X_{i_t}^{n_{i_t}} \\ &= a_{n_1} X_1^{n_1} X_{i_1}^{n_{i_1}} \cdots X_{i_t}^{n_{i_t}} + a_{n_1-1} X_1^{n_1-1} X_{i_1}^{n_{i_1}} \cdots X_{i_t}^{n_{i_t}} + \cdots \\ &+ a_0 X_{i_1}^{n_{i_1}} \cdots X_{i_t}^{n_{i_t}}. \end{aligned}$$

Since $X'_1 X_{i_1}^{n_{i_1}} \cdots X_{i_t}^{n_{i_t}}$ is one of the longest monomials of this form in Q, we conclude that $a_{n_1} X_1^{n_1} X_{i_1}^{n_{i_1}} \cdots X_{i_t}^{n_{i_t}}$ cannot be canceled out in Q'. Hence, $Q'(X_1, \ldots, X_m)$ is non-trivial and there is a linear dependency between the monomials in $P^d(N_1, \ldots, N_m)$ with variables in $\{X_1, \ldots, X_m\}$. Therefore, C cannot be VCD_Ψ-maximum. □

The next lemma extends Lemma 4.5 and states that if we also map more than one column to binary values and keep the other columns unchanged, the resulting class is still maximum of the same dimension.

Lemma 4.6. *Let $\Psi = \Psi_1 \times \cdots \times \Psi_m$, where each Ψ_i, for $i \in [m]$, is a spanning family of mappings on X_i, and C be VCD_Ψ-maximum. Let $\overline{\varphi} = (\varphi_1, \ldots, \varphi_m)$ be a tuple of non-constant mappings where $\varphi_i \in (\Psi_i \cup \{\mathrm{id}_i\})$, for all $i \in [m]$. Then $\overline{\varphi}(C)$ is also a VCD_Ψ-maximum class of dimension $\mathrm{VCD}_\Psi(C)$.*

Proof. Choose $k \in [m]$. W.l.o.g., let $\varphi_i \in \Psi_\mathrm{i}$, for all $i \in \{1, \ldots, k\}$, and φ_i, $k+1 \leq i \leq m$, be the identity mapping on X_i. In other words, $\overline{\varphi} = (\varphi_1, \ldots, \varphi_k, \mathrm{id}_{k+1}, \ldots, \mathrm{id}_m)$.

Also, let $\overline{\varphi_t} = (\mathrm{id}_1, \ldots, \mathrm{id}_{t-1}, \varphi_t, \mathrm{id}_{t+1}, \ldots, \mathrm{id}_m)$, for $1 \leq t \leq k$. It is easy to see that $\overline{\varphi}(C) = \overline{\varphi_k}(\cdots \overline{\varphi_1}(C))$. Applying Lemma 4.5 to each φ_t repeatedly from $t = 1$ to $t = k$ proves the claim. □

Recall that Ψ^* is based on the family of all label mappings including constant mappings on X_i, for all $i \in [m]$. It is obvious that if one of the φ_i's is a constant mapping, then $\overline{\varphi}(C)$ is not maximum because it contains a constant column of 0s or 1s. Thus we obtain the following corollaries.

Corollary 4.7. *Let $\Psi = \Psi_1 \times \cdots \times \Psi_m$, where each Ψ_i, for $i \in [m]$, is a spanning family of mappings on X_i, and C be VCD_Ψ-maximum. Let $\overline{\varphi} = (\varphi_1, \ldots, \varphi_m)$ be a tuple of mappings where $\varphi_i \in \Psi_i$, for all $i \in [m]$. Then $\overline{\varphi}(C)$ is* VCD*-maximum of dimension $\mathrm{VCD}_\Psi(C)$ iff φ_i is non-constant for all $i \in [m]$.*

Corollary 4.8. *Let C be VCD_{Ψ^*}-maximum and $\overline{\varphi} = (\varphi_1, \ldots, \varphi_m)$ a tuple of mappings $\varphi_i : X_i \to \{0, 1\}$. Then $\overline{\varphi}(C)$ is* VCD*-maximum of dimension $\mathrm{VCD}_{\Psi^*}(C)$ iff φ_i is non-constant for all $i \in [m]$.*

Lemma 4.5 may be of interest beyond the study of VCD_{Ψ_G}, as it applies to a broad class of notions of VC-dimension. The following two lemmas are the immediate corollaries we obtain from Lemma 4.5 and Lemma 4.6, respectively.

Lemma 4.9. *Let C be VCD_{Ψ_G}-maximum. Let $\varphi_t \in \Psi_{G_t}$, for some $t \in [m]$, and $\overline{\varphi_t} = (\mathrm{id}_1, \ldots, \mathrm{id}_{t-1}, \varphi_t, \mathrm{id}_{t+1}, \ldots, \mathrm{id}_m)$. Then $\overline{\varphi_t}(C)$ is VCD_{Ψ_G}-maximum of dimension $\mathrm{VCD}_{\Psi_G}(C)$.*

Lemma 4.10. *Let C be a VCD_{Ψ_G}-maximum class and let $\overline{\varphi} = (\varphi_1, \ldots, \varphi_m)$ be a tuple of mappings such that $\varphi_i \in (\Psi_{G_i} \cup \{\mathrm{id}_i\})$, for all $i \in [m]$. Then $\overline{\varphi}(C)$ is also a VCD_{Ψ_G}-maximum class of dimension $\mathrm{VCD}_{\Psi_G}(C)$.*

In the binary case, restrictions and reductions of maximum classes are again maximum [Wel87]. Theorem 3.14 implies that the restriction of a VCD_{Ψ_G}-maximum class of VCD_{Ψ_G} d is also maximum of VCD_{Ψ_G} $d-1$. Our core result here is that any reduction of a VCD_{Ψ_G}-maximum class is also VCD_{Ψ_G}-maximum. To show this, we first claim that for any VCD_{Ψ_G}-maximum class C, each concept $c \in C - X_t$, for all $t \in [m]$, has either a unique extension in C or all possible extensions in C. A challenging part of the proof of this claim is to establish the following crucial lemma.

Lemma 4.11. *Let $X_i = \{0,1\}$, for $i \in [m-1]$, $X_m = \{0,\ldots,N_m\}$, $N_m \geq 2$. Let $\Psi = \mathrm{id}_1 \times \cdots \times \mathrm{id}_{m-1} \times \Psi_G$ and $C \subseteq \prod_{i=1}^{m} X_i$ be VCD_Ψ-maximum with $\mathrm{VCD}_\Psi(C) = m-1$. Then for all $\overline{c} \in C - X_m$, $|\{c \in C \mid c - X_m = \overline{c}\}| \in \{1, N_m+1\}$.*

Proof. Note that $|C| = \Phi_{m-1}(1,\ldots,1,N_m)$ and $C - X_m = \{0,1\}^{m-1}$. We show that if some $\overline{c} \in C - X_m$ has more than one but fewer than N_m+1 extensions in C, then $\mathrm{VCD}_\Psi(C) = m$. To do this, we first partition C into N_m+1 classes $C_i = \{c \in C \mid c(X_m) = i\}$, for $0 \leq i \leq N_m$. Clearly, $C_i \cap C_j = \emptyset$, for $i \neq j$, and $C = \bigcup_{i=0}^{N_m} C_i$. We claim that

$$2^{m-1} - 1 \leq |C_i| \leq 2^{m-1}, \quad \text{for all } i \in \{0,\ldots,N_m\}. \tag{4.1}$$

$|C_i| \leq |C - X_m| = 2^{m-1}$ yields the upper bound. For the lower bound, assume $|C_t| = 2^{m-1} - k$, $k \geq 2$, for some $t \in X_m$. Then one can show $|C \setminus C_t| \geq (2^{m-1}-1)N_m + 2$, as follows.

$$\begin{aligned} |C \setminus C_t| = |C| - |C_t| &= 2^{m-1} + 2^{m-1}N_m - N_m - (2^{m-1} - k) \\ &= 2^{m-1} + 2^{m-1}N_m - N_m - 2^{m-1} + k = (2^{m-1}-1)N_m + k \\ &\geq (2^{m-1}-1)N_m + 2. \end{aligned}$$

So, by the pigeonhole principle and by $|C_i| \leq 2^{m-1}$, at least two $C_l, C_{l'} \subseteq (C \setminus C_t)$ satisfy $|C_l| = |C_{l'}| = 2^{m-1}$ and $C_l - X_m = C_{l'} - X_m = \{0,1\}^{m-1}$. Thus, any tuple in Ψ_G that maps l and l' to different values, i.e., $\psi_{l \neq l'}$ as in Remark 3.8, makes C shatter $\{X_1, \ldots, X_m\}$ — a contradiction. Hence, for all $i \in \{0, \ldots, N_m\}$, $|C_i| \geq 2^{m-1} - 1$. We claim

(a) There exists some $t \in X_m$, such that $|C_t| = 2^{m-1}$.

(b) $|C_i| = 2^{m-1} - 1$ for all $i \in X_m \setminus \{t\}$.

Assume that for all $i \in X_m$, $|C_i| = 2^{m-1} - 1$. Then $|C| = \sum_{i=0}^{N_m} |C_i| = (N_m + 1)(2^{m-1} - 1) = 2^{m-1} + 2^{m-1} N_m - N_m - 1 < 2^{m-1} + 2^{m-1} N_m - N_m = \Phi_{m-1}(1, \ldots, 1, N_m)$. So, there is at least one concept class $C_t \subseteq C$ such that $|C_t| > 2^{m-1} - 1$, that is, $|C_t| = 2^{m-1}$ from (4.1), which proves (a). Consequently, $\sum_{i=0,\ i \neq t}^{N_m} |C_i| = |C| - |C_t| = 2^{m-1} + 2^{m-1} N_m - N_m - 2^{m-1} = 2^{m-1} N_m - N_m = (2^{m-1} - 1) N_m$. Since $|C_i| \geq 2^{m-1} - 1$, for all $0 \leq i \leq N_m$, we conclude that $|C_i| = 2^{m-1} - 1$, for all $i \in X_m \setminus \{t\}$, i.e., we have proven (b).

Now let $1 \leq k < N_m$. Suppose there is a $\overline{c} \in C - X_m$ with $|\{c \in C \mid c - X_m = \overline{c}\}| = k + 1$. Let $c_0, \ldots, c_k \in C$ with $c_i \neq c_j$ and $c_i - X_m = c_j - X_m = \overline{c}$, for all $i, j \in \{0, \ldots, k\}$, $i \neq j$. W.l.o.g., $c_i(X_m) = i$ for $i \in \{0, \ldots, k\}$. On the one hand,

$$c_i = \overline{c} \times \{i\} \in C_i \quad \text{for each } i \in \{0, \ldots, k\}. \tag{4.2}$$

On the other hand, for $c \in C$ with $c - X_m = \overline{c}$, $c(X_m) \neq l$, for all $l \in \{k + 1, \ldots, N_m\}$. Thus, for all $l \in \{k + 1, \ldots, N_m\}$, $\overline{c} \times \{l\} \notin C$ and $\overline{c} \times \{l\} \notin C_l$. So, $C_l \subseteq (\{0,1\}^{m-1} \times \{l\}) \setminus \{\overline{c} \times \{l\}\}$, for $l \in \{k + 1, \ldots, N_m\}$ and thus, from (4.1), $|C_l| = 2^{m-1} - 1$ and $C_l = (\{0,1\}^{m-1} \times \{l\}) \setminus \{\overline{c} \times \{l\}\}$, for $l \in \{k + 1, \ldots, N_m\}$. Consequently, from (a), for some $t \in \{0, \ldots, k\}$, $|C_t| = 2^{m-1}$.

We show $\mathrm{VCD}_\Psi(C) = m$. Let $\overline{\psi} = (\mathrm{id}_1, \ldots, \mathrm{id}_{m-1}, \psi_m)$, where $\psi_m(x) = 1$ if $x = t$, else $\psi_m(x) = 0$. First, $\overline{\psi}(C_t) = \{0,1\}^{m-1} \times \{1\}$. Second, $\overline{c} \times \{k+1\} \notin C_{k+1}$, so $\overline{\psi}(C_{k+1}) = (\{0,1\}^{m-1} \times \{0\}) \setminus \{\overline{c} \times \{0\}\}$. Hence, $\{0,1\}^m \setminus \{\overline{c} \times \{0\}\} \subseteq \overline{\psi}(C)$. By (4.2), $\overline{c} \times \{0\} \in \overline{\psi}(C_i)$, for all $i \in \{0, \ldots, k\} \setminus \{t\}$, so $\overline{\psi}(C) = \{0,1\}^m$. □

We now generalize Lemma 4.11 and come back to the main theorem of this section, which claims that $\mathrm{VCD}_{\Psi_\mathrm{G}}$ fulfills the reduction property.

Proof of Theorem 4.4. Let C be a $\mathrm{VCD}_{\Psi_\mathrm{G}}$-maximum class with $\mathrm{VCD}_{\Psi_\mathrm{G}}(C) = d$. Let $t \in [m]$ and $\overline{c} \in C - X_t$. By Definition 4.1 we need to show that $|\{c \in C \mid c - X_t = \overline{c}\}| \in \{1, N_t + 1\}$.

Note that, by definition, $m \geq d$. For $m = d$, we obtain $\mathrm{VCD}_{\Psi_\mathrm{G}}(C) = m$ and thus $C = \prod_{i=1}^m X_i$. So, for any $t \in [m]$, and any concept $c \in C - X_t$, c has all possible extensions to concepts in C. For $m = d + 1$, the statement of the theorem coincides with Lemma 4.11 and is thus proven. So suppose $m > d + 1$.

Consider a $\mathrm{VCD}_{\Psi_\mathrm{G}}$-maximum class $C \subseteq \prod_{i=1}^m X_i$ with $\mathrm{VCD}_{\Psi_\mathrm{G}}(C) = d$. It suffices to prove the statement of the theorem for $t = 1$. So, let $1 \leq k < N_1$, and suppose there is some $\overline{c} \in C - X_1$ such that $|\{c \in C \mid c - X_1 = \overline{c}\}| = k + 1$. Let $c_0, \ldots, c_k \in C$ such that $c_i \neq c_j$ and $c_i - X_1 = c_j - X_1 = \overline{c}$, for all $i, j \in \{0, \ldots, k\}$ with $i \neq j$. W.l.o.g., let $c_i(X_1) = i$ for $i \in \{0, \ldots, k\}$.

Let $c_\mathrm{new} = \overline{c} \cup \{(X_1, k+1)\}$ and $C_\mathrm{new} = C \cup \{c_\mathrm{new}\}$. C is $\mathrm{VCD}_{\Psi_\mathrm{G}}$-maximum of dimension d, so C_new shatters a subset of the instance space of size $d + 1$, including X_1. W.l.o.g., let $\{X_1, \ldots, X_{d+1}\}$ be shattered by C_new. That is, there is a tuple of mappings $\overline{\psi} = (\psi_1, \ldots, \psi_m)$ where $\psi_i : X_i \to \{0,1\}$, for all $i \in [m]$ and $\overline{\psi}(C_\mathrm{new})|_{\{X_1,\ldots,X_{d+1}\}} = \{0,1\}^{d+1}$.

We show that $\{X_1, \ldots, X_{d+1}\}$ is shattered by C as well. By Theorem 3.14, $C|_{\{X_1,\ldots,X_{d+1}\}}$ is $\mathrm{VCD}_{\Psi_\mathrm{G}}$-maximum of dimension d. Since, $c_i|_{\{X_1,\ldots,X_{d+1}\}} \in C|_{\{X_1,\ldots,X_{d+1}\}}$,

for all $i \in \{0, \ldots, k\}$, by Lemma 4.11, $c_i|_{\{X_2,\ldots,X_{d+1}\}}$ has either a unique or all extensions to concepts in $C|_{\{X_1,\ldots,X_{d+1}\}}$. Since $\overline{c}$ has more than one extension to concepts in C, we obtain that $\overline{c}|_{\{X_2,\ldots,X_{d+1}\}}$ has more than one extension — and thus all possible extensions — to concepts in $C|_{\{X_1,\ldots,X_{d+1}\}}$. In particular, there is a concept $c' \in C|_{\{X_1,\ldots,X_{d+1}\}}$, such that $c'|_{\{X_2,\ldots,X_{d+1}\}} = \overline{c}|_{\{X_2,\ldots,X_{d+1}\}}$, and $c'(X_1) = k+1$. Equivalently, $c_{\text{new}}|_{\{X_1,\ldots,X_{d+1}\}} \in C|_{\{X_1,\ldots,X_{d+1}\}}$, and therefore $C|_{\{X_1,\ldots,X_{d+1}\}} = C_{\text{new}}|_{\{X_1,\ldots,X_{d+1}\}}$. Hence, $\overline{\psi}(C|_{\{X_1,\ldots,X_{d+1}\}}) = \overline{\psi}(C_{\text{new}}|_{\{X_1,\ldots,X_{d+1}\}}) = \{0,1\}^{d+1}$ and C shatters $\{X_1, \ldots, X_{d+1}\}$ in contradiction to $\text{VCD}_{\Psi_G}(C) = d$. □

Hence, for a VCD_{Ψ_G}-maximum class C, $[C]^{X_t}_{\geq 2} = C^{X_t}$, for all $t \in [m]$. More precisely, for a VCD_{Ψ_G}-maximum class C, it does not make any difference whether the reduction C^{X_t} is defined as the set of all concepts in $C - X_t$ that have more than one extension in C, or the set of all concepts in $C - X_t$ that have all $N_t + 1$ extensions in C.

Now, the following statement is an obvious corollary of Theorem 4.2 and Theorem 4.4.

Corollary 4.12. *Let C be a VCD_{Ψ_G}-maximum class with $\text{VCD}_{\Psi_G}(C) = d$. Then C^{X_t} is VCD_{Ψ_G}-maximum with $\text{VCD}_{\Psi_G}(C^{X_t}) = d - 1$, for any $t \in [m]$.*

We remind the reader that $\Psi^* \supseteq \Psi_G$ and also, any VCD_{Ψ^*}-maximum class C is also VCD_{Ψ_G}-maximum with $\text{VCD}_{\Psi_G}(C) = \text{VCD}_{\Psi^*}(C)$. So, all the statements and proofs in this section can be applied to VCD_{Ψ^*} as well. In particular, we have the following corollaries of Theorem 4.4 and Theorem 4.2, respectively.

Corollary 4.13. VCD_{Ψ^*} *fulfills the reduction property.*

Corollary 4.14. *For any VCD_{Ψ^*}-maximum class C with $\text{VCD}_{\Psi^*}(C) = d$ and for any $t \in [m]$, C^{X_t} is also VCD_{Ψ^*}-maximum with $\text{VCD}_{\Psi^*}(C^{X_t}) = d - 1$.*

4.2 Pollard's pseudo-dimension

For $\mathrm{VCD}_{\Psi_\mathrm{P}}$, we give a counterexample to the reduction property.

Proposition 4.15. *There is a* $\mathrm{VCD}_{\Psi_\mathrm{P}}$*-maximum class* C *with* $\mathrm{VCD}_{\Psi_\mathrm{P}}(C) = 2$ *such that, for some* $X_t \in X$ *and some* $\overline{c} \in C - X_t$, $|\{c \in C \mid c - X_t = \overline{c}\}| = 2 \leq N_t$.

Proof. Consider the concept class C in Table 4.1. Note that the mapping $\psi_{P,0}$ maps all values on X_i, $i \in \{1,2,3\}$, to 1 and therefore is useless in finding the VCD_{Ψ_P} of C. Also, for any choice of k_1 and k_2 with $k_1, k_2 \in \{1,2\}$ and thus the tuple of mappings $\overline{\psi} = (\psi_{P,k_1}, \psi_{P,k_2}, \mathrm{id}_3)$, $\overline{\psi}(C) = C'$ where C' is the concept class in Table 4.3. As shown in Table 4.3, applying any mapping ψ_{P,k_3}, $k_3 \in \{1,2\}$, on X_3 results in a VCD-maximum class C^i of VCD 2. Since $|C| = \Phi_2(2,2,2)$, we conclude that C is $\mathrm{VCD}_{\Psi_\mathrm{P}}$-maximum of dimension 2. As shown in bold in Table 4.1, there are two different choices for $\overline{c}$. For example, for $\overline{c} = (0,0)$, $|\{c \in C \mid c - X_3 = \overline{c}\}| = |\{c_1, c_2\}| = 2$, and for $\overline{c} = (1,2)$, $|\{c \in C \mid c - X_3 = \overline{c}\}| = |\{c_{18}, c_{19}\}| = 2$. □

The class in Table 4.1 does not stay $\mathrm{VCD}_{\Psi_\mathrm{P}}$-maximum when applying either definition of reduction w.r.t. X_3.

Corollary 4.16. *There is a* $\mathrm{VCD}_{\Psi_\mathrm{P}}$*-maximum class* C *such that for some* $X_t \in X$*, neither* $[C]^{X_t}_{\geq 2}$ *nor* C^{X_t} *is* $\mathrm{VCD}_{\Psi_\mathrm{P}}$*-maximum.*

Proof. Consider the concept class C in Table 4.1. As shown in Table 4.2, $[C]^{X_3}_{\geq 2}$ is of $\mathrm{VCD}_{\Psi_\mathrm{P}}$ 2 with $\Phi_1(2,2) < |C^{X_3}| < \Phi_2(2,2)$, and C^{X_3} is of VCD_{Ψ_P} 1 with $|C^{X_3}| < \Phi_1(2,2)$. So, in either case, the reduction of C w.r.t. X_3 is not $\mathrm{VCD}_{\Psi_\mathrm{P}}$-maximum. □

c	X_1	X_2	X_3
$\mathbf{c_1}$	**0**	**0**	**0**
$\mathbf{c_2}$	**0**	**0**	**1**
c_3	0	1	0
c_4	0	1	1
c_5	1	0	0
c_6	1	0	1
c_7	1	1	1
c_8	0	2	0
c_9	0	2	1
c_{10}	0	2	2
c_{11}	2	0	0
c_{12}	2	0	1
c_{13}	2	0	2
c_{14}	2	1	1
c_{15}	2	1	2
c_{16}	2	2	1
c_{17}	2	2	2
$\mathbf{c_{18}}$	**1**	**2**	**1**
$\mathbf{c_{19}}$	**1**	**2**	**2**

Table 4.1: Maximum class C of $\mathrm{VCD}_{\Psi_{\mathrm{P}}}$ 2 used in the proof of Proposition 4.15.

$c \in [C]^{X_3}_{\geqslant 2}$	X_1	X_2
c_1	0	0
c_2	0	1
c_3	1	0
c_4	0	2
c_5	2	0
c_6	2	1
c_7	2	2
c_8	1	2

$c \in C^{X_3}$	X_1	X_2
c_1	0	2
c_2	2	0

Table 4.2: Both reductions of C where C is the $\mathrm{VCD}_{\Psi_{\mathrm{P}}}$-maximum class from Table 4.1.

$c \in C'$	$\psi_{P,k_1}(X_1)$	$\psi_{P,k_2}(X_2)$	X_3
$\mathbf{c_1}$	**0**	**0**	**0**
$\mathbf{c_2}$	**0**	**0**	**1**
c_3	0	1	0
c_4	0	1	1
c_5	0	1	2
c_6	1	0	0
c_7	1	0	1
c_8	1	0	2
$\mathbf{c_9}$	**1**	**1**	**1**
$\mathbf{c_{10}}$	**1**	**1**	**2**

$c \in C^2$	$\psi_{P,k_1}(X_1)$	$\psi_{P,k_2}(X_2)$	$\psi_{P,1}(X_3)$
c_1	0	0	0
c_2	0	0	1
c_3	0	1	0
c_4	0	1	1
c_5	1	0	0
c_6	1	0	1
c_7	1	1	1

$c \in C^1$	$\psi_{P,k_1}(X_1)$	$\psi_{P,k_2}(X_2)$	$\psi_{P,2}(X_3)$
c_1	0	0	0
c_2	0	1	0
c_3	0	1	1
c_4	1	0	0
c_5	1	0	1
c_6	1	1	0
c_7	1	1	1

Table 4.3: Mappings of the concept class C from Table 4.1.

4.3 The Natarajan Dimension

We provide the same result for the Natarajan-dimension as for Pollard's pseudo-dimension. That is, we give a counterexample to the reduction property for $\mathrm{VCD}_{\Psi_{\mathrm{N}}}$.

Proposition 4.17. *There is a* $\mathrm{VCD}_{\Psi_{\mathrm{N}}}$*-maximum class* C *with* $\mathrm{VCD}_{\Psi_{\mathrm{N}}}(C) = 1$ *such that, for some* $X_t \in X$ *and some* $\overline{c} \in C - X_t$, $|\{c \in C \mid c - X_t = \overline{c}\}| = 2 \leq N_t$.

Proof. Consider the concept class C in Table 4.4. Obviously, C cannot be of $\mathrm{VCD}_{\Psi_{\mathrm{N}}}$ 2 as there is no occurrence of the combinations $\{aa, ab, ba, bb\}$, for all $a, b \in \{0, 1, 2\}$, $a \neq b$. As shown in bold in Table 4.4, there are two choices for $\overline{c}$. For $\overline{c} = (0, 0)$, $|\{c \in C \mid c - X_3 = \overline{c}\}| = |\{c_1, c_2\}| = 2$, and for $\overline{c} = (2, 0)$, $|\{c \in C \mid c - X_3 = \overline{c}\}| = |\{c_9, c_{10}\}| = 2$. □

The reduction of the class in Table 4.4 is not $\mathrm{VCD}_{\Psi_{\mathrm{N}}}$-maximum under either definition of reduction.

$c \in C$	X_1	X_2	X_3
$\mathbf{c_1}$	**0**	**0**	**0**
$\mathbf{c_2}$	**0**	**0**	**1**
c_3	0	1	0
c_4	1	0	0
c_5	1	2	2
c_6	2	1	2
c_7	2	2	1
c_8	2	2	2
$\mathbf{c_9}$	**2**	**0**	**0**
$\mathbf{c_{10}}$	**2**	**0**	**2**

Table 4.4: Maximum class C of $\mathrm{VCD}_{\Psi_\mathrm{N}}$ 1 used in the proof of Proposition 4.17.

Corollary 4.18. *There is a* $\mathrm{VCD}_{\Psi_\mathrm{N}}$*-maximum class* C *such that for some* $X_t \in X$*, neither* $[C]^{X_t}_{\geq 2}$ *nor* C^{X_t} *is* $\mathrm{VCD}_{\Psi_\mathrm{N}}$*-maximum.*

Proof. Consider the $\mathrm{VCD}_{\Psi_\mathrm{N}}$-maximum class C in Table 4.4 and the Natarajan family of mappings Ψ_N. Clearly, C^{X_3} is the empty set and also as shown in Table 4.5, $[C]^{X_3}_{\geq 2}$ is of $\mathrm{VCD}_{\Psi_\mathrm{N}}$ 1 with $\sum_{i=0}^{0} \binom{3}{i}\binom{3}{2}^i < \mathrm{size}([C]^{X_3}_{\geq 2}) < \sum_{i=0}^{1} \binom{3}{i}\binom{3}{2}^i$. So, in either case, the reduction of C w.r.t. X_3 is not $\mathrm{VCD}_{\Psi_\mathrm{N}}$-maximum. □

$c \in [C]^{X_3}_{\geqslant 2}$	X_1	X_2
c_1	0	0
c_2	2	2
c_3	2	0

Table 4.5: Reduction of C where C is the $\mathrm{VCD}_{\Psi_\mathrm{N}}$-maximum class from Table 4.4.

Chapter 5

Sample Compression Schemes

This chapter discusses sample compression schemes for multi-label concept classes. In the case of binary concept classes, the main motivation for studying sample compression schemes is that the smallest possible size of a sample compression scheme yields sample bounds for PAC-learning [LW86]. In Section 5.1, we show that the same statement holds in the multi-label case, i.e., Littlestone and Warmuth's PAC bounds in the size of an SCS are immediately transferred to the multi-label case.

We henceforth focus on VCD_Ψ-maximum classes where VCD_Ψ fulfills the reduction property and Ψ is based on a spanning family of mappings. In Section 5.2, we extend Floyd and Warmuth's compression scheme for binary VCD-maximum classes [FW95] to VCD_Ψ-maximum classes in the multi-label case and continuing in Section 5.3, we introduce and study the notion of a *tight* compression scheme for VCD_Ψ-maximum classes which is in fact a generalization of Kuzmin and Warmuth's unlabeled compression scheme [KW07]. Section 5.4 explores a connection between the tight compression schemes and the one-inclusion hypergraph of VCD_Ψ-maximum classes.

Finally, in Section 5.5, we show that any $\mathrm{VCD}_{\Psi_\mathrm{G}}$ 1 concept class has also a sample compression scheme of size 1. Although a similar result exists for binary classes of

VCD 1, our approach here is not a straightforward translation of the proof in the binary case. In fact, as opposed to the binary case, VCD_{Ψ_G} 1 classes may not be contained in VCD_{Ψ_G}-maximum classes of dimension 1, which is the crucial property used for showing the corresponding result in the binary case.

The notion of sample compression can be easily generalized to the multi-label case:

Definition 5.1 (**sample compression scheme**). [LW86] *A sample compression scheme for C is a pair (f, g) of mappings as follows. Given any C-realizable sample S, one requires (i) $f(S) \subseteq S$, and (ii) $g(f(S)) = (l_1, \ldots, l_m)$, where $(X_i, \ell_i) \in S$ implies $\ell_i = l_i$, for all $i \in [m]$. The size of (f, g) is the maximum cardinality of a set $f(S)$, taken over all C-realizable samples.*

Throughout this chapter we assume that $\Psi = \Psi_1 \times \cdots \times \Psi_m$ and $C \subseteq \prod_{1 \leq i \leq m} X_i$, where each Ψ_i is a spanning family of mappings on X_i, for all $i \in [m]$, unless stated otherwise.

5.1 PAC-learnability of Multi-label Classes

Valiant introduced the probably approximately correct (PAC for short) learning model for binary concept classes [Val84]. Let $X_i = \{0, \ldots, N_i\}$ for all $i \geq 1$, X be the set of instances, and as in the binary case, the goal of a PAC-learning algorithm in the multi-label case is to learn a good approximation of any target concept in the concept class with a high probability.

The learning algorithm is thus given ϵ (accuracy parameter), δ (confidence parameter) and a set of examples (sample) that are labeled consistently with some concept c^* (target concept) in C. The number of examples is called the *sample size* of the

algorithm. We make an assumption that there is a probability distribution $\mathcal{D}$ over the instance space X and the instances of the input examples for the algorithm are randomly drawn i.i.d. from X w.r.t. $\mathcal{D}$. The learning algorithm then makes a prediction about the target concept c^* by returning the hypothesis h. The error of the hypothesis h is defined as the total probability w.r.t. $\mathcal{D}$ of the instances for which h and c^* do not agree. That is,

$$\mathrm{err}_{\mathcal{D}}(c^*, h) = \mathcal{D}\{X_i \in X \mid c^*(X_i) \neq h(X_i)\}.$$

We call h ϵ-accurate if the error of h is at most ϵ.

A concept class C is PAC-learnable if there exists a learning algorithm L and a polynomial $p(\frac{1}{\epsilon}, \frac{1}{\delta})$ such that, for any $\epsilon, \delta > 0$, any concept $c^* \in C$ and any probability distribution $\mathcal{D}$ on X, if L draws $X_1, \ldots, X_k$, i.i.d. from X w.r.t. $\mathcal{D}$ for $k \geq p(\frac{1}{\epsilon}, \frac{1}{\delta})$, and L is given $\{(X_1, c^*(X_1)), \ldots, (X_k, c^*(X_k))\}$ then, with probability at least $1 - \delta$, L returns an ϵ-accurate hypothesis. That is,

$$Pr_{\mathcal{D}}(\mathrm{err}_{\mathcal{D}}(c^*, h) \leq \epsilon) > 1 - \delta.$$

For binary classes, the smallest possible *size of a sample compression scheme* yields sample bounds for PAC-learning [LW86, FW95] and an open question is whether this parameter is linear in the VC-dimension. The proof that (in the binary case) a sample compression scheme yields a successful PAC-learner with bounds expressed in terms of its size [LW86] can be immediately generalized to the multi-label case.

The following lemma is a straightforward extension of Littlestone and Warmuth's theorem. We just rephrased their proof [LW86] in our notation.

Lemma 5.2. *Let $\mathcal{D}$ be any probability distribution over X, $d \in [m]$, and g be any*

function mapping sets of at most d labeled examples consistent with some concept in C to hypotheses in $\prod_{i=1}^{m} X_i$. *Then the probability that* $k \geq d$ *labeled examples that are randomly drawn i.i.d. according to* $\mathcal{D}$ *contain a subset of at most d examples that map via g to a hypothesis that is both consistent with all k examples and has error greater than* ϵ *is at most* $\sum_{i=1}^{d} \binom{k}{i}(1-\epsilon)^{k-i}$.

Proof. Let $c \in C$ and let S be a set of k distinct instances that are randomly drawn i.i.d. according to $\mathcal{D}$ and labeled consistently w.r.t. c. We call any set $S_{\leq d} \subseteq S$ of size at most d a compression set for S if $g(S_{\leq d})$ is consistent with S. The lemma states that the probability of having a sample set S that contains a compression set $S_{\leq d}$, such that $g(S_{\leq d})$ has error greater than ϵ is at most $\sum_{i=1}^{d} \binom{k}{i}(1-\epsilon)^{k-i}$.

Choose $t \leq d$. Let $S_t = \{(X_{i_1}, c(X_{i_1})), \ldots, (X_{i_t}, c(X_{i_t}))\} \subseteq S$ and $h = g(S_t)$ such that S_t is a compression set for S and $\mathrm{err}_{\mathcal{D}}(c, h) > \epsilon$. Obviously, by having S_t, we can immediately produce a hypothesis $g(S_t)$ and find the error of this hypothesis.

