Sample Compression for Multi-label Concept Classes

Rahim Samei samei20r@cs.uregina.ca
Department of Computer Science
University of Regina
Regina, SK S4S 0A2, Canada

Pavel Semukhin pavel.semukhin@liverpool.ac.uk
Department of Computer Science
University of Liverpool
Liverpool, L69 3BX, UK

Boting Yang boting@cs.uregina.ca
Department of Computer Science
University of Regina
Regina, SK S4S 0A2, Canada

Sandra Zilles zilles@cs.uregina.ca
Department of Computer Science
University of Regina
Regina, SK S4S 0A2, Canada

Editor: Editor Name

Abstract

This paper studies sample compression of multi-label concept classes for various notions of VC-dimension. It formulates a sufficient condition for a notion of VC-dimension to yield labeled compression schemes for maximum classes of dimension d in which the compression sets have size at most d. The same condition also yields a so-called tight sample compression scheme, which we define to generalize the unlabeled binary scheme by Kuzmin and Warmuth (2007) to the multi-label case. The well-known Graph-dimension satisfies our sufficient condition, while neither Pollard's pseudo-dimension nor the Natarajan dimension does. As was previously done for the binary case, we connect our tight compression schemes to a recently introduced teaching notion by Zilles et al. (2011), namely the recursive teaching dimension, and to the one-inclusion hypergraph, a natural extension of the one-inclusion graph to the multi-label case. We further show that every multi-label class of Graph-dimension 1 has a sample compression scheme using only sets of size at most 1. As opposed to the binary case, the latter result is not immediately implied by the compression results on maximum classes, since there are multi-label concept classes of dimension 1 that are not contained in maximum classes of dimension 1.

Keywords: multi-label concept class, sample compression, VC-dimension

1. Introduction

In the context of the theory of concept learning, a long-standing open problem is the *sample compression conjecture* (Littlestone and Warmuth, 1986; Floyd and Warmuth, 1995). A concept is usually modeled as a subset $c \subseteq X$ of an instance space X and a concept class

C is a set of such concepts. When learning a concept $c \in C$, information about c can be provided in the form of a set of labeled examples, i.e., a set S of pairs (x, l), where $x \in X$ is an instance and $l \in \{0, 1\}$ is a label indicating whether or not x belongs to c. A *sample compression scheme (SCS)* for C is a pair (f, g) of mappings with the following constraints. The compression mapping f compresses each set S of labeled examples for a concept $c \in C$ to a subset $f(S) \subseteq S$ of size at most d. When uncompressing $f(S)$ to some concept $g(f(S))$ over X, the label information contained in S must be correctly reconstructed by the decompression mapping g. The size of the SCS is the cardinality of its largest compressed set $f(S)$, where S is any set of examples for some concept in C. The sample compression conjecture states that every concept class has an SCS whose size is at most the VC-dimension of the class. Just like the VC-dimension, the size of an SCS is a combinatorial parameter that yields sample bounds for a PAC-learner for C (Littlestone and Warmuth, 1986)—thus proving the conjecture would establish a connection between the known sample bounds.

Floyd and Warmuth (1995) proved the conjecture for maximum C, that is, any C meeting Sauer's upper bound on the size of classes with a given VCD (Sauer, 1972). This result was recently extended to so-called *extremal classes* (Moran and Warmuth, 2015). 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 VC-dimension 1 has an SCS of size 1. This is due to the fact that every concept class of VC-dimension 1 is contained in a maximum class of VC-dimension 1 over the same instance space (Welzl and Woeginger, 1987). An astonishing observation was made by Kuzmin and Warmuth (2007), who proved that each maximum class of VCD d even has an *unlabeled* SCS of size d, i.e., an SCS in which the compression sets have no label information. Concerning infinite concept classes in general, it was recently shown that classes of finite VC-dimension have sample compression schemes of size exponential in the VC-dimension (Moran and Yehudayoff, 2015), but no size bound that is linear in the VC-dimension has been established yet.

To the best of our knowledge, the notion of SCS has been studied exclusively for the notion of concept class introduced above. Such concept classes are called binary concept classes, since they correspond to subsets of the power set of $\{0, 1\}^{|X|}$. This paper extends the study of sample compression to multi-label concept classes, i.e., subsets of the product $\{0, \ldots, N_1\} \times \cdots \times \{0, \ldots, N_m\}$, where the set of possible labels for an instance $X_i \in X = \{X_1, \ldots, X_m\}$ is $\{0, \ldots, N_i\}$. Since a vast number of applications in Machine Learning deal with multi-class classification, the study of multi-label concept classes on a formal level certainly deserves the attention of the learning theory community. As we will explain in Section 5, Littlestone and Warmuth's proof (1986) that (in the binary case) an SCS of size d yields a successful PAC-learner with bounds expressed in terms of d can be immediately transferred to the multi-label case. Hence, it is natural to extend also the study of SCS to the multi-label case, which is the focus of this paper.

Most prior work on multi-label classes concerns the combinatorial structure of such classes, and in particular various options for defining analogues of the VC-dimension (Alon, 1983; Natarajan, 1989; Vapnik, 1989; Pollard, 1990; Gurvits, 1997) that coincide with the VC-dimension in the binary case. Haussler and Long (1995) generalize Sauer's bound to multi-label classes for a variety of such analogues of VC-dimension. As in the binary case, classes meeting this generalized Sauer bound are called "maximum" with respect to the underlying notion of VC-dimension. It turns out that, as in the binary case, the finiteness of most of the dimensions studied is sufficient and necessary for the PAC-learnability of multi-label classes (Ben-David et al., 1995). More recent studies show that results relating the VC-dimension to the density of the so-called one-inclusion graph of a concept class can also be extended to some of the multi-label analogues (Rubinstein et al., 2009; Simon and Szörényi, 2010) and provide sample bounds for various learning models and strategies (Rubinstein et al., 2009; Daniely et al., 2011). In our work, we generalize over notions of VC-dimension, but we illustrate our findings in particular with the Graph-dimension (Natarajan, 1989), Pollard's pseudo-dimension (Pollard, 1990), and the Natarajan dimension (Natarajan, 1989).

The main contributions of this paper are the following:

1. We identify a crucial property of notions of VC-dimension in the multi-label case, which we henceforth call the *reduction property*. 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? A VC-dimension notion 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. The reduction property is discussed in Section 4, and later in Section 6 we prove that while the Graph-dimension has the reduction property, neither Pollard's pseudo-dimension nor the Natarajan dimension fulfill it.

2. We generalize both Floyd and Warmuth's compression scheme (1995) and Kuzmin and Warmuth's unlabeled SCSs (2007) to the multi-label case. In particular, we show that Kuzmin and Warmuth's result (2007) on unlabeled compression for maximum classes finds a natural extension to the multi-label case. This is not trivial, since unlabeled SCSs of size VCD cannot exist for maximum multi-label C for any known notion of VCD—simply because the size of C is larger than the number of unlabeled sets of size VCD. To generalize Kuzmin and Warmuth's unlabeled SCSs for maximum classes, we observe that they fulfill a property we call *tightness*. As opposed to the Floyd-Warmuth scheme and its extension to the multi-label case, a tight SCS uses exactly as many compression sets as there are concepts in C (trivially, it is impossible to use fewer sets, since each concept needs a different compression set—hence the term "tight"). Our main result is the following: for every notion VCD_Ψ in a broad and natural category of VC-dimension notions, the reduction property is sufficient for proving that each maximum multi-label class of VCD_Ψ d has a (tight) SCS of size d.

3. We connect tight compression schemes to *recursive teaching*, a recently introduced teaching model (Zilles et al., 2011). In the original teaching model, which was introduced by Goldman and Kearns (1991) and Shinohara and Miyano (1991), the learner is provided with a set of well-chosen examples (*teaching set*) by a teacher and is re-

quired to identify the target concept exactly—in contrast to the PAC-learning model in which the learner is given randomly chosen examples in order to approximate the target concept. The recursive teaching model (Zilles et al., 2011) is a variation in which teacher and learner use fewer examples than in the original teaching model. The so-called recursive teaching sets are computed as follows. One first identifies a concept c in the class C whose teaching sets with respect to C have the smallest size. The corresponding smallest teaching set is the so-called recursive teaching set for this concept. One then removes the concept from the class C and proceeds recursively with the class $C' = C \setminus \{c\}$ of the remaining concepts. Note that the recursive teaching sets depend on the particular choice of c among all concepts that possess teaching sets of the same size as c does. The sequence in which concepts are removed from C is called a canonical teaching plan. We show that for any VCD_Ψ-maximum class C where VCD_Ψ fulfills the reduction property, there is a canonical teaching plan for C in which the recursive teaching sets coincide with the compression sets resulting from a tight compression scheme. This extends the corresponding result for the binary case, which was proven by Doliwa et al. (2014). In particular, this strengthens recent results that indicate that teaching and sample compression (and thus also complexity parameters in teaching and complexity parameters in PAC-learning) may be closely related (Doliwa et al., 2014; Darnstädt et al., 2013).

4. We also establish a connection between tight compression schemes and the one-inclusion hypergraph, which is a natural extension of the one-inclusion graph to the multi-label case.[1] Furthermore, we show that the one-inclusion hypergraph of maximum classes of a given VCD_Ψ is shortest-path closed, which is the extension of Kuzmin and Warmuth's (2007) result for one-inclusion graphs in the binary case.

5. We show that every class of Graph-dimension 1 has an SCS of size 1. The reasoning used in the binary case does not apply here; in particular, we provide a class of Graph-dimension 1 that is not contained in a maximum class of Graph-dimension 1 over the same instance space. Any such class cannot trivially inherit an SCS of size 1 from a maximum class of dimension 1, as it would in the binary case. Thus we give an independent constructive proof that provides an SCS of size 1 for each class whose Graph-dimension equals 1.

This paper is an extension of two conference papers (Samei et al., 2014b,c).

2. Preliminaries

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 $i \in [m]$. 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

[1] In the one-inclusion graph of a binary concept class C, the concepts of C are the vertices; two vertices have an edge between them if they differ in exactly one instance in X. The density of the one-inclusion graph of a class is known to provide a lower bound on its VC-dimension (Haussler et al., 1994).

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.

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, $C|_Y= \{c|_Y|\ c \in C\}$ denotes the restriction of C to Y. We use $\text{size}(C|_Y)$ 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.

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., $C^{X_t} = \{c \in C - X_t \mid c \cup \{(X_t, 0)\},\ c \cup \{(X_t, 1)\} \in C\}$. 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}.

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 $\text{VCD}(C)$. The literature offers a variety of VCD notions for the non-binary case (Alon, 1983; Natarajan, 1989; Vapnik, 1989; Pollard, 1990; Gurvits, 1997). Gurvits' framework (Gurvits, 1997) generalizes over many of these notions. We first need to introduce the notion of a *label mapping* and the required 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 $\{0, 1\}^m$.

Definition 1 (Gurvits, 1997) *Let Ψ_i, $1 \leq i \leq m$, 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* $\text{VCD}_\Psi(C)$ *and define it by* $\text{VCD}_\Psi(C) = \max_{\overline{\psi} \in \Psi} \text{VCD}(\overline{\psi}(C))$.