Consider all sample sequences S' with $X(S') = X(S)$ and $|S'| = k$ that contain all labeled examples of S. Since the examples of S' are drawn independently from the distribution $\mathcal{D}$, we can assume that the t instances occurring in $X(S_t)$ are drawn first and $h = g(S_t)$ can be calculated at the time. Next, the $k - t$ instances of $X(S') \setminus X(S_t)$ are drawn. Since S_t is a compression set for S, h is consistent with S and consequently with the remaining $k - t$ examples of S. Moreover, $\mathrm{err}_{\mathcal{D}}(c, h) > \epsilon$ by assumption. That is, $\mathcal{D}\{X_i \in X \mid c(X_i) \neq h(X_i)\} > \epsilon$ and consequently,

$$\mathcal{D}\{X_i \in X \mid c(X_i) = h(X_i)\} \leq 1 - \epsilon. \tag{5.1}$$

So, from (5.1), the probability to draw a single instance X_i for which both h and c have the same label is at most $1 - \epsilon$. Consequently, the probability that $k - t$ instances drawn from X receive the same label from h and c is at most $(1-\epsilon)^{k-t}$, as

the examples are drawn i.i.d. with respect to $\mathcal{D}$.

For the set of k examples S, there are at most $\binom{k}{t}$ compression sets S_t of size t. So, the probability that k labeled examples that are randomly drawn i.i.d. with respect to $\mathcal{D}$, contain a compression set S_t such that $g(S_t)$ has error greater than ϵ is at most $\binom{k}{t}(1-\epsilon)^{k-t}$. t was chosen arbitrarily and can have any value from 0 to d. Hence, the probability that $k \geq d$ labeled examples that are randomly drawn i.i.d. with respect to $\mathcal{D}$, contain a subset of at most d examples that map via g to a hypothesis that is both consistent with all k examples and has error greater than ϵ is at most $\sum_{i=1}^{d}\binom{k}{i}(1-\epsilon)^{k-i}$. □

Lemma 5.3. [LW86] *For $0 \leq \epsilon, \delta \leq 1$ and any $0 < \beta < 1$, if $k \geq \frac{1}{1-\beta}(\frac{1}{\epsilon}\ln\frac{1}{\delta} + d + \frac{d}{\epsilon}\ln\frac{1}{\beta\epsilon})$, then $\sum_{i=1}^{d}\binom{k}{i}(1-\epsilon)^{k-i} \leq \delta$.*

Littlestone and Warmuth [LW86] proved the following theorem for any $C \subseteq 2^X$. It turned out that the theorem is correct for multi-label concept classes as well.

Theorem 5.4. *Let C be a multi-label concept class with a sample compression scheme of size at most d. Then for $0 < \epsilon, \delta < 1$ and any $0 < \beta < 1$, the learning algorithm using this sample compression scheme PAC-learns C with sample size*

$$k \geq \frac{1}{1-\beta}(\frac{1}{\epsilon}\ln\frac{1}{\delta} + d + \frac{d}{\epsilon}\ln\frac{1}{\beta\epsilon}).$$

Proof. Follows immediately from Lemma 5.2 and Lemma 5.3. □

Therefore, the existence of a sample compression scheme of size VC-dimension of the class yields a PAC-learning algorithm for the class that requires at most $p(\frac{1}{\epsilon}, \frac{1}{\delta})$ examples, where $p(\frac{1}{\epsilon}, \frac{1}{\delta})$ is the polynomial in Theorem 5.4.

5.2 Generalizing Floyd and Warmuth's Compression Scheme

In this section, we show that every VCD_Ψ-maximum class of dimension d has a sample compression scheme of size d. We closely follow the technique that Floyd and Warmuth [FW95] used to show that any binary VCD-maximum class has a sample compression scheme of the size of its VC-dimension.

The results presented in this section have been published in the Proceedings of the 27th Annual Conference on Learning Theory [SSYZ14b].

The objective of this section is to prove the following theorem:

Theorem 5.5. *Let Ψ_i, $1 \leq i \leq m$, be a spanning family of mappings. Let $\Psi = \Psi_1 \times \cdots \times \Psi_m$. If VCD_Ψ fulfills the reduction property then any VCD_Ψ-maximum class C has a labeled sample compression scheme of size $\mathrm{VCD}_\Psi(C)$.*

The scheme and also parts of the proof follow the work on the so-called VC Compression Scheme for binary maximum classes, as introduced by Floyd and Warmuth [FW95]. However, there are some technical difficulties that need to be overcome in order to adapt Floyd and Warmuth's technique. For the rest of this section, let $C \subseteq \prod_{1 \leq i \leq m} X_i$ be a VCD_Ψ-maximum class of dimension d, where VCD_Ψ has the reduction property and Ψ is the direct product of spanning families of mappings.

We use the following example to illustrate our statements and arguments. Recall that as shown in Chapter 4, $\mathrm{VCD}_{\Psi_\mathrm{G}}$ fulfills the reduction property.

Example 5.6. *Consider the concept class $C \subseteq \{0,1,2\}^4$ in Table 5.1 and the compression sets for the concepts in C. One can verify that $\mathrm{VCD}_{\Psi_\mathrm{G}}(C) = 2$ and $|C| = \Phi_d(2,2,2,2)$, so C is $\mathrm{VCD}_{\Psi_\mathrm{G}}$-maximum of $\mathrm{VCD}_{\Psi_\mathrm{G}}$ 2.*

c	X_1	X_2	X_3	X_4	compression sets
c_1	0	0	0	0	$\{(X_1,0),(X_2,0)\},\{(X_1,0),(X_3,0)\},\{(X_2,0),(X_3,0)\}$
c_2	0	0	1	0	$\{(X_1,0),(X_3,1)\},\{(X_1,0),(X_4,0)\},\{(X_2,0),(X_3,1)\},$ $\{(X_2,0),(X_4,0)\}$
c_3	0	0	2	0	$\{(X_1,0),(X_3,2)\},\{(X_2,0),(X_3,2)\}$
c_4	0	0	1	1	$\{(X_1,0),(X_4,1)\},\{(X_2,0),(X_4,1)\}$
c_5	0	0	1	2	$\{(X_1,0),(X_4,2)\},\{(X_2,0),(X_4,2)\}$
c_6	0	1	0	0	$\{(X_1,0),(X_2,1)\},\{(X_2,1),(X_3,0)\},\{(X_3,0),(X_4,0)\}$
c_7	0	2	0	0	$\{(X_1,0),(X_2,2)\},\{(X_2,2),(X_3,0)\}$
c_8	1	0	0	0	$\{(X_1,1),(X_2,0)\},\{(X_1,1),(X_3,0)\}$
c_9	2	0	0	0	$\{(X_1,2),(X_2,0)\},\{(X_1,2),(X_3,0)\}$
c_{10}	1	0	1	0	$\{(X_1,1),(X_3,1)\},\{(X_1,1),(X_4,0)\}$
c_{11}	1	0	2	0	$\{(X_1,1),(X_3,2)\}$
c_{12}	2	0	1	0	$\{(X_1,2),(X_3,1)\},\{(X_1,2),(X_4,0)\}$
c_{13}	2	0	2	0	$\{(X_1,2),(X_3,2)\}$
c_{14}	1	0	1	1	$\{(X_1,1),(X_4,1)\}$
c_{15}	2	0	1	1	$\{(X_1,2),(X_4,1)\}$
c_{16}	1	0	1	2	$\{(X_1,1),(X_4,2)\}$
c_{17}	2	0	1	2	$\{(X_1,2),(X_4,2)\}$
c_{18}	1	1	0	0	$\{(X_1,1),(X_2,1)\}$
c_{19}	1	2	0	0	$\{(X_1,1),(X_2,2)\}$
c_{20}	2	1	0	0	$\{(X_1,2),(X_2,1)\}$
c_{21}	2	2	0	0	$\{(X_1,2),(X_2,2)\}$
c_{22}	0	1	0	1	$\{(X_3,0),(X_4,1)\}$
c_{23}	0	1	0	2	$\{(X_3,0),(X_4,2)\}$
c_{24}	0	1	1	0	$\{(X_2,1),(X_3,1)\},\{(X_2,1),(X_4,0)\},\{(X_3,1),(X_4,0)\}$
c_{25}	0	1	2	0	$\{(X_2,1),(X_3,2)\},\{(X_3,2),(X_4,0)\}$
c_{26}	0	2	1	0	$\{(X_2,2),(X_3,1)\},\{(X_2,2),(X_4,0)\}$
c_{27}	0	2	2	0	$\{(X_2,2),(X_3,2)\}$
c_{28}	0	1	1	1	$\{(X_2,1),(X_4,1)\},\{(X_3,1),(X_4,1)\}$
c_{29}	0	2	1	1	$\{(X_2,2),(X_4,1)\}$
c_{30}	0	1	2	1	$\{(X_3,2),(X_4,1)\}$
c_{31}	0	1	1	2	$\{(X_2,1),(X_4,2)\},\{(X_3,1),(X_4,2)\}$
c_{32}	0	2	1	2	$\{(X_2,2),(X_4,2)\}$
c_{33}	0	1	2	2	$\{(X_3,2),(X_4,2)\}$

Table 5.1: VCD_{Ψ_G}-maximum class C and the extension of Floyd and Warmuth's compression scheme.

Proposition 5.7. *Let C be a VCD_Ψ-maximum class with $\mathrm{VCD}_\Psi(C) = d < m$ and let $Y \subseteq \{X_1, \ldots, X_m\}$ with $|Y| = d$. Then $\mathrm{VCD}_\Psi(C^Y) = 0$ and C^Y consists of a single concept.*

Proof. The proof is the same as that in the binary case [FW95]. Let $Y = \{X_{i_1}, \ldots, X_{i_d}\}$. By applying Theorem 4.2 to $C^Y = ((C^{X_{i_1}}) \cdots)^{X_{i_d}}$ repeatedly, C^Y is a VCD_Ψ-maximum

class of dimension 0. So, $|C^Y| = 1$. □

Example 5.8. *Consider the concept class C in Table 5.1 and let $Y = \{X_1, X_2\}$. As shown in Table 5.2, C^{X_1} is VCD_{Ψ_G}-maximum of VCD_{Ψ_G} 1, and $(C^{X_1})^{X_2}$ contains the single concept $c_1'' = \{(X_3, 0)\}$.*

$c \in C^{X_1}$	X_2	X_3	X_4
c_1'	0	0	0
c_2'	1	0	0
c_3'	2	0	0
c_4'	0	1	0
c_5'	0	2	0
c_6'	0	1	1
c_7'	0	1	2

$c \in (C^{X_1})^{X_2}$	X_3	X_4
c_1''	0	0

Table 5.2: C^{X_1} and $(C^{X_1})^{X_2}$ where C is the VCD_{Ψ_G}-maximum class from Table 5.1.

For any VCD_Ψ-maximum class C with $\mathrm{VCD}_\Psi(C) = d < m$ and any subset $Y \subseteq X$ with $|Y| = d$, we denote by $c_{Y,C}$ the single concept in C^Y. For $Y = \{X_{i_1}, \ldots, X_{i_d}\}$, the concept $c_{Y,C} \in C^Y$ can be extended in $\prod_{j=1}^{d}(N_{i_j} + 1)$ ways to concepts in C, that is, $c_{Y,C} \times \prod_{j=1}^{d} X_{i_j} \subseteq C$. In particular, for any tuple $(n_{i_1}, \ldots, n_{i_d}) \in \prod_{j=1}^{d} X_{i_j}$, $c_{Y,C} \cup \{(X_{i_1}, n_{i_1}), \ldots, (X_{i_d}, n_{i_d})\} \in C$. Thus, any set $S = \{(X_{i_1}, n_{i_1}), \ldots, (X_{i_d}, n_{i_d})\}$ with $X(S) = Y$ corresponds to the unique concept $c_{Y,C} \cup S = c_{X(S),C} \cup S$ in C.

Example 5.9. *Consider the concept class C in Table 5.1 and let $Y = \{X_1, X_2\}$. Then, as shown in Table 5.2, $c_{Y,C} = c_1'' = \{(X_3, 0), (X_4, 0)\}$. One can easily verify that $\{(X_3, 0), (X_4, 0)\} \times X_2 = \{c_1', c_2', c_3'\} \subseteq C^{X_1}$ and $\{(X_3, 0), (X_4, 0)\} \times X_1 \times X_2 \subseteq C$. Also, the samples $S = \{(X_1, 0), (X_2, 1)\}$ and $S' = \{(X_1, 2), (X_2, 2)\}$ correspond to the concepts $c_6 = c_{Y,C} \cup \{(X_1, 0), (X_2, 1)\}$ and $c_{21} = c_{Y,C} \cup \{(X_1, 2), (X_2, 2)\}$ in C, respectively.*

Definition 5.10. *Let C be a VCD_Ψ-maximum class with $\mathrm{VCD}_\Psi(C) = d < m$. Let S with $|S| = d$ be a C-realizable sample and $c_{X(S),C}$ be the single concept in $C^{X(S)}$. S*

is called a compression set *for the concept* $c_{S,C} \in C$ *where* $c_{S,C} = (c_{X(S),C}) \cup S$. *The concept* $c_{S,C}$ *is called the* decompression set *for the sample* S *in the class* C.

Example 5.11. *Consider the concept class* C *in Table 5.1. As discussed in Example 5.9,* $S = \{(X_1, 0), (X_2, 1)\}$ *is a compression set for* $c_6 \in C$ *and also* $S' = \{(X_1, 2), (X_2, 2)\}$ *is a compression set for* $c_{21} \in C$. *So,* $c_{S,C} = c_6$ *and* $c_{S',C} = c_{21}$.

The following lemma is useful in the proof of the two upcoming lemmas.

Lemma 5.12. *Let* C *be a* VCD_Ψ*-maximum class with* $\mathrm{VCD}_\Psi(C) = d < m$, *and* S *be a* C*-realizable sample with* $|X(S)| = d$ *and* $X(S) = \{X_{i_1}, \ldots, X_{i_d}\}$. *Let* $X_t \in X \setminus X(S)$ *and* $c_{S,C}(X_t) = p$, *for some* $p \in X_t$. *Let* $\overline{\psi} = (\psi_{i_1}, \ldots, \psi_{i_d})$ *be a tuple of non-constant mappings* $\psi_{i_j} \in \Psi_{i_j}$, *for all* $j \in \{1, \ldots, d\}$, $\psi_t : X_t \to \{0, 1\}$ *with* $\psi_t(p) = l$, *and* $\overline{\psi'} = (\psi_{i_1}, \ldots, \psi_{i_d}, \psi_t)$. *Then* $\{\{0,1\}^d \times \{l\}\} \subseteq \overline{\psi'}(C|_{\{X_{i_1}, \ldots, X_{i_d}, X_t\}})$.

Proof. W.l.o.g., assume that $S = \{(X_1, l_1), \ldots, (X_d, l_d)\}$ and $t = d + 1$. From Theorem 3.14, $C|_{\{X_1, \ldots, X_d\}}$ is VCD_Ψ-maximum of dimension d and by Corollary 4.7, $\overline{\psi}(C|_{\{X_1, \ldots, X_d\}}) = \{0, 1\}^d$.

Since $c_{S,C}(X_{d+1}) = p$, for each labeling $((X_1, n_1), \ldots, (X_d, n_d))$ of $X(S)$, there is a concept $c \in C$ that is consistent with that labeling and fulfills $c(X_{d+1}) = p$. That is, for each $(n_1, \ldots, n_d) \in C|_{\{X_1, \ldots, X_d\}}$, there is a concept $c \in C$, such that $c|_{\{X_1, \ldots, X_d\}} = (n_1, \ldots, n_d)$ and $c(X_{d+1}) = p$. Consequently, for each tuple $(\psi_1(n_1), \ldots, \psi_d(n_d)) \in \overline{\psi}(C|_{\{X_1, \ldots, X_d\}}) = \{0, 1\}^d$, there is a concept $c \in C$, such that $\overline{\psi}(c|_{\{X_1, \ldots, X_d\}}) = (\psi_1(n_1), \ldots, \psi_d(n_d))$ and $c(X_{d+1}) = p$. Therefore, $\{\{0,1\}^d \times \{l\}\} \subseteq \overline{\psi'}(C|_{\{X_1, \ldots, X_{d+1}\}})$. □

In order to have a compression scheme of size d for a class C, any C-realizable sample of size at least d should have a compression set of size at most d. In other words, we need to show that any concept in $C|_Y$ has a compression set of size at most

d, where $Y \subseteq X$ with $|Y| > d$. Since C is VCD_Ψ-maximum, by Theorem 3.14, $C|_Y$ is VCD_Ψ-maximum and Definition 5.10 applies to $C|_Y$, too.

To prove that each concept in a VCD_Ψ-maximum class can be compressed to a subset of d examples, we need two lemmas. Although we have to deal with label mappings here, the proof ideas are similar to those in [FW95]. We first show that any sample S of size d over Y yields the same set when considering the concept class C and restricting the compression set corresponding to S to the domain Y, as when considering the concept class $C|_Y$ and taking the compression set corresponding to S. The proof is a translation of that in the binary case [FW95].

Lemma 5.13. *Let C be a VCD_Ψ-maximum class with $\mathrm{VCD}_\Psi(C) = d < m$. Let S be a C-realizable with $X(S) \subseteq Y \subseteq X$, and $|X(S)| = d$. Then $(c_{S,C})|_Y = c_{S,C|_Y}$.*

Proof. W.l.o.g., assume that $X(S) = \{X_1, \ldots, X_d\}$. Clearly, $c_{S,C}$ and $c_{S,C|_Y}$ agree on $X(S)$. Assume that $c_{S,C}$ and $c_{S,C|_Y}$ differ on some $X_t \in Y \setminus X(S)$. W.l.o.g., let $c_{S,C}(X_{d+1}) = 0$ and $c_{S,C|_Y}(X_{d+1}) = 1$. We show that then $\{X_1, \ldots, X_{d+1}\}$ is shattered by C, in contradiction to $\mathrm{VCD}_\Psi(C) = d$.

Let $\overline{\psi^{1,d}} = (\psi_1, \ldots, \psi_d)$ be a tuple of non-constant mappings $\psi_i \in \Psi_i$. From Theorem 3.14, $C|_{\{X_1,\ldots,X_d\}}$ is VCD_Ψ-maximum of dimension d and by Corollary 4.7, $\overline{\psi^{1,d}}(C|_{\{X_1,\ldots,X_d\}}) = \{0,1\}^d$. Let $\psi_{d+1} : X_{d+1} \to \{0,1\}$ be $\psi_{0\neq 1}$ as discussed in Remark 3.8, i.e.,

$$\psi_{d+1}(x) = \begin{cases} 0 & \text{if } x = 0 \\ 1 & \text{if } x = 1 \\ 0 \text{ or } 1 & \text{otherwise,} \end{cases}$$

and $\overline{\psi^{1,d+1}} = (\psi_1, \ldots, \psi_d, \psi_{d+1})$. Lemma 5.12 yields

$$\{\{0,1\}^d \times \{0\}\} \subseteq \overline{\psi^{1,d+1}}(C|_{\{X_1,\ldots,X_{d+1}\}}).$$

Moreover, $c_{S,C|_Y}(X_{d+1}) = 1$ implies that for each labeling $((X_1, n_1), \ldots, (X_d, n_d))$ of $X(S)$, there is a concept $c \in C|_Y$, that is consistent with that labeling and fulfills $c(X_{d+1}) = 1$. That is, for each $(n_1, \ldots, n_d) \in C|_{\{X_1,\ldots,X_d\}}$, there is a concept $c \in C|_Y$, such that $c|_{\{X_1,\ldots,X_d\}} = (n_1, \ldots, n_d)$ and $c(X_{d+1}) = 1$. So, for each tuple $(\psi_1(n_1), \ldots, \psi_d(n_d)) \in \overline{\psi^{1,d}}(C|_{\{X_1,\ldots,X_d\}}) = \{0,1\}^d$, there is a concept $c \in C|_Y$, such that $\overline{\psi^{1,d}}(c|_{\{X_1,\ldots,X_d\}}) = (\psi_1(n_1), \ldots, \psi_d(n_d))$ and $c(X_{d+1}) = 1$. Thus, $\{\{0,1\}^d \times \{1\}\} \subseteq \overline{\psi^{1,d+1}}(C|_{\{X_1,\ldots,X_{d+1}\}})$.

Hence, $\overline{\psi^{1,d+1}}(C|_{\{X_1,\ldots,X_{d+1}\}}) = \{0,1\}^{d+1}$ and C shatters a set of $d+1$ instances. □

Example 5.14. *Consider the concept class C in Table 5.1. Assume that $Y = \{X_1, X_2, X_3\}$ and $S = \{(X_1, 0), (X_2, 1)\}$. As discussed in Example 5.11, $c_{S,C} = c_6$, so, $c_{S,C}|_Y = c_6|_{\{X_1,X_2,X_3\}} = \{(X_1, 0), (X_2, 1), (X_3, 0)\}$. Now, we need to verify that S corresponds to $\{(X_1, 0), (X_2, 1), (X_3, 0)\}$ in $C|_{\{X_1,X_2,X_3\}}$. One can see in Table 5.3, that $c_{X(S),C|_Y} = c_1'' = \{(X_3, 0)\}$. So, $c_{S,C|_Y} = c_{X(S),C|_Y} \cup S = \{(X_3, 0)\} \cup \{(X_1, 0), (X_2, 1)\} = \{(X_1, 0), (X_2, 1), (X_3, 0)\}$.*

Next, one needs to establish that, for any sample S of size $d-1$ and any instance X_t not occurring in S, the decompression set for the sample S in the class C^{X_t} equals the restriction of the decompression set for the sample $S \cup \{(X_t, i)\}$ in the class C, to $X \setminus X_t$. This statement is not easy to see here, as opposed to the binary case. We thus need to prove it in a separate lemma.

Lemma 5.15. *Let C be a VCD_Ψ-maximum class with $\mathrm{VCD}_\Psi(C) = d < m$. Let $t \in [m]$, $c \in C^{X_t}$, S be a sample consistent with c, such that $|X(S)| = d - 1$ and $S_i = S \cup \{(X_t, i)\}$, for all $i \in X_t$. Then $c_{S_i,C} - X_t = c_{S,C^{X_t}}$.*

$c \in C\|_Y$	X_1	X_2	X_3
c_1	0	0	0
c_2	0	0	1
c_3	0	0	2
c_4	0	1	0
c_5	0	2	0
c_6	1	0	0
c_7	2	0	0
c_8	1	0	1
c_9	1	0	2
c_{10}	2	0	1
c_{11}	2	0	2
c_{12}	1	1	0
c_{13}	1	2	0
c_{14}	2	1	0
c_{15}	2	2	0
c_{16}	0	1	1
c_{17}	0	1	2
c_{18}	0	2	1
c_{19}	0	2	2

$c \in (C\|_Y)^{X_1}$	X_2	X_3
c'_1	0	0
c'_2	0	1
c'_3	0	2
c'_4	1	0
c'_5	2	0

$c \in ((C\|_Y)^{X_1})^{X_2}$	X_3
c''_1	0

Table 5.3: $C|_Y$, $(C|_Y)^{X_1}$ and $((C|_Y)^{X_1})^{X_2}$ where $Y = \{X_1, X_2, X_3\}$ and C is the VCD_{Ψ_G}-maximum class from Table 5.1.

Proof. W.l.o.g, let $t = d$ and $X(S) = \{X_1, \ldots, X_{d-1}\}$. Since S is consistent with $c \in C^{X_d}$, the reduction property implies that S_i is consistent with some concept in C, for all $i \in \{0, \ldots, N_d\}$. Clearly, $c_{S_i,C}$ and $c_{S,C^{X_d}}$ agree on $X(S)$. Assume that $c_{S_i,C}$ and $c_{S,C^{X_d}}$ differ on some $X_j \in X \setminus \{X_1, \ldots, X_d\}$. W.l.o.g., let $c_{S_i,C}(X_{d+1}) = 0$ and $c_{S,C^{X_d}}(X_{d+1}) = 1$.

We show that $\{X_1, \ldots, X_{d+1}\}$ is shattered by C, which contradicts the fact that $\mathrm{VCD}_\Psi(C) = d$. Let $\overline{\psi^{1,d+1}} = (\psi_1, \ldots \psi_{d+1})$ be a tuple of non-constant mappings $\psi_i \in \Psi_i$, where $\psi_{d+1} : X_{d+1} \to \{0, 1\}$ is $\psi_{0\neq 1}$, as discussed in Remark 3.8. From Theorem 3.14, we obtain that $C|_{\{X_1,\ldots,X_d\}}$ is VCD_Ψ-maximum of dimension d and by Corollary 4.7, $\overline{\psi^{1,d}}(C|_{\{X_1,\ldots,X_d\}}) = \{0, 1\}^d$.

On the one hand, $\{\{0, 1\}^d \times \{0\}\} \subseteq \overline{\psi^{1,d+1}}(C|_{\{X_1,\ldots,X_{d+1}\}})$, from Lemma 5.12.

On the other hand, because $c_{S,C^{X_d}}(X_{d+1}) = 1$, we conclude that for each labeling $((X_1, n_1), \ldots, (X_{d-1}, n_{d-1}))$ of $X(S)$, there is a concept $c \in C^{X_d}$ that is consistent

with that labeling and fulfills $c(X_{d+1}) = 1$. That is, for each tuple $(n_1, \ldots, n_{d-1}) \in C|_{\{X_1,\ldots,X_{d-1}\}}$, there is a concept $c \in C^{X_d}$, such that $c|_{\{X_1,\ldots,X_{d-1}\}} = (n_1, \ldots, n_{d-1})$ and $c(X_{d+1}) = 1$. Also, by Definition 4.1, for each $c \in C^{X_d}$, $c \cup \{(X_d, i)\} \in C$, for all $0 \leq i \leq N_d$. Consequently, for each tuple $(\psi_1(n_1), \ldots, \psi_{d-1}(n_{d-1})) \in \overline{\psi^{1,d-1}}(C|_{\{X_1,\ldots,X_{d-1}\}}) = \{0,1\}^{d-1}$, there is a $c \in C^{X_d}$, such that $\overline{\psi^{1,d-1}}(c|_{\{X_1,\ldots,X_{d-1}\}}) = (\psi_1(n_1), \ldots, \psi_{d-1}(n_{d-1}))$, $c \cup \{(X_d, 0)\} \in C$, $c \cup \{(X_d, 1)\} \in C$, and $c(X_{d+1}) = 1$. So, $\{\{0,1\}^{d-1} \times \{0,1\} \times \{1\}\} = \{\{0,1\}^d \times \{1\}\} \subseteq \overline{\psi^{1,d+1}}(C|_{\{X_1,\ldots,X_{d+1}\}})$.

Hence, $\overline{\psi^{1,d+1}}(C|_{\{X_1,\ldots,X_{d+1}\}}) = \{0,1\}^{d+1}$ and C shatters a set of $d+1$ instances. □

Example 5.16. *Consider the concept class C in Table 5.1, and let $S = \{(X_1, 1)\}$ and $t = 2$. Then $S_0 = \{(X_1, 1)\} \cup \{(X_2, 0)\}$, $S_1 = \{(X_1, 1)\} \cup \{(X_2, 1)\}$ and $S_2 = \{(X_1, 1)\} \cup \{(X_2, 2)\}$. Note that $X(S_0) = X(S_1) = X(S_2) = \{X_1, X_2\}$ and $c_{S_i,C} = c_{X(S_i),C} \cup S_i$, for all $i \in \{1, 2, 3\}$. As shown in Table 5.2 and discussed in Example 5.9, $c_{\{X_1,X_2\},C} = \{(X_3, 0), (X_4, 0)\}$, so*

$$c_{S_0,C} = \{(X_3, 0), (X_4, 0)\} \cup \{(X_1, 1), (X_2, 0)\} = c_8$$
$$c_{S_1,C} = \{(X_3, 0), (X_4, 0)\} \cup \{(X_1, 1), (X_2, 1)\} = c_{18}$$
$$c_{S_2,C} = \{(X_3, 0), (X_4, 0)\} \cup \{(X_1, 1), (X_2, 2)\} = c_{19}$$
and
$$c_{S_0,C} - X_2 = c_8 - X_2 = \{(X_1, 1), (X_3, 0), (X_4, 0)\}$$
$$c_{S_1,C} - X_2 = c_{18} - X_2 = \{(X_1, 1), (X_3, 0), (X_4, 0)\}$$
$$c_{S_2,C} - X_2 = c_{19} - X_2 = \{(X_1, 1), (X_3, 0), (X_4, 0)\}.$$

On the other hand, as shown in Table 5.4, $c_{X(S),C^{X_2}} = c_{\{X_1\},C^{X_2}} = \{(X_3, 0), (X_4, 0)\}$. So,

$$\begin{aligned} c_{S,C^{X_2}} &= c_{X(S),C^{X_2}} \cup S \\ &= \{(X_3, 0), (X_4, 0)\} \cup \{(X_1, 1)\} = \{(X_1, 1), (X_3, 0), (X_4, 0)\}. \end{aligned}$$

Now, we are ready to show that for each concept in a VCD_Ψ-maximum class, there exists a compression set whose size is equal to the VCD_Ψ-dimension of the class.

$c \in C^{X_2}$	X_1	X_3	X_4
c'_1	0	0	0
c'_2	0	1	0
c'_3	0	2	0
c'_4	1	0	0
c'_5	2	0	0
c'_6	0	1	1
c'_7	0	1	2

$c \in (C^{X_2})^{X_1}$	X_3	X_4
c''_1	0	0

Table 5.4: C^{X_2} and $(C^{X_2})^{X_1}$ where C is the VCD_{Ψ_G}-maximum class from Table 5.1.

Theorem 5.17. *Let C be a VCD_Ψ-maximum class with $\mathrm{VCD}_\Psi(C) = d$. Then for each concept $c \in C$, there is a compression set S of exactly d examples such that $c = c_{S,C}$.*

Proof. The proof is a straightforward translation of that in the binary case [FW95] and is by double induction on m and d.

If $d = m$, then each concept has exactly d examples and is a compression set for itself.

For any $m \geq 1$, if $d = 0$, the empty set compresses the single concept in C.

For the induction step, assume that the theorem holds for all $d' \leq d$ and $m' < m$. If $m = d$, we know that the theorem holds. So we suppose that $m > d$. Let $c \in C - X_m$. To show that all extensions of c to concepts in C have a compression set as claimed, we need to consider two possible cases.

Case 1: c has a unique extension to a concept in C (and is thus not contained in C^{X_m}.)

W.l.o.g., let $c \cup \{(X_m, 0)\} \in C$, and for all $i \in \{1, \ldots, N_m\}$, $c \cup \{(X_m, i)\} \notin C$.

By Theorem 3.14, $C - X_m$ is VCD_Ψ-maximum of dimension d. So, by induction hypothesis, for each $c \in C - X_m$ there is a compression set S, such that $c = c_{S,C-X_m}$. By Proposition 5.7, S also represents the concept $c_{S,C} = c_{X(S),C} \cup S$

because $c_{X(S),C}$ is the single concept in $C^{X(S)}$. We show that S is a compression set for $c \cup \{(X_m, 0)\}$, too. From Lemma 5.13, $c_{S,C} - X_m = c_{S,C-X_m}$, i.e, $c_{S,C} - X_m = c$. If $c_{S,C}(X_m) = i$, for some $1 \leq i \leq N_m$, then $c \cup \{(X_m, i)\} \in C$ which contradicts the condition of Case 1. Hence, $c_{S,C}(X_m) = 0$, and consequently S is a compression set for $c_{S,C} = c \cup \{(X_m, 0)\}$.

Case 2: c has all $N_m + 1$ extensions to concepts in C. Clearly, $c \in C^{X_m}$.

By Theorem 4.2, C^{X_m} is VCD_Ψ-maximum of dimension $d-1$. So, by induction hypothesis, for each $c \in C^{X_m}$ there is a compression set S of $d-1$ examples, such that $c = c_{S,C^{X_m}}$. Let $S_i = S \cup \{(X_m, i)\}$, for all $0 \leq i \leq N_t$. By Proposition 5.7, S_i represents the concept $c_{S_i,C} = c_{X(S_i),C} \cup S_i$ because $c_{X(S_i),C}$ is the single concept in $C^{X(S_i)}$.

We show that S_i is a compression set for $c \cup \{(X_m, i)\}$, too. From Lemma 5.15, $c_{S_i,C} - X_m = c_{S,C^{X_m}}$, i.e, $c_{S_i,C} - X_m = c$. So, $c_{S_i,C}$ and $c_{S,C^{X_m}}$ assign the same labels to all instances in $X \setminus \{X_m\}$. Consequently S_i is a compression set for $c_{S_i,C} = c \cup \{(X_m, i)\}$.

□

We have everything required to prove the main theorem of this section, which states that every VCD_Ψ-maximum class has a compression scheme of size VCD_Ψ of the class if the reduction property is fulfilled by VCD_Ψ.

Proof of Theorem 5.5. The compression function f on the input of a sample S of size at least d, where S agrees with at least one concept in C, works as follows: S is a concept $c \in C|_{X(S)}$. Since $C|_{X(S)}$ is VCD_Ψ-maximum with $\text{VCD}_\Psi(C) = d$, Theorem 5.17 yields a compression set $S' \subseteq S$ for S such that $|S'| = d$. In particular, $c = c_{S',C|_{X(S)}}$. Any such compression set is returned by the compression function, that

is, $f(S) = S'$.

The decompression function, given a compression set S' of size d and an $X_i \in X$, returns as a hypothesis the concept $c_{S',C} = c_{X(S'),C} \cup S'$ on X from the class C and thus predicts $c_{S',C}(X_i)$ as the label of X_i. □

An inspection of the proof will show that Theorem 5.5 also holds if X is infinite. In that case, a class is called VCD_Ψ-maximum of dimension d, if all of its restrictions to finite subsets of X of size at least d are VCD_Ψ-maximum of dimension d.

For an infinite instance space and for a C-realizable sample S with $X(S) \subseteq X' \subset X$, such that X' is finite and $|S| = d$, we define $c_{X(S),C}$ on the instances in $X' \setminus X(S)$ as $c_{X(S),C|_{X'}}$. Consequently, $c_{S,C}$ is defined as $c_{S,C|_{X'}}$. Note that X' can contain finitely many instances from X and since C is maximum, $C|_{X'}$ is also maximum. So, by Lemma 5.13, $c_{X(S),C}$ assigns a unique label to each instance $X_i \in X \setminus X(S)$. That is, the concept $c_{S',C}$ on X is consistent with the original sample set $c_{S',C|_{X(S)}}$. So, Theorem 5.5 works for infinite instance spaces as well.