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 $\psi_i : X_i \to \{0, 1\}$.

The term *Graph-dimension* (Natarajan, 1989) refers to VCD_{Ψ_G}, where $\Psi_G = \Psi_{G_1} \times \cdots \times \Psi_{G_m}$ and for all $i \in [m]$, $\Psi_{G_i} = \{\psi_{G,k} : k \in N_i\}$ and $\psi_{G,k}(x) = 1$ if $x = k$, $\psi_{G,k}(x) = 0$ if $x \neq k$.

Example 1 *The class C on the left of Table 1 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*

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

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'. (No binary class resulting from C can shatter X, since C has only 5 concepts.) 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.

$c \in C$	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 (d)	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 (d)	0	0	0
c''_5 (d)	1	1	0

Table 1: 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 (d).

By *Pollard's pseudo-dimension* (Pollard, 1990) we refer to $\mathrm{VCD}_{\Psi_\mathrm{P}}$, where $\Psi_P = \Psi_{P_1} \times \cdots \times \Psi_{P_m}$ and for all $i \in [m]$, $\Psi_{P_i} = \{\psi_{P,k} \mid k \in X_i\}$ and $\psi_{P,k}(x) = 1$ if $x \geq k$, $\psi_{P,k}(x) = 0$ if $x < k$. The term *Natarajan-dimension* (Natarajan, 1989) refers to $\mathrm{VCD}_{\Psi_\mathrm{N}}$, where $\Psi_N = \Psi_{N_1} \times \cdots \times \Psi_{N_m}$ and for all $i \in [m]$, $\Psi_{N_i} = \{\psi_{N,k,k'} \mid k, k' \in X_i, k \neq k'\}$ and $\psi_{N,k,k'}(x) = 1$ if $x = k$, $\psi_{N,k,k'}(x) = 0$ if $x = k'$, $\psi_{N,k,k'}(x) = *$, otherwise. (Here technically, ψ_i maps to $\{0, 1, *\}$, where $*$ is a null element to be ignored when computing the VC-dimension.)

Clearly, VCD_{Ψ^*} upper-bounds all VCD notions. Also, $\mathrm{VCD}_{\Psi_\mathrm{P}} \geq \mathrm{VCD}_{\Psi_\mathrm{N}}$ and $\mathrm{VCD}_{\Psi_\mathrm{G}} \geq \mathrm{VCD}_{\Psi_\mathrm{N}}$ (Haussler and Long, 1995). However, $\mathrm{VCD}_{\Psi_\mathrm{P}}$ and $\mathrm{VCD}_{\Psi_\mathrm{G}}$ are incomparable (Ben-David et al., 1995).

As in the binary case (Floyd and Warmuth, 1995), 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 $\mathrm{Forb}(C, Y) = \mathrm{Forb}(C) = \emptyset$.

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 2 (Alon et al., 1987) *The* one-inclusion graph $G(C)$ *of a concept class C is the labeled graph G with $V(G) = C$ and $E(G) = \{\{c, c'\} \mid |c \triangle c'| = 1\}$. Every edge $\{c, c'\} \in E(G)$ is labeled by the instance from $c \triangle c'$.*

3. Generalized Sauer Bound

Let Ψ_i be a family of mappings $\psi_i : X_i \to \{0, 1\}$. The statement "$\Psi_i$ spans $\mathbb{R}^{N_i+1}$" or "Ψ_i is spanning on X_i" means that any real-valued function on X_i can be expressed as a linear combination of mappings from Ψ_i. 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, $\Psi_i = \{\psi_1, \ldots, \psi_m\}$ is spanning on X_i iff any vector in $\mathbb{R}^{N_i+1}$ (real-valued function on X_i) can be expressed as a linear combination of the vectors $\boldsymbol{\psi}_j = (\psi_j(0), \ldots, \psi_j(N_i))$ for $j \in [m]$.

Remark 3 *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*

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

We will make use of some results by Gurvits (1997).

Definition 4 *Let $C = \{c_1, \ldots, c_n\}$, $|C| = n$, and let $p(X_1, \ldots, X_m) \in \mathbb{R}[X_1, \ldots, X_m]$ be a polynomial. We identify p with a vector $\mathbf{p} = (p_1, \ldots, p_n) \in \mathbb{R}^{|C|}$ via $p_i = p(c_i(X_1), \ldots, c_i(X_m))$. The phrase "$p(X_1, \ldots, X_m) = 0$ on C" means that $\mathbf{p}$ corresponds to the zero vector in $\mathbb{R}^{|C|}$.*

If $\mathcal{P}$ is a collection of polynomials from $\mathbb{R}[X_1, \ldots, X_m]$, then we say that $\mathcal{P}$ spans *$\mathbb{R}^{|C|}$ if the set of vectors that correspond to polynomials from $\mathcal{P}$ spans $\mathbb{R}^{|C|}$.*

To make the proofs in the paper easier to follow, we make use of the following notation. 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

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

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,

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

Gurvits (1997) and Smolensky (1997) took advantage of the linear algebraic method for the first time in proving Sauer's lemma.

Theorem 5 (Gurvits, 1997; Smolensky, 1997) *Let $X_i = \{0, 1\}$ for all $i \in [m]$. If $\mathrm{VCD}(C) = d$ then the set of monomials $\{X_{i_1} \cdots X_{i_k} \mid 1 \leq i_1 < \cdots < i_k \leq m,\ k \leq d\}$ spans $\mathbb{R}^{|C|}$.*

The immediate consequence of Theorem 5 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, $|C| \leq \Phi_d(m)$.

This approach was also exploited by Gurvits (1997) in generalizing Sauer's bound to the multi-label case. To prove the generalized Sauer bound (Theorem 7), Gurvits first observed the following.

Lemma 6 (Gurvits, 1997) *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 7 (Gurvits, 1997) *Let Ψ_i, $1 \le i \le m$, 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 6, 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 1, $\mathrm{VCD}(C') \le \mathrm{VCD}_\Psi(C) = d$. By Theorem 5, the monomial $X'_1 \cdot \ldots \cdot X'_m$ can be expressed as a linear combination of short products

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

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

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

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

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 8 (Generalized Sauer bound) *Let Ψ_i, $1 \le i \le m$, 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 $|C| \le \Phi_d(N_1, \ldots, N_m)$.*

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 9 ($\mathbf{VCD}_\Psi$**-maximum class)** *Let Ψ_i, $1 \le i \le m$, 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 $\mathrm{VCD}_\Psi(C) = d$ 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.

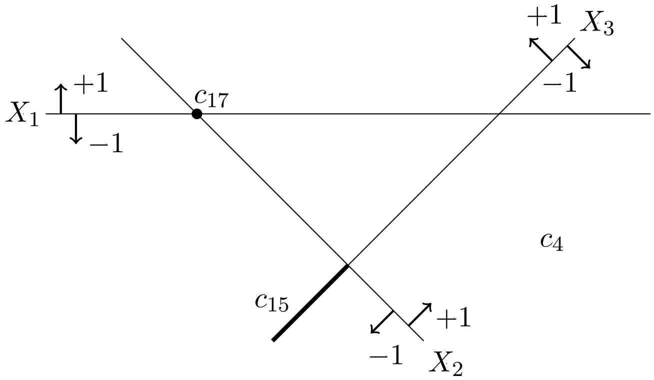

Figure 1: The geometric class described in Example 2 for $m = 3$.

$c \in C$	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 2: The $\mathrm{VCD}_{\Psi_{\mathrm{G}}}$-maximum class obtained from Figure 1.

Example 2 *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 1 illustrates such a class for $m = 3$, and Table 2 shows the corresponding concepts.*

The spanning property allows us to establish some interesting statements about VCD_Ψ-maximum classes. 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 10 *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, $\mathrm{VCD}_\Psi(C') = \max_{\overline{\psi} \in \Psi} \mathrm{VCD}(\overline{\psi}(C')) \leq \max_{\overline{\psi} \in \Psi} \mathrm{VCD}(\overline{\psi}(C)) = d$. By Theorem 7, 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 $\mathbb{R}^{|C'|}$. If C' is not VCD_Ψ-maximum of dimension d, then the monomials in $P^d(1, N_2 \ldots, N_m)$ are 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 $P^d(N_1, \ldots, N_m)$ are 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 10 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 11 *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 10 to each φ_t repeatedly from $t = 1$ to $t = k$ proves the claim. ■

It is obvious that if one of the φ_i's in $\overline{\varphi} = (\varphi_1, \ldots, \varphi_m)$ 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 corollary.

Corollary 12 *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 $\overline{\varphi} = (\varphi_1, \ldots, \varphi_m)$ be a tuple of mappings where $\varphi_i \in \Psi_i$, for all $i \in [m]$. Then $\overline{\varphi}(C)$ is VCD_Ψ-maximum of dimension $\mathrm{VCD}_\Psi(C)$ iff φ_i is non-constant for all $i \in [m]$.*

Recall that Ψ^* is based on the family of all label-mappings including constant mappings on X_i, for all $i \in [m]$.

Corollary 13 *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* $i \in [m]$*.*

We now generalize Corollary 13 as follows.

Corollary 14 *Let C be* VCD_{Ψ^*}*-maximum and* $\Psi = \Psi_1 \times \cdots \times \Psi_m$*, where each* Ψ_i*, for* $i \in [m]$*, is a spanning family of mappings on* X_i*. Then C is* VCD_{Ψ}*-maximum of dimension* $\mathrm{VCD}_{\Psi^*}(C)$*.*

In the binary case, restrictions and reductions of maximum classes are again maximum (Welzl, 1987). For the multi-label case, the corresponding result is known for restrictions.

Theorem 15 (Gurvits, 1997) *Let* Ψ_i*,* $1 \le i \le m$*, 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* $|Y| \ge d$*. Then* $C|_Y$ *is* VCD_{Ψ}*-maximum with* $\mathrm{VCD}_{\Psi}(C|_Y) = d$*.*

4. The Reduction Property

We next define a core notion of our work, namely, the reduction property. It provides a sufficient condition for maximum classes of VCD_{Ψ} d to have sample compression schemes of size d, provided that Ψ is based on spanning families. First, we define the notion of sample compression for binary concept classes.

Definition 16 (sample compression scheme) (Littlestone and Warmuth, 1986) *A sample compression scheme of size k, $k \in \mathbb{N}$, for a binary concept class C is a pair (f, g) of mappings with the following properties: (i) the compression function f compresses any C-realizable sample S to a set $S' \subseteq S$ of size at most k, that is, $f(S) = S' \subseteq S$; (ii) for any C-realizable sample S, the decompression function g decompresses $f(S)$ to a sample $g(f(S)) \supseteq S$ of size m, where for each $i \in \{1, \ldots, m\}$ there is exactly one $l_i \in \{0, 1\}$ such that $(X_i, l_i) \in g(f(S))$.*

A long-standing open question is whether every concept class has a sample compression scheme of the size of its VC-dimension (Floyd and Warmuth, 1995). To the best of our knowledge, this question has so far been addressed only in the binary case. This paper extends the previous studies to the multi-label case.