5.3 Generalizing the Kuzmin-Warmuth Unlabeled Scheme

The reduction property is also useful for extending the Kuzmin-Warmuth unlabeled compression scheme, as we will see next. To show this, we first generalize the definition of an unlabeled scheme to a "tight" labeled compression scheme for the multi-label case.

Obviously, for all notions of VCD_Ψ introduced in Chapter 2, unlabeled compression schemes of size d for a VCD_Ψ-maximum class C of VCD_Ψ d cannot exist, as the number of concepts in C is larger than the number of subsets of the instance space

of size at most $\mathrm{VCD}_\Psi(C)$, i.e., $\Phi_d(N_1, \ldots, N_m) > \Phi_d(m) = \Phi_d(1, \ldots, 1)$. Here, we generalize the unlabeled compression scheme for VCD-maximum classes by Kuzmin and Warmuth [KW07] to VCD_Ψ-maximum classes, where VCD_Ψ fulfills the reduction property and is based on spanning families of mappings, by first observing its *tightness*.

Definition 5.18 (**tight compression scheme**). *Let C be a VCD_Ψ-maximum class with $\mathrm{VCD}_\Psi(C) = d$. A sample compression scheme (f, g) of size d for C is tight iff:*

(i) $|\{f(S) \mid S \text{ is } C\text{-realizable}\}| = |C|$.

(ii) If S is C-realizable, then there is exactly one set $T \in \{f(S') \mid S' \text{ is } C\text{-realizable}\}$ such that $S \supseteq T$ and $g(T)$ is consistent with S.

Note that both conditions are necessary for the tightness of the compression scheme. For illustration, consider the class C and the representatives shown in Table 5.5, which yield a tight scheme. As required in (i), no concept can have more than one compression set. Without condition (ii), one might map c_2 to $(X_4, 0)$ instead of $(X_3, 1)$ and the scheme would still satisfy (i), while the sample $\{(X_1, 0), (X_4, 0)\}$ could be compressed to either $(X_4, 0)$ or $\emptyset$.

The critical point exploited in the tight scheme is the property of *missing labelings* in the compression sets, that is, for each set of at most $\mathrm{VCD}_\Psi(C)$ instances $\{X_{i_1}, \ldots, X_{i_k}\}$, there is a tuple of labels $(l_{i_1}, \ldots, l_{i_k}) \in \prod_{1 \leqslant j \leqslant k} X_{i_j}$, such that for each compression set S with $X(S) = \{X_{i_1}, \ldots, X_{i_k}\}$ and for all $j \in \{1, \ldots, k\}$, $(X_{i_j}, l_{i_j}) \notin S$. Indeed, $(l_{i_1}, \ldots, l_{i_k})$ *induces* all missing labelings for the compression sets of size k on $\{X_{i_1}, \ldots, X_{i_k}\}$. For example, consider the class C and the compression sets in Table 5.5, and notice the compression sets S with $X(S) = \{X_1, X_3\}$. As one can verify, any such compression set does not contain $(X_1, 0)$ or $(X_3, 0)$. That is,

$(0,0) \in X_1 \times X_3$ is the tuple that induces all missing labelings for the compression sets of size 2 on $\{X_1, X_3\}$.

In the binary case, our scheme coincides precisely with the Kuzmin-Warmuth scheme, which also exploits the non-trivial property of missing labelings. If one adds labels to the compression sets in the Kuzmin-Warmuth scheme, each set $S \subseteq X$ of size $k \in \{1, \ldots, \mathrm{VCD}(C)\}$ has exactly one missing labeling, and thus $2^k - 1$ assignments of 0 and 1 to the k instances in S are not used as compression sets. But then there is only one possible assignment of labels to the instances in S left, which is why the scheme is in fact unlabeled.

c	X_1	X_2	X_3	X_4	$r(c)$
c_1	0	0	0	0	$\emptyset$
c_2	0	0	1	0	$(X_3, 1)$
c_3	0	0	2	0	$(X_3, 2)$
c_4	0	0	1	1	$(X_4, 1)$
c_5	0	0	1	2	$(X_4, 2)$
c_6	0	1	0	0	$(X_2, 1)$
c_7	0	2	0	0	$(X_2, 2)$
c_8	1	0	0	0	$(X_1, 1)$
c_9	2	0	0	0	$(X_1, 2)$
c_{10}	1	0	1	0	$(X_1, 1), (X_3, 1)$
c_{11}	1	0	2	0	$(X_1, 1), (X_3, 2)$
c_{12}	2	0	1	0	$(X_1, 2), (X_3, 1)$
c_{13}	2	0	2	0	$(X_1, 2), (X_3, 2)$
c_{14}	1	0	1	1	$(X_1, 1), (X_4, 1)$
c_{15}	2	0	1	1	$(X_1, 2), (X_4, 1)$
c_{16}	1	0	1	2	$(X_1, 1), (X_4, 2)$
c_{17}	2	0	1	2	$(X_1, 2), (X_4, 2)$
c_{18}	1	1	0	0	$(X_1, 1), (X_2, 1)$
c_{19}	1	2	0	0	$(X_1, 1), (X_2, 2)$
c_{20}	2	1	0	0	$(X_1, 2), (X_2, 1)$
c_{21}	2	2	0	0	$(X_1, 2), (X_2, 2)$
c_{22}	0	1	0	1	$(X_3, 0), (X_4, 1)$
c_{23}	0	1	0	2	$(X_3, 0), (X_4, 2)$
c_{24}	0	1	1	0	$(X_2, 1), (X_3, 1)$
c_{25}	0	1	2	0	$(X_2, 1), (X_3, 2)$
c_{26}	0	2	1	0	$(X_2, 2), (X_3, 1)$
c_{27}	0	2	2	0	$(X_2, 2), (X_3, 2)$
c_{28}	0	1	1	1	$(X_2, 1), (X_4, 1)$
c_{29}	0	2	1	1	$(X_2, 2), (X_4, 1)$
c_{30}	0	1	2	1	$(X_3, 2), (X_4, 1)$
c_{31}	0	1	1	2	$(X_2, 1), (X_4, 2)$
c_{32}	0	2	1	2	$(X_2, 2), (X_4, 2)$
c_{33}	0	1	2	2	$(X_3, 2), (X_4, 2)$

Table 5.5: VCD_{Ψ_G}-maximum class and representatives resulting from Algorithm 2.

Our goal here is to justify the following theorem and the remainder of this section is devoted to its proof.

Theorem 5.19. *Let Ψ_i, $1 \leq i \leq m$, be a spanning family of mappings. Let $\Psi = \Psi_1 \times \cdots \times \Psi_m$. If VCD_Ψ fulfills the reduction property then any VCD_Ψ-maximum class C has a tight sample compression scheme of size $\mathrm{VCD}_\Psi(C)$.*

Our proof has the same structure as the one by Kuzmin and Warmuth for the

binary case. However, various technical barriers have to be overcome for the multi-label case. For the rest of this section, let $C \subseteq \prod_{1 \leq i \leq m} X_i$ be a VCD_Ψ-maximum class of dimension d, where VCD_Ψ has the reduction property and Ψ is the direct product of spanning families of mappings.

In [KW07] a *representation mapping* r for a VCD-maximum class $C \subseteq 2^X$ is a bijection between C and the set of all subsets of X of size at most $\mathrm{VCD}(C)$ such that for any $c, c' \in C$, $c|_{(r(c) \cup r(c'))} \neq c'|_{(r(c) \cup r(c'))}$, that is, c and c' do not *clash* w.r.t. r. The non-clashing property for a representation mapping is equivalent to having a unique representative for each C-realizable sample [KW07]. Kuzmin and Warmuth showed that, given a representation mapping r for a class C, for any sample S of a concept from C with $|S| \geq \mathrm{VCD}(C)$, there is some concept $c \in C$ that is consistent with S for which S can be mapped to $r(c) \subseteq X(S)$ and for any $c' \in C$, $c' \neq c$, consistent with S, $r(c') \not\subseteq X(S)$.

As we need to use labels in the compression sets, we modify the definition of representation mapping. For a set $Y = \{X_{i_1}, \ldots, X_{i_k}\} \subseteq X$, let L^Y always denote a tuple of labels $L^Y = (l_1^Y, \ldots, l_k^Y) \in \prod_{1 \leq j \leq k} X_{i_j}$. Consider the set $\mathrm{Rep}_{\leq d}(X) = \{Y \subseteq X \mid 0 \leq |Y| \leq d\}$. We construct a set of labeled representatives $\mathrm{LRep}_{\leq d}(X)$ from $\mathrm{Rep}_{\leq d}(X)$ using Algorithm 1.

For each $Y = \{X_{i_1}, \ldots, X_{i_k}\}$ with $k \leq d$, $C|_Y = \prod_{1 \leq j \leq k} X_{i_j}$. So, for any output $\mathrm{LRep}_{\leq d}(X)$ from Algorithm 1, and for any representative $S \in \mathrm{LRep}_{\leq d}(X)$, there is a $c \in C$ with $S \subseteq c$. Further,

$$\begin{aligned} |\mathrm{LRep}_{\leq d}(X)| &= 1 + \sum_{1 \leq i \leq m} N_i + \sum_{1 \leq i_1 < i_2 \leq m} N_{i_1} N_{i_2} + \cdots + \sum_{1 \leq i_1 < \cdots < i_d \leq m} N_{i_1} \cdots N_{i_d} \\ &= \Phi_d(N_1, \ldots, N_m) \\ &= |C|, \quad \text{for each } \mathrm{LRep}_{\leq d}(X) \text{ resulting from Algorithm 1.} \end{aligned}$$

Labeled Representatives Construction Algorithm
Input: the set $\mathrm{Rep}_{\leq d}(X) = \{Y \subseteq X \mid 0 \leq |Y| \leq d\}$
Output: a set of labeled representatives from $\mathrm{Rep}_{\leq d}(X)$

1. **Set** $\mathrm{LRep}_{\leq d}(X) \leftarrow \{\emptyset\}$.
2. **For each** $Y = \{X_{i_1}, \ldots, X_{i_k}\} \in \mathrm{Rep}_{\leq d}(X) \setminus \{\emptyset\}$ **do**
   Set $\mathrm{Rep}_{\leq d}(X) \leftarrow \mathrm{Rep}_{\leq d}(X) \setminus \{Y\}$
   Pick some $L^Y = (l_1^Y, \ldots, l_k^Y) \in \prod_{1 \leq j \leq k} X_{i_j}$
   Set $\mathrm{LabeledRep}(Y, L^Y) \leftarrow \prod_{1 \leq j \leq k} (X_{i_j} \setminus \{l_j^Y\})$
   Set $\mathrm{LRep}_{\leq d}(X) \leftarrow \mathrm{LRep}_{\leq d}(X) \cup \mathrm{LabeledRep}(Y, L^Y)$.

Algorithm 1: Constructing a set of representatives.

We say that a bijection r between C and some $\mathrm{LRep}_{\leq d}(X)$ is *consistent*, if for each $c \in C$, $r(c) \subseteq c$. We also say that the concepts $c, c' \in C$, $c \neq c'$, clash w.r.t. a consistent bijection r, if $r(c) \subseteq c'$ and $r(c') \subseteq c$.

Definition 5.20. *A* representation mapping *for C is a consistent bijection r between C and some representative set* $\mathrm{LRep}_{\leq d}(X)$ *in which no two concepts clash.*

For example, the bijection r for the class C in Table 5.5 is a representation mapping, because one can see that no two concepts clash w.r.t. r.

Essentially, we want to find a representation mapping for VCD_Ψ-maximum classes with a fixed VCD_Ψ. As in the binary case [KW07], the following lemma shows how the non-clashing property is useful for finding unique labeled representatives for samples in the multi-label case.

Lemma 5.21. *Let r be a consistent bijection between C and a set of labeled representatives* $\mathrm{LRep}_{\leq d}(X)$*. Then the following two statements are equivalent:*

1. *No two concepts clash w.r.t. r.*
2. *For any sample S that is consistent with at least one concept in C, there is exactly one concept $c \in C$ that is consistent with S and $r(c) \subseteq S$.*

Proof. The proof is by contradiction (analogous to the proof of Lemma 1 in [KW07]).

$2 \Rightarrow 1$: Assume $\neg 1$. That is, there are two concepts $c, c' \in C$, such that $r(c) \subseteq c'$ and $r(c') \subseteq c$. Let $S = r(c) \cup r(c')$. Then it is obvious that both c and c' are consistent with S, $r(c) \subset S$ and $r(c') \subset S$, which negates 2.

$1 \Rightarrow 2$: Assume $\neg 2$. We need to consider two cases. First, assume that there is a sample S for which there are at least two consistent concepts $c, c' \in C$ such that $r(c) \subseteq S$ and $r(c') \subseteq S$. Since $S \subseteq c$ and $S \subseteq c'$, it is obvious that $r(c) \subseteq c'$ and also $r(c') \subseteq c$, which negates 1. Second, assume that there is a sample S for which there is no consistent concept $c \in C$ with $r(c) \subseteq S$. Let $X(S) = \{X_{i_1}, \ldots, X_{i_k}\}$, for some $k \in [m]$. Then

$$\begin{aligned} \text{size}(C|_{X(S)}) &= \Phi_d(N_{i_1}, \ldots, N_{i_k}) = |\text{LRep}_{\leq d}(X(S))| \\ &= |\{c \in C \mid r(c) \in \text{LRep}_{\leq d}(X(S))\}| \end{aligned} \tag{5.2}$$

and thus by the pigeonhole principle, there must be a sample $S' \neq S$ with $X(S') = X(S)$ for which there are two such concepts, which again negates 1. □

Lemma 5.21 helps us to construct a compression scheme of size VCD_Ψ for a VCD_Ψ-maximum class C from a representation mapping r. For compression, a sample S is compressed to $r(c) \subseteq S$, where c is consistent with S. For reconstruction, $r(c)$ is mapped to $c \supseteq S$, as r is a consistent bijective mapping.

We demonstrate that a representation mapping r can be used as a compression-decompression function for the concepts in a VCD_Ψ-maximum class C. In the next corollary, we use such a mapping to derive a compression scheme of size d for $C|_Y$, for any $Y \subseteq X$ with $|Y| > d$. For any $\bar{c} \in C|_Y$, define $r_Y(\bar{c}) := r(c)$ where c is the unique concept in C with $c|_Y = \bar{c}$ and $r(c) \subseteq \bar{c}$.

Corollary 5.22. *Let r be a representation mapping for C. Let $Y \subseteq X$ with $|Y| > d$. Then r_Y is a representation mapping for $C|_Y$.*

Proof. (Partially analogous to the proof of Corollary 2 in [KW07]) As it is clear from the statement, we are treating a concept in the restricted class $C|_Y$ as a sample of the original class C. So, by Lemma 5.21, r_Y is uniquely defined. We need to show that r_Y is a representation mapping. First, we consider the non-clashing property. Assume that there are concepts $\bar{c_1}, \bar{c_2} \in C|_Y$, such that $r(\bar{c_1}) \subseteq \bar{c_2}$ and $r(\bar{c_2}) \subseteq \bar{c_1}$. Then there are concepts $c_1, c_2 \in C$ where $\bar{c_1} = c_1|_Y$, $\bar{c_2} = c_2|_Y$ and c_1 and c_2 clash w.r.t. r. Second, we verify the bijectivity of r_Y. By replacing $X(S)$ with Y in (5.2), and applying the same counting argument as in the second part of the proof of Lemma 5.21, we conclude that r_Y is bijective. □

At this point, the crucial notion of *tail* comes into play. As in the binary case, we define the *tail* of a concept class C on an instance $X_t \in X$ as the set of all concepts $c \in C$ such that $c - X_t \in (C - X_t) \setminus C^{X_t}$ [KW07]. This corresponds to the set of concepts in $C - X_t$ that do not have all extensions onto X, or equivalently (by the reduction property), that have a unique extension onto X. That is, for any $c \in \text{tail}_{X_t}(C)$, there exists only one label $l \in \{0, 1, \ldots, N_t\}$ such that $(c - X_t) \cup \{(X_t, l)\} \in C$. Note that $C = (C^{X_t} \times X_t) \cup \text{tail}_{X_t}(C)$.

Recall that a forbidden labeling of C with $\text{VCD}_\Psi(C) = d < |X|$, is a set of $d + 1$ examples that is inconsistent with all concepts in C. As in the binary case, we establish a connection between tail concepts and forbidden labelings. By assumption, for $X_t \in X$, every concept in $C - X_t$ has either a unique or all possible extensions to concepts in C. So, each concept in $\text{tail}_{X_t}(C)$ corresponds to a concept in $C - X_t$ that has only one extension onto X_t. That is, $|\text{tail}_{X_t}(C)| = |\text{tail}_{X_t}(C) - X_t|$. Further, $C - X_t = C^{X_t} \cup (\text{tail}_{X_t}(C) - X_t)$ where C^{X_t} and $(\text{tail}_{X_t}(C) - X_t)$ are disjoint. By

Theorem 3.14 and Theorem 4.2, for $d < m$, $C - X_t$ and C^{X_t} are VCD_Ψ-maximum of dimensions d and $d-1$, respectively. So,

$$\begin{aligned}
|\text{tail}_{X_t}(C)| &= |\text{tail}_{X_t}(C) - X_t| = |C - X_t| - |C^{X_t}| \\
&= \Phi_d(N_1, \ldots, N_{t-1}, N_{t+1}, \ldots, N_m) - \Phi_{d-1}(N_1, \ldots, N_{t-1}, N_{t+1}, \ldots, N_m) \\
&= \sum\nolimits_{1 \le i_1 < \cdots < i_d \le m,\ i_j \ne t} N_{i_1} \cdots N_{i_d}.
\end{aligned}$$

For $Y = \{X_{i_1}, \ldots, X_{i_{d+1}}\}$, $C|_Y$ is VCD_Ψ-maximum of dimension d and thus

$$|\text{Forb}(C, Y)| = (N_{i_1} + 1) \cdots (N_{i_{d+1}} + 1) - \Phi_d(N_{i_1}, \ldots, N_{i_{d+1}}) = N_{i_1} \cdots N_{i_{d+1}}.$$

As in the binary case, it is easy to see that every concept in $\text{tail}_{X_t}(C)$ contains some forbidden labeling of C^{X_t} of size d and each such forbidden labeling occurs in at least one tail concept. Note that C^{X_t} is a VCD_Ψ-maximum class of dimension $d-1$ and for each set of d instances $Y = \{X_{i_1}, \ldots, X_{i_d}\} \subseteq (X \setminus \{X_t\})$, $|\text{Forb}(C^{X_t}, Y)| = N_{i_1} \cdots N_{i_d}$. So,

$$\begin{aligned}
|\text{Forb}(C^{X_t})| &= \sum_{\substack{Y \subseteq (X \setminus \{X_t\}) \\ |Y| = d}} |\text{Forb}(C^{X_t}, Y)| \\
&= \sum\nolimits_{1 \le i_1 < \cdots < i_d \le m,\ i_j \ne t} N_{i_1} \cdots N_{i_d} \\
&= |\text{tail}_{X_t}(C)|.
\end{aligned}$$

First, adding any concept in $\text{tail}_{X_t}(C) - X_t$ to C^{X_t} increases the VCD_Ψ of C^{X_t} due to the maximum size property of C^{X_t}. So, each concept in $\text{tail}_{X_t}(C)$ contains at least one forbidden labeling of C^{X_t}. Second, $C - X_t = C^{X_t} \cup (\text{tail}_{X_t}(C) - X_t)$ where the reduction class and the tail class are disjoint. Next, for each set of d instances

$Y \subseteq (X \setminus \{X_t\})$, $(C - X_t)|_Y = \prod_{X_i \in Y} X_i$, since C is a VCD_Ψ-maximum class of dimension d. That is,

$$C^{X_t}|_Y \cup (\text{tail}_{X_t}(C) - X_t)|_Y = \prod_{X_i \in Y} X_i$$

and

$$(\text{tail}_{X_t}(C) - X_t)|_Y \supseteq \prod_{X_i \in Y} X_i \setminus C^{X_t}|_Y = \text{Forb}(C^{X_t}, Y).$$

In other words, all forbidden labelings of C^{X_t} on Y are in $(\text{tail}_{X_t}(C) - X_t)|_Y$. Since Y was chosen arbitrarily, we conclude that all forbidden labelings of C^{X_t} appear in $\text{tail}_{X_t}(C)$.

Example 5.23. *Consider the class C in Table 5.5 and C^{X_1} and $\text{tail}_{X_1}(C)$ in Table 5.6. Then for $Y = \{X_2, X_3\}$,*

$$\begin{aligned}
\text{Forb}(C^{X_1}, Y) &= \{\{(X_2,1),(X_3,1)\},\{(X_2,1),(X_3,2)\},\{(X_2,2),(X_3,1)\},\\
&\quad \{(X_2,2),(X_3,2)\}\}\\
&\subseteq \text{tail}_{X_1}(C) - X_1)|_Y\\
&= \{\{(X_2,1),(X_3,0)\},\{(X_2,1),(X_3,1)\},\{(X_2,1),(X_3,2)\},\\
&\quad \{(X_2,2),(X_3,1)\},\{(X_2,2),(X_3,2)\}\}.
\end{aligned}$$

The Kuzmin-Warmuth scheme [KW07] finds representatives for C by partitioning C into $C^{X_i} \times X_i$ and $\text{tail}_{X_i}(C)$ for some $X_i \in X$. It identifies the representatives for C^{X_i} recursively, and extends them to representatives for C. That is, for any concept $c \in C^{X_i}$ with a representative $r(c)$, $r(c \cup (X_i, 0)) := r(c)$ and $r(c \cup (X_i, 1)) := r(c) \cup X_i$. Next, it finds representatives for the remaining concepts, i.e., those in $\text{tail}_{X_i}(C)$ by assigning each of them a forbidden labeling of the class C^{X_i} of size d. Since the

representative for each concept in $\text{tail}_{X_i}(C)$ is a forbidden labeling of the class C^{X_i}, the non-clashing property between $\text{tail}_{X_i}(C)$ and C^{X_i} is guaranteed.

$c \in C^{X_1}$	X_2	X_3	X_4	$r(c')$
c'_1	0	0	0	$\emptyset$
c'_2	1	0	0	$(X_2, 1)$
c'_3	2	0	0	$(X_2, 2)$
c'_4	0	1	0	$(X_3, 1)$
c'_5	0	2	0	$(X_3, 2)$
c'_6	0	1	1	$(X_4, 1)$
c'_7	0	1	2	$(X_4, 2)$

$c \in \text{tail}_{X_1}(C)$	X_1	X_2	X_3	X_4	$r(c'')$
c''_1	0	1	0	1	$(X_3, 0), (X_4, 1)$
c''_2	0	1	0	2	$(X_3, 0), (X_4, 2)$
c''_3	0	1	1	0	$(X_2, 1), (X_3, 1)$
c''_4	0	1	2	0	$(X_2, 1), (X_3, 2)$
c''_5	0	2	1	0	$(X_2, 2), (X_3, 1)$
c''_6	0	2	2	0	$(X_2, 2), (X_3, 2)$
c''_7	0	1	1	1	$(X_2, 1), (X_4, 1)$
c''_8	0	2	1	1	$(X_2, 2), (X_4, 1)$
c''_9	0	1	2	1	$(X_3, 2), (X_4, 1)$
c''_{10}	0	1	1	2	$(X_2, 1), (X_4, 2)$
c''_{11}	0	2	1	2	$(X_2, 2), (X_4, 2)$
c''_{12}	0	1	2	2	$(X_3, 2), (X_4, 2)$

Table 5.6: C^{X_1} and $\text{tail}_{X_1}(C)$ where C is the VCD_{Ψ_G}-maximum class from Table 5.5.

As in the Kuzmin-Warmuth scheme, we establish a recursive structure for tails by proving the next lemma. Note that such structure in the multi-label case is not as simple as the one in the binary case and it cannot be presented in a single statement. However, our proof follows a similar reasoning as the one in the binary case [KW07].

We introduce some notation, first. For $s, t \in [m]$, with $s < t$ and a concept $\bar{c} \in C|_{X \setminus \{X_s, X_t\}}$, let $i\bar{c}$, $\bar{c}j$ and $i\bar{c}j$ denote $\bar{c} \cup \{(X_s, i)\}$, $\bar{c} \cup \{(X_t, j)\}$ and $\bar{c} \cup \{(X_s, i), (X_t, j)\}$, respectively.

Lemma 5.24. *Let $s, t \in [m]$ with $s \neq t$. Then the following statements are true.*

1. *For each $c \in \text{tail}_{X_s}(C^{X_t})$ there are at least N_t distinct labels $l_1, \ldots, l_{N_t} \in X_t$ such that $c \times \{l_1, \ldots, l_{N_t}\} \subseteq \text{tail}_{X_s}(C)$. If $c \in \text{tail}_{X_s}(C^{X_t}) \setminus \text{tail}_{X_s}(C - X_t)$, then there are exactly N_t such labels.*

2. *For each $c \in \text{tail}_{X_s}(C - X_t)$ there is at least one label $l \in X_t$ such that $c \times \{l\} \in \text{tail}_{X_s}(C)$. If $c \in \text{tail}_{X_s}(C - X_t) \cap \text{tail}_{X_s}(C^{X_t})$, then $c \times X_t \subseteq \text{tail}_{X_s}(C)$.*

3. *Each concept in* $\text{tail}_{X_s}(C)$ *is an extension of either a concept in* $\text{tail}_{X_s}(C^{X_t})$ *or a concept in* $\text{tail}_{X_s}(C - X_t)$.

Proof. W.l.o.g., assume $s < t$.

1. W.l.o.g., let $c = 0\bar{c} \in \text{tail}_{X_s}(C^{X_t})$. We show that for some set $\{l_1, \ldots, l_{N_t}\} \subset X_t$, $0\bar{c}j \in \text{tail}_{X_s}(C)$, for all $j \in \{l_1, \ldots, l_{N_t}\}$. Clearly, $\text{tail}_{X_s}(C^{X_t}) \subseteq C^{X_t}$, so $0\bar{c} \in C^{X_t}$ and thus $0\bar{c}0, \ldots, 0\bar{c}N_t \in C$. We need to show that N_t concepts $0\bar{c}j$, $j \in \{l_1, \ldots, l_{N_t}\}$, belong to $\text{tail}_{X_s}(C)$. For the purpose of contradiction, assume that $0\bar{c}0, 0\bar{c}1 \notin \text{tail}_{X_s}(C)$, that is, $\bar{c}0, \bar{c}1 \in C^{X_s}$. Since C^{X_s} is VCD_Ψ-maximum, $\bar{c}$ has $N_t + 1$ extensions to concepts in C^{X_s}. Therefore,

$$\bar{c}0, \bar{c}1, \ldots, \bar{c}N_t \in C^{X_s} \Rightarrow \begin{cases} 0\bar{c}0, & 1\bar{c}0, & \ldots, & N_s\bar{c}0 & \in C \\ 0\bar{c}1, & 1\bar{c}1, & \ldots, & N_s\bar{c}1 & \in C \\ \vdots & & & & \\ 0\bar{c}N_t, & 1\bar{c}N_t, & \ldots, & N_s\bar{c}N_t & \in C \end{cases}$$

i.e., $0\bar{c}, \ldots, N_s\bar{c} \in C^{X_t}$ and $\bar{c} \in (C^{X_t})^{X_s}$. So, $0\bar{c} \notin \text{tail}_{X_s}(C^{X_t})$ — a contradiction.

We need to show that if $0\bar{c} \in \text{tail}_{X_s}(C^{X_t}) \setminus \text{tail}_{X_s}(C - X_t)$, there is an $l \in X_t$ for which $0\bar{c}l \notin \text{tail}_{X_s}(C)$. Assume that for all $j \in X_t$, $0\bar{c}j \in \text{tail}_{X_s}(C)$, i.e., $\bar{c}j \notin C^{X_s}$. That is, for all $j \in X_t$, $0\bar{c}j \in C$ and $\bar{c}j$ has only one extension on X_s to concepts in C, namely with $(X_s, 0)$. So, for all $i \in X_s \setminus \{0\}$ and all $j \in X_t$, $i\bar{c}j \notin C$, and thus $i\bar{c} \notin C - X_t$. This implies $0\bar{c} \in \text{tail}_{X_s}(C - X_t)$.

2. Let $0\bar{c} \in \text{tail}_{X_s}(C - X_t)$. We show that for each $j \in X_t$ with $0\bar{c}j \in C$, we have $0\bar{c}j \in \text{tail}_{X_s}(C)$. W.l.o.g., assume that $0\bar{c}0 \in C$, but $0\bar{c}0 \notin \text{tail}_{X_s}(C)$. That is, $\bar{c}0 \in C^{X_s}$, and consequently, $i\bar{c}0 \in C$, for all $i \in X_s$. So, $i\bar{c} \in C - X_t$, for all $i \in X_s$ and thus $\bar{c} \in (C - X_t)^{X_s}$. Hence, $0\bar{c} \notin \text{tail}_{X_s}(C - X_t)$ — a contradiction.

For a concept $0\bar{c} \in \text{tail}_{X_s}(C - X_t) \cap \text{tail}_{X_s}(C^{X_t})$, and thus $0\bar{c} \in C^{X_t}$, we have

$0\bar{c}j \in C$, for all $j \in X_t$. According to the previous paragraph, we conclude that $0\bar{c}j \in \text{tail}_{X_s}(C)$, for all $j \in X_t$.

3. First one can show that $|\text{tail}_{X_s}(C)| = N_t|\text{tail}_{X_s}(C^{X_t})| + |\text{tail}_{X_s}(C - X_t)|$ as follows.

$$\begin{aligned}
|\text{tail}_{X_s}(C)| &= \sum_{\substack{1\le i_1<\cdots<i_d\le m \\ i_j\neq s}} N_{i_1}\cdots N_{i_d} \\
&= N_t \sum_{\substack{1\le i_1<\cdots<i_{d-1}\le m \\ i_j\neq s,\ i_j\neq t}} N_{i_1}\cdots N_{i_{d-1}} + \sum_{\substack{1\le i_1<\cdots<i_d\le m \\ i_j\neq s,\ i_j\neq t}} N_{i_1}\cdots N_{i_d} \\
&= N_t|\text{tail}_{X_s}(C^{X_t})| + |\text{tail}_{X_s}(C - X_t)|.
\end{aligned}$$

Second, from Statements 1 and 2, any concept in $\text{tail}_{X_s}(C^{X_t})$ can be mapped to N_t concepts in $\text{tail}_{X_s}(C)$, and any concept in $\text{tail}_{X_s}(C - X_t)$ can be mapped to a single concept in $\text{tail}_{X_s}(C)$. Hence, each concept in $\text{tail}_{X_s}(C)$ is an extension of either a concept in $\text{tail}_{X_s}(C^{X_t})$ or a concept in $\text{tail}_{X_s}(C - X_t)$. □

Example 5.25. *Consider the class* C^{X_1} *in Table 5.7 and* $\text{tail}_{X_2}(C^{X_1})$, $\text{tail}_{X_2}((C^{X_1})^{X_3})$ *and* $\text{tail}_{X_2}(C^{X_1} - X_3)$ *in Tabel 5.8. Then*

$$c_1^1 \in \text{tail}_{X_2}((C^{X_1})^{X_3}) \setminus \text{tail}_{X_2}(C^{X_1} - X_3) \Rightarrow\ c_1^1 \times \{1,2\} = \{c_1^0, c_2^0\} \subset \text{tail}_{X_2}(C^{X_1}),$$

$$c_1^2 \in \text{tail}_{X_2}(C^{X_1} - X_3) \text{ and } c_1^2 \notin \text{tail}_{X_2}((C^{X_1})^{X_3}) \ \Rightarrow\ c_1^2 \times \{1\} = c_3^0 \in \text{tail}_{X_2}(C^{X_1})$$

and

$$c_2^2 \in \text{tail}_{X_2}(C^{X_1} - X_3) \text{ and } c_2^2 \notin \text{tail}_{X_2}((C^{X_1})^{X_3}) \ \Rightarrow\ c_2^2 \times \{2\} = c_4^0 \in \text{tail}_{X_2}(C^{X_1}).$$

So, each concept in $\text{tail}_{X_2}(C^{X_1})$ *is an extension of a concept either in* $\text{tail}_{X_2}((C^{X_1})^{X_3})$ *or in* $\text{tail}_{X_2}(C^{X_1} - X_3)$.

The next lemma states that the reduction and restriction operations are interchangeable in the order in which they are applied.

Lemma 5.26. *For any $s, t \in [m]$, with $s \neq t$, $C^{X_s} - X_t = (C - X_t)^{X_s}$.*

Proof. (Analogous to the proof of Lemma 7 in [KW07]) W.l.o.g., assume that $s < t$. On the one hand, we show that $C^{X_s} - X_t \subseteq (C - X_t)^{X_s}$. Let $\bar{c} \in C^{X_s} - X_t$. So, there is at least one label $j \in X_t$, such that $\bar{c}j \in C^{X_s}$, and thus $i\bar{c}j \in C$, for all $i \in X_s$, since C^{X_s} is a VCD_Ψ-maximum class. Therefore, $i\bar{c} \in C - X_t$, for all $i \in X_s$, and consequently, $\bar{c} \in (C - X_t)^{X_s}$. On the other hand, it is easy to see that $C^{X_s} - X_t$ and $(C - X_t)^{X_s}$ are of the same size, since they are both VCD_Ψ-maximum classes on the same instance space and have the same VCD_Ψ-dimension. Hence, $C^{X_s} - X_t = (C - X_t)^{X_s}$. □

$c \in C^{X_1}$	X_2	X_3	X_4	$r(c)$
c'_1	0	0	0	$r(c_1) = \emptyset$
c'_2	1	0	0	$r(c''_1) \cup \{(X_2, 1)\} = \{(X_2, 1)\}$
c'_3	2	0	0	$r(c''_1) \cup \{(X_2, 2)\} = \{(X_2, 2)\}$
c'_4	0	1	0	$r(c^0_1) = \{(X_3, 1)\}$
c'_5	0	2	0	$r(c^0_2) = \{(X_3, 2)\}$
c'_6	0	1	1	$r(c^0_3) = \{(X_4, 2)\}$
c'_7	0	1	2	$r(c^0_4) = \{(X_4, 1)\}$

$c \in (C^{X_1})^{X_2}$	X_3	X_4	$r(c)$
c''_1	0	0	$\emptyset$

Table 5.7: C^{X_1} and $(C^{X_1})^{X_2}$ where C is the VCD_{Ψ_G}-maximum class from Table 5.5.

$c \in tail_{X_2}(C^{X_1})$	X_2	X_3	X_4	$r(c)$
c^0_1	0	1	0	$r(c^1_1) \cup \{(X_3, 1)\} = \{(X_3, 1)\}$
c^0_2	0	2	0	$r(c^1_1) \cup \{(X_3, 2)\} = \{(X_3, 2)\}$
c^0_3	0	1	1	$r(c^2_1) = \{(X_4, 1)\}$
c^0_4	0	1	2	$r(c^2_2) = \{(X_4, 2)\}$

$c \in tail_{X_2}((C^{X_1})^{X_3})$	X_2	X_4	$r(c)$
c^1_1	0	0	$\emptyset$

$c \in tail_{X_2}(C^{X_1} - X_3)$	X_2	X_4	$r(c)$
c^2_1	0	1	$\{(X_4, 1)\}$
c^2_2	0	2	$\{(X_4, 2)\}$

Table 5.8: $\mathrm{tail}_{X_2}(C^{X_1})$, $\mathrm{tail}_{X_2}((C^{X_1})^{X_3})$ and $\mathrm{tail}_{X_2}(C^{X_1} - X_3)$ where C is the VCD_{Ψ_G}-maximum class from Table 5.5.

Lemma 5.26 has the following corollary. The proof is the same as that of Corollary 8 in [KW07] and is presented here for the sake of completeness.

Corollary 5.27. $\text{Forb}((C - X_t)^{X_s}) \subseteq \text{Forb}(C^{X_s})$.

Proof. By Lemma 5.26, $(C - X_t)^{X_s}$ and $C^{X_s} - X_t$ have the same forbidden labelings. Moreover, the forbidden labelings for $C^{X_s} - X_t$ are all forbidden labelings of C^{X_s} that are restricted to $X \setminus \{X_t\}$. □

$c \in C - X_4$	X_1	X_2	X_3
c_1	0	0	0
c_2	0	0	1
c_3	0	0	2
c_4	0	1	0
c_5	0	2	0
c_6	1	0	0
c_7	2	0	0
c_8	1	0	1
c_9	1	0	2
c_{10}	2	0	1
c_{11}	2	0	2
c_{12}	1	1	0
c_{13}	1	2	0
c_{14}	2	1	0
c_{15}	2	2	0
c_{16}	0	1	1
c_{17}	0	1	2
c_{18}	0	2	1
c_{19}	0	2	2

$c \in C^{X_1}$	X_2	X_3	X_4
c''_1	0	0	0
c''_2	1	0	0
c''_3	2	0	0
c''_4	0	1	0
c''_5	0	2	0
c''_6	0	1	1
c''_7	0	1	2

$c \in (C - X_4)^{X_1}$	X_2	X_3
c'_1	0	0
c'_2	0	1
c'_3	0	2
c'_4	1	0
c'_5	2	0

$c \in C^{X_1} - X_4$	X_2	X_3
c'_1	0	0
c'_2	0	1
c'_3	0	2
c'_4	1	0
c'_5	2	0

Table 5.9: $C - X_4$, $(C - X_4)^{X_1}$, C^{X_1} and $C^{X_1} - X_4$ where C is the VCD_{Ψ_G}-maximum class from Table 5.5.

Example 5.28. *As shown in Table 5.9,* $(C - X_4)^{X_1} = C^{X_1} - X_4$, *for the* VCD_{Ψ_G}-*maximum class* C *from Table 5.5. Moreover, all forbidden labelings for* $(C - X_4)^{X_1}$,

which are $\{\{(X_2,1),(X_3,1)\},\{(X_2,1),(X_3,2)\},\{(X_2,2),(X_3,1)\},\{(X_2,2),(X_3,2)\}\}$, *are also forbidden labelings for* C^{X_1}.

The next lemma connects the forbidden labelings for $(C^{X_s})^{X_t}$ to the ones for C^{X_s} where $X_s, X_t \in X$ and $s \neq t$. Although we follow the logic of the proof of Lemma 9 in [KW07], our proof here is substantially more extensive, because the statement is more complicated to validate in the multi-label case.

Lemma 5.29. *Any forbidden labeling for* $(C^{X_s})^{X_t}$ *can be extended to* N_t *forbidden labelings for* C^{X_s}.

Proof. Let $\mathrm{VCD}_\Psi(C) = d$. We show that for any set of d instances $Y \subseteq X \setminus \{X_s\}$ with $X_t \in Y$, there are N_t forbidden labelings $S_i = S \cup \{(X_t, l_i)\}$, $1 \leq i \leq N_t$ and $l_i \in X_t$, for C^{X_s} such that $X(S_i) = Y$, $X(S) = Y \setminus X_t$, and S is a forbidden labeling of size $d-1$ for $(C^{X_s})^{X_t}$.

Let $Y = \{X_{i_1}, \ldots, X_{i_{d-1}}, X_t\} \subseteq X \setminus \{X_s\}$, $X(S) = \{X_{i_1}, \ldots, X_{i_{d-1}}\}$, and let $S_1 = S \cup \{(X_t, l_1)\}$ be a forbidden labeling for C^{X_s}. We first prove by contradiction that S is a forbidden labeling for $(C^{X_s})^{X_t}$. Assume that S is not a forbidden labeling for $(C^{X_s})^{X_t}$, and thus is consistent with some concept $c \in (C^{X_s})^{X_t}$. Since $c \times X_t \subseteq C^{X_s}$, we conclude that each sample $S \cup \{(X_t, j)\}$, $j \in X_t$, is consistent with some concept in C^{X_s}. Thus, $S \cup \{(X_t, l_1)\}$ is not a forbidden labeling for C^{X_s} — a contradiction.

W.l.o.g., assume that $N_t \geq 2$. We next show that there are $N_t - 1$ more forbidden labels $S_i = S \cup \{(X_t, l_i)\}$, $2 \leq i \leq N_t$, $l_i \in X_t$ for C^{X_s}, i.e., for any concept $\bar{c} \in C^{X_s}$ with $\bar{c}|_{\{X_{i_1},\ldots,X_{i_{d-1}}\}} = S$, $\bar{c}(X_t) = l$ for some $l \in X_t \setminus \{l_1, \ldots, l_{N_t}\}$. Note that C^{X_s} is VCD_Ψ-maximum of dimension $d-1$ so that $C^{X_s}|_{\{X_{i_1},\ldots,X_{i_{d-1}}\}} = \prod_{j \in \{1,\ldots,d-1\}} X_{i_j}$, and thus $S \in C^{X_s}|_{\{X_{i_1},\ldots,X_{i_{d-1}}\}}$. For any $\bar{c} \in C^{X_s}$ with $\bar{c}|_{\{X_{i_1},\ldots,X_{i_{d-1}}\}} = S$, it is clear that $\bar{c}(X_t) \neq l_1$, as $S \cup \{(X_t, l_1)\}$ is a forbidden labeling for C^{X_s}. That is, for any $c' \in C^{X_s}|_Y$ with $c' - X_t = S$, $c'(X_t) \neq l_1$. So, $C^{X_s}|_Y$ does not have all extensions of S and thus,

$C^{X_s}|_Y$ has a unique extension of S on X_t, as $C^{X_s}|_Y$ is a VCD_Ψ-maximum class of dimension $d-1$ on Y. So, there is only one concept $c' \in C^{X_s}|_Y$ with $c' - X_t = S$ and $c'(X_t) = l$, for some $l \in X_t \setminus \{l_1, \ldots, l_{N_t}\}$.

Now, we need to show that C^{X_s} has a unique extension of S on X_t, namely $S \cup \{(X_t, l)\}$. In other words, we need to prove that whenever S occurs in a concept $\bar{c} \in C^{X_s}$, $\bar{c}$ could only have the label l on X_t. For the purpose of contradiction, assume that there are concepts $\bar{c}_1, \bar{c}_2 \in C^{X_s}$ with $\bar{c}_1|_{\{X_{i_1},\ldots,X_{i_{d-1}}\}} = \bar{c}_2|_{\{X_{i_1},\ldots,X_{i_{d-1}}\}} = S$ and $\bar{c}_1(X_t) \neq \bar{c}_2(X_t)$. Let $\bar{c}_1(X_t) = l$ and $\bar{c}_2(X_t) = l'$. Since $\bar{c}_1|_Y \neq \bar{c}_2|_Y$ and $\bar{c}_1|_Y, \bar{c}_2|_Y \in C^{X_s}|_Y$, we conclude that $C^{X_s}|_Y$ has two extensions of S with (X_t, l) and (X_t, l') — a contradiction. So, for any $\bar{c} \in C^{X_s}$ with $\bar{c}|_{\{X_{i_1},\ldots,X_{i_{d-1}}\}} = S$, $\bar{c}(X_t) = l$. In other words, each sample $S \cup \{(X_t, l_i)\}$, $1 \leq i \leq N_t$, is a forbidden labeling for C^{X_s}.

Since C is VCD_Ψ-maximum of dimension d, by Theorem 3.14 and Theorem 4.2, C^{X_s} and $(C^{X_s})^{X_t}$ are both VCD_Ψ-maximum of dimension $d-1$ and $d-2$, respectively. One can then show that $|\text{Forb}(C^{X_s})| = N_t|\text{Forb}((C^{X_s})^{X_t})| + |\text{Forb}((C - X_t)^{X_s})|$ as follows.

$$\begin{aligned}
|\text{Forb}(C^{X_s})| &= \sum_{\substack{1 \leq i_1 < \cdots < i_d \leq m \\ i_j \neq s}} N_{i_1} \cdots N_{i_d} \\
&= N_t \sum_{\substack{1 \leq i_1 < \cdots < i_{d-1} \leq m \\ i_j \neq s \\ i_j \neq t}} N_{i_1} \cdots N_{i_{d-1}} + \sum_{\substack{1 \leq i_1 < \cdots < i_d \leq m \\ i_j \neq s \\ i_j \neq t}} N_{i_1} \cdots N_{i_d} \\
&= N_t|\text{Forb}((C^{X_s})^{X_t})| + |\text{Forb}((C - X_t)^{X_s})|. \qquad (5.3)
\end{aligned}$$

So,

$$|\text{Forb}((C^{X_s})^{X_t})| = \frac{1}{N_t}|\text{Forb}(C^{X_s}, Y)| \qquad (5.4)$$

for all $Y \subseteq X \setminus \{X_s\}$ with $|Y| = d$ and $X_t \in Y$.

Therefore, any set of N_t forbidden labelings $S_i = S \cup \{(X_t, l_i)\}$, $1 \leq i \leq N_t$ for C^{X_s}

Labeled Tail Matching Function (LTMF)
Input: a VCD_Ψ-maximum multi-label concept class C and X with $|X| \geq 1$
Output: a mapping r assigning representatives to all concepts in C

$r =$ **LTMF**(C,X)
  If $\text{VCD}_\Psi(C) = 0$ **then** $r(c) := \emptyset$; (since $C = \{c\}$)
  Else { pick any $X_s \in X$; $\tilde{r} =$ **LTMF**$(C^{X_s}, X \setminus \{X_s\})$;
    For each $\bar{c} \in C^{X_s}$ **do** {
      For $i = 1$ **to** N_s **do**
        $r(\bar{c} \cup \{(X_s, i)\}) := \tilde{r}(\bar{c}) \cup \{(X_s, i)\}$;
      $r(\bar{c} \cup \{(X_s, 0)\}) := \tilde{r}(\bar{c})$; }
    Set $r \leftarrow r \cup$ **LTS**(C, X, X_s);} (see Algorithm 3 for **LTS**)
  return r;

Algorithm 2: Recursively constructing labeled compression sets for concepts.

can be mapped to one forbidden labeling S for $(C^{X_s})^{X_t}$. By counting the number of forbidden labelings for C^{X_s} that contain X_t (as shown in (5.4)), we conclude that any forbidden labeling for $(C^{X_s})^{X_t}$ can be extended to N_t forbidden labelings for C^{X_s}. □

Example 5.30. *Consider the classes C^{X_1} and $(C^{X_1})^{X_2}$ in Table 5.7. $\{(X_3, 1)\}$ is a forbidden labeling for $(C^{X_1})^{X_2}$ on $\{X_3\}$ that can be extended to two forbidden labelings for C^{X_1} on $\{X_2, X_3\}$, namely $\{(X_2, 1), (X_3, 1)\}$ and $\{(X_2, 2), (X_3, 1)\}$.*

The next lemma is now obvious.

Lemma 5.31. *Each forbidden labeling of C^{X_s} is an extension of either a forbidden labeling of $(C^{X_s})^{X_t}$ or a forbidden labeling of $C^{X_s} - X_t$.*

Proof. This follows immediately from Corollary 5.27, Lemma 5.29 and (5.3). □

The following lemma is crucial in connecting the set of forbidden labelings to a labeled set of representatives. While its statement is obvious in the binary case, it is not trivial in the multi-label case. We first establish the statement for the special case when $\text{VCD}_\Psi(C) = |X| - 1$, and then proceed to the general case.

Lemma 5.32. *For any set* $Y = \{X_{i_1}, \ldots, X_{i_d}\} = X \setminus \{X_s\}$ *with* $|Y| = d = |X|-1$, *there is a tuple* $(l_1, \ldots, l_d) \in \prod_{1 \le j \le d} X_{i_j}$ *such that* $\mathrm{Forb}(C^{X_s}, Y) = \prod_{1 \le j \le d}(X_{i_j} \setminus \{l_j\})$.

Proof. Let $m = |X|$, C be a VCD_Ψ-maximum class on m instances and $\mathrm{VCD}_\Psi(C) = m - 1$. The proof is by induction on m. The base case, $m = 1$ ($d = 0$), is obvious. Assume that $m > 1$ and the claim is true for any $m' < m$. Pick $X_t \in X \setminus \{X_s\}$. By Lemma 5.31, each forbidden labeling of C^{X_s} is an extension of a forbidden labeling of either $(C^{X_s})^{X_t}$ or $C^{X_s} - X_t$.

$C^{X_s} - X_t$ is VCD_Ψ-maximum on $m-2$ instances and of VCD_Ψ $m-2$. By definition, $\mathrm{Forb}(C^{X_s} - X_t) = \emptyset$ and consequently, all forbidden labelings of C^{X_s} are the extensions of the forbidden labelings for $(C^{X_s})^{X_t}$.

$\mathrm{Forb}((C^{X_s})^{X_t}) = \mathrm{Forb}((C^{X_t})^{X_s})$, because $(C^{X_s})^{X_t} = (C^{X_t})^{X_s}$. C^{X_t} is VCD_Ψ-maximum on $m - 1$ instances and of VCD_Ψ $m - 2$. So, by induction hypothesis, for each set $Y = \{X_{i_1}, \ldots, X_{i_{m-2}}\} = X \setminus \{X_s, X_t\}$, there is a tuple $(l_1, \ldots, l_{m-2}) \in \prod_{1 \le j \le m-2} X_{i_j}$ such that $\mathrm{Forb}((C^{X_t})^{X_s}, Y) = \prod_{1 \le j \le m-2}(X_{i_j} \setminus \{l_j\})$, and hence

$$\mathrm{Forb}((C^{X_s})^{X_t}, Y) = \prod_{1 \le j \le m-2} (X_{i_j} \setminus \{l_j\}).$$

By Lemma 5.29, any forbidden labeling on Y for $(C^{X_s})^{X_t}$ is extended to N_t forbidden labelings on $Y \cup \{X_t\}$ for C^{X_s}. That is, for some $l_t \in X_t$, (X_t, l_t) never occurs in a forbidden labeling on $Y \cup \{X_t\}$. Therefore, for each $Y' = \{X_{i_1}, \ldots, X_{i_{m-2}}, X_t\} = X \setminus \{X_s\}$, there is a tuple $(l_1, \ldots, l_{m-2}, l_t) \in (\prod_{1 \le j \le m-2} X_{i_j}) \times X_t$ such that

$$\mathrm{Forb}(C^{X_s}, Y') = (\prod\nolimits_{1 \le j \le m-2}(X_{i_j} \setminus \{l_j\})) \times (X_t \setminus \{l_t\}).$$

□

Lemma 5.33. *For any set* $Y = \{X_{i_1}, \ldots, X_{i_d}\} \subseteq X \setminus \{X_s\}$ *with* $|Y| = d < |X|$, *there*

is a tuple $(l_1, \ldots, l_d) \in \prod_{1 \le j \le d} X_{i_j}$ *such that* $\text{Forb}(C^{X_s}, Y) = \prod_{1 \le j \le d}(X_{i_j} \setminus \{l_j\})$.

Proof. Let $m = |X|$. We need to prove the claim for the general case, i.e. a VCD_Ψ-maximum class on m instances with $\text{VCD}_\Psi(C) = d < m$. The proof is an induction on m. The base case is $m = d + 1$ or equivalently $d = m - 1$, which is proved in Lemma 5.32. Assume that the claim is true for any $m' < m$. Pick $X_t \in X \setminus \{X_s\}$. By Lemma 5.31, each forbidden labeling of C^{X_s} is an extension of a forbidden labeling of either $(C^{X_s})^{X_t}$ or $C^{X_s} - X_t$.

By Lemma 5.26, $C^{X_s} - X_t = (C - X_t)^{X_s}$ and thus $\text{Forb}(C^{X_s} - X_t) = \text{Forb}((C - X_t)^{X_s})$. $C - X_t$ is VCD_Ψ-maximum on $m - 1$ instances and of VCD_Ψ d. So, by induction hypothesis, for any set $Y = \{X_{i_1}, \ldots, X_{i_d}\} \subseteq X \setminus \{X_s, X_t\}$, there is a tuple $(l_1, \ldots, l_d) \in \prod_{1 \le j \le d} X_{i_j}$ such that $\text{Forb}((C - X_t)^{X_s}, Y) = \prod_{1 \le j \le d}(X_{i_j} \setminus \{l_j\})$ and hence $\text{Forb}(C^{X_s} - X_t, Y) = \prod_{1 \le j \le d}(X_{i_j} \setminus \{l_j\})$. Forbidden labelings of $C^{X_s} - X_t$ are exactly all forbidden labelings of C^{X_s} that do not contain X_t. Therefore, for each $Y = \{X_{i_1}, \ldots, X_{i_d}\} \subseteq X \setminus \{X_s, X_t\}$, there is a tuple $(l_1, \ldots, l_d) \in \prod_{1 \le j \le d} X_{i_j}$ with

$$\text{Forb}(C^{X_s}, Y) = \prod\nolimits_{1 \le j \le d}(X_{i_j} \setminus \{l_j\}). \tag{5.5}$$

Furthermore, $(C^{X_s})^{X_t} = (C^{X_t})^{X_s}$, so $\text{Forb}((C^{X_s})^{X_t}) = \text{Forb}((C^{X_t})^{X_s})$. C^{X_t} is VCD_Ψ-maximum on $m-1$ instances and of VCD_Ψ $d-1$. So, by induction hypothesis, for each set $Y = \{X_{i_1}, \ldots, X_{i_{d-1}}\} \subseteq X \setminus \{X_s, X_t\}$, there is a tuple $(l_1, \ldots, l_{d-1}) \in \prod_{1 \le j \le d-1} X_{i_j}$ such that $\text{Forb}((C^{X_t})^{X_s}, Y) = \prod_{1 \le j \le d-1}(X_{i_j} \setminus \{l_j\})$, which implies that $\text{Forb}((C^{X_s})^{X_t}, Y) = \prod_{1 \le j \le d-1}(X_{i_j} \setminus \{l_j\})$. By Lemma 5.29, any forbidden labeling on Y for $(C^{X_s})^{X_t}$ is extended to N_t forbidden labelings on $Y \cup \{X_t\}$ for C^{X_s}. That is, for some $l_t \in X_t$, (X_t, l_t) never occurs in a forbidden labeling on $Y \cup \{X_t\}$. Therefore, for each $Y' = \{X_{i_1}, \ldots, X_{i_{d-1}}, X_t\} \subseteq X \setminus \{X_s\}$, there is a tuple $(l_1, \ldots, l_{d-1}, l_t) \in$

$(\prod_{1\leq j\leq d-1} X_{i_j}) \times X_t$ such that

$$\text{Forb}(C^{X_s}, Y') = (\prod\nolimits_{1\leq j\leq d-1}(X_{i_j} \setminus \{l_j\})) \times (X_t \setminus \{l_t\}). \tag{5.6}$$

Now, we need to show that if the claim holds for $C^{X_s} - X_t$ and $(C^{X_t})^{X_s}$ then it also holds for C^{X_s}. Note that $\text{Forb}(C^{X_s})$ can be partitioned into the set of forbidden labelings on $Y \subseteq X \setminus \{X_s, X_t\}$, and the set of forbidden labelings on $Y' \subseteq X \setminus \{X_s\}$, with $X_t \in Y'$. By combining this fact with (5.5) and (5.6), we conclude that for each $Y = \{X_{i_1}, \ldots, X_{i_d}\} \subseteq X \setminus \{X_s\}$, there is a tuple $(l_1, \ldots, l_d) \in \prod_{1\leq j\leq d} X_{i_j}$ such that $\text{Forb}(C^{X_s}, Y) = \prod_{1\leq j\leq d}(X_{i_j} \setminus \{l_j\})$. □

Example 5.34. *Consider the class* C^{X_1} *in Table 5.7 of* VCD_{Ψ_G} 1. *For* $Y = \{X_2, X_3\}$,

$$\begin{aligned}\text{Forb}(C^{X_1}, Y) &= \{\{(X_2,1),(X_3,1)\},\{(X_2,1),(X_3,2)\},\{(X_2,2),(X_3,1)\},\\ &\quad \{(X_2,2),(X_3,2)\}\}\\ &= (X_2 \setminus \{0\}) \times (X_3 \setminus \{0\}),\end{aligned}$$

and for $Y' = \{X_3, X_4\}$,

$$\begin{aligned}\text{Forb}(C^{X_1}, Y') &= \{\{(X_3,0),(X_4,1)\},\{(X_3,0),(X_4,2)\},\{(X_3,2),(X_4,1)\},\\ &\quad \{(X_3,2),(X_4,2)\}\}\\ &= (X_3 \setminus \{1\}) \times (X_4 \setminus \{0\}).\end{aligned}$$

The final step of connecting tail concepts to forbidden labelings is accomplished in the next theorem.

Theorem 5.35. *For any* $X_s \in X$, *there is a bipartite graph between the set* $\text{tail}_{X_s}(C)$

Labeled Tail Subroutine (LTS)
Input: a VCD_Ψ-maximum multi-label concept class C over X and $X_s \in X$
Output: a mapping r assigning representatives to all concepts in $\mathrm{tail}_{X_s}(C)$
r=**LTS**(C,X,X_s)

1. **If** $\mathrm{VCD}_\Psi(C) = 0$ **then** $r(c) := \emptyset$; (since $C = \mathrm{tail}_{X_s}(C) = \{c\}$)
   Else if $\mathrm{VCD}_\Psi(C) = |X|$ **then** $r := \emptyset$; (since $C = \prod_{X_i \in X} X_i$ and $\mathrm{tail}_{X_s}(C) = \emptyset$)
   $(*)$ **Else** {pick $t \neq s$; r_1 =**LTS**(C^{X_t},$X \setminus \{X_t\}$,X_s);
   r_2 =**LTS**($C - X_t$,$X \setminus \{X_t\}$,X_s);

2. **For each** $\bar{c} \in \mathrm{tail}_{X_s}(C^{X_t}) \setminus \mathrm{tail}_{X_s}(C - X_t)$ **do**
   For each $c \in \mathrm{tail}_{X_s}(C)$ **do**
   For $i = 0$ **to** N_t **do**
   If $c = \bar{c} \cup \{(X_t, i)\}$ **then** $r(c) := r_1(\bar{c}) \cup \{(X_t, i)\}$;

3. **For each** $\bar{c} \in \mathrm{tail}_{X_s}(C - X_t) \setminus \mathrm{tail}_{X_s}(C^{X_t})$ **do**
   For each $c \in \mathrm{tail}_{X_s}(C)$ **do** { **If** $c - X_t = \bar{c}$ **then** $r(c) := r_2(\bar{c})$; }

4. **For each** $\bar{c} \in \mathrm{tail}_{X_s}(C^{X_t}) \cap \mathrm{tail}_{X_s}(C - X_t)$ **do**
   For each $c \in \mathrm{tail}_{X_s}(C)$ **do**
   For $i = 0$ **to** N_t **do**
   If $c = \bar{c} \cup \{(X_t, i)\}$ **then**
   If $r_1(\bar{c}) \cup \{(X_t, i)\}$ inconsistent with all $\hat{c} \in C^{X_s} \setminus \{c\}$ **then**
   $r(c) := r_1(\bar{c}) \cup \{(X_t, i)\}$;
   Else $r(c) := r_2(\bar{c})$; } (end of $(*)$ Else)

return r;

Algorithm 3: Recursively finding representatives for the tail concepts.

and the set $\mathrm{Forb}(C^{X_s})$*, with an edge between a concept and a forbidden labeling if and only if this forbidden labeling is contained in the concept. All such graphs have a unique matching.*

Proof. (Analogous to the proof of Theorem 10 in [KW07]) Let $m = |X|$ and $\mathrm{VCD}_\Psi(C) = d$. The proof is by double induction on m and d. For $m = d$, there is nothing to prove as $\mathrm{tail}_{X_s}(C) = \mathrm{Forb}(C^{X_s}) = \emptyset$, for all $s \in \{1, \ldots, m\}$. Also, for $d = 0$, C contains a single concept which is always in the tail and gets matched to the empty set.

Suppose that the claim is true for all d' and m' such that $d' \leq d$, $m' \leq m$ and

$m' + d' < m + d$. Pick $X_s, X_t \in X$. First, by Lemma 5.31, each forbidden labeling of C^{X_s} is an extension of a forbidden labeling of either $(C^{X_s})^{X_t}$ or $C^{X_s} - X_t$. Second, by Lemma 5.24(3), any concept in $\text{tail}_{X_s}(C)$ is an extension of either a concept in $\text{tail}_{X_s}(C^{X_t})$ or a concept in $\text{tail}_{X_s}(C - X_t)$. Also, $\text{tail}_{X_s}(C^{X_t})$ is a VCD_Ψ-maximum class of dimension $d-1$ and $\text{tail}_{X_s}(C - X_t)$ is a VCD_Ψ-maximum class of dimension d; both on the instance space $X \setminus \{X_t\}$. So, by induction hypothesis there exists a unique matching between $\text{tail}_{X_s}(C - X_t)$ and $\text{Forb}((C - X_t)^{X_s})$, and also, between $\text{tail}_{X_s}(C^{X_t})$ and $\text{Forb}((C^{X_s})^{X_t})$. We combine these two matchings to form a matching for $\text{tail}_{X_s}(C)$. This is done in steps 2, 3 and 4 in Algorithm 3, as described in the following paragraphs.

Concepts in $\text{tail}_{X_s}(C^{X_t}) \setminus \text{tail}_{X_s}(C - X_t)$ are matched to the forbidden labelings for $(C^{X_s})^{X_t}$ of size $d-1$. Consider a concept $\bar{c} \in \text{tail}_{X_s}(C^{X_t}) \setminus \text{tail}_{X_s}(C - X_t)$ which gets matched to a forbidden labeling F for $(C^{X_s})^{X_t}$. On the one hand, by Lemma 5.24(1), there are N_t concepts $c_i \in \text{tail}_{X_s}(C)$ such that $c_i - X_t = \bar{c}$, for $i \in \{1, \ldots, N_t\}$. W.l.o.g., assume that for $i \in \{1, \ldots, N_t\}$, $c_i = \bar{c} \cup \{(X_t, i)\}$, that is, $c_0 = \bar{c} \cup \{(X_t, 0)\}$ is not in $\text{tail}_{X_s}(C)$ and thus, $c_0 \in C^{X_s}$. Since F is contained in $\bar{c}$, it is contained in c_i, $i \in \{0, \ldots, N_t\}$, too. On the other hand, by Lemma 5.29, F can be extended to N_t forbidden labelings for C^{X_s}. Clearly, $F \cup \{(X_t, 0)\}$ is not a forbidden labeling for C^{X_s}, as it is contained in c_0 and $c_0 \in C^{X_s}$. So, for $i \in \{1, \ldots, N_t\}$, $F \cup \{(X_t, i)\}$ is a forbidden labeling for C^{X_s} and thus can be matched to c_i. Therefore, any matching of $\text{tail}_{X_s}(C^{X_t}) \setminus \text{tail}_{X_s}(C - X_t)$ can be transferred to N_t matchings in $\text{tail}_{X_s}(C)$ (Step 2 of Algorithm 3).

Concepts in $\text{tail}_{X_s}(C - X_t) \setminus \text{tail}_{X_s}(C^{X_t})$ are matched to the forbidden labelings for $(C - X_t)^{X_s}$ of size d. Consider a concept $\bar{c} \in \text{tail}_{X_s}(C - X_t) \setminus \text{tail}_{X_s}(C^{X_t})$ which gets matched to a forbidden labeling F for $(C - X_t)^{X_s}$. By Lemma 5.24(2), $\bar{c}$ corresponds to a concept $\bar{c} \cup \{(X_t, l)\}$ in $\text{tail}_{X_s}(C)$, for some $l \in X_t$. Since F gets matched to $\bar{c}$, F

is contained in $\bar{c}$ and thus is contained in $\bar{c} \cup \{(X_t, l)\}$. Moreover, by Corollary 5.27, any forbidden labeling of $(C - X_t)^{X_s}$ is also a forbidden labeling of C^{X_s}, that is, F is also a forbidden labeling for $(C - X_t)^{X_s}$. So, $\bar{c} \cup \{(X_t, l)\}$ and F are matched in $\text{tail}_{X_s}(C)$ and consequently, each matching of $\text{tail}_{X_s}(C - X_t) \setminus \text{tail}_{X_s}(C^{X_t})$ can be transferred to a matching in $\text{tail}_{X_s}(C)$ (Step 3 of Algorithm 3).

Each concept $\bar{c} \in \text{tail}_{X_s}(C^{X_t}) \cap \text{tail}_{X_s}(C - X_t)$ is matched to a forbidden labeling F for $(C^{X_s})^{X_t}$ of size $d-1$ in one setting and also, is matched to a forbidden labeling F' for $(C - X_s)^{X_t}$ of size d in another setting. By Lemma 5.29, F can be extended to N_t forbidden labelings for C^{X_s}. W.l.o.g., assume that $F_i = F \cup \{(X_t, i)\}$, for all $i \in \{1, \ldots, N_t\}$, is a forbidden labeling for C^{X_s}. Clearly, F does not belong to $\{F_1, \ldots, F_{N_t}\}$, as it is a sample on $X \setminus \{X_s, X_t\}$. As discussed in the previous paragraph, there are also N_t concepts $c_i \in \text{tail}_{X_s}(C)$ such that $c_i = \bar{c} \cup \{(X_t, i)\}$ and F_i is matched to c_i, for $i \in \{1, \ldots, N_t\}$. On the other hand, by Corollary 5.27, F' is a forbidden labeling for C^{X_s} and thus gets matched to $c_0 = \bar{c} \cup \{(X_t, 0)\}$ in $\text{tail}_{X_s}(C)$ as explained before. (Step 4 of Algorithm 3).

Finally, we need to verify that the proposed perfect matching is also unique. To do this, we will show that any matching for $\text{tail}_{X_s}(C)$ can be used to construct matchings for $\text{tail}_{X_s}(C^{X_t})$ and $\text{tail}_{X_s}(C - X_t)$ with the property that two different matchings for $\text{tail}_{X_s}(C)$ will result in two different matchings for $\text{tail}_{X_s}(C^{X_t})$ or two different matchings for $\text{tail}_{X_s}(C - X_t)$, which contradicts the induction hypothesis.

First, consider any concept $c \in \text{tail}_{X_s}(C)$, such that $c - X_t \in \text{tail}_{X_s}(C^{X_t}) \setminus \text{tail}_{X_s}(C - X_t)$. That is, $c - X_t \in C - X_t$, but $c - X_t \notin \text{tail}_{X_s}(C - X_t)$, and thus $(c - X_t) - X_s \in (C - X_t)^{X_s}$. So, $(c - X_t) - X_s$ cannot contain a forbidden labeling for $(C - X_t)^{X_s}$ and consequently, $c - X_t$ contains no forbidden labeling for $(C - X_t)^{X_s}$. We claim that any forbidden labeling for C^{X_s} that is a subset of c must contain X_t. More precisely, consider any $Y \subseteq X \setminus \{X_s\}$ with $|Y| = d$, such that $c|_Y$ is

a forbidden labeling for C^{X_s}. We claim that $X_t \in Y$. Otherwise, $c|_Y$ is a forbidden labeling for $C^{X_s} - X_t$ and, by Lemma 5.26, is a forbidden labeling for $(C - X_t)^{X_s}$, which contradicts the fact that $c - X_t$ contains no forbidden labeling for $(C - X_t)^{X_s}$.

Second, consider any concept $c \in \text{tail}_{X_s}(C)$, such that $c - X_t \in \text{tail}_{X_s}(C - X_t) \setminus \text{tail}_{X_s}(C^{X_t})$. That is, $c - X_t \in C - X_t$, but $c - X_t \notin \text{tail}_{X_s}(C^{X_t})$, and thus $(c - X_t) - X_s \in (C^{X_s})^{X_t}$. So, $(c - X_t) - X_s$ cannot contain a forbidden labeling for $(C^{X_s})^{X_t}$ and consequently, $c - X_t$ contains no forbidden labeling for $(C^{X_s})^{X_t}$. We claim that any forbidden labeling for C^{X_s} that is a subset of c cannot contain X_t. In fact, any forbidden labeling for C^{X_s} of size d that contains X_t can also be a forbidden labeling of size $d - 1$ for $(C^{X_s})^{X_t}$ by removing (X_t, l) from it. So, our claim follows from the fact that $c - X_t$ contains no forbidden labeling for $(C^{X_s})^{X_t}$.