As shown by Floyd and Warmuth (1995), 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 $\mathrm{VCD}(C)$ (Kuzmin and Warmuth, 2007). 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 (Welzl, 1987).

Theorem 15 shows that restrictions of VCD_{Ψ}-maximum classes are still maximum in the multi-label case. However, the definition of reduction in the multi-label case is not as straightforward as in the binary case. In particular, for any instance $X_t \in X$ and a concept class C, we have two definitions of reduction: $[C]^{X_t}_{\ge 2}$ and C^{X_t}, respectively. We thus define the core notion of our work, namely, the reduction property.

Definition 17 (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., $[C]^{X_t}_{\geq 2} = C^{X_t}$).*

In Section 5 (Theorems 22 and 30), we will show that when Ψ is based on spanning families of mappings, the reduction property is a sufficient condition for VCD_Ψ-maximum classes of VCD_Ψ d to have a sample compression scheme of size d.

When VCD_Ψ fulfills the reduction property, by a *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 18 *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 See the appendix. ∎

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 by Welzl (1987) in the binary case, and we include it in the appendix for the sake of completeness.

Proposition 19 *For any X_i, X_j with $i \neq j$, $(C^{X_i})^{X_j} = (C^{X_j})^{X_i}$.*

Proof See the appendix. ∎

5. Sample Compression Schemes for Maximum Classes

This section discusses sample compression schemes for multi-label concept classes.

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

Definition 20 *A sample compression scheme for C is a pair (f, g) of mappings with the following properties. 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.*

Littlestone and Warmuth (1986) 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.

Theorem 21 (Littlestone and Warmuth, 1986) *Let C be a 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*

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

It turns out that Theorem 21 is correct for multi-label concept classes using the same proof as the one by Littlestone and Warmuth (1986) (details are omitted). Therefore, the existence of a sample compression scheme of size VCD_Ψ of a multi-label 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 21. This motivates the extension of the study of sample compression schemes to the multi-label case.

Throughout this section, we assume that $\Psi = \Psi_1 \times \cdots \times \Psi_m$ and $C \subseteq \prod_{1 \leq i \leq m} X_i$ is a VCD_Ψ-maximum class of dimension d, where each Ψ_i is a spanning family of mappings on X_i, for all $i \in [m]$, and VCD_Ψ fulfills the reduction property.

5.1 Generalizing Floyd and Warmuth's Compression Scheme

We first show that every VCD_Ψ-maximum class of dimension d has a sample compression scheme of size d. Our objective here is to prove the following theorem:

Theorem 22 *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 $\mathrm{VCD}_\Psi(C)$.*

We will later prove a much stronger result, but our first proof of Theorem 22 is interesting in that it demonstrates how Floyd and Warmuth's technique can be extended to the multi-label case. The scheme we present and also parts of the proof closely follow their so-called VC-Compression Scheme for binary maximum classes. However, there are some technical difficulties that need to be overcome in order to adapt Floyd and Warmuth's technique.

Proposition 23 *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 (Analogous to the proof of Corollary 2 in (Floyd and Warmuth, 1995)). Let $Y = \{X_{i_1}, \ldots, X_{i_d}\}$. By applying Theorem 18 to $C^Y = ((C^{X_{i_1}}) \cdots)^{X_{i_d}}$ repeatedly, C^Y is a VCD_Ψ-maximum class of dimension 0. So, $|C^Y| = 1$. ■

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.

Definition 24 *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.*

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

Lemma 25 *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 $X(S) = \{X_{i_1}, \dots, X_{i_d}\}$. Let $X_t \in X \setminus X(S)$ and $c_{S,C}(X_t) = p$, for some $p \in X_t$. Let $\overline{\psi} = (\psi_{i_1}, \dots, \psi_{i_d})$ be a tuple of non-constant mappings $\psi_{i_j} \in \Psi_{i_j}$, for all $j \in \{1, \dots, d\}$, $\psi_t : X_t \to \{0, 1\}$ with $\psi_t(p) = l$, and $\overline{\psi'} = (\psi_{i_1}, \dots, \psi_{i_d}, \psi_t)$. Then $\{\{0, 1\}^d \times \{l\}\} \subseteq \overline{\psi'}(C|_{\{X_{i_1},\dots,X_{i_d},X_t\}})$.*

Proof W.l.o.g., assume that $S = \{(X_1, l_1), \dots, (X_d, l_d)\}$ and $t = d + 1$. From Theorem 15, $C|_{\{X_1,\dots,X_d\}}$ is VCD_Ψ-maximum of dimension d and by Corollary 12, $\overline{\psi}(C|_{\{X_1,\dots,X_d\}}) = \{0, 1\}^d$.

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

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 15, $C|_Y$ is VCD_Ψ-maximum and Definition 24 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-mapping here, the proof ideas are similar to those in (Floyd and Warmuth, 1995). 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.

Lemma 26 *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 See the appendix. ■

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 27 *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 $c_{S_i,C} - X_t = c_{S,C^{X_t}}$.*

Proof See the appendix. ■

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.

Theorem 28 *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 See the appendix. ■

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 22. 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 $\mathrm{VCD}_\Psi(C) = d$, Theorem 28 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. ■

Table 3 illustrates a $\mathrm{VCD}_{\Psi_\mathrm{G}}$-maximum class of $\mathrm{VCD}_{\Psi_\mathrm{G}}$ 2 and the compression sets obtained from our compression scheme.

An inspection of the proof will show that Theorem 22 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. By Lemma 26 , $c_{X(S),C}$ assigns a unique label to each $X_i \in X \setminus X(S)$. That is, the concept $c_{S',C}$ on X is consistent with the original sample set $c_{S',C|_{X(S)}}$. So, Theorem 22 works for infinite instance spaces as well.