To summarize the last two paragraphs, we showed that if a concept $c \in \text{tail}_{X_s}(C)$ is matched to a forbidden labeling containing X_t, then $c - X_t \in \text{tail}_{X_s}(C^{X_t})$, and if it is matched to a forbidden labeling not containing X_t, then $c - X_t \in \text{tail}_{X_s}(C - X_t)$. Hence, a matching for $\text{tail}_{X_s}(C)$ splits into a matching for $\text{tail}_{X_s}(C^{X_t})$ and a matching for $\text{tail}_{X_s}(C - X_t)$, and consequently, the existence of two matchings for $\text{tail}_{X_s}(C)$ implies the existence of two matchings for $\text{tail}_{X_s}(C^{X_t})$ or two matchings for $\text{tail}_{X_s}(C - X_t)$. □

Let $\text{LRep}_d(X) \subset \text{LRep}_{\leq d}(X)$ denote the set of labeled representatives of size d that are constructed from Algorithm 1. The following corollary shows that there is a representation mapping between $\text{tail}_{X_s}(C)$ and $\text{LRep}_d(X \setminus \{X_s\})$.

Corollary 5.36. *Algorithm 3 returns a representation mapping between* $\text{tail}_{X_s}(C)$ *and some* $\text{LRep}_d(X \setminus \{X_s\})$.

Proof. By Theorem 5.35, there is a unique consistent bijection r between $\text{tail}_{X_s}(C)$

and $\mathrm{Forb}(C^{X_s})$. From

$$\mathrm{Forb}(C^{X_s}) = \bigcup_{Y \subseteq X \setminus \{X_s\},\ |Y|=d} \mathrm{Forb}(C^{X_s}, Y)$$

and Lemma 5.33, we conclude that $\mathrm{Forb}(C^{X_s})$ equals some $\mathrm{LRep}_d(X \setminus \{X_s\})$, and thus r is a consistent bijection between $\mathrm{tail}_{X_s}(C)$ and $\mathrm{LRep}_d(X \setminus \{X_s\})$. To finish the proof, we need to show that there is no clash between the concepts in $\mathrm{tail}_{X_s}(C)$ w.r.t. r. Assume that there exist two concepts $c_1, c_2 \in \mathrm{tail}_{X_s}(C)$ that clash w.r.t. r, that is, $r(c_1) = r_1$, $r(c_2) = r_2$, $r_1 \subseteq c_2$ and $r_2 \subseteq c_1$. Then we can swap the representatives of c_1 and c_2 and set $r(c_1) = r_2$, $r(c_2) = r_1$ and create a new valid matching. This contradicts the uniqueness of the matching in Theorem 5.35. □

Theorem 5.37. *Algorithm 2 returns a representation mapping between the* VCD_Ψ*-maximum class* C *on* X *with* $\mathrm{VCD}_\Psi(C) = d$ *and some* $\mathrm{LRep}_{\leq d}(X)$.

Proof. (Analogous to the proof of Theorem 11 in [KW07]) Proof by induction on d. For $d = 0$, the class has a single concept which is mapped to the empty set. Otherwise, Algorithm 2 first finds the representatives for C^{X_s}, for some $X_s \in X$, and extends them to the representatives for C. The algorithm then finds the representatives for $\mathrm{tail}_{X_s}(C)$ by calling Algorithm 3.

For the induction step, assume that Algorithm 2 finds a representation mapping $\tilde{r}$ between C^{X_s} and $\mathrm{LRep}_{\leq d-1}(X \setminus \{X_s\})$.

Bijection condition: As shown in step 2 of Algorithm 2, $\tilde{r}$ extends to a bijective mapping between $C^{X_s} \times \{i\}$ and the set of all labeled representatives of size d that contain (X_s, i), for all $i \in \{1, \ldots N_s\}$, and between $C^{X_s} \times \{0\}$ and the set of all labeled representatives of size $d-1$ on $X \setminus \{X_s\}$. By Corollary 5.36, Algorithm 3 returns a bijection between $\mathrm{tail}_{X_s}(C)$ and the set of all labeled representatives of size d on

$X \setminus \{X_s\}$. Hence, Algorithm 2 returns a bijection between C and some $\mathrm{LRep}_{\leq d}(X)$.

Non-clashing condition: By the induction hypothesis there cannot be a clash between the concepts in C^{X_s}, and therefore, there cannot be a clash internally within the concepts in $C^{X_s} \times \{i\}$, for each $i \in X_s$. On the one hand, clashes between concepts $c_i \in C^{X_s} \times \{i\}$ and $c_j \in C^{X_s} \times \{j\}$, for $i, j \in \{1, \ldots, N_s\}$, $i \neq j$, cannot occur as $(X_s, i) \in r(c_i)$ and $(X_s, j) \in r(c_j)$, and consequently, $r(c_i) \not\subseteq c_j$ and $r(c_j) \not\subseteq c_i$. On the other hand, clashes between the concepts $c_i \in C^{X_s} \times \{i\}$, $i \in \{1, \ldots, N_s\}$ and $c_0 \in C^{X_s} \times \{0\}$ cannot occur as $(X_s, i) \in r(c_i)$ and thus, $r(c_i) \not\subseteq c_0$. Also, no clashes occur between $\mathrm{tail}_{X_s}(C)$ and $C^{X_s} \times X_s$, since the concepts in $\mathrm{tail}_{X_s}(C)$ are mapped to forbidden labels for C^{X_s}. Finally, by Corollary 5.36, no clashes occur between the concepts in $\mathrm{tail}_{X_s}(C)$. □

Now we have all the pieces in place for verifying Theorem 5.19, which states that if VCD_Ψ fulfills the reduction property then any VCD_Ψ-maximum class C has a tight sample compression scheme of size $\mathrm{VCD}_\Psi(C)$.

Proof of Theorem 5.19. By Theorem 5.37, there exists a representation mapping r for C, i.e., a consistent bijection between C and some $\mathrm{LRep}_{\leq d}(X)$ in which no two concepts clash. Condition (i) of Definition 5.18 is then obvious as $|\mathrm{LRep}_{\leq d}(X)| = |C|$, and condition (ii) follows from the non-clashing property of r and Lemma 5.21. □

Example 5.38. *Table 5.5 shows the representatives computed for a* VCD_{Ψ_G}*-maximum class. Notice the missing labeling property and the tightness of the scheme. Table 5.6 shows the representatives for* C^{X_1} *and* $\mathrm{tail}_{X_1}(C)$ *separately; the table on the left is computed using Algorithm 3.*

*To illustrate the "**for each**" block of Algorithm 2, see Table 5.7.* c_1'' *is the only*

concept in $(C^{X_1})^{X_2}$ *with* $r(c_1'') = \emptyset$*. So,*

$$r(c_1'' \cup \{(X_2, 1)\}) = r(c_2') := r(c_1'') \cup \{(X_2, 1)\} = \{(X_2, 1)\}$$

and

$$r(c_1'' \cup \{(X_2, 2)\}) = r(c_3') := r(c_1'') \cup \{(X_2, 2)\} = \{(X_2, 2)\}.$$

To illustrate Steps 2 *and* 3 *of Algorithm 3, see Table 5.8. Assume we want representatives for* $\mathrm{tail}_{X_2}(C^{X_1})$ *and we recursively found the representative* $r(c_1^1) = \emptyset$ *for the (only) concept in* $\mathrm{tail}_{X_2}((C^{X_1})^{X_3})$*.* $r(c_1^1)$ *is extended to* $r(c_1^0)$ *for* $c_1^0 \in \mathrm{tail}_{X_2}(C^{X_1})$*, that is,* $r(c_1^0) = r(c_1^1) \cup \{(X_3, 1)\} = \{(X_3, 1)\}$ *because* $c_1^0 = c_1^1 \cup \{(X_3, 1)\}$*. Similarly,* $r(c_1^1)$ *is extended to* $r(c_2^0)$ *for* $c_2^0 \in \mathrm{tail}_{X_2}(C^{X_1})$*, that is,* $r(c_2^0) = r(c_1^1) \cup \{(X_3, 2)\} = \{(X_3, 2)\}$ *because* $c_2^0 = c_1^1 \cup \{(X_3, 2)\}$*. Next assume that we recursively found the representatives for* $\mathrm{tail}_{X_2}(C^{X_1} - X_3)$*.* $r(c_1^2)$ *for* $c_1^2 \in \mathrm{tail}_{X_2}(C^{X_1} - X_3)$ *is extended to* $r(c_3^3)$ *for* $c_3^0 \in \mathrm{tail}_{X_2}(C^{X_1})$ *because* $c_3^0 - X_3 = c_1^2$*. Similarly,* $r(c_2^2)$ *for* $c_2^2 \in \mathrm{tail}_{X_2}(C^{X_1} - X_3)$ *is extended to* $r(c_4^0)$ *for* $c_4^0 \in \mathrm{tail}_{X_2}(C^{X_1})$ *because* $c_4^0 - X_3 = c_2^2$*.*

Remark 5.39. *In Chapter 4 (Table 4.1), we presented an example of a* $\mathrm{VCD}_{\Psi_\mathrm{P}}$*-maximum class that does not fulfill the reduction property. This class does have a tight compression scheme, as shown in Table 5.10. Hence, the reduction property for* VCD_Ψ *is not a necessary condition for the existence of a tight compression scheme for* VCD_Ψ*-maximum classes.*

Remark 5.40. *We provided an example in Chapter 4 (Table 4.4) of a* $\mathrm{VCD}_{\Psi_\mathrm{N}}$*-maximum class that does not fulfill the reduction property. Interestingly, this class has* no *tight compression scheme. Note that the Natarajan-dimension violates both premises of Theorems 5.5 and 5.19 — it violates the reduction property, and it is not based on a spanning family.*

c	X_1	X_2	X_3	compression set
$\mathbf{c_1}$	**0**	**0**	**0**	$\emptyset$
$\mathbf{c_2}$	**0**	**0**	**1**	$(X_3, 1)$
c_3	0	1	0	$(X_2, 1)$
c_4	0	1	1	$(X_2, 1), (X_3, 1)$
c_5	1	0	0	$(X_1, 1)$
c_6	1	0	1	$(X_1, 1), (X_3, 1)$
c_7	1	1	1	$(X_1, 1), (X_2, 1)$
c_8	0	2	0	$(X_2, 2)$
c_9	0	2	1	$(X_2, 2), (X_3, 1)$
c_{10}	0	2	2	$(X_1, 0), (X_3, 2)$
c_{11}	2	0	0	$(X_1, 2)$
c_{12}	2	0	1	$(X_1, 2), (X_3, 1)$
c_{13}	2	0	2	$(X_3, 2)$
c_{14}	2	1	1	$(X_1, 2), (X_2, 1)$
c_{15}	2	1	2	$(X_2, 1), (X_3, 2)$
c_{16}	2	2	1	$(X_1, 2), (X_2, 2)$
c_{17}	2	2	2	$(X_2, 2), (X_3, 2)$
$\mathbf{c_{18}}$	**1**	**2**	**1**	$(X_1, 1), (X_2, 2)$
$\mathbf{c_{19}}$	**1**	**2**	**2**	$(X_1, 1), (X_3, 2)$

Table 5.10: Maximum class C of $\mathrm{VCD}_{\Psi_{\mathrm{P}}}$ 2 from Table 4.1 with a tight compression scheme.

5.4 Connection to the One-inclusion Hypergraph

Assume that VCD_{Ψ} fulfills the reduction property. As in the binary case [KW07], we also explore a connection between the one-inclusion hypergraph and a representation mapping r for a VCD_{Ψ}-maximum class C. We show that every representation mapping for a VCD_{Ψ}-maximum class C maps any concept $c \in C$ to a sample set $S \subseteq c$, such that the instances appearing in S label the incident hyperedges to c in the one-inclusion hypergraph for C.

Following Kuzmin and Warmuth, for a hyperedge e labeled with an instance X_t, we say that e *charges* a concept $c \in e$ iff $X_t \in X(r(c))$, i.e., $r(c)$ contains an example (X_t, l), for some $l \in X_t$.

The next proposition connects any hyperedge to the representatives of its incident concepts. The corresponding result for the binary case is that for any representation

mapping r for a VCD-maximum class C, any edge $e = (c, c')$ labeled with X_t, for some $X_t \in X$, in the one-inclusion graph of C lies exactly in one of the representatives $r(c)$ or $r(c')$ [KW07]. Note that in the multi-label case, because of the reduction property, every hyperedge for a VCD_Ψ-maximum class contains exactly $N_t + 1$ concepts (where X_t is the label of the hyperedge).

Proposition 5.41. *Let C be VCD_Ψ-maximum of dimension d and r be a representation mapping between C and some $\text{LRep}_{\leq d}(X)$. Let $G(C) = (V, E)$ be the one-inclusion hypergraph for C. Then for any hyperedge $e = \{c_0, c_1, \ldots, c_{N_t}\}$ labeled with X_t in $E(G)$, $t \in [m]$, e charges exactly N_t incident concepts to e.*

Proof. The proof is a straightforward extension from the similar result in the binary case. First, we show that for any hyperedge e labeled with X_t, $t \in [m]$, there are at least N_t concepts in e that are charged with e. For purposes of contradiction, assume that $X_t \notin X(r(c_p))$ and $X_t \notin X(r(c_q))$, for $c_p, c_q \in e$, $p \neq q$. Then $r(c_p) \subseteq c_q$ and $r(c_q) \subseteq c_p$, since $c_p - X_t = c_q - X_t$. This contradicts the non-clashing property of r. So, there are at least N_t concepts $c_{i_1}, \ldots, c_{i_{N_t}} \in e$ for which $X_t \in X(r(c_{i_j}))$, $j \in \{1, \ldots, N_t\}$. Next, we show that there are exactly N_t such concepts in e.

Let $\text{Chg}(e, X_t)$ denote the set of all incident concepts to e that are charged by e, where e is a hyperedge with the label X_t, $t \in [m]$. So far, we know that $\text{Chg}(e, X_t) \geq N_t$. Since C is VCD_Ψ-maximum, there are $|C^{X_t}|$ hyperedges labeled with X_t. First, for any pair of hyperedges $e, e' \in E$ with the label X_t, $e \neq e'$, $e \cap e' = \emptyset$. Second, each concept in C corresponds to a unique representative and no two concepts in C have the same representatives. So, for each $t \in [m]$, the total number of charges by all the hyperedges labeled with X_t, $\sum_{e \in E} \text{Chg}(e, X_t)$, is lower-bounded by

$$\sum_{e \in E} \text{Chg}(e, X_t) \geq N_t |C^{X_t}| = N_t \Phi_{d-1}(N_1, \ldots, N_{t-1}, N_{t+1}, \ldots, N_m).$$

Consequently, the total number of charges by all hyperedges in E is lower-bounded as follows:

$$\begin{aligned}
\sum_{1\le t\le m}\sum_{e\in E}\mathrm{Chg}(e,X_t) &\ge \sum_{1\le t\le m} N_t\Phi_{d-1}(N_1,\ldots,N_{t-1},N_{t+1},\ldots,N_m)\\
&= N_1(1+\sum_{2\le i\le m} N_i+\sum_{2\le i_1<i_2\le m} N_{i_1}N_{i_2}+\cdots\\
&\quad+\sum_{2\le i_1<\cdots<i_{d-1}\le m} N_{i_1}\cdots N_{i_{d-1}})+\cdots\\
&\quad+N_m(1+\sum_{1\le i\le m-1} N_i+\sum_{1\le i_1<i_2\le m-1} N_{i_1}N_{i_2}+\cdots\\
&\quad+\sum_{1\le i_1<\cdots<i_{d-1}\le m-1} N_{i_1}\cdots N_{i_{d-1}})\\
&= \sum_{1\le i\le m} N_i+2\sum_{1\le i_1<i_2\le m} N_{i_1}N_{i_2}+\cdots\\
&\quad+d\sum_{1\le i_1<\cdots<i_d\le m} N_{i_1}\cdots N_{i_d}.
\end{aligned}$$

To obtain an upper-bound, note that each concept $c\in C$ can be charged at most $|r(c)|$ times by $|r(c)|$ many different edges with different labels. So, the total number of charges by all hyperedges in E is upper bounded by the total size of all representatives in $\mathrm{LRep}_{\le d}(X)$. That is,

$$\begin{aligned}
\sum_{1\le t\le m}\sum_{e\in E}\mathrm{Chg}(e,X_t) &\le \sum_{S\in\mathrm{LRep}_{\le d}(X)}|S|\\
&= \sum_{1\le i\le m} N_i+2\sum_{1\le i_1<i_2\le m} N_{i_1}N_{i_2}+\cdots\\
&\quad+d\sum_{1\le i_1<\cdots<i_d\le m} N_{i_1}\cdots N_{i_d}.
\end{aligned}$$

Hence,

$$\begin{aligned}\sum_{1\leq t\leq m}\sum_{e\in E}\mathrm{Chg}(e,X_t) &= \sum_{S\in \mathrm{LRep}_{\leq d}(X)}|S| \\ &= \sum_{1\leq t\leq m} N_t\Phi_{d-1}(N_1,\ldots,N_{t-1},N_{t+1},\ldots,N_m)\end{aligned} \tag{5.7}$$

and consequently,

$$\sum_{e\in E}\mathrm{Chg}(e,X_t) = N_t\Phi_{d-1}(N_1,\ldots,N_{t-1},N_{t+1},\ldots,N_m) = N_t|C^{X_t}|, \quad \text{for each } t\in[m].$$

Therefore, $\mathrm{Chg}(e,X_t) = N_t$ and each hyperedge e labeled with X_t charges exactly N_t incident concepts to e. □

Corollary 5.42. *Let r be a representation mapping for the VCD_{Ψ_G}-maximum class C. Then for each $c\in C$, $X(r(c))$ is a subset of the set of labels of incident hyperedges on c.*

Proof. Proposition 5.41 along with (5.7) shows that for each concept c and for each example $(X_t,l)\in r(c)$, $t\in[m]$ and $l\in X_t$, there exists a hyperedge incident to c and labeled with X_t which charges c. □

5.5 Sample compression for classes of VCD_{Ψ} 1

For binary concept classes, compression schemes of size d for maximum classes of VC-dimension d, like the VC Scheme proposed by Floyd and Warmuth [FW95], immediately yield compression schemes of size 1 for all classes of VC-dimension 1. This is because every binary class of VC-dimension 1 is contained in a binary VCD-maximum class of VC-dimension 1 [WW87]. In other words, in the binary case, every maximal

class of VC-dimension 1 is VCD-maximum. The term "maximal" refers to a class whose VC-dimension increases if any concept is added to it. In the multi-label case, a concept class is called VCD_Ψ-maximal w.r.t. a family of mappings $\Psi = \Psi_1 \times \cdots \times \Psi_m$ if adding any new concept to the class increases its VCD_Ψ-dimension.

$c \in \hat{C}$	X_1	X_2
c_0	0	0
c_1	1	1
c_3	2	2

Table 5.11: A $\mathrm{VCD}_{\Psi_\mathrm{G}}$-maximal class of $\mathrm{VCD}_{\Psi_\mathrm{G}}$ 1 that is not $\mathrm{VCD}_{\Psi_\mathrm{G}}$-maximum.

An obvious idea for proving that compression schemes of size 1 exist for multi-label classes C with $\mathrm{VCD}_\Psi(C) = 1$ would be to prove that the latter are contained in VCD_Ψ-maximum classes of dimension 1, and then to apply Theorem 5.5 or Theorem 5.19. However, this approach is fruitless, since it does not work for all VCD_Ψ 1 classes, where Ψ is the direct product of spanning families of mappings, even if VCD_Ψ fulfills the reduction property. In particular, there is a $\mathrm{VCD}_{\Psi_\mathrm{G}}$-maximal class C such that $\mathrm{VCD}_{\Psi_\mathrm{G}}(C) = 1$ and C is not $\mathrm{VCD}_{\Psi_\mathrm{G}}$-maximum. As an example, consider the class $\hat{C} \subseteq \{0,1,2\} \times \{0,1,2\}$ in Table 5.11. Clearly, $\mathrm{VCD}_{\Psi_\mathrm{G}}(\hat{C}) = \mathrm{VCD}_{\Psi^*}(\hat{C}) = 1$ and $\hat{C}$ is too small to be $\mathrm{VCD}_{\Psi_\mathrm{G}}$-maximum. However, it is $\mathrm{VCD}_{\Psi_\mathrm{G}}$-maximal.

One can see that the class $\hat{C}$ in Table 5.11 is not $\mathrm{VCD}_{\Psi_\mathrm{P}}$ or $\mathrm{VCD}_{\Psi_\mathrm{N}}$-maximal. This means, for different family of mappings, we need to study VCD_Ψ 1 classes separately.

5.5.1 The Graph Dimension

We will prove that, despite the changes in structural properties when compared to the binary case, every multi-label class C with $\mathrm{VCD}_{\Psi_\mathrm{G}}(C) = 1$ has a sample compression scheme of size 1.

Remark 5.43. *Our approach for* VCD_{Ψ_G} 1 *classes is an alternative proof for the existence of compression schemes of size* 1 *for* VCD 1 *classes in the binary case.*

Recall that a sample S is a teaching set for a concept c in a class C, if c is the only concept from C that is consistent with S, and the teaching dimension of c in C is the size of the smallest teaching set for c.

Lemma 5.44. *Let* $\mathrm{VCD}_{\Psi_G}(C) = 1$. *Then for any* $X_i, X_j \in X$ *with* $i \neq j$, *there is at most one concept in* $C|_{\{X_i,X_j\}}$ *with teaching dimension* 2 *w.r.t.* $C|_{\{X_i,X_j\}}$.

Proof. If there is no concept in $C|_{\{X_i,X_j\}}$ with teaching dimension 2, we are done. Assume some $c \in C|_{\{X_i,X_j\}}$ fulfills $\mathrm{TD}(c, C|_{\{X_i,X_j\}}) = 2$. W.l.o.g., $c = \{(X_i, 0), (X_j, 0)\}$ and $\mathrm{TS}(c, C|_{\{X_i,X_j\}}) = \{\{(X_i, 0), (X_j, 0)\}\}$. Since no sample of size 1 can be a minimal teaching set for c in $C|_{\{X_i,X_j\}}$, there must exist concepts $c_\alpha, c_\beta \in C|_{\{X_i,X_j\}}$ with $c(X_i) = c_\beta(X_i)$ and $c(X_j) = c_\alpha(X_j)$. That is, $c_\alpha = \{(X_i, a), (X_j, 0)\}$ and $c_\beta = \{(X_i, 0), (X_j, b)\}$ for some nonzero $a \in X_i$ and $b \in X_j$.

$c \in C\|_{\{X_i,X_j\}}$	X_i	X_j
c	0	0
c_α	a	0
c_β	0	b
$\vdots$		

Table 5.12: Illustration of the proof of Lemma 5.44.

Now, we consider all other possible concepts $c' = \{(X_i, a'), (X_j, b')\}$ that can exist in $C|_{\{X_i,X_j\}}$. Based on the possible values for a' and b', we consider three groups of concepts:

Group 1 : $a' \in X_i \setminus \{0\}$ and $b' \in X_j \setminus \{0\}$. Let $\psi_1 : X_i \to \{0, 1\}$, $\psi_2 : X_j \to \{0, 1\}$ and $\overline{\psi} = (\psi_1, \psi_2)$ such that $\psi_1(x) = \psi_2(x) = 0$ if $x = 0$, and $\psi_1(x) = \psi_2(x) = 1$ if $x \neq 0$. Having $c, c_a, c_b, c' \in C|_{\{X_i,X_j\}}$, it is easy to see that $\{(0, 0), (1, 0), (0, 1), (1, 1)\} \subseteq$

$\overline{\psi}(C|_{\{X_i,X_j\}})$. This contradicts the assumption that $\mathrm{VCD}_{\Psi_G}(C) = 1$. So, this case cannot occur.

Group 2 : $a' = 0$ and $b' \in X_j \setminus \{0, b\}$. Since case 1 is not possible, any such concept has teaching dimension 1. In particular, $\{(X_j, b')\} \in \mathrm{TS}(c', C|_{\{X_i,X_j\}})$.

Group 3 : $a' \in X_i \setminus \{0, a\}$ and $b' = 0$. Again, since case 1 is not possible, any such concept has teaching dimension 1. In particular, $\{(X_i, a')\} \in \mathrm{TS}(c', C|_{\{X_i,X_j\}})$.

Since Group 1 is empty, we conclude that for any concept $c' \in C|_{\{X_i,X_j\}} \setminus \{c, c_\alpha, c_\beta\}$, $c'(X_i) \neq a$ and $c'(X_j) \neq b$. Thus, $\{(X_i, a)\} \in \mathrm{TS}(c_\alpha, C|_{\{X_i,X_j\}})$ and $\{(X_j, b)\} \in \mathrm{TS}(c_\beta, C|_{\{X_i,X_j\}})$.

Hence, there is no other concept in $C|_{\{X_i,X_j\}}$ with teaching dimension 2. □

This result does not generalize to the case when $\mathrm{VCD}_{\Psi_G}(C) = 2$, not even for binary classes. For example, as shown in Table 5.13, the VCD-maximum class of VC-dimension 2 over 3 instances that corresponds to the class of all sets of size at most 2 has 4 concepts $c_1, c_2, c_3, c_4 \in C|_{\{X_1,X_2,X_3\}} = C$ of teaching dimension 3, namely the empty concept $(0, 0, 0)$ and the singletons $(1, 0, 0)$, $(0, 1, 0)$, and $(0, 0, 1)$. For $\mathrm{VCD}_{\Psi_G}(C) = 1$, Lemma 5.44 will help us to compress a set of two examples to one example.

$c \in C$	X_1	X_2	X_3
c_1	0	0	0
c_2	0	0	1
c_3	0	1	0
c_4	1	0	0
c_5	0	1	1
c_6	1	0	1
c_7	1	1	0

Table 5.13: A concept class of VCD 2 with 4 concepts of teaching dimension 3.

Definition 5.45. *Let C be a concept class and let S be a C-realizable sample. For*

$X_i, X_j \in X(S)$ *with* $i \neq j$*, we say*

(1) $(X_i, l_i) \in S$ explicitly implies $(X_j, l_j) \in S$ *if* $\{(X_i, l_i)\} \in \mathrm{TS}(S|_{\{X_i,X_j\}}, C|_{\{X_i,X_j\}})$.

(2) $(X_i, l_i) \in S$ implicitly implies $(X_j, l_j) \in S$ *if* $\mathrm{TS}(S|_{\{X_i,X_j\}}, C|_{\{X_i,X_j\}}) = \{S|_{\{X_i,X_j\}}\}$.

$(X_i, l_i) \in S$ implies $(X_j, l_j) \in S$ *if it explicitly or implicitly implies* (X_j, l_j)*. Moreover,* (X_i, l_i) uniquely *implies* (X_j, l_j) *if for any sample* $S' \supseteq \{(X_i, l_i), (X_j, l')\}$, $l' \neq l_j$, *consistent with some concept in* C, (X_i, l_i) *does not imply* $(X_j, l') \in S'$*. An example* $(X_i, l_i) \in S$ *is called a* representative *for* S*, if every example in* S *is uniquely implied by* (X_i, l_i)*.*

$c \in C$	X_1	X_2	X_3	X_4	representatives
c_1	2	2	0	0	$\{(X_2, 2)\}, \{(X_3, 0)\}$
c_2	2	0	2	0	$\{(X_1, 2)\}, \{(X_2, 0)\}, \{(X_3, 2)\}$
c_3	2	1	1	1	$\{(X_4, 1)\}$
c_4	2	1	1	2	$\{(X_4, 2)\}$
c_5	1	0	2	0	$\{(X_1, 1)\}$

Table 5.14: A concept class of VCD_{Ψ_G} 1 and its compression sets of size 1.

Example 5.46. *Consider the class in Table 5.14. For* $S = \{(X_1, 2), (X_2, 2), (X_4, 0)\}$ *consistent with* c_1*,* $(X_2, 2)$ *explicitly implies* $(X_1, 2)$ *and* $(X_1, 2)$ *implicitly implies* $(X_4, 0)$*. For* $S' = \{(X_1, 2), (X_2, 1), (X_3, 1)\}$ *consistent with* c_4*,* $(X_2, 1)$ *explicitly implies* $(X_1, 2)$ *and* $(X_2, 1)$ *explicitly implies* $(X_3, 1)$*.* $(X_2, 1)$ *uniquley implies* $(X_1, 2)$ *and* $(X_3, 1)$*, indeed, so* $(X_2, 1)$ *is a representative for* S'*.*

Using Definition 5.45, we obtain a simple lemma.

Lemma 5.47. *Let* S *be a* C*-realizable sample and* $(X_i, l_i), (X_j, l_j) \in S$*, such that* (X_i, l_i) *implies* (X_j, l_j)*. If* $\mathrm{VCD}_{\Psi_G}(C) = 1$ *then* (X_i, l_i) *uniquely implies* (X_j, l_j)*.*

Proof. Let $e_i = (X_i, l_i)$, and $e_j = (X_j, l_j)$. First, we consider the case when e_i explicitly implies e_j. Then $\{e_i\} \in \mathrm{TS}(\{e_i, e_j\}, C|_{\{X_i,X_j\}})$ and thus there is no sample

$S' \supseteq \{(X_i, l_i), (X_j, l')\}$, with $l' \neq l_j$, consistent with some concept in C. Hence, e_i uniquely implies e_j.

Second, we consider the case when e_i implicitly implies e_j. That is, none of $\{e_i\}$ or $\{e_j\}$ is a minimal teaching set for $\{e_i, e_j\}$ in $C|_{\{X_i,X_j\}}$. So, for every sample $S' \supseteq \{(X_i, l_i), (X_j, l')\}$ consistent with some concept in C, (X_i, l_i) does not explicitly imply (X_j, l'). Further, by Lemma 5.44, $\{e_i, e_j\}$ is the only sample in $C|_{\{X_i,X_j\}}$ that has teaching dimension 2 and all other samples in $C|_{\{X_i,X_j\}}$ have a minimal teaching set of size 1. So, (X_i, l_i) cannot imply any example other than (X_j, l_j), or equivalently, e_i uniquely implies e_j. □

Corollary 5.48. *Let C be a concept class and let S be a C-realizable sample and $(X_i, l_i), (X_j, l_j) \in S$. If $\mathrm{VCD}_{\Psi_G}(C) = 1$, then at least one of the following statements is true:*

1. *(X_i, l_i) explicitly implies (X_j, l_j).*
2. *(X_j, l_j) explicitly implies (X_i, l_i).*
3. *(X_i, l_i) implicitly implies (X_j, l_j) and (X_j, l_j) implicitly implies (X_i, l_i).*

Proof. Let $e_i = (X_i, l_i)$, and $e_j = (X_j, l_j)$. If $\{e_i\} \in \mathrm{TS}(\{e_i, e_j\}, C|_{\{X_i,X_j\}})$ then e_i explicitly implies e_j. If $\{e_j\} \in \mathrm{TS}(\{e_i, e_j\}, C|_{\{X_i,X_j\}})$ then e_j explicitly implies e_i.

If $\mathrm{TS}(\{e_i, e_j\}, C|_{\{X_i,X_j\}}) = \{\{e_i, e_j\}\}$, then e_i implicitly implies e_j and also e_j implicitly implies e_i. By Lemma 5.47 e_i uniquely implies e_j and e_j uniquely implies e_i. □

So far, we can compress two examples to one example by using unique implication. However, we need a compression set for any sample consistent with some concept in a concept class. To do so, we first show that the relation of implication is "partially transitive".

Lemma 5.49. *Let* $\mathrm{VCD}_{\Psi_G}(C) = 1$*, and let* S *be a* C*-realizable sample with* $e_1, e_2, e_3 \in S$*. If* e_1 *explicitly implies* e_2 *and* e_2 *explicitly implies* e_3*, then* e_1 *explicitly implies* e_3*. If* e_1 *explicitly implies* e_2 *and* e_2 *implicitly implies* e_3*, then* e_1 *implies* e_3*. In particular, in either case,* e_1 *uniquely implies* e_3*.*

Proof. Proof of the first statement: W.l.o.g., suppose $e_1 = (X_1, l_1)$, $e_2 = (X_2, l_2)$, $e_3 = (X_3, l_3)$. By the definition of explicit implication, every $c \in C$ with $c(X_1) = l_1$ satisfies $c(X_2) = l_2$, and every $c \in C$ with $c(X_2) = l_2$ satisfies $c(X_3) = l_3$. Thus every $c \in C$ with $c(X_1) = l_1$ satisfies $c(X_3) = l_3$, i.e., e_1 explicitly implies e_3.

Proof of the second statement: W.l.o.g., let $e_1 = (X_1, 0)$, $e_2 = (X_2, 0)$, $e_3 = (X_3, 0)$. So, $(0, 0) \in C|_{\{X_1, X_2\}}$ and $(0, 0) \in C|_{\{X_1, X_3\}}$.

e_2 implicitly implies e_3, so $\mathrm{TS}(\{e_2, e_3\}, C|_{\{X_2, X_3\}}) = \{\{(X_2, 0), (X_3, 0)\}\}$. That is, there are some concepts $c_1, c_2 \in C|_{\{X_2, X_3\}}$ such that $c_1(X_2) = 0$, $c_1(X_3) = l_3$, for some nonzero $l_3 \in N_3$, and $c_2(X_2) = l_2$, $c_2(X_3) = 0$, for some nonzero $l_2 \in N_2$. Now, we discuss the possible values for c_2 on X_1.