5.2 Generalizing Kuzmin and Warmuth's Unlabeled Scheme

The reduction property is also useful for extending the Kuzmin-Warmuth unlabeled compression scheme (Kuzmin and Warmuth, 2007), as we will see next. To this end, we first generalize the definition of an unlabeled scheme to a "tight" labeled compression scheme for the multi-label case.y Theorem 7

Obviously, for all notions of VCD_Ψ studied in the literature, unlabeled compression schemes of size d for a VCD_Ψ-maximum class C of VCD_Ψ d cannot exist, as the number

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

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 (2007) to VCD_Ψ-maximum classes, where VCD_Ψ fulfills the reduction property and Ψ is based on spanning families of mappings, by first observing its *tightness*.

Definition 29 (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'.

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

Both conditions are necessary for the tightness of the compression scheme. For illustration, consider the class C and the representatives shown in Table 4, 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 4, and notice the compression sets S with $X(S) = \{X_1, X_3\}$. As one can verify, any such compression set does not contain

```
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.

$(X_1, 0)$ 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 exactly coincides 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.

Our goal here is to justify the following theorem.

Theorem 30 *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 that by Kuzmin and Warmuth (2007) for the binary case. However, various technical barriers have to be overcome for the multi-label case.

In (Kuzmin and Warmuth, 2007) 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 (Kuzmin and Warmuth, 2007). Kuzmin and Warmuth (2007) 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, $r(c') \nsubseteq X(S)$.

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,

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

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 31 *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 4 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 (Kuzmin and Warmuth, 2007), the following lemma shows how the non-clashing property is useful for finding unique labeled representatives for samples in the multi-label case.

Lemma 32 *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 See the appendix. ■

Lemma 32 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 showed that a representation mapping can be used as a compression-reconstruction 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 33 *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 See the appendix. ■

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 $c - X_t \in (C - X_t) \setminus C^{X_t}$ (Kuzmin and Warmuth, 2007). 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 $C = (C^{X_t} \times X_t) \cup \text{tail}_{X_t}(C)$.

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 15 and Theorem 18, for $d < m$, $C - X_t$ and C^{X_t} are VCD$_\Psi$-maximum of dimensions d and $d - 1$, respectively. So,

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

For $Y = \{X_{i_1}, \ldots, X_{i_{d+1}}\}$, $C|_Y$ is VCD$_\Psi$-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$_\Psi$-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,

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

First, adding any concept in $\text{tail}_{X_t}(C) - X_t$ to C^{X_t} increases the VCD$_\Psi$ 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$_\Psi$-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 $(\mathrm{tail}_{X_t}(C) - X_t)|_Y$. Since Y was chosen arbitrarily, we conclude that all forbidden labelings of C^{X_t} appear in $\mathrm{tail}_{X_t}(C)$.

Kuzmin and Warmuth (2007) find representatives for C by partitioning C into $C^{X_i} \times X_i$ and $\mathrm{tail}_{X_i}(C)$ for some $X_i \in X$. Their scheme 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 $\mathrm{tail}_{X_i}(C)$ by assigning each of them a forbidden labeling of the class C^{X_i} of size d. Since the representative for each concept in $\mathrm{tail}_{X_i}(C)$ is a forbidden labeling of the class C^{X_i}, the non-clashing property between $\mathrm{tail}_{X_i}(C)$ and C^{X_i} is guaranteed.

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 has similar reasoning to the one in the binary case (Kuzmin and Warmuth, 2007).

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

1. *For each $c \in \mathrm{tail}_{X_s}(C^{X_t})$ there are at least N_t labels $l_1, \ldots, l_{N_t} \in X_t$ such that $c \times \{l_1, \ldots, l_{N_t}\} \subseteq \mathrm{tail}_{X_s}(C)$. If $c \in \mathrm{tail}_{X_s}(C^{X_t}) \setminus \mathrm{tail}_{X_s}(C - X_t)$, then there are exactly N_t such labels.*
2. *For each $c \in \mathrm{tail}_{X_s}(C - X_t)$ there is at least one label $l \in X_t$ such that $c \times \{l\} \in \mathrm{tail}_{X_s}(C)$. If $c \in \mathrm{tail}_{X_s}(C - X_t) \cap \mathrm{tail}_{X_s}(C^{X_t})$, then $c \times X_t \subseteq \mathrm{tail}_{X_s}(C)$.*
3. *Each concept in $\mathrm{tail}_{X_s}(C)$ is an extension of either a concept in $\mathrm{tail}_{X_s}(C^{X_t})$ or a concept in $\mathrm{tail}_{X_s}(C - X_t)$.*

Proof See the appendix. ■

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

Lemma 35 *For any $s, t \in [m]$, with $s \neq t$, $C^{X_s} - X_t = (C - X_t)^{X_s}$.*

Proof See the appendix. ■

Lemma 35 has the following corollary. The proof is the same as that of Corollary 8 in (Kuzmin and Warmuth, 2007).

Corollary 36 $\mathrm{Forb}((C - X_t)^{X_s}) \subseteq \mathrm{Forb}(C^{X_s})$.

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 (Kuzmin and Warmuth, 2007), our proof here is substantially more extensive, because the statement is more complicated to validate in the multi-label case.

```
Labeled Tail Matching Function (LTMF)
Input: a VCD$_\Psi$-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 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.

Lemma 37 *Any forbidden labeling for $(C^{X_s})^{X_t}$ can be extended to N_t forbidden labelings for C^{X_s}.*

Proof See the appendix. ■

The next lemma is now obvious.

Lemma 38 *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 36, Lemma 37 and (8). ■

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 $\mathrm{VCD}_\Psi(C) = |X|-1$ and then proceed to the general case.