If $c_2(X_1) = 0$, then $(0, l_2) \in C|_{\{X_1, X_2\}}$ and $(X_1, 0)$ is not a minimal teaching set for $\{e_1, e_2\} = \{(X_1, 0), (X_2, 0)\}$ in $C|_{\{X_1, X_2\}}$. So, $c_2(X_1) = l_1$, for some nonzero $l_1 \in N_1$. This means that $(l_1, 0) \in C|_{\{X_1, X_3\}}$ and $(X_3, 0)$ is not a minimal teaching set for $\{e_1, e_3\} = \{(X_1, 0), (X_3, 0)\}$ in $C|_{\{X_1, X_2\}}$. So, e_3 does not explicitly imply e_1. Now, if $e_1 \in \mathrm{TS}(\{e_1, e_3\}, C|_{\{X_1, X_3\}})$ then e_1 explicitly implies e_3. Otherwise, $\mathrm{TS}(\{e_1, e_3\}, C|_{\{X_1, X_3\}}) = \{\{(X_1, 0), (X_3, 0)\}\}$ and e_1 implicitly implies e_3. So, in any case, e_1 implies e_3 and since $\mathrm{VCD}_{\Psi_G}(C) = 1$, e_1 uniquely implies e_3 by Lemma 5.47. □

Example 5.50. *Consider the class in Table 5.14. For* $S = \{(X_2, 1), (X_3, 1), (X_4, 2)\}$ *consistent with* c_4*,* $(X_4, 2)$ *explicitly implies* $(X_2, 1)$ *and* $(X_2, 1)$ *explicitly implies* $(X_3, 1)$*. So, as it can be verified in Table 5.14,* $(X_4, 2)$ *explicitly implies* $(X_3, 1)$*. For*

$S' = \{(X_1, 2), (X_2, 0), (X_3, 2)\}$ *consistent with* c_2, $(X_3, 2)$ *explicitly implies* $(X_2, 0)$ *and* $(X_2, 0)$ *implicitly implies* $(X_1, 2)$. *Therefore,* $(X_3, 2)$ *implicitly implies* $(X_1, 2)$. *In particular,* $(X_4, 2)$ *is a representative for* S *and* $(X_3, 2)$ *is a representative for* S'.

The next theorem shows that the existence of a representative for the samples S and S' in the previous example is not by accident.

Theorem 5.51. *Let* $\mathrm{VCD}_{\Psi_G}(C) = 1$. *Then any* C*-realizable sample* S *has a representative.*

Proof. For $|S| = 1$, there is nothing to show, and for $|S| = 2$, Corollary 5.48 proves the claim.

Let $S = \{e_1, \ldots, e_k\}$, with $k \geq 3$. We find a representative r of S inductively as follows. In step 1, let $r = e_1$. In step i, for $2 \leq i \leq k$, test whether r implies e_i in $C|_{\{X(r), X(e_i)\}}$. If yes, don't change r. If no, then, if e_i explicitly implies r in $C|_{\{X(r), X(e_i)\}}$ then $r = e_i$.

Consider step i for $i \geq 2$. By Corollary 5.48, either r implies e_i or e_i explicitly implies r. If r implies e_i, then r uniquely implies e_i and thus r is still a representative for $\{e_1, \ldots, e_i\}$. Let e_i explicitly imply r. Let $1 \leq j < i$. If r explicitly implies e_j, then by Lemma 5.49, e_i explicitly and thus uniquely implies e_j. If r implicitly implies e_j, then by Lemma 5.49, e_i uniquely implies e_j. So, e_i uniquely implies any example in $\{e_1, \ldots, e_i\}$, i.e., e_i is a representative for $\{e_1, \ldots, e_i\}$. □

Table 5.14 shows the representatives for concepts in a VCD_{Ψ_G} 1 concept class. One can see that no two concepts share the same representative, so, each concept can be compressed to one of its representatives.

Theorem 5.51 now allows us to define a compression scheme of size 1 for any VCD_{Ψ_G} 1 class.

Corollary 5.52. *Let* $\mathrm{VCD}_{\Psi_G}(C) = 1$*. Then* C *has a sample compression scheme of size* 1*.*

Proof. The compression function, given a sample S that is labeled consistently with some concept in C, outputs a representative r for s, which exists by Theorem 5.51.

The decompression function, on input of an example r and an instance $X_t \in X$, works as follows. If $X_t = X(r)$, then $r = (X_t, l_t)$ and the output is l_t. If $X_t \neq X(r)$, the decompression function looks for a label $l_t \in X_t$ such that r uniquely implies (X_t, l_t). If l_t exists, it is output. Else the output is 0. □

Example 5.53. *Consider the class in Table 5.14. One can see that* $(X_4, 2)$ *is a representative for* $S = \{(X_2, 1), (X_3, 1), (X_4, 2)\}$ *as it explicitly implies* $(X_3, 1)$ *and* $(X_2, 1)$*. Decompression of* $\{(X_4, 2)\}$ *would yield* c_4*, since* $(X_4, 2)$ *explicitly implies* $(X_1, 2)$ *as well. For* $S' = \{(X_1, 2), (X_2, 1), (X_3, 1)\}$ *consistent with* c_3 *and* c_4*,* $(X_3, 1)$ *explicitly implies* $(X_2, 1)$ *and* $(X_2, 0)$*, i.e.,* $(X_3, 1)$ *is a representative for* S'*. However, decompression of* $\{(X_3, 1)\}$ *would result in* $\{(X_1, 2), (X_2, 1), (X_3, 1), (X_4, 0)\} \notin C$*, because* $(X_3, 1)$ *does not imply* (X_4, l)*, for any* $l \in \{0, 1, 2\}$*.*

The assumption that X is finite is not used in the proof of Corollary 5.52, so that the latter applies also to infinite concept classes of VCD_{Ψ_G}-dimension 1.

5.5.2 Pollard's Pseudo-dimension

Although we could not prove that VCD_{Ψ_P} 1 classes have compression schemes of size 1, we show that the approach that we used for classes of VCD_{Ψ_G} 1 does not work here. In particular, we illustrate that Lemma 5.44 does not hold for VCD_{Ψ_P} 1 classes.

Proposition 5.54. *There is a multi-label class* C *of* VCD_{Ψ_P} 1 *with a sample compression scheme of size* 1 *in which for some* $X_i, X_j \in X$ *with* $i \neq j$*, there is more than one concept in* $C|_{\{X_i, X_j\}}$ *with teaching dimension* 2 *w.r.t.* $C|_{\{X_i, X_j\}}$*.*

Proof. Consider the class $C \subseteq \{0,1,2\}^2$ in Table 5.15. It is easy to see that C is of $\mathrm{VCD}_{\Psi_\mathrm{P}}$ 1, while $\mathrm{TD}(c_1, C) = \mathrm{TD}(c_2, C) = 2$. However, there exists a sample compression scheme of size 1 for this class. For instance, one can compress c_1, c_2, c_3, c_4 to $\{(X_1, 1)\}$, $\{(X_2, 0)\}$, $\{(X_3, 0)\}$ and $\{(X_4, 2)\}$, respectively. □

$c \in C$	X_1	X_2
c_1	1	1
c_2	0	1
c_3	1	0
c_4	0	2

Table 5.15: A concept class of $\mathrm{VCD}_{\Psi_\mathrm{P}}$ 1 with 2 concepts of teaching dimension 2.

Note that the concept class in Table 5.11 is not $\mathrm{VCD}_{\Psi_\mathrm{P}}$-maximal. In fact, we could neither find a proper $\mathrm{VCD}_{\Psi_\mathrm{P}}$-maximal class (a $\mathrm{VCD}_{\Psi_\mathrm{P}}$-maximal class that is not $\mathrm{VCD}_{\Psi_\mathrm{P}}$-maximum) of $\mathrm{VCD}_{\Psi_\mathrm{P}}$ 1 nor prove that $\mathrm{VCD}_{\Psi_\mathrm{P}}$ 1 classes can be embedded in $\mathrm{VCD}_{\Psi_\mathrm{P}}$-maximum classes of $\mathrm{VCD}_{\Psi_\mathrm{P}}$ 1.

5.5.3 The Natarajan Dimension

As for Pollard's pseudo-dimension, we could not prove that $\mathrm{VCD}_{\Psi_\mathrm{N}}$ 1 classes have compression schemes of size 1. We are still able to show that Lemma 5.44 does not hold for $\mathrm{VCD}_{\Psi_\mathrm{N}}$ 1 classes though.

Proposition 5.55. *There is a multi-label class C of* $\mathrm{VCD}_{\Psi_\mathrm{N}}$ 1 *with a sample compression scheme of size* 1 *in which for some $X_i, X_j \in X$ with $i \neq j$, there is more than one concept in $C|_{\{X_i, X_j\}}$ with teaching dimension* 2 *w.r.t. $C|_{\{X_i, X_j\}}$.*

Proof. Consider the class $C \subseteq \{0,1,2\}^2$ in Table 5.16. One can simply verify that C is of $\mathrm{VCD}_{\Psi_\mathrm{N}}$ 1, while $\mathrm{TD}(c_1, C) = \mathrm{TD}(c_2, C) = \mathrm{TD}(c_3, C) = 2$. By the way, C does have a sample compression scheme of size 1, because one can map c_1, c_2, c_3, c_4, c_5 to $\{(X_1, 1)\}$, $\{(X_1, 2)\}$, $\{(X_2, 2)\}$, $\{(X_2, 0)\}$ and $\{(X_1, 0)\}$, respectively. □

$c \in C$	X_1	X_2
c_1	1	1
c_2	2	1
c_3	2	2
c_4	1	0
c_5	0	2

Table 5.16: A concept class of VCD_{Ψ_N} 1 with 3 concepts of teaching dimension 2.

Note that the concept class in Table 5.11 is not VCD_{Ψ_N}-maximal. In fact, we could neither find a proper VCD_{Ψ_N}-maximal class (a VCD_{Ψ_N}-maximal class that is not VCD_{Ψ_N}-maximum) of VCD_{Ψ_N} 1 nor prove that VCD_{Ψ_N} 1 classes can be embedded in VCD_{Ψ_N}-maximum classes of VCD_{Ψ_N} 1.

Chapter 6

Recursive Teaching Dimension

In this chapter we study a recently introduced teaching notion, namely the recursive teaching dimension (RTD) [ZLHZ11], for multi-label concept classes. Recent work by Doliwa et al. [DSZ10, DFSZ14] indicates connections between the VC-dimension and the RTD in the binary case. In this chapter, we establish a further connection between RTD and VC-dimension in both the multi-label and binary cases.

We first prove a statement that connects $\mathrm{VCD}_{\Psi_{\mathrm{G}}}$ 1 classes to the classes of RTD 1. Next in Section 6.1, we show that a Sauer-type function upper-bounds the size of classes of a given RTD. We then focus on RTD-maximum classes, classes of a given RTD whose size meets the Sauer-type bound, and compare them to VCD_{Ψ}-maximum classes. In fact, we examine some of the most interesting properties of VCD-maximum classes and try to generalize them to VCD_{Ψ}-maximum classes along with RTD-maximum classes. We further prove some nice results on the teaching plans of RTD-maximum classes.

In Section 6.2, we discuss RTD-maximal classes, which are classes whose RTD increases when adding any new concept. Although it is trivial that any RTD-maximum class is also RTD-maximal, we show that the other direction is not true in general,

i.e., there exist some proper RTD-maximal classes.

We start this chapter with an observation about VCD_{Ψ_G} 1 classes and show that any VCD_{Ψ_G} 1 class is of RTD 1. This is an extension of the analogous result in the binary case [DSZ10, DFSZ14]. Our proof here is completely different from that in [DSZ10, DFSZ14], as we prove our claim directly and without using any other teaching complexity notions.

Proposition 6.1. *Let* $\mathrm{VCD}_{\Psi_G}(C) = 1$. *Then there is a concept* $c \in C$ *such that* $\mathrm{TD}(c, C) = 1$.

Proof. The proof is by induction on m. For $m = 1$ the claim is obviously true.

Let $m > 1$ and $C \subseteq \prod_{i=1}^{m} X_i$ with $\mathrm{VCD}_{\Psi_G}(C) = 1$. Then $\mathrm{VCD}_{\Psi_G}(C - X_m) \leq 1$. If $\mathrm{VCD}_{\Psi_G}(C - X_m) = 0$ then $(X_m, c(X_m))$ is a teaching set for c w.r.t. C for any $c \in C$, and thus $\mathrm{TD}(c, C) = 1$ for all $c \in C$. If $\mathrm{VCD}_{\Psi_G}(C - X_m) = 1$, by induction hypothesis, there is a concept $\overline{c} \in C - X_m$ with $\mathrm{TD}(\overline{c}, C - X_m) = 1$. That is, there is an $X_t \in \{X_1, \ldots, X_{m-1}\}$ for which $\overline{c}(X_t) \neq \overline{c'}(X_t)$, for all $\overline{c'} \in C - X_m \setminus \{\overline{c}\}$. W.l.o.g., let $\overline{c}(X_{m-1}) = 0$ and for all $\overline{c'} \in (C - X_m) \setminus \{\overline{c}\}$, $\overline{c'}(X_{m-1}) \neq 0$. Now, there are two possible cases for c in the class C to consider:

Case 1: $\overline{c}$ has a unique extension onto the concepts in C. That is, there is only one concept $c \in C$ such that $c - X_m = \overline{c}$. So, c is the only concept in C for which $c(X_{m-1}) = 0$, or equivalently, for all concepts $c' \in C \setminus \{c\}$, $c'(X_{m-1}) \neq 0$. Hence, $\mathrm{TD}(c, C) = 1$.

Case 2: $\overline{c}$ does not have a unique extension onto the concepts in C. That is, there are at least two concepts $c_1, c_2 \in C$ such that $c_1 - X_m = c_2 - X_m = \overline{c}$ and $c_1(X_m) \neq c_2(X_m)$. Let $c_1(X_m) = l_1$ and $c_2(X_m) = l_2$, for some $l_1, l_2 \in X_m$, with $l_1 \neq l_2$. So, $c_1|_{\{X_{m-1}, X_m\}} = (0, l_1)$ and $c_2|_{\{X_{m-1}, X_m\}} = (0, l_2)$. Again, we have two cases to consider:

Case 2.a: For some $z \in \{1,2\}$, c_z is the only concept in C with $c_z(X_m) = l_z$. Then $\mathrm{TD}(c_z, C) = 1$.

Case 2.b: There are concepts $c'_1, c'_2 \in C \backslash \{c_1, c_2\}$ with $c'_1(X_m) = l_1$ and $c'_2(X_m) = l_2$. Clearly, $c'_1 - X_m \neq c_1 - X_m$ and $c'_2 - X_m \neq c_2 - X_m$. Since for all $c_i \in C - X_m \setminus \{c\}$, $c_i(X_{m-1}) \neq 0$, we conclude that for all concepts $c' \in C$ with $c' - X_m \neq c$, $c'(X_{m-1}) \neq 0$. So, $c'_1(X_{m-1}) \neq 0$ and $c'_2(X_{m-1}) \neq 0$. Let $c'_1(X_{m-1}) = l'_1$ and $c'_2(X_{m-1}) = l'_2$, with $l'_1 \neq 0$ and $l'_2 \neq 0$. So, $c'_1|_{\{X_{m-1}, X_m\}} = (l'_1, l_1)$ and $c'_2|_{\{X_{m-1}, X_m\}} = (l'_2, l_2)$.

Having $c_1, c_2, c'_1, c'_2 \in C$, implies

$$\{(0, l_1), (0, l_2), (l'_1, l_1), (l'_2, l_2)\} \subseteq C|_{\{X_{m-1}, X_m\}}.$$

Now, consider $\psi_{m-1} : X_{m-1} \to \{0,1\}$ and $\psi_m : X_m \to \{0,1\}$ such that

$$\psi_{m-1}(x) = \begin{cases} 0 & \text{if } x = 0 \\ 1 & \text{if } x \in \{l'_1, l'_2\} \\ 1 & \text{otherwise.} \end{cases}$$

and

$$\psi_m(x) = \begin{cases} 0 & \text{if } x = l_1 \\ 1 & \text{if } x = l_2 \\ 1 & \text{otherwise.} \end{cases}$$

Let $\overline{\psi} = (\psi_{m-1}, \psi_m)$. It is easy to see that $\{(0,0), (0,1), (1,0), (1,1)\} \subseteq \overline{\psi}(C|_{\{X_{m-1}, X_m\}})$ and consequently, $\{X_{m-1}, X_m\}$ is shattered by C, which contradicts the fact that $\mathrm{VCD}_{\Psi_G}(C) = 1$. So, the case 2.$b$ cannot occur. □

The following corollary is now obvious.

Corollary 6.2. *Let C be of* VCD_{Ψ_G} 1. *Then* $\mathrm{RTD}(C) = 1$.

6.1 RTD-maximum Classes

In this section we present a Sauer-type bound on the size of a concept class with a given RTD. Having established this tight bound, we then define RTD-maximum classes and prove various properties of RTD-maximum classes and compare them with their analogs for VC-dimension. All our results on the RTD in the binary case have been published in the Proceedings of the 23rd International Conference on Algorithmic Learning Theory and in its special issue in the Journal *Theoretical Computer Science* [SSYZ12, SSYZ14a].

We now proceed to the main result of this section. Lemma 3.16, which provides an algebraic characterization of the teaching sets for a concept c in a concept class C, has a central role in the proof of our main result. We need a flashback to Chapter 3 first. Recall that, as described in (3.1), for each $i \in \{1, \ldots, m\}$ and $k \in \{0, \ldots, N_i\}$, $p_{i,k} : \mathbb{R} \to \{0, 1\}$ is a polynomial of degree N_i with

$$p_{i,k}(X_i) = \begin{cases} 1 & \text{if } X_i = k \\ 0 & \text{otherwise.} \end{cases}$$

Theorem 6.3. *Let* $C \subseteq \prod_{i=1}^{m} \{0, \ldots, N_i\}$. *If* $\mathrm{RTD}(C) = r$ *then the monomials from* $P^r(N_1, \ldots, N_m)$ *span the vector space* $\mathbb{R}^{|C|}$.

Proof. Let $c_1, c_2, \ldots, c_n$ be all the concepts from C listed in the same order as they appear in some teaching plan for C of order r. In particular, for every $s = 1, \ldots, n$, we have $\mathrm{TD}(c_s, \{c_s, \ldots, c_n\}) \leq r$.

To prove that $\mathbb{R}^n$ is spanned by the vectors corresponding to the monomials from $P^r(N_1, \ldots, N_m)$, it suffices to show that every standard basis vector, that is every $c_1, \ldots, c_n$, is in the span of $P^r(N_1, \ldots, N_m)$.

We show by induction that each c_s lies in the span of $P^r(N_1, \ldots, N_m)$. By assumption, we have $\mathrm{TD}(c_1, C) \leq r$. Let $\{(X_{i_1}, n_{i_1}), \ldots, (X_{i_k}, n_{i_k})\}$ be a teaching set for c_1 where $k \leq r$. By Lemma 3.16, c_1 lies in the span of $P^k(N_{i_1}, \ldots, N_{i_k})$. In particular, since $k \leq r$, c_1 lies in the span of $P^r(N_{i_1}, \ldots, N_{i_k})$ and consequently, c_1 is in the span of $P^r(N_1, \ldots, N_m)$.

Now suppose that $c_1, \ldots, c_s$ are in the span of $P^r(N_1, \ldots, N_m)$. Let $\mathbb{R}^{s,0}$ be the subspace of $\mathbb{R}^n$ consisting of the vectors whose last $n-s$ coordinates are zeros. Similarly, let $\mathbb{R}^{0,n-s}$ be the subspace of $\mathbb{R}^n$ consisting of the vectors whose first s coordinates are zeros. Also, let $(\mathbf{v})_{s,0}$ and $(\mathbf{v})_{0,n-s}$ be the projections of a vector $\mathbf{v} \in \mathbb{R}^n$ to the subspaces $\mathbb{R}^{s,0}$ and $\mathbb{R}^{0,n-s}$, respectively. In particular, we have $\mathbf{v} = (\mathbf{v})_{s,0} + (\mathbf{v})_{0,n-s}$.

Since $\mathrm{RTD}(C) \leq r$, we have that $\mathrm{TD}(c_{s+1}, \{c_{s+1}, \ldots, c_n\}) \leq r$. Let $\{(X_{i_1}, n_{i_1}), \ldots, (X_{i_k}, n_{i_k})\}$, where $k \leq r$, be a teaching set for c_{s+1} in the class $\{c_{s+1}, \ldots, c_n\}$. Let $p(X_{i_1}, \ldots, X_{i_k}) = p_{i_1,n_{i_1}}(X_{i_1}) \cdots p_{i_k,n_{i_k}}(X_{i_k})$, where $p_{i_j,n_{i_j}}$ is the polynomial introduced in (3.1), i.e.,

$$p_{i_j,n_{i_j}}(X_{i_j}) = \begin{cases} 1 & \text{if } X_{i_j} = n_{i_j} \\ 0 & \text{otherwise,} \end{cases}$$

and let $\mathbf{p} \in \mathbb{R}^n$ be the vector that corresponds to $p(X_{i_1}, \ldots, X_{i_k})$.

On the one hand, $p(c_{s+1}) = 1$. On the other hand, $\{(X_{i_1}, n_{i_1}), \ldots, (X_{i_k}, n_{i_k})\}$ is not consistent with any concept in $\{c_{s+2}, \ldots, c_n\}$ and thus $p(c_{s+2}) = \cdots = p(c_n) = 0$. So,

$$(\mathbf{c_{s+1}})_{0,n-s} = (\mathbf{p})_{0,n-s}.$$

In other words, $(\mathbf{c_{s+1}} - \mathbf{p})_{0,n-s} = \mathbf{0}$, which means that $(\mathbf{c_{s+1}} - \mathbf{p})$ belongs to the subspace $\mathbb{R}^{s,0}$.

Since each $p_{i_j,n_{i_j}}(X_{i_j})$ is a polynomial of degree N_{i_j}, we can write $p(X_{i_1}, \ldots, X_{i_k})$ as a linear combination of monomials from $P^k(N_{i_1}, \ldots, N_{i_k})$. So $p(X_{i_1}, \ldots, X_{i_k})$ lies

in the span of $P^k(N_{i_1}, \ldots, N_{i_k})$, and since $k \leq r$, p lies in the span of $P^r(N_{i_1}, \ldots, N_{i_k})$. In particular, $p(X_{i_1}, \ldots, X_{i_k})$ lies in the span of $P^r(N_1, \ldots, N_m)$ and thus $\mathbf{p}$ is in the span of $P^r(N_1, \ldots, N_m)$. Moreover, by the induction hypothesis, the vectors $\mathbf{c_1}, \ldots, \mathbf{c_s}$ are in the span of $P^r(N_1, \ldots, N_m)$, and hence the subspace $\mathbb{R}^{s,0}$ is contained in the span of $P^r(N_1, \ldots, N_m)$. Therefore, c_{s+1} is in the span of $P^r(N_1, \ldots, N_m)$. □

Remark 6.4. *Theorem 6.3 states that if* $\mathrm{RTD}(C) = r$*, then the monomials from* $P^r(N_1, \ldots, N_m)$ *span the vector space* $\mathbb{R}^{|C|}$*. As pointed out in Chapter 3, an analogous idea was used previously by Smolensky [Smo97] and Gurvits [Gur97] in the binary case to prove Sauer's bound for* VCD*. Namely, they showed that if* $\mathrm{VCD}(C) = d$*, then the monomials from* $P^d(1, \ldots, 1)$ *span* $\mathbf{F}_2^{|C|}$*, where* $\mathbf{F}_2^{|C|}$ *is a vector space of dimension* $|C|$ *over the field* $\mathbf{F}_2$ *(i.e., the field consisting of 2 elements). Gurvits exploited the same technique to generalize this result to the multi-label case, as discussed in Theorem 3.10.*

Note that the technique we used to prove that $P^r(N_1, \ldots, N_m)$ *spans* $\mathbb{R}^{|C|}$ *is different from the one used in the literature. In particular, our technique is based on the algebraic characterization of teaching sets provided in Lemma 3.16.*

As a corollary from Theorem 6.3, we obtain a generalized Sauer-type bound for RTD.

Corollary 6.5. *Let* $C \subseteq \prod_{i=1}^{m} \{0, \ldots, N_i\}$*. If* $\mathrm{RTD}(C) = r$ *then*

$$|C| \leq \Phi_r(N_1, \ldots, N_m).$$

The Sauer-type bound shown above is tight for any r and m. In particular, it is met by the standard VCD_Ψ-maximum class of VCD_Ψ r, namely the class of concepts that have at most r non-zero elements, where Ψ is the direct product of spanning families of mappings. This suggests the following definition.

Definition 6.6. *Let* $C \subseteq \prod_{i=1}^{m} \{0, \ldots, N_i\}$ *with* $\mathrm{RTD}(C) = r$. *C is* RTD-*maximum if* $|C| = \Phi_r(N_1, \ldots, N_m)$. *$C$ is called* RTD-*maximal if* $\mathrm{RTD}(C \cup \{c\}) > r$ *for any concept* $c \notin C$.

The above results suggest that the notions of RTD and VCD are related, at least for certain types of concept classes. We will study in the following, which structural properties VCD-maximum and RTD-maximum classes have in common. For convenience, the main results of Chapter 6 are summarized in Table 6.1.

Property	**K = VCD**		**K = RTD**
	binary	multi-label	
C is K-maximum $\Rightarrow C\|_{X'}$ is K-maximum	Yes	Yes	No (Prop. 6.17)
C is K-maximum $\Rightarrow \overline{C}$ is K-maximum	Yes	No	No (Prop. A.2 and Table A.1)
C is K-maximum $\Rightarrow K(C) + K(\overline{C}) = \|X\|-1$	Yes	No (Cor. 6.12)	No (Cor. 6.14)
C is K-maximum $\Rightarrow$ C is shortest-path closed	Yes	Yes (Thm. 3.23)	No (Prop. 6.18)
C is K-maximum $\Rightarrow$ C has a sample compression scheme	Yes	Yes (Thm. 5.19)	No (Cor. 6.16)
C is K-maximal $\Rightarrow$ C shatters all subsets of size $K(C)$	No (Table 6.4)	No (Table 6.4)	Yes (Prop. 6.23)
$K(C) = 1$ and C is K-maximal $\Rightarrow$ C is K-maximum	Yes	?	Yes (Prop. 6.26)
There is a K-maximal class that is not K-maximum	Yes	Yes	Yes (Prop. 6.27)

Table 6.1: Summary of the main results of Sections 6.1 and 6.2, in the context of known results on VCD.

In the binary case, Doliwa et al. proved that for every binary VCD-maximum class, RTD and VCD are equal [DSZ10, DFSZ14]. Here, we present a generalization of that result for the multi-label case, which is in fact an alternative proof for the result in the binary case. The new contribution of our proof is that it establishes a connection between tight compression schemes and recursive teaching plans for VCD_Ψ-maximum classes in the multi-label case.

We show that if VCD_Ψ fulfills the reduction property then any VCD_Ψ-maximum class C is also RTD-maximum. Our idea is to recursively teach the concepts in a multi-label concept class using their compression sets resulting from the tight compression scheme. On the one hand, any VCD_Ψ-maximum class with VCD_Ψ fulfilling the reduction property has such a scheme of size VCD_Ψ of the class, and thus $\text{RTD}(C) \leq d$. On the other hand, C is VCD_Ψ-maximum and $|C| = \Phi_d(N_1, \ldots, N_m)$, so by Definition 6.6, C is also RTD-maximum of RTD d.

We first overview the idea at a high level and then proceed to the formal proof. Since VCD_Ψ fulfills the reduction property, for any $t \in [m]$, C can be partitioned into the classes $C^{X_t} \times X_t$ and $\text{tail}_{X_t}(C)$. Recall that when C is VCD_Ψ-maximum with $\text{VCD}_\Psi(C) = d$, then C^{X_t} is VCD_Ψ-maximum of VCD_Ψ $d-1$ and $\text{Forb}(C^{X_t})$ denotes the set of forbidden labelings of size d for C^{X_t}. We have already shown that there is a bipartite graph between $\text{tail}_{X_t}(C)$ and $\text{Forb}(C^{X_t})$ with a unique perfect matching such that there is an edge between $c \in \text{tail}_{X_t}(C)$ and $S \in \text{Forb}(C^{X_t})$ iff S is consistent with c, i.e., $S \subseteq c$ (see Theorem 5.35). We first teach each tail concept with its matched forbidden labeling. After teaching and removing the tail concepts, we next teach every concept $c \in C^{X_t} \times X_t$ using its corresponding compression set, which is an extension of the compression set for $c - X_t \in C^{X_t}$ on X_t.

The following lemma allows us to conclude that there is a forbidden labeling in $\text{Forb}(C^{X_t})$ that is consistent with only one concept in $\text{tail}_{X_t}(C)$.

Lemma 6.7. [LP86, ZZ13] *Let $G = (U \cup V, E)$ be a bipartite graph with two parts U and V. If G has a unique perfect matching then it must contain two degree-1 vertices $u \in U$ and $v \in V$.*

Now, we are ready to prove the main theorem leading to the aforementioned connection between VCD_Ψ-maximum classes and RTD-maximum classes. Doliwa et

al. revealed a strong relationship between sample compression schemes and recursive teaching sets [DSZ10]. In particular, they showed that for the case of VCD-maximum classes, there exists a teaching plan in which there is a one-to-one correspondence between the recursive teaching sets and the compression sets used in the Kuzmin and Warmuth unlabeled compression scheme for those classes. Here, we generalize that result to the multi-label case.

Theorem 6.8. *Let C be a VCD_Ψ-maximum class of VCD_Ψ d where VCD_Ψ fulfills the reduction property. Then there is a teaching plan $\mathcal{P} = \{(c_1, r(c_1)), \ldots, (c_{|C|}, r(c_{|C|}))\}$ where $r(c_i)$, $i \in \{1, \ldots, |C|\}$, is the compression set for c_i resulting from Algorithm 2 of Chapter 5.*

Proof. We need to find a teaching plan for C in which each concept in C is taught by its compression set obtained from Algorithm 2 in Chapter 5. The proof is an induction on d. The base case, $d = 0$, is obvious. Assume that the claim is true for any $d' < d$. Pick $s \in [m]$ and partition C into $C^{X_s} \times X_s$ and $\mathrm{tail}_{X_s}(C)$. C^{X_s} is VCD_Ψ-maximum of VCD_Ψ $d-1$, so by induction hypothesis, there is a teaching plan $\mathcal{P}^1 = \{(c_1, \tilde{r}(c_1)), \ldots, (c_l, \tilde{r}(c_l))\}$ for C^{X_s}, where $l = |C^{X_s}|$ and $\tilde{r}(c_i)$ is the compression set for c_i returned by Algorithm 2, for all $i \in \{1, \ldots, l\}$. We use $\mathcal{P}^1$ to construct a teaching plan $\mathcal{P}^2$ for $C^{X_s} \times X_s$. Let $c_i^k = c_i \cup \{(X_s, k)\}$, for all $i \in \{1, \ldots, l\}$ and $k \in X_s$. In particular, $C^{X_s} \times X_s = \{c_i^k \mid 1 \leq i \leq l \text{ and } 0 \leq k \leq N_s\}$. Let the teaching plan $\mathcal{P}^2$ for $C^{X_s} \times X_s$ be as follows:

$$\mathcal{P}^2 = \{(c_1^{N_s}, r(c_1^{N_s})), \ldots, (c_1^0, r(c_1^0)), (c_2^{N_s}, r(c_2^{N_s})), \ldots, (c_2^0, r(c_2^0)), \ldots,$$
$$(c_l^{N_s}, r(c_l^{N_s})), \ldots, c_l^0, r(c_l^0))\},$$

where for all $i \in \{1, \ldots, l\}$, as in the Else block of Algorithm 2 (for each $\bar{c} \in C^{X_s}$),

$$r(c_i^k) = \begin{cases} \tilde{r}(c_i) \cup \{(X_s, k)\}, & \text{if } \ 1 \leq k \leq N_s \\ \tilde{r}(c_i), & \text{if } \ k = 0. \end{cases}$$

That is, for all $i \in \{1, \ldots, l\}$ and $k \in \{0, \ldots, N_s\}$, $r(c_i^k)$ is the same as the compression set for $c_i \cup \{(X_s, k)\}$ that is constructed from the compression set for c_i in Algorithm 2. We claim that $\mathcal{P}^2$ is in fact a valid teaching plan for $C^{X_s} \times X_s$ of order d. In particular, $r(c_i^k) \in \mathrm{TS}(c_i^k, \{c_i^k, c_i^{k-1}, \ldots, c_i^0, \ldots, c_l^{N_s}, \ldots, c_l^0\})$, for all $i \in \{1, \ldots, l\}$ and $k \in X_s$. We prove our claim by examining three different cases for i and k:

Case 1: $i = l$, $k \in \{0, \ldots, N_s\}$.

Clearly, for all $k \in \{1, \ldots, N_s\}$, $\mathrm{TS}(c_l^k, \{c_l^k, c_l^{k-1}, \ldots, c_l^0\}) = \{(X_s, k)\}$. Since $r(c_l^k) = \emptyset \cup \{(X_s, k)\}$, for all $k \in \{1, \ldots, N_s\}$, and $r(c_l^0) = \tilde{r}(c_l^0) = \emptyset$, we have $r(c_l^k) \in \mathrm{TS}(c_l^k, \{c_l^k, c_l^{k-1}, \ldots, c_l^0\})$, for all $k \in \{0, \ldots, N_s\}$.

Case 2: $i \in \{1, \ldots, l-1\}$ and $k = 0$.

According to $\mathcal{P}^1$, $\tilde{r}(c_i) \in \mathrm{TS}(c_i, \{c_i, c_{i+1}, \ldots, c_l\})$ and thus,

$$r(c_i^0) = \tilde{r}(c_i^0) \in \mathrm{TS}(c_i^0, \{c_i^0, c_{i+1}^{N_s}, \ldots, c_{i+1}^0, \ldots, c_l^{N_s}, \ldots, c_l^0\}).$$

Case 3: $i \in \{1, \ldots, l-1\}$ and $k = \{1, \ldots, N_s\}$.

Since $\tilde{r}(c_i) \in \mathrm{TS}(c_i, \{c_i, c_{i+1}, \ldots, c_l\})$,

$$r(c_i^k) = \tilde{r}(c_i) \cup \{(X_s, k)\} \in \mathrm{TS}(c_i^k, \{c_i^k, c_{i+1}^{N_s}, \ldots, c_{i+1}^0, \ldots, c_l^{N_s}, \ldots, c_l^0\}).$$

Also, $\{(X_s, k)\} \in \mathrm{TS}(c_i^k, \{c_i^k, c_i^{k-1}, \ldots, c_i^0\})$ and thus,

$$r(c_i^k) \in \mathrm{TS}(c_i^k, \{c_i^k, \ldots, c_i^0, c_{i+1}^{N_s}, \ldots, c_{i+1}^0, \ldots, c_l^{N_s}, \ldots, c_l^0\}).$$

Now, we move to the tail concepts and show that there is a teaching plan $\mathcal{P}$ for C of order d in which the concepts in $\text{tail}_{X_s}(C)$ are taught by their corresponding compression sets before the concepts in $C^{X_s} \times X_s$. For simplicity, let $C' = \text{tail}_{X_s}(C)$ and $l' = |C'|$. As proven before, each tail concept is compressed to a forbidden labeling of size d for C^{X_s}. By definition, for each $\overline{f} \in \text{Forb}(C^{X_s})$, $\overline{f}$ is not consistent with any concept in C^{X_s}, and consequently, with any concept in $C^{X_s} \times X_s$. In other words, for each concept $c' \in C'$, $r(c') \in \text{TS}(c', \{c'\} \cup C^{X_s} \times X_s)$. So to accomplish the proof, we only need to find an ordering for the concepts in C', such that $C' = \{c'_1, \ldots, c'_{l'}\}$ and $r(c'_i) \in \text{TS}(c'_i, \{c'_i, \ldots, c'_{l'}\})$, for all $i \in \{1, \ldots, l'\}$. Such an ordering along with $\mathcal{P}^2$ yields the teaching plan

$$\mathcal{P} = \{(c'_1, r(c'_1)), \ldots, (c'_{l'}, r(c'_{l'})), (c_1^{N_s}, r(c_1^{N_s})), \ldots, (c_1^0, r(c_1^0)), \ldots,$$
$$(c_l^{N_s}, r(c_l^{N_s})), \ldots, c_l^0, r(c_l^0))\}$$

of order d for C.

By Theorem 5.35, there is a bipartite graph $G = (C' \cup \text{Forb}(C^{X_s}), E')$ with a unique perfect matching between C' and $\text{Forb}(C^{X_s})$, where there is an edge between a concept in C' and a forbidden labeling in $\text{Forb}(C^{X_s})$ if and only if this forbidden labeling is contained in the concept. By Lemma 6.7, there is a forbidden labeling $\overline{f_1} \in \text{Forb}(C^{X_s})$ that is consistent with only one tail concept $c'_1 \in C'$. In particular, $r(c'_1) = \overline{f_1}$ and $\overline{f_1} \not\subseteq c'$, for all $c' \in C' \setminus \{c'_1\}$, that is, $r(c'_1) \in \text{TS}(c'_1, \{c'_1, \ldots, c'_{l'}\})$. The subgraph G_1 induced by $C' \setminus \{c'_1\} \cup \text{Forb}(C^{X_s}) \setminus \{\overline{f_1}\}$ also has a unique perfect matching, because otherwise $G = (C', \text{Forb}(C^{X_s}))$ cannot have a unique perfect matching. Similarly, by using Lemma 6.7, we conclude that there is a concept $c'_2 \in C' \setminus \{c'_1\}$ such that $r(c'_2) \in \text{TS}(c'_2, \{c'_2, \ldots, c'_{l'}\})$. Following the same procedure, we find the desired ordering for the concepts in C'. □

The following corollary is now obvious. Although it has already been shown that any binary VCD-maximum class is also RTD-maximum [DSZ10, SSYZ12, SSYZ14a], Theorem 6.8 establishes this result with a completely different approach from the one in the literature.

Corollary 6.9. *Let C be a VCD_Ψ-maximum class of VCD_Ψ d where VCD_Ψ fulfills the reduction property. Then C is* RTD*-maximum with $\mathrm{RTD}(C) = d$.*

Proof. On the one hand, $|C| = \Phi_d(N_1, \ldots, N_m)$, so by Corollary 6.5, $\mathrm{RTD}(C) \geq d$. On the other hand, by Theorem 6.8, there is a teaching plan for C of order d. Hence, $\mathrm{RTD}(C) = d$. □

In Appendix A (Proposition A.3), we show that the other direction of the above corollary is not always true. In fact, we present a particular binary RTD-maximum class that is not VCD-maximum.

For a concept class $C \subseteq \prod_{1 \leq i \leq m} X_i$, the complement of C is defined as $\overline{C} = \prod_{1 \leq i \leq m} X_i \setminus C$. One of the fascinating properties of VCD-maximum classes in the binary case is that, whenever $C \subseteq \{0,1\}^m$ is VCD maximum of VCD $d \leq m$, $\overline{C}$ is also VCD maximum of dimension $m - d - 1$ [Flo89]. As shown in Appendix A (Proposition A.1), this result does not hold for binary RTD-maximum classes that are not VCD-maximum.

We now demonstrate that the structure of $\overline{C}$ for VCD_Ψ-maximum classes in the multi-label case is not as interesting as in the binary case. To do this, we first introduce some notation and prove a lemma. For each $k \in [m]$, let $T_k = \sum_{1 \leq i_1 < \cdots < i_k \leq m} N_{i_1} \cdots N_{i_k}$, $A(d) = 1 + \sum_{k=1}^{m-d-1} T_k$ and $B(d) = \sum_{k=d+1}^{m} T_k$.

Lemma 6.10. *Let $N \in \mathbb{N}^+$ with $N > 1$ and $N_i = N$, for all $i \in [m]$. Then $A(d) < B(d)$.*

Proof. Since $N_i = N$, for all $i \in [m]$,

$$A(d) = 1 + \binom{m}{1}N + \cdots + \binom{m}{m-d-1}N^{m-d-1}$$

and

$$B(d) = \binom{m}{d+1}N^{d+1} + \cdots + \binom{m}{m}N^m.$$

First, let $d+1 \leq \frac{m}{2}$. Then $m-d-1 \geq \frac{m}{2}$ and thus

$$A(d) = 1+\binom{m}{1}N+\cdots+\binom{m}{d+1}N^{d+1}+\cdots+\binom{m}{m/2}N^{m/2}+\cdots+\binom{m}{m-d-1}N^{m-d-1}$$

and

$$B(d) = \binom{m}{d+1}N^{d+1}+\cdots+\binom{m}{m/2}N^{m/2}+\cdots+\binom{m}{m-d-1}N^{m-d-1}+\cdots+\binom{m}{m}N^m.$$

So,

$$\begin{aligned} A(d) - B(d) &= 1 + \binom{m}{1}N + \cdots + \binom{m}{d}N^d - \binom{m}{m-d}N^{m-d} - \cdots - \binom{m}{m}N^m \\ &= \sum_{i=0}^{d}(\binom{m}{d-i}N^{d-i} - \binom{m}{m-d+i}N^{m-d+i}) \\ &= \sum_{i=0}^{d}\binom{m}{d-i}(N^{d-i} - N^{m-d+i}). \ \text{(since } \binom{m}{d-i} = \binom{m}{m-d+i}\text{)} \end{aligned}$$

Since $d+1 \leq \frac{m}{2}$, we have

$$2d+2 \leq m \Rightarrow d+2 \leq m-d,$$

and thus for all $i \in \{0, \ldots, d\}$, $d + 2 - i \leq m - d - i \leq m - d + i$. So,

$$d - i < m - d + i, \quad \text{for all } i \in \{0, \ldots, d\} \tag{6.1}$$

and $A(d) - B(d) < 0$. Hence, $A(d) < B(d)$.

Second, let $d + 1 > \frac{m}{2}$. Then

$$\begin{aligned}
A(d) - B(d) &= 1 + \binom{m}{1} N + \cdots + \binom{m}{m-d-1} N^{m-d-1} \\
&\quad - \binom{m}{d+1} N^{d+1} - \cdots - \binom{m}{m} N^{m} \\
&= \sum_{i=0}^{m-d-1} \binom{m}{m-d-1-i} N^{m-d-1-i} - \sum_{i=0}^{m-(d+1)} \binom{m}{d+1+i} N^{d+1+i} \\
&= \sum_{i=0}^{m-d-1} \left(\binom{m}{m-d-1-i} N^{m-d-1-i} - \binom{m}{d+1+i} N^{d+1+i} \right) \\
&= \sum_{i=0}^{m-d-1} \binom{m}{m-d-1-i} (N^{m-d-1-i} - N^{d+1+i}).
\end{aligned}$$

The last equality follows from $\binom{m}{m-d-1-i} = \binom{m}{d+1+i}$, for all $i \in \{0, \ldots, m-d-1\}$.

Since $d + 1 > m/2$, we have

$$2d + 2 > m \Rightarrow d + 1 > m - d - 1.$$

So, for all $i \in \{0, \ldots, m - d - 1\}$,

$$d + 1 + i > m - d - 1 + i > m - d - 1 - i$$

and thus $A(d) - B(d) < 0$. So, $A(d) < B(d)$. □

As shown in [Flo89] in the binary case, when C is VCD-maximum of VCD d, then

$\overline{C}$ is also VCD-maximum with $\mathrm{VCD}(\overline{C}) = m - d - 1$. The next theorem shows that this cannot be extended to the multi-label case.

Theorem 6.11. *Let $N \in \mathbb{N}^+$ with $N > 1$ and $X_i = \{0, \ldots, N\}$, for all $i \in [m]$. Let Ψ_i, $1 \leq i \leq m$, be a spanning family of mappings $\psi_i : X_i \to \{0, 1\}$, and $\Psi = \Psi_1 \times \cdots \times \Psi_m$. If $\mathrm{VCD}_\Psi(C) = d$ then $\mathrm{VCD}_\Psi(\overline{C}) > m - d - 1$.*

Proof. By Corollary 3.11, $|C| \leq \Phi_d(N, \ldots, N)$. So,

$$\begin{aligned} |\overline{C}| &\geq (N+1)^m - \Phi_d(N, \ldots, N) \\ &= \binom{m}{d+1} N^{d+1} - \cdots - \binom{m}{m} N^m \\ &> 1 + \binom{m}{1} N + \cdots + \binom{m}{m-d-1} N^{m-d-1} \text{ (by Lemma 6.10)} \end{aligned}$$

or equivalently,

$$|\overline{C}| > \Phi_{m-d-1}(N, \ldots, N) \tag{6.2}$$

By Corollary 3.11, $\mathrm{VCD}_\Psi(\overline{C}) \leq m - d - 1$ implies $|\overline{C}| \leq \Phi_{m-d-1}(N, \ldots, N)$ which contradicts (6.2). So, $\mathrm{VCD}_\Psi(\overline{C}) > m - d - 1$. □

In the binary case, a class C is VCD-maximum if and only if $\mathrm{VCD}(C) + \mathrm{VCD}(\overline{C}) = |X| - 1$. Necessity of the condition was proven by [RBR09]. For sufficiency, suppose C with $\mathrm{VCD}(C) = d$ is not VCD-maximum. Then $|C| < \Phi_d(|X|)$ and thus $|\overline{C}| > 2^{|X|} - \Phi_d(|X|) = \Phi_{|X|-d-1}(|X|)$, which implies $\mathrm{VCD}(\overline{C}) > |X| - d - 1$[1]. The next corollary shows that at least one direction of this result cannot be generalized to the multi-label case.

[1]The same reasoning yields that in the binary case, if $\mathrm{RTD}(C) + \mathrm{RTD}(\overline{C}) = |X| - 1$ then C is RTD-maximum.

Corollary 6.12. *Let $N \in \mathbb{N}^+$ with $N > 1$ and $X_i = \{0, \ldots, N\}$, for all $i \in [m]$. Let Ψ_i, $1 \leq i \leq m$, be a spanning family of mappings $\psi_i : X_i \to \{0, 1\}$, and $\Psi = \Psi_1 \times \cdots \times \Psi_m$. Let C be VCD_Ψ-maximum with $\mathrm{VCD}_\Psi(C) = d$. Then $\mathrm{VCD}_\Psi(C) + \mathrm{VCD}_\Psi(\overline{C}) > m - 1$. In particular, $\overline{C}$ is not VCD_Ψ-maximum of VCD_Ψ $m - d - 1$.*

Proof. By Theorem 6.11, $\mathrm{VCD}_\Psi(\overline{C}) > m - d - 1$, so, $\overline{C}$ is not VCD_Ψ-maximum of VCD_Ψ $m - d - 1$. Obviously, $\mathrm{VCD}_\Psi(C) + \mathrm{VCD}_\Psi(\overline{C}) > m - 1$. □

Theorem 6.11 can be easily modified to work for RTD-maximum classes as follows.

Theorem 6.13. *Let $N \in \mathbb{N}^+$ with $N > 1$ and $X_i = \{0, \ldots, N\}$, for all $i \in [m]$. Let Ψ_i, $1 \leq i \leq m$, be a spanning family of mappings $\psi_i : X_i \to \{0, 1\}$, and $\Psi = \Psi_1 \times \cdots \times \Psi_m$. If $\mathrm{RTD}(C) = r$ then $\mathrm{RTD}(\overline{C}) > m - r - 1$.*

Proof. By Corollary 6.5, $|C| \leq \Phi_r(N, \ldots, N)$. So,

$$\begin{aligned} |\overline{C}| &\geq (N+1)^m - \Phi_r(N, \ldots, N) \\ &= \binom{m}{r+1} N^{r+1} - \cdots - \binom{m}{m} N^m \\ &> 1 + \binom{m}{1} N + \cdots + \binom{m}{m-r-1} N^{m-r-1} \text{ (by Lemma 6.10)} \end{aligned}$$

or equivalently,

$$|\overline{C}| > \Phi_{m-r-1}(N, \ldots, N) \tag{6.3}$$

By Corollary 6.5, $\mathrm{RTD}(\overline{C}) \leq m - r - 1$ implies $|\overline{C}| \leq \Phi_{m-r-1}(N, \ldots, N)$ which contradicts (6.3). So, $\mathrm{RTD}(\overline{C}) > m - r - 1$. □

The same reasoning as that of Corollary 6.12 implies the following corollary.

Corollary 6.14. *Let $N \in \mathbb{N}^+$ with $N > 1$ and $X_i = \{0, \ldots, N\}$, for all $i \in [m]$. Let Ψ_i, $1 \leq i \leq m$, be a spanning family of mappings $\psi_i : X_i \to \{0, 1\}$, and $\Psi = \Psi_1 \times \cdots \times \Psi_m$. Let C be* RTD*-maximum with* $\mathrm{RTD}(C) = r$. *Then* $\mathrm{RTD}(C) + \mathrm{RTD}(\overline{C}) > m - 1$. *In particular, $\overline{C}$ is not* RTD*-maximum of* RTD $m - d - 1$.

Since for RTD-maximum classes the VC-dimension can exceed the recursive teaching dimension, it is natural to ask how large the difference between these two parameters can be. The following proposition answers this question.

Proposition 6.15. *For any two integers i and d with $1 \leq i < d$, there is an* RTD*-maximum binary class C such that* $\mathrm{RTD}(C) = i$ *and* $\mathrm{VCD}(C) \geq d$.

Proof. Fix positive integers d and i with $1 \leq i \leq d$. Choose any integer m such that $\binom{m}{i} \geq 2^d$. Let C' be a concept class on $X = Y \cup Z$ with $|Y| = m$ and $|Z| = d$ such that C' contains all subsets of X of size at most i. So, $|C'| = \Phi_i(|X|)$ and for any concept $c' \in C'$, $|c'| = |\{X_j \in X : c'(X_j) = 1\}| \leq i$. Note that $\mathrm{RTD}(C') = i$; in particular, there is a teaching plan $((c'_1, S_1), \ldots, (c'_{|C'|}, S_{|C'|}))$ of order i for C' for which the concepts c'_1, c'_2, $\ldots$, c'_{2^d} are of size i and are subsets of Y.

We construct an RTD-maximum concept class C with $\mathrm{VCD}(C) = d$ and $\mathrm{RTD}(C) = i$ in the following way. First, we fix an order over all subsets of Z. Then, we define new concepts c_k over X, for $1 \leq k \leq |C'|$, as follows.

- For $1 \leq k \leq 2^d$, let $c_k = c'_k \cup b_k$ where b_k is the kth subset of Z in the given order.
- For $k > 2^d$, let $c_k = c'_k$.

Let $C = \{c_1, \ldots, c_{|C'|}\}$. In particular, $|C| = |C'| = \Phi_i(|X|)$. On the one hand, $((c_1, S_1), \ldots, (c_{|C'|}, S_{|C'|}))$ is a teaching plan of order i for C and thus $\mathrm{RTD}(C) \leq i$. On the other hand, since $|C| = \Phi_i(|X|)$, Corollary 6.5 implies $\mathrm{RTD}(C) \geq i$. Hence,

we obtain $\mathrm{RTD}(C) = i$ and C is RTD-maximum. Furthermore, $\mathrm{VCD}(C) \geq d$ since C shatters the set Z. □

An interesting consequence of Proposition 6.15 is that, even for RTD-maximum classes, the recursive teaching dimension is not an upper bound on the smallest possible size of a sample compression scheme.

Corollary 6.16. *There is an* RTD*-maximum class C for which no (labeled or unlabeled) sample compression scheme of size* $\mathrm{RTD}(C)$ *exists.*

Proof. Floyd and Warmuth [FW95] showed that no concept class of VC-dimension d has a sample compression scheme of size at most $d/5$. By Proposition 6.15, for any $d \geq 5$, there are binary RTD-maximum classes C with $\mathrm{VCD}(C) \geq d$ for which $\mathrm{RTD}(C) \leq d/5$; these cannot have sample compression schemes of size $\mathrm{RTD}(C)$. □

Due to [Wel87], we know that restricting a VCD-maximum class to a subset of its instance space yields another VCD-maximum class. But this property does not in general hold for RTD-maximum classes, as we show next.

Proposition 6.17. *There is an* RTD*-maximum class which has a restriction that is not* RTD*-maximum. Furthermore, there is an* RTD*-maximum class C that has an* RTD*-maximum restriction C' such that* $\mathrm{RTD}(C') > \mathrm{RTD}(C)$.

Proof. Consider C_2 in Table 6.2. It is easy to see that C_2 is RTD-maximum and $\mathrm{RTD}(C_2) = 1$. However, $\mathrm{RTD}(C_2 - X_4) = 2$ and $C_2 - X_4$ is not RTD-maximum. Furthermore, consider the RTD-maximum class C_1 in Table A.2. Clearly, $C_1 - X_4$ is RTD-maximum and $\mathrm{RTD}(C_1) = 2 < \mathrm{RTD}(C_1 - X_4) = 3$. □

$c \in C_2$	X_1	X_2	X_3	X_4
c_1	$\underline{1}$	0	0	0
c_2	0	1	1	$\underline{1}$
c_3	0	$\underline{1}$	0	0
c_4	0	0	$\underline{1}$	0
c_5	0	0	0	0

Table 6.2: C_2 is RTD-maximum but $C_2 - X_4$ is not. Recursive teaching sets are underlined.

Another difference between VCD-maximum classes and RTD-maximum classes is that the former are always shortest-path closed [KW07], while the latter are not in general.

Proposition 6.18. *There is an* RTD*-maximum class that is not shortest-path closed.*

Proof. The RTD-maximum class C_1 in Table A.2 is not shortest-path closed because the concept c_{11} has Hamming distance at least 2 from any other concept in the class. □

Note that shortest-path closedness is not the distinguishing property between RTD-maximum and VCD-maximum classes.

Proposition 6.19. *There is an* RTD*-maximum class that is shortest-path closed but not* VCD*-maximum.*

Proof. The RTD-maximum class C in Table 6.3 is of RTD 2 and also shortest-path closed. However, the VC-dimension of C is 3 which means that it is not VCD-maximum. This class was found by exhaustive enumeration of all concept classes over instance spaces of size at most 5. □

6.1.1 Teaching Plans of RTD-maximum Classes

In this section, we analyze the structure of teaching plans of RTD-maximum classes, which is motivated by the rich structure of teaching plans of VCD-maximum classes.

$c \in C$	X_1	X_2	X_3	X_4
c_1	$\underline{0}$	0	0	$\underline{1}$
c_2	$\underline{0}$	0	$\underline{1}$	0
c_3	$\underline{0}$	$\underline{0}$	0	0
c_4	$\underline{0}$	1	0	0
c_5	1	$\underline{0}$	$\underline{0}$	1
c_6	1	$\underline{0}$	1	$\underline{0}$
c_7	1	$\underline{0}$	1	1
c_8	1	1	$\underline{0}$	$\underline{0}$
c_9	1	1	$\underline{0}$	1
c_{10}	1	1	1	$\underline{0}$
c_{11}	1	1	1	1

Table 6.3: C is RTD-maximum and shortest-path closed (found by computer experiments) but not VCD-maximum ($\{X_2, X_3, X_4\}$ is shattered). Recursive teaching sets are underlined.

In particular, as discussed before, Doliwa et al. found a strong relationship between the recursive teaching sets for VCD-maximum classes and the unlabeled compression sets [DSZ10]. Further motivation is that understanding the structure of teaching plans of RTD-maximum classes can help us to understand the structure of these classes themselves.

We begin by showing that, for any teaching plan of an RTD-maximum class C, all instance sets of size $\mathrm{RTD}(C)$ are used as recursive teaching sets. This result is a consequence of the proof of Theorem 6.3.

Proposition 6.20. *Let C be* RTD*-maximum and* $\mathrm{RTD}(C) = r$*. Let $X' \subseteq X$ be any subset of size r. Then for any teaching plan $\mathcal{P}$ for C of order r, there is a concept $c \in C$ and a recursive teaching set S for c with respect to $\mathcal{P}$, such that $X(S) = X'$.*

Proof. Let $X' = \{X_{i_1}, \ldots, X_{i_r}\}$, and $\mathcal{P}$ be a teaching plan for C of order r such that $c_1, c_2, \ldots, c_n$ are all concepts from C listed in the same order as they appear in $\mathcal{P}$. Assume that X' does not appear as a recursive teaching set in the plan $\mathcal{P}$. Then, in the proof of Theorem 6.3 we can always represent the concept c_{s+1} inside the class $\{c_{s+1}, \ldots, c_n\}$ as a polynomial $p(Z_1, \ldots, Z_r)$ over $\mathbb{R}$ such that $\{Z_1, \ldots, Z_r\} \neq$

$\{X_{i_1}, \ldots, X_{i_r}\}$. (This follows from Lemma 3.16 and the fact that X' is not used as a recursive teaching set.) As a consequence, we can span $\mathbb{R}^{|C|}$ without using the monomials in $\mathcal{M} = \{X_{i_1}^{n_1} \cdots X_{i_r}^{n_r} \mid 1 \leq n_j \leq N_{i_j}, \text{ for all } j \in \{1, \ldots, r\}\}$. This implies that $|C| = \dim(\mathbb{R}^{|C|}) \leq \Phi_r(N_1, \ldots, N_m) - |\mathcal{M}|$. Hence C is not RTD-maximum, which is a contradiction. □

Another consequence of Theorem 6.3 is that, for an RTD-maximum class, teaching sets of size 1 cannot be used too early in any teaching plan.

Proposition 6.21. *Let C be* RTD*-maximum with* $\mathrm{RTD}(C) = r > 1$. *For an arbitrary teaching plan for C, let $(c_1, c_2, \ldots, c_n)$ be the sequence of all concepts of C listed in the plan. Then for any $t \in [m]$ and any integer $0 < j < \Phi_{r-1}(N_1, \ldots, N_{t-1}, N_{t+1}, \ldots, N_m)$, we have* $\mathrm{TD}(c_j, \{c_j, \ldots, c_n\}) > 1$.

Proof. W.l.o.g, let $t = 1$ and assume that $\mathrm{TD}(c_j, \{c_j, \ldots, c_n\}) = 1$ for some $j < \Phi_{r-1}(N_2, \ldots, N_m)$. W.l.o.g, let $(X_1, 1) \in \mathrm{TS}(c_j, \{c_j, \ldots, c_n\})$. Then we have $c(X_1) \neq 1$ for any $c \in \{c_{j+1}, \ldots, c_n\}$. So, $\{c_{j+1}, \ldots, c_n\} \subseteq X_1 \setminus \{1\} \times \prod_{i=2}^{m} X_i$. Consequently,

$$\begin{aligned} |C| &= |\{c_1, \ldots, c_j\}| + |\{c_{j+1}, \ldots, c_n\}| = j + |\{c_{j+1}, \ldots, c_n\}| \\ &\leq j + \Phi_r(N_1 - 1, N_2, \ldots, N_m) \quad \text{(by Corollary 6.5)} \\ &< \Phi_{r-1}(N_2, \ldots, N_m) + \Phi_r(N_1 - 1, N_2, \ldots, N_m) \\ &= \Phi_r(N_1, \ldots, N_m). \end{aligned}$$

Thus C is not RTD-maximum, which is a contradiction. □

Note that Proposition 6.21 is tight in the sense that there is a binary RTD-maximum class of RTD r that possesses an optimal teaching plan in which the concept at position $i = \Phi_{r-1}(m-1)$ has a recursive teaching set of size 1. In particular, the

standard binary VCD-maximum class, the class of all binary concepts of size at most r, fulfills this property. The witnessing optimal teaching plan begins with all concepts of size r that contain an arbitrary but fixed instance $X_t \in X$, followed by those of size $r-1$ that contain X_t and so on. The concept c containing only X_t occurs at position $\binom{m-1}{r-1} + \cdots + \binom{m-1}{1} + 1 = \Phi_{r-1}(m-1)$ in this plan. After that point, no concept in the plan contains X_t, and hence $\{(X_t, 1)\}$ is a recursive teaching set of size one for c in the chosen plan.

As a generalization of Proposition 6.21, we observe that, in any teaching plan for any RTD-maximum class C, recursive teaching sets S of size less than $\mathrm{RTD}(C)$ can only be used after all instance sets $X' \supset X(S)$ of size $\mathrm{RTD}(C)$ have been used as recursive teaching sets.

Proposition 6.22. *Let C be* RTD*-maximum of* RTD *r and $\mathcal{P} = ((c_1, S_1), \ldots, (c_n, S_n))$ be a teaching plan of order r for C. Suppose there is some $j \in \{1, \ldots, n\}$ such that $|X(S_j)| \leq r-1$. Then for all sets X' with $X(S_j) \subset X' \subseteq X$ and $|X'| = r$, there is an index $k \in \{1, \ldots, j-1\}$ such that $X(S_k) = X'$.*

Proof. W.l.o.g., assume that there is an index j such that $S_j = \{(X_1, n_1), \ldots, (X_t, n_t)\}$, where $t \leq r-1$, but $X(S_k) \neq \{X_1, \ldots, X_t, X_{t+1}, \ldots, X_r\}$, for every $k < j$. That is, $c_1, \ldots, c_{j-1}$ do not contain the monomial $X_1^{N_1} X_2^{N_2} \cdots X_r^{N_r}$ in their expression as a linear combination of $P^r(N_1, \ldots, N_m)$. In particular, $\mathbf{c_1}, \ldots, \mathbf{c_{j-1}}$ do not contain the vector $\mathbf{v}$ that corresponds to the monomial $X_1^{N_1} X_2^{N_2} \cdots X_r^{N_r}$. We will use the same notation as in the proof of Theorem 6.3.

Note that $P^r(N_1, \ldots, N_m)$ is a basis for $\mathbb{R}^{|C|}$ because $P^r(N_1, \ldots, N_m)$ spans $\mathbb{R}^{|C|}$ and $|C| = \Phi_r(N_1, \ldots, N_m)$. Thus every concept $c \in C$ can be expressed as a linear combination of monomials from $P^r(N_1, \ldots, N_m)$ in a unique way. We now express c_j as a linear combination of monomials from $P^r(N_1, \ldots, N_m)$ in two different ways:

one will contain the monomial $X_1^{N_1}X_2^{N_2}\cdots X_r^{N_r}$ and the other will not.

Since $S_j = \{(X_1, n_1), \ldots, (X_t, n_t)\}$, by Lemma 3.16, $(\mathbf{c_j})_{0,n-j} = \mathbf{p}$ such that $\mathbf{p}$ corresponds to the polynomial

$$\begin{aligned} p(X_1, \ldots, X_t) &= p_{1,n_1}(X_1)\cdots p_{t,n_t}(X_t) \\ &= X_1^{N_1}X_2^{N_2}\cdots X_t^{N_t} + L(X_1, \ldots, X_t), \end{aligned}$$

where $L(X_1, \ldots, X_t)$ is a linear combination of monomials in

$$P^t(N_1, \ldots, N_t) \setminus \{X_1^{N_1}X_2^{N_2}\cdots X_t^{N_t}\}.$$

In particular, the linear combination for $(\mathbf{c_j})_{0,n-j}$ does not contain $\mathbf{v}$. Note that $\mathbf{c_j}$ is equal to $(\mathbf{c_j})_{0,n-j}$ plus a linear combination of $\mathbf{c_1}, \ldots, \mathbf{c_{j-1}}$. By assumption, none of $\mathbf{c_1}, \ldots, \mathbf{c_{j-1}}$ contains $\mathbf{v}$. Therefore, c_j can be expressed as a linear combination of monomials from $P^r(N_1, \ldots, N_m)$ in which the monomial $X_1^{N_1}X_2^{N_2}\cdots X_r^{N_r}$ does not occur.

Notice that any superset of $X(S_j)$, and in particular the set $\{X_1, \ldots, X_r\}$, is also a recursive teaching set for c_j according to $\mathcal{P}$. Let $c_j|_{\{X_1,\ldots,X_r\}} = (n_1, \ldots, n_t, n_{t+1}, \ldots, n_r)$. Again, by Lemma 3.16, we can write $(\mathbf{c_j})_{0,n-j} = \mathbf{p'}$ such that $\mathbf{p'}$ corresponds to the polynomial

$$\begin{aligned} p'(X_1, \ldots, X_r) &= p_{1,n_1}(X_1)\cdots p_{r,n_r}(X_r) \\ &= X_1^{N_1}X_2^{N_2}\cdots X_r^{N_r} + L(X_1, \ldots, X_r), \end{aligned}$$

where $L(X_1, \ldots, X_r)$ is a linear combination of monomials in

$$P^r(N_1, \ldots, N_r) \setminus \{X_1^{N_1}X_2^{N_2}\cdots X_r^{N_r}\}.$$

Hence $\mathbf{v}$ appears in the linear combination for $(\mathbf{c_j})_{0,n-j}$. As before, $\mathbf{c_j}$ is equal to $(\mathbf{c_j})_{0,n-j}$ plus a linear combination of $\mathbf{c_1}, \ldots, \mathbf{c_{j-1}}$ and, by assumption, none of $\mathbf{c_1}, \ldots, \mathbf{c_{j-1}}$ contains $\mathbf{v}$. So, $\mathbf{c_1}, \ldots, \mathbf{c_{j-1}}$ cannot cancel out $\mathbf{v}$ from $(\mathbf{c_j})_{0,n-j}$. Therefore, c_j can be expressed as a linear combination of monomials from $P^r(N_1, \ldots, N_m)$ which contains $X_1^{N_1} X_2^{N_2} \cdots X_r^{N_r}$.

Thus we have expressed c_j as a linear combination of monomials in $P^r(N_1, \ldots, N_m)$ in two different ways. This contradicts the fact that $P^r(N_1, \ldots, N_m)$ is a basis. □

6.2 RTD-maximal Classes

In this section we present some properties of RTD-maximal classes. We first need to introduce the notion of fully shattering. The set $Y = \{X_{i_1}, \ldots, X_{i_k}\} \subseteq X$ is *fully shattered* by $C \subseteq \prod_{1 \leq i \leq m} X_i$ iff $C|_Y = X_{i_1} \times \cdots \times X_{i_k}$. In particular, $\text{size}(C|_Y) = (N_{i_1} + 1) \times \cdots \times (N_{i_k} + 1)$.

The next proposition shows that an RTD-maximal class fully shatters each subset of the instance space whose size is equal to RTD.

Proposition 6.23. *Let C be* RTD*-maximal with* $\text{RTD}(C) = r$*. Then, for any subset $X' \subseteq X$ with $|X'| = r$, C fully shatters X'.*

Proof. Assume $X' = \{X_{i_1}, \ldots, X_{i_r}\}$ is not fully shattered by C. Then $\text{size}(C|_{X'}) < (N_{i_1} + 1) \times \cdots \times (N_{i_r} + 1)$ and we can add a new concept c_{new} to C such that $c_{\text{new}}|_{X'} \notin C|_{X'}$ and thus, $\text{TD}(c_{\text{new}}, C \cup \{c_{\text{new}}\}) \leq r$. Since $\text{RTD}(C) = r$, C has a teaching plan of order r. So, $C \cup \{c_{\text{new}}\}$ also has a teaching plan of order r, which starts with c_{new} and then continues with any teaching plan for C of order r. Therefore, $\text{RTD}(C \cup \{c_{\text{new}}\}) \leq r$ and C is not RTD-maximal. □

Note that Proposition 6.23 is not true in general for VCD-maximal classes. Table

6.4 contains an example of a VCD-maximal class, found by computer experiments, that does not shatter all subsets of the instance space whose size is equal to the VC-dimension.

$c \in C$	X_1	X_2	X_3	X_4	X_5
c_1	0	1	0	1	1
c_2	0	1	1	0	1
c_3	0	1	1	1	0
c_4	1	0	0	0	0
c_5	1	0	0	0	1
c_6	1	0	0	1	0
c_7	1	0	1	0	0
c_8	1	1	0	0	0
c_9	1	1	0	0	1
c_{10}	1	1	0	1	0
c_{11}	1	1	0	1	1
c_{12}	1	1	1	0	0
c_{13}	1	1	1	0	1
c_{14}	1	1	1	1	0

Table 6.4: VCD-maximal class of VCD 2 that does not shatter the subset $\{X_1, X_2\}$ (found by computer experiments).

As a corollary of Proposition 6.23 we can show that, for any RTD-maximal class, the minimal teaching dimension and the recursive teaching dimension coincide.

Corollary 6.24. *For any* RTD*-maximal class* C, $\mathrm{TD}_{min}(C) = \mathrm{RTD}(C)$.