Lemma 39 *For any set $Y = \{X_{i_1}, \dots, X_{i_d}\} \subseteq X \setminus \{X_s\}$ with $|Y| = d = |X|-1$, there is a tuple $(l_1, \dots, l_d) \in \prod_{1\leq j\leq d} X_{i_j}$ such that $\mathrm{Forb}(C^{X_s}, Y) = \prod_{1\leq j\leq d}(X_{i_j} \setminus \{l_j\})$.*

Proof Let $m = |X|$ and C be a VCD$_\Psi$-maximum class on m instances with $\mathrm{VCD}_\Psi(C) = m-1$. 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 38, 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$_\Psi$-maximum on $m-2$ instances and of VCD$_\Psi$ $m-2$. 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 $(C^{X_s})^{X_t}$.

$\mathrm{Forb}((C^{X_s})^{X_t}) = \mathrm{Forb}((C^{X_t})^{X_s})$, since $(C^{X_s})^{X_t} = (C^{X_t})^{X_s}$. C^{X_t} is VCD$_\Psi$-maximum on $m-1$ instances and of VCD$_\Psi$ $m-2$. 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 $(l_1, \ldots, l_{m-2}) \in \prod_{1 \leq j \leq m-2} X_{i_j}$ such that $\mathrm{Forb}((C^{X_t})^{X_s}, Y) = \prod_{1 \leq j \leq m-2}(X_{i_j} \setminus \{l_j\})$, and hence

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

By Lemma 37, 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 $(l_1, \ldots, l_{m-2}, l_t) \in (\prod_{1 \leq j \leq m-2} X_{i_j}) \times X_t$ such that

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

■

Lemma 40 *For any set* $Y = \{X_{i_1}, \ldots, X_{i_d}\} \subseteq X \setminus \{X_s\}$ *with* $|Y| = d < |X|$, *there is a tuple* $(l_1, \ldots, l_d) \in \prod_{1 \leq j \leq d} X_{i_j}$ *such that* $\mathrm{Forb}(C^{X_s}, Y) = \prod_{1 \leq j \leq d}(X_{i_j} \setminus \{l_j\})$.

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 $\mathrm{VCD}_\Psi(C) = d < m$. The proof is an induction on m. The base case is $m = d+1$ or equivalently $d = m-1$, which is proved in Lemma 39. Assume that the claim is true for any $m' < m$. Pick $X_t \in X \setminus \{X_s\}$. By Lemma 38, 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 35, $C^{X_s} - X_t = (C - X_t)^{X_s}$ and thus $\mathrm{Forb}(C^{X_s} - X_t) = \mathrm{Forb}((C - X_t)^{X_s})$. $C - X_t$ 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 $(l_1, \ldots, l_d) \in \prod_{1 \leq j \leq d} X_{i_j}$ such that $\mathrm{Forb}((C - X_t)^{X_s}, Y) = \prod_{1 \leq j \leq d}(X_{i_j} \setminus \{l_j\})$ and hence $\mathrm{Forb}(C^{X_s} - X_t, Y) = \prod_{1 \leq j \leq d}(X_{i_j} \setminus \{l_j\})$. 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 $(l_1, \ldots, l_d) \in \prod_{1 \leq j \leq d} X_{i_j}$ with

$$\mathrm{Forb}(C^{X_s}, Y) = \textstyle\prod_{1 \leq j \leq d}(X_{i_j} \setminus \{l_j\}). \quad (1)$$

Furthermore, $(C^{X_s})^{X_t} = (C^{X_t})^{X_s}$, so $\mathrm{Forb}((C^{X_s})^{X_t}) = \mathrm{Forb}((C^{X_t})^{X_s})$. 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 $(l_1, \ldots, l_{d-1}) \in \prod_{1 \leq j \leq d-1} X_{i_j}$ such that $\mathrm{Forb}((C^{X_t})^{X_s}, Y) = \prod_{1 \leq j \leq d-1}(X_{i_j} \setminus \{l_j\})$, and hence $\mathrm{Forb}((C^{X_s})^{X_t}, Y) = \prod_{1 \leq j \leq d-1}(X_{i_j} \setminus \{l_j\})$. By Lemma 37, 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