Proof. First, note that $\mathrm{TD}_{min}(C) \leq \mathrm{RTD}(C)$. Now assume $\mathrm{TD}_{min}(C) < \mathrm{RTD}(C)$. In this case, there is a concept $c \in C$ for which $\{X_{i_1}, \ldots, X_{i_k}\}$ is a teaching set, for some $k < \mathrm{RTD}(C)$. Consider any subset $X' \subseteq X$ such that $|X'| = \mathrm{RTD}(C)$ and $\{X_{i_1}, \ldots, X_{i_k}\} \subset X'$. Then C does not shatter X', since otherwise there would exist at least one more concept $c' \in C$ with $c'|_{\{X_{i_1},\ldots,X_{i_k}\}} = c|_{\{X_{i_1},\ldots,X_{i_k}\}}$. This is impossible because $\{X_{i_1}, \ldots, X_{i_k}\}$ is a teaching set for c in C. Hence, by Proposition 6.23, C cannot be RTD-maximal — a contradiction. □

The next lemma demonstrates that for any RTD-maximal class C of RTD 1 and

any teaching plan $\mathcal{P} = \{(c_1, S_1), \ldots, (c_n, S_n)\}$ of order 1 for C, each concept c_i, $i \in \{1, \ldots, n-1\}$, has only one minimal teaching set in the class $\{c_i, c_{i+1}, \ldots, c_n\}$.

Lemma 6.25. *Let C be* RTD*-maximal with* $\mathrm{RTD}(C) = 1$, $n = |C|$ *and* $\mathcal{P} = \{(c_1, S_1), \ldots, (c_n, S_n)\}$ *be a teaching plan of order* 1 *for* C. *Let* $S'_t \in TS(c_t, \{c_t, \ldots, c_n\})$, $t \in \{1, \ldots, n-1\}$. *Then either* $S'_t = S_t$ *or* $|S'_t| > 1$.

Proof. For the purpose of contradiction, assume that for some $t \in \{1, \ldots, n-1\}$, there is $S'_t \in TS(c_t, \{c_t, \ldots, c_n\})$ such that $S'_t \neq S_t$ and $|S'_t| = 1$. W.l.o.g., let $S_t = \{(X_1, 1)\}$ and $S'_t = \{(X_2, 1)\}$, that is, $c_t(X_1) = c_t(X_2) = 1$ and for all $i \in \{t+1, \ldots, n\}$, $c_i(X_1) \neq 1$ and $c_i(X_2) \neq 1$. W.l.o.g., let $0 \in \{c_i(X_1) \mid i \in \{t+1, \ldots, n\}\}$, which implies

$$\{(X_1, 0)\} \notin \mathrm{TS}(c_i, \{c_i, \ldots, c_{t-1}, c_t, \ldots, c_n\}),$$

for all $i \in \{1, \ldots, t-1\}$. We now show that a concept $c_{\mathrm{new}} \notin C$ can be added to C such that $\mathrm{RTD}(C \cup \{c_{\mathrm{new}}\}) = \mathrm{RTD}(C)$. This contradicts the fact that C is RTD-maximal.

Let $c_{\mathrm{new}} - X_1 = c_t - X_1$ and $c_{\mathrm{new}}(X_1) = 0$, in particular, $c_{\mathrm{new}}(X_1) \neq c_t(X_1)$ and $c_{\mathrm{new}}(X_1) = c_k(X_1)$, for some $k \in \{t+1, \ldots, n\}$. We first show that $c_{\mathrm{new}} \notin C$. On the one hand, $(X_2, c_t(X_2)) \in \mathrm{TS}(c_t, \{c_t, \ldots, c_n\})$, or equivalently, $c_{\mathrm{new}}(X_2) = c_t(X_2) \neq c_i(X_2)$, for all $i \in \{t+1, \ldots, n\}$, and thus $c_{\mathrm{new}} \notin \{c_{t+1}, \ldots, c_n\}$. On the other hand, $c_{\mathrm{new}} \notin \{c_1, \ldots, c_{t-1}\}$, because $\mathrm{TD}(c_{\mathrm{new}}, \{c_{\mathrm{new}}, c_t, \ldots, c_n\}) > 1$ while $\mathrm{TD}(c_i, \{c_i, c_{i+1}, \ldots, c_t, c_{t+1}, \ldots, c_n\}) = 1$, for all $i \in \{1, \ldots, t-1\}$.

We next prove that $\mathrm{RTD}(C \cup \{c_{\mathrm{new}}\}) = \mathrm{RTD}(C)$. In fact, we show that the plan

$$\mathcal{P}' = \{(c_1, S_1), \ldots, (c_t, S_t), (c_{\mathrm{new}}, S'_t), (c_{t+1}, S_{t+1}), \ldots, (c_n, S_n)\}$$

is a teaching plan of order 1 for $C \cup \{c_{\mathrm{new}}\}$. By the definition of c_{new}, it is easy to see

that for all $i \in \{1, \ldots, t-1\}$,

$$S_i \in \mathrm{TS}(c_i, \{c_i, \ldots, c_{t-1}, c_t, c_{\mathrm{new}}, c_{t+1} \ldots, c_n\}).$$

Moreover,

$$S_t = \{(X_1, 1)\} \in \mathrm{TS}(c_t, \{c_t, c_{\mathrm{new}}, c_{t+1}, \ldots, c_n\}),$$

since $c_{\mathrm{new}}(X_1) = 0$, and by assumption,

$$S'_t = \{(X_2, 1)\} \in \mathrm{TS}(c_{\mathrm{new}}, \{c_{\mathrm{new}}, c_{t+1}, \ldots, c_n\}).$$

Therefore, $C \cup \{c_{\mathrm{new}}\}$ is of RTD 1, i.e, $\mathrm{RTD}(C \cup \{c_{\mathrm{new}}\}) = \mathrm{RTD}(C)$. □

In the binary case, it is not hard to see that VCD-maximal classes of VC-dimension 1 are VCD-maximum [WW87]. We now show that the same holds for RTD-maximal classes, even in the multi-label case.

Proposition 6.26. *Let C be* RTD*-maximal. If* $\mathrm{RTD}(C) = 1$*, then C is* RTD*-maximum.*

Proof. Let $n = |C|$ and $\mathcal{P} = \{(c_1, S_1), \ldots, (c_n, S_n)\}$ be a teaching plan of order 1 for C. Consider the concept c_n and the set

$$E = \{(X_j, l) \mid 1 \leq j \leq m \text{ and } l \in (X_j \setminus \{c_n(X_j)\})\}.$$

Our goal is to show that $S_1 \cup S_2 \cup \ldots \cup S_{n-1} = E$ and thus

$$\begin{aligned} |C| &= |\{c_1, \ldots, c_n\}| = |S_1 \cup S_2 \cup \ldots \cup S_{n-1}| = |S_1 \cup S_2 \cup \ldots \cup S_{n-1}| + 1 \\ &= |E| + 1 = N_1 + \cdots + N_m + 1 \\ &= \Phi_1(N_1, \ldots, N_m). \end{aligned}$$

Obviously, for all $i \in \{1, \ldots, n-1\}$, S_i is not consistent with c_n. That is, $S_i \subseteq E$, for all $i \in \{1, \ldots, n-1\}$, or equivalently $S_1 \cup S_2 \cup \ldots \cup S_{n-1} \subseteq E$.

Assume that $E \nsubseteq S_1 \cup S_2 \cup \ldots \cup S_{n-1}$, in particular, there is $e = (X', l') \in E$ such that $e \notin S_1 \cup S_2 \cup \ldots \cup S_{n-1}$. By Proposition 6.23, C must contain a concept consistent with e. We show that $\{e\}$ is in fact a teaching set for a concept c_t, $t \in \{1, \ldots, n-1\}$, in $\{c_t, \ldots, c_{n-1}\}$, which contradicts Lemma 6.25. Note that $\{e\}$ is not consistent with c_n, since $l' \in X' \setminus \{c_n(X')\}$. So, e must be consistent with a concept in $\{c_1, \ldots, c_{n-1}\}$. Let $t \in \{1, \ldots, n-1\}$, such that $c_t(X') = l'$ and $c_i(X') \neq l'$, for all $i \in \{t+1, \ldots, n-1\}$. Obviously, $\{e\} \in \mathrm{TS}(c_t, \{c_t, \ldots, c_n\})$. □

Surprisingly, not all RTD-maximal classes are RTD-maximum.

Proposition 6.27. *There is an* RTD*-maximal class that is not* RTD*-maximum.*

Proof. Consider the RTD-maximal class C in Table 6.5. Since $\mathrm{RTD}(C) = 2$ and $|C| = 13 < \Phi_2(5)$, C is not RTD-maximum. This class was found by exhaustive enumeration of all concept classes over instance spaces of size at most 5. □

$c \in C$	X_1	X_2	X_3	X_4	X_5
c_1	$\underline{0}$	$\underline{0}$	1	1	1
c_2	$\underline{1}$	$\underline{1}$	1	1	1
c_3	0	1	0	$\underline{1}$	$\underline{1}$
c_4	$\underline{0}$	1	$\underline{0}$	1	0
c_5	0	$\underline{1}$	1	0	$\underline{1}$
c_6	$\underline{0}$	1	1	$\underline{0}$	0
c_7	$\underline{0}$	1	1	1	0
c_8	1	0	$\underline{0}$	0	$\underline{1}$
c_9	1	0	1	0	$\underline{1}$
c_{10}	1	0	$\underline{0}$	$\underline{0}$	0
c_{11}	1	0	$\underline{0}$	1	0
c_{12}	1	0	1	$\underline{0}$	0
c_{13}	1	0	1	1	0

Table 6.5: C is RTD-maximal but not RTD-maximum (the class was found by computer experiments). Recursive teaching sets are underlined.

Chapter 7

Conclusions

In conclusion, we summarize the original contributions of this thesis, and discuss potential future research directions.

7.1 Linear Algebraic Approach

In Chapter 3, we broadly discussed the application of Linear Algebra for establishing results on combinatorial parameters in Computational Learning Theory. In particular, we introduced our algebraic characterization of teaching sets in Section 3.3, which turned out to be essential in proving our Sauer-type upper bound on the size of concept classes of a given RTD. Another consequence of our characterization was shown in Section 3.4, where we proved that the hypergraph of VCD_Ψ-maximum classes is shortest-path closed. Our proof is applied to the binary case, i.e., our algebraic approach is an alternative proof of shortest-path closedness of one-inclusion graphs for VCD-maximum classes [KW07].

7.2 The Reduction Property

In Chapter 4, we introduced the reduction property for VCD_Ψ and classified the known analogues of the VCD according to whether or not they fulfill the reduction property. To the best of our knowledge, this property has been never studied in the literature. The proof procedure of Theorem 4.4 shows that the fulfillment of the reduction property is not trivial at all. We showed that if VCD_Ψ fulfills the reduction property, then the reduction of any VCD_Ψ-maximum class of VCD_Ψ d is VCD_Ψ-maximum of VCD_Ψ $d-1$. In Section 4.1, we proved that the Graph-dimension fulfills the reduction property. However, in Section 4.2 and Section 4.3, we illustrated that neither Pollard's pseudo-dimension nor the Natarajan-dimension fulfill the reduction property.

7.3 Sample Compression Schemes

In Chapter 5, we extended the study of sample compression schemes to multi-label concept classes. We first in Section 5.1 showed that, as in the binary case, the smallest possible size of a sample compression scheme yields sample bounds for PAC-learning in the multi-label case. We then studied sample compression schemes for VCD_Ψ-maximum classes, where VCD_Ψ fulfills the reduction property. In Section 5.2, we extended Floyd and Warmuth's compression scheme to the multi-label case and afterwards, in Section 5.3 we introduced the notion of a tight compression scheme and proved that any VCD_Ψ-maximum class has a tight compression scheme of size of its VCD_Ψ. In Section 5.4, we showed that every tight SCS for a VCD_Ψ-maximum class C maps any concept $c \in C$ to a sample set $S \subseteq c$, such that the instances appearing in S label the incident hyperedges to c in the one-inclusion hypergraph for C. In

Section 5.5, we focused on classes of VCD_Ψ 1 and proved that any concept class of Graph-dimension 1 has a sample compression scheme of size 1.

7.4 Recursive Teaching Dimension

In Chapter 6, we studied the recursive teaching dimension for multi-label concept classes (all results are applied to the binary case) and established a further connection between RTD and VCD in both the multi-label and binary cases. We first proved that classes of Graph-dimension 1 have RTD 1. Then in Section 6.1, we established a Sauer-type upper bound on the size of concepts classes of a given RTD by utilizing our algebraic characterization of teaching sets (Theorem 6.3). We defined RTD-maximum classes to be the classes of a fixed RTD whose size meets our Sauer-type bound, and we also made a close connection between RTD-maximum classes and VCD_Ψ-maximum classes. Some properties of the teaching plans of RTD-maximum classes were identified and discussed as well. In Section 6.2, we studied RTD-maximal classes, classes of a given RTD in which the RTD increases by adding any new concept, and we connected RTD-maximality to the notion of (fully) shattering. We identified some properties of RTD-maximal classes and finally showed that proper RTD-maximal classes exist.

7.5 Future Research Directions

In this thesis, we encountered many interesting problems that are still unresolved. We discuss these problems in an order that follows the structure of the thesis.

Open Problem 1. Is there any alternative definition for the one-inclusion hypergraph for multi-label concept classes that is as interesting and useful as the one by

Rubinstein et al. [RBR09]?

The one-inclusion hypergraph can be defined in different ways for multi-label concept classes and each might have interesting characterizations and properties.

Open Problem 2. What properties of the one-inclusion graph can be generalized to the one-inclusion hypergraph?

In this thesis, we discussed two properties: (i) the shortest-path closedness property, (ii) connection between the tight compression sets for each concept and the instances of the incident edges on each concept.

Open Problem 3. What properties does a family of mappings Ψ have in order for VCD_{Ψ} to fulfill the reduction property?

Although we could successfully identify the reduction property for VCD_{Ψ}, we have not been able to find the root of this property in the mappings. It is an interesting combinatorial problem to see what necessary or sufficient conditions for the mappings in Ψ result in the reduction property for VCD_{Ψ}.

Open Problem 4. What are other interesting geometric examples of $\text{VCD}_{\Psi_{\text{G}}}$-maximum classes?

We described a geometric example of a $\text{VCD}_{\Psi_{\text{G}}}$-maximum class in Example 3.13. There might exist more such examples that can illustrate some interesting properties of $\text{VCD}_{\Psi_{\text{G}}}$-maximum classes.

Open Problem 5. Is there any interesting geometric example of a $\text{VCD}_{\Psi_{\text{P}}}$-maximum class?

We believe so, but we could not find any.

Open Problem 6. Is there any proper $\mathrm{VCD}_{\Psi_\mathrm{P}}$-maximal ($\mathrm{VCD}_{\Psi_\mathrm{N}}$-maximal) class of $\mathrm{VCD}_{\Psi_\mathrm{P}}$ ($\mathrm{VCD}_{\Psi_\mathrm{N}}$) 1?

If the answer is 'no', this would imply that every class of $\mathrm{VCD}_{\Psi_\mathrm{P}}$ ($\mathrm{VCD}_{\Psi_\mathrm{N}}$) 1 is contained in a $\mathrm{VCD}_{\Psi_\mathrm{P}}$-maximum ($\mathrm{VCD}_{\Psi_\mathrm{N}}$-maximum) class of $\mathrm{VCD}_{\Psi_\mathrm{P}}$ ($\mathrm{VCD}_{\Psi_\mathrm{N}}$) 1 and thus has an SCS of size 1; consequently, the answer to the next two problems would be 'yes'. If the answer is 'yes', then the existence of SCSs for $\mathrm{VCD}_{\Psi_\mathrm{P}}$-maximum ($\mathrm{VCD}_{\Psi_\mathrm{N}}$-maximum) classes would not imply the existence of SCSs for every class of $\mathrm{VCD}_{\Psi_\mathrm{P}}$ ($\mathrm{VCD}_{\Psi_\mathrm{N}}$) 1.

Open Problem 7. Does every class of $\mathrm{VCD}_{\Psi_\mathrm{P}}$ 1 have an SCS of size 1?

Open Problem 8. Does every class of $\mathrm{VCD}_{\Psi_\mathrm{N}}$ 1 have an SCS of size 1?

Open Problem 9. Does every RTD-maximum binary class have a repetition-free teaching plan?

A teaching plan $P = ((c_1, S_1), \ldots, (c_n, S_n))$ is called *repetition-free*, if $X(S_i) \neq X(S_j)$ for all $i, j \in \{1, \ldots, n\}$ with $i \neq j$. We know that every VCD-maximum class C possesses a repetition-free teaching plan of order $\mathrm{RTD}(C) = \mathrm{VCD}(C)$, but it remains open whether the same is true for RTD-maximum binary classes in general.

Open Problem 10. Can every binary class with a repetition-free teaching plan be extended to an RTD-maximum binary class without increasing the RTD?

If the answer is 'yes', this would imply that any RTD-maximal binary class with a repetition-free teaching plan must already be RTD-maximum. If the answer is 'no', this would immediately lead to the question under which conditions on the repetition-free teaching plan of a binary class C one could conclude that C is contained in an RTD-maximum binary class of the same RTD.

References

[AHW87] N. Alon, D. Haussler, and E. Welzl. Partitioning and geometric embedding of range spaces of finite Vapnik-Chervonenkis dimension. In *Proceedings of the Third Annual Symposium on Computational Geometry (SCG)*, pages 331–340, 1987.

[Alo83] N. Alon. On the density of sets of vectors. *Discrete Mathematics*, 46(2):199–202, 1983.

[Bal08] F. J. Balbach. Measuring teachability using variants of the teaching dimension. *Theoretical Computer Science*, 397(1–3):94 – 113, 2008.

[BCHL95] S. Ben-David, N. Cesa-Bianchi, D. Haussler, and P. M. Long. Characterizations of learnability for classes of $\{0, ..., n\}$-valued functions. *Journal of Computer and System Sciences*, 50(1):74–86, 1995.

[BEHW89] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36(4):929–965, 1989.

[BL98] S. Ben-David and A. Litman. Combinatorial variability of Vapnik-Chervonenkis classes with applications to sample compression schemes. *Discrete Applied Mathematics*, 86(1):3–25, 1998.

[DFSZ14] T. Doliwa, G. Fan, H. U. Simon, and S. Zilles. Recursive teaching dimension, VC-dimension, and sample compression. *Journal of Machine Learning Research*, 15:3107–3131, 2014.

[DS14] A. Daniely and S. Shalev-Shwartz. Optimal learners for multiclass problems. In *Proceedings of the 27th Annual Conference on Learning Theory (COLT)*, pages 287–316, 2014.

[DSBS11] A. Daniely, S. Sabato, S. Ben-David, and S. Shalev-Shwartz. Multiclass learnability and the ERM principle. In *Proceedings of the 24th Annual Conference on Learning Theory (COLT)*, pages 207–232, 2011.

[DSZ10] T. Doliwa, H. U. Simon, and S. Zilles. Recursive teaching dimension, learning complexity, and maximum classes. In *Proceedings of the 21st International Conference on Algorithmic Learning Theory (ALT)*, volume 6331 of *Lecture Notes in Artificial Intelligence*, pages 209–223. Springer, 2010.

[Flo89] S. Floyd. *On space-bounded learning and the Vapnik-Chervonenkis dimension*. PhD thesis, International Computer Science Institute, Berkeley, CA, 1989.

[FS12] Z. Füredi and A. Sali. Optimal multivalued shattering. *SIAM Journal on Discrete Mathematics*, 26:737–744, 2012.

[FW95] S. Floyd and M. K. Warmuth. Sample compression, learnability, and the Vapnik-Chervonenkis dimension. *Machine Learning*, 21(3):269–304, 1995.

[GK91] S. A. Goldman and M. J. Kearns. On the complexity of teaching. In

Proceedings of the Fourth Annual Workshop on Computational Learning Theory (COLT), pages 303–314, 1991.

[GK95] S. A. Goldman and M. J. Kearns. On the complexity of teaching. *Journal of Computer and System Sciences*, 50:20–31, 1995.

[Gur97] L. Gurvits. Linear algebraic proofs of VC-dimension based inequalities. In *Proceedings of the Third European Conference on Computational Learning Theory (COLT)*, pages 238–250, London, UK, 1997. Springer-Verlag.

[HL95] D. Haussler and P. M. Long. A generalization of Sauer's lemma. *Journal of Combinatorial Theory, Series A*, 71(2):219–240, 1995.

[KW07] D. Kuzmin and M. K. Warmuth. Unlabeled compression schemes for maximum classes. *Journal of Machine Learning Research*, 8:2047–2081, 2007.

[LP86] L. Lovász and M. D. Plummer. *Matching Theory*, volume 121, page 139. North-Holland Mathematics Studies, North-Holland Publishing, Amsterdam, 1986.

[LW86] N. Littlestone and M. Warmuth. Relating data compression and learnability. Unpublished notes, 1986.

[Nat89] B. K. Natarajan. On learning sets and functions. *Machine Learning*, 4(1):67–97, 1989.

[Pol90] D. Pollard. Empirical Processes: Theory and Applications. *NSF-CBMS Regional Conference Series in Probability and Statistics*, 2:pp. i–iii+v+vii–viii+1–86, 1990.

[RBR09] B. I. P. Rubinstein, P. L. Bartlett, and J. H. Rubinstein. Shifting: one-inclusion mistake bounds and sample compression. *Journal of Computer and System Sciences*, 75(1):37–59, 2009.

[RR08] B. I. P. Rubinstein and J. H. Rubinstein. Geometric & topological representations of maximum classes with applications to sample compression. In *Proceedings of the 21st Annual Conference on Learning Theory (COLT)*, pages 299–310, 2008.

[RR12] B. I. P. Rubinstein and J. H. Rubinstein. A geometric approach to sample compression. *Journal of Machine Learning Research*, 13(1):1221–1261, 2012.

[Sau72] N. Sauer. On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13(1):145–147, 1972.

[She72] S. Shelah. A combinatorial problem: stability and order for models and theories in infinitary languages. *Pacific Journal of Mathematics*, 4:247–261, 1972.

[SM91] A. Shinohara and S. Miyano. Teachability in computational learning. *New Generation Computing*, 8(4):337–347, 1991.

[Smo97] R. Smolensky. Well-known bound for the VC-dimension made easy. *Computational Complexity*, 6(4):299–300, 1997.

[SS10] H. U. Simon and B. Szörényi. One-inclusion hypergraph density revisited. *Information Processing Letters*, 110(8-9):341–344, 2010.

[SSYZ12] R. Samei, P. Semukhin, B. Yang, and S. Zilles. Sauer's bound for a

notion of teaching complexity. In *Proceedings of the 23rd International Conference on Algorithmic Learning Theory (ALT)*, pages 96–110, 2012.

[SSYZ14a] R. Samei, P. Semukhin, B. Yang, and S. Zilles. Algebraic methods proving sauer's bound for teaching complexity. *Theoretical Computer Science*, 558:35–50, 2014.

[SSYZ14b] R. Samei, P. Semukhin, B. Yang, and S. Zilles. Sample compression for multi-label concept classes. In *Proceedings of the 27th Annual Conference on Learning Theory (COLT)*, pages 371–393, 2014.

[SYZ14] R. Samei, B. Yang, and S. Zilles. Generalizing labeled and unlabeled sample compression to multi-label concept classes. In *Proceedings of the 25th International Conference on Algorithmic Learning Theory (ALT)*, pages 275–290, 2014.

[Val84] L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[Vap89] V. N. Vapnik. Inductive principles of the search for empirical dependences (methods based on weak convergence of probability measures). In *Proceedings of the 2nd Annual Workshop on Computational Learning Theory (COLT)*, pages 3–21, 1989.

[VC71] V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16:264–280, 1971.

[Wel87] E. Welzl. Complete range spaces. Unpublished notes, 1987.

[WW87] E. Welzl and G. Woeginger. On Vapnik-Chervonenkis dimension one. Unpublished notes, 1987.

[ZLHZ11] S. Zilles, S. Lange, R. Holte, and M. Zinkevich. Models of cooperative teaching and learning. *Journal of Machine Learning Research*, 12:349–384, 2011.

[ZZ13] C. Zhongyuana and C. Zhibob. Conjugated circuits and forcing edges. *Communications in Mathematical and in Computer Chemistry / MATCH*, 69(3):721–732, 2013.

Appendix A

Binary RTD-maximum Classes

As mentioned in Section 6.1, the complement of any binary VCD-maximum class is VCD-maximum. But RTD-maximum classes may not possess an analogous property.

Proposition A.1. *There is an* RTD*-maximum class whose complement is not* RTD*-maximum.*

Proof. Consider the RTD-maximum class C with $\mathrm{RTD}(C) = 3$ in Table A.1. $\overline{C}$ is not RTD-maximum because $\mathrm{RTD}(\overline{C}) = 2$ and $6 < \Phi_2(5) = 16$. □

By contrast with Proposition A.1, we can show that the complement of a binary RTD-maximum class of RTD 1 is still RTD-maximum. As shown in Corollary 6.14, this result cannot be extended to the multi-label case.

Proposition A.2. *Let* $m \in \mathbb{N}^+$ *with* $m \geq 2$ *and* $X_i = \{0, 1\}$*, for all* $i \in \{1, \ldots, m\}$*. Let* $X = \{X_1, \ldots, X_m\}$ *and* $C \subseteq \prod_{i=1}^{m} X_i$ *be* RTD*-maximum. If* $\mathrm{RTD}(C) = 1$*, then* $\overline{C}$ *is* RTD*-maximum and* $\mathrm{RTD}(\overline{C}) = |X|-2$*.*

Proof. For $|X| = 2$, $|C| = \Phi_1(2) = 3$ and thus $\overline{C}$ contains only one concept. So, $\mathrm{RTD}(\overline{C}) = 0$ and $\overline{C}$ is RTD-maximum with $\mathrm{RTD}(\overline{C}) = |X|-2$.

$c \in C$	X_1	X_2	X_3	X_4	X_5
c_1	$\underline{1}$	$\underline{1}$	$\underline{1}$	1	1
c_2	1	1	$\underline{0}$	$\underline{1}$	$\underline{1}$
c_3	$\underline{1}$	$\underline{1}$	0	$\underline{1}$	0
c_4	$\underline{1}$	$\underline{1}$	0	0	$\underline{1}$
c_5	0	$\underline{1}$	1	$\underline{1}$	$\underline{1}$
c_6	$\underline{1}$	0	1	$\underline{1}$	$\underline{1}$
c_7	0	0	1	$\underline{1}$	$\underline{1}$
c_8	$\underline{1}$	$\underline{1}$	0	0	0
c_9	$\underline{1}$	0	$\underline{1}$	0	$\underline{1}$
c_{10}	$\underline{1}$	0	0	0	$\underline{1}$
c_{11}	0	$\underline{1}$	$\underline{1}$	$\underline{1}$	0
c_{12}	0	$\underline{1}$	0	$\underline{1}$	0
c_{13}	0	$\underline{1}$	$\underline{1}$	0	$\underline{1}$

$c \in C$	X_1	X_2	X_3	X_4	X_5
c_{14}	0	$\underline{1}$	0	0	$\underline{1}$
c_{15}	$\underline{1}$	0	$\underline{1}$	$\underline{1}$	0
c_{16}	$\underline{1}$	0	0	$\underline{1}$	0
c_{17}	0	$\underline{1}$	$\underline{1}$	0	0
c_{18}	0	$\underline{1}$	0	0	0
c_{19}	0	0	$\underline{1}$	$\underline{1}$	0
c_{20}	0	0	0	$\underline{1}$	0
c_{21}	$\underline{1}$	0	$\underline{1}$	0	0
c_{22}	$\underline{1}$	0	0	0	0
c_{23}	0	0	$\underline{1}$	0	$\underline{1}$
c_{24}	0	0	$\underline{1}$	0	0
c_{25}	0	0	0	0	$\underline{1}$
c_{26}	0	0	0	0	0

$c \in \overline{C}$	X_1	X_2	X_3	X_4	X_5
c_1	$\underline{0}$	$\underline{0}$	0	1	1
c_2	$\underline{0}$	1	0	1	1
c_3	1	0	$\underline{0}$	1	1
c_4	1	1	1	0	$\underline{1}$
c_5	1	1	1	$\underline{0}$	0
c_6	1	1	1	1	0

Table A.1: C is RTD-maximum but $\overline{C}$ is not. Recursive teaching sets are underlined.

Suppose the claim is true for $|X| < m$. Now consider the case $|X| = m > 2$. Let $c_1 \in C$ with $\mathrm{TD}(c_1, C) = 1$. W.l.o.g., let $\{(X_1, 1)\}$ be a teaching set for c_1 in C. Then we can write C as a disjoint union of $\{c_1\}$ and $\{0\} \times C_1$, where $C_1 = (C \setminus \{c_1\}) - X_1$ is an RTD-maximum class with $\mathrm{RTD}(C_1) = 1$ on $X \setminus \{X_1\}$. So, the complement of C is equal to the disjoint union $\overline{C} = (\{0\} \times \overline{C_1}) \cup (\{1\} \times C_2)$, where $C_2 = 2^{X \setminus \{X_1\}} \setminus \{c_1 - X_1\}$ is a class of size $2^{m-1} - 1$ on $X \setminus \{X_1\}$.

By the induction hypothesis, there is a teaching plan of order $m - 3$ for $\overline{C_1}$. Take such a plan and extend every recursive teaching set S from this plan to $S \cup \{(X_1, 0)\}$. As a result, we obtain a teaching plan for $\{0\} \times \overline{C_1}$ of order $m - 2$, which we call P_1. Note that C_2 is a VCD-maximum class with $\mathrm{VCD}(C_2) = |X \setminus \{X_1\}| - 1 = m - 2$, and hence $\mathrm{RTD}(C_2) = m - 2$. Since $\mathrm{RTD}(\{1\} \times C_2) = \mathrm{RTD}(C_2)$, there is a teaching plan of order $m - 2$ for $\{1\} \times C_2$, which we call P_2.

Every recursive teaching set from P_1 contains $(X_1, 0)$, which distinguishes the concepts in $\{0\} \times \overline{C_1}$ from those in $\{1\} \times C_2$. So, P_1 and P_2 can be merged into a teaching plan for $\overline{C}$ of order $m-2$. Thus $\mathrm{RTD}(\overline{C}) \leq m-2$. Furthermore, $|\overline{C}| = 2^m - |C| = 2^m - (m+1) = \Phi_{m-2}(m)$. By Corollary 6.5, we have $\mathrm{RTD}(\overline{C}) \geq m-2$, and hence $\overline{C}$ is RTD-maximum. □

In Corollary 6.9, we prove that when VCD_Ψ fulfills the reduction property, any VCD_Ψ-maximum class is RTD-maximum. The following proposition verifies that the converse of that statement does not hold in general.

Proposition A.3. *There is a binary class C for which both C and $\overline{C}$ are* RTD*-maximum, but neither C nor $\overline{C}$ is* VCD*-maximum. In particular, there are* RTD*-maximum classes that are not* VCD*-maximum.*

Proof. C_1 in Table A.2 is RTD-maximum with $\mathrm{RTD}(C_1) = 2$, and $\overline{C_1}$ is RTD-maximum with $\mathrm{RTD}(\overline{C_1}) = 1$. As $\mathrm{VCD}(C_1) = 3$ and $\mathrm{VCD}(\overline{C_1}) = 2$, neither C_1 nor $\overline{C_1}$ is VCD-maximum. □

$c_i \in C_1$	X_1	X_2	X_3	X_4
c_1	$\underline{1}$	1	1	$\underline{1}$
c_2	0	0	$\underline{1}$	$\underline{1}$
c_3	0	$\underline{1}$	0	$\underline{1}$
c_4	0	$\underline{1}$	$\underline{1}$	0
c_5	$\underline{1}$	0	$\underline{1}$	0
c_6	$\underline{1}$	$\underline{1}$	0	0
c_7	0	0	0	$\underline{1}$
c_8	0	0	$\underline{1}$	0
c_9	0	$\underline{1}$	0	0
c_{10}	$\underline{1}$	0	0	0
c_{11}	0	0	0	0

$c_i \in \overline{C_1}$	X_1	X_2	X_3	X_4
c_1	$\underline{0}$	1	1	1
c_2	1	1	1	$\underline{0}$
c_3	1	0	$\underline{1}$	1
c_4	1	$\underline{1}$	0	1
c_5	1	0	0	1

Table A.2: C_1 and $\overline{C_1}$ are RTD-maximum but neither C_1 ($\{X_1, X_2, X_3\}$ is shattered) nor $\overline{C_1}$ ($\{X_2, X_3\}$ is shattered) is VCD-maximum. Recursive teaching sets are underlined.

As discussed in Section 6.1, if C is a binary class with $\mathrm{RTD}(C) + \mathrm{RTD}(\overline{C}) = |X| - 1$, then C is RTD-maximum. The converse of this statement is false in general: Table

A.1 contains an RTD-maximum class C on 5 instances for which $\mathrm{RTD}(C)+\mathrm{RTD}(\overline{C}) = 3+2 > 5-1$.

As mentioned in Section 6.2, binary VCD-maximal classes of VC-dimension 1 are VCD-maximum [WW87]. As shown in Proposition 6.26, the same holds for RTD-maximal classes, even in the multi-label case. Here, we provide an alternative proof for Proposition 6.26 in the binary case.

Proposition A.4. *Let $m \in \mathbb{N}^+$ and $X_i = \{0,1\}$, for all $i \in \{1,\ldots,m\}$. Let $X = \{X_1,\ldots,X_m\}$ and $C \subseteq \prod_{i=1}^{m} X_i$ be* RTD*-maximal. If $\mathrm{RTD}(C) = 1$, then C is* RTD*-maximum.*

Proof. For $|X| = 1$ there is only one RTD-maximal class with two concepts which is clearly RTD-maximum. Suppose that the proposition holds when $|X| = m$. Now we consider the case that $|X| = m+1$ and C is an RTD-maximal class on X with $\mathrm{RTD}(C) = 1$. Since $\mathrm{RTD}(C) = 1$, there is a concept $c \in C$ such that $\mathrm{TD}(c, C) = 1$. W.l.o.g, let $\{(X_1, 1)\}$ be a teaching set for c. Then, for any $c' \in C\backslash\{c\}$, $(X_1, 1) \notin c'$ or equivalently, $(X_1, 0) \in c'$, which implies that $|C\backslash\{c\}| = |(C\backslash\{c\}) - X_1|$. Clearly, $(C\backslash\{c\}) - X_1$ is RTD-maximal, otherwise C would not be RTD-maximal. So, by the induction hypothesis, $|(C\backslash\{c\}) - X_1| = \Phi_1(m)$. Therefore, $|C| = \Phi_1(m)+1 = \Phi_1(m+1)$ and C is RTD-maximum. □