$$\mathrm{Forb}(C^{X_s}, Y') = (\textstyle\prod_{1 \leq j \leq d-1}(X_{i_j} \setminus \{l_j\})) \times (X_t \setminus \{l_t\}). \quad (2)$$

```
Labeled Tail Subroutine (LTS)
Input: a $\mathrm{VCD}_\Psi$-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.

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 $\mathrm{Forb}(C^{X_s})$ 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 (1) and (2), 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 $\mathrm{Forb}(C^{X_s}, Y) = \prod_{1 \leq j \leq d}(X_{i_j} \setminus \{l_j\})$. ■

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

Theorem 41 *For any $X_s \in X$, there is a bipartite graph between the set $\mathrm{tail}_{X_s}(C)$ and the set $\mathrm{Forb}(C^{X_s})$, with an edge between a concept and a forbidden labeling if this forbidden labeling is contained in the concept. All such graphs have a unique matching.*

Proof See the appendix. ■

Let $\mathrm{LRep}_d(X) \subset \mathrm{LRep}_{\leq d}(X)$ 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 $\mathrm{tail}_{X_s}(C)$ and $\mathrm{LRep}_d(X \setminus \{X_s\})$.

Corollary 42 *Algorithm 3 returns a representation mapping between* $\mathrm{tail}_{X_s}(C)$ *and some* $\mathrm{LRep}_d(X \setminus \{X_s\})$*.*

Proof By Theorem 41, there is a unique consistent bijection r between $\mathrm{tail}_{X_s}(C)$ 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 40, 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 41. ■

Theorem 43 *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 See the appendix. ■

Now we have all the pieces in place for verifying Theorem 30, 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 30. By Theorem 43, 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 29 is then obvious as $|\mathrm{LRep}_{\leq d}(X)| = |C|$, and condition (ii) follows from the non-clashing property of r and Lemma 32. ■

6. Which notions of VCD fulfill the reduction property?

This section examines 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.

6.1 The Graph-Dimension

Our main objective here is to justify the following theorem.

Theorem 44 VCD_{Ψ_G} *fulfills the reduction property.*

To prove Theorem 44 we need a sequence of lemmas and theorems. Recall that as shown in Lemma 10, 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 10 may be of interest beyond the study of $\mathrm{VCD}_{\Psi_{\mathrm{G}}}$, as it applies to a broad class of notions of VC-dimension. The following two lemmas are immediate corollaries from Lemma 10 and Lemma 11, respectively.

Lemma 45 *Let C be $\mathrm{VCD}_{\Psi_{\mathrm{G}}}$-maximum. Let $\varphi_t \in \Psi_{\mathrm{G_t}}$, 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 $\mathrm{VCD}_{\Psi_{\mathrm{G}}}$-maximum of dimension $\mathrm{VCD}_{\Psi_{\mathrm{G}}}(C)$.*

Lemma 46 *Let C be a $\mathrm{VCD}_{\Psi_{\mathrm{G}}}$-maximum class and let $\overline{\varphi} = (\varphi_1, \ldots, \varphi_m)$ be a tuple of mappings such that $\varphi_i \in (\Psi_{\mathrm{G_i}} \cup \{\mathrm{id}_i\})$, for all $i \in [m]$. Then $\overline{\varphi}(C)$ is also a $\mathrm{VCD}_{\Psi_{\mathrm{G}}}$-maximum class of dimension $\mathrm{VCD}_{\Psi_{\mathrm{G}}}(C)$.*

In the binary case, restrictions and reductions of maximum classes are again maximum (Welzl, 1987). Theorem 15 implies that the restriction of a $\mathrm{VCD}_{\Psi_{\mathrm{G}}}$-maximum class of $\mathrm{VCD}_{\Psi_{\mathrm{G}}}$ d is also maximum of $\mathrm{VCD}_{\Psi_{\mathrm{G}}}$ $d-1$. Our core result here is that any reduction of a $\mathrm{VCD}_{\Psi_{\mathrm{G}}}$-maximum class is also $\mathrm{VCD}_{\Psi_{\mathrm{G}}}$-maximum. To show this, we first claim that for any $\mathrm{VCD}_{\Psi_{\mathrm{G}}}$-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. To prove this claim we first establish the following crucial lemma.

Lemma 47 *Let $X_i = \{0,1\}$, for $i \in [m-1]$, $X_m = \{0, \ldots, N_m\}$, $N_m \geq 2$. Let $\Psi = \mathrm{id}_1 \times \cdots \times \mathrm{id}_{m-1} \times \Psi_{\mathrm{G_m}}$ 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

$$2^{m-1} - 1 \leq |C_i| \leq 2^{m-1}, \quad \text{for all } i \in \{0, \ldots, N_m\}. \tag{3}$$

$|C_i| \leq |C - X_m| = 2^{m-1}$ 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, 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 (3), 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,

$$c_i = \overline{c} \times \{i\} \in C_i \quad \text{for each } i \in \{0, \ldots, k\}. \tag{4}$$

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 (3), $|C_l|= 2^{m-1}-1$ 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), $\overline{c} \times \{0\} \in \overline{\psi}(C_i)$, for all $i \in \{0, \ldots, k\} \setminus \{t\}$, so $\overline{\psi}(C) = \{0,1\}^m$. ■

We now generalize Lemma 48 and come back to the main theorem of this section. We first prove Theorem 44 for a special case.

Lemma 48 *Let C be $\mathrm{VCD}_{\Psi_\mathrm{G}}$-maximum with $\mathrm{VCD}_{\Psi_\mathrm{G}}(C) = m-1$. Then for all $\overline{c} \in C-X_m$, $|\{c \in C \mid c - X_m = \overline{c}\}|\in \{1, N_m + 1\}$.*

Proof For the purpose of contradiction, assume that for some $\overline{c} \in C - X_m$, $|\{c \in C \mid c - X_m = \overline{c}\}|\in \{2, \ldots, N_m\}$. Consider the mapping $\varphi_i \in \Psi_{\mathrm{G_i}}$, for $i \in \{1, \ldots, m - 1\}$, with

$$\varphi_i(x) = \begin{cases} 1 & \text{if } x = \overline{c}(X_i) \\ 0 & \text{otherwise.} \end{cases}$$

and let $\overline{\varphi} = (\varphi_1, \ldots, \varphi_{m-1}, \mathrm{id}_m)$. By Lemma 46, $\overline{\varphi}(C)$ is $\mathrm{VCD}_{\Psi_\mathrm{G}}$-maximum of $\mathrm{VCD}_{\Psi_\mathrm{G}}$ $m-1$. Let $c' \in C$ with $c' - X_m = \overline{c}$. By the definition of φ_i, for each $c \in C$ with $c - X_m \neq \overline{c}$, $\overline{\varphi}(c) - X_m \neq \overline{\varphi}(c') - X_m$. That is, $|\{c \in C \mid c - X_m = \overline{c}\}|\in \{2, \ldots, N_m\}$ implies that $|\{c'' \in \overline{\varphi}(C) \mid c'' - X_m = \overline{\varphi}(c') - X_m\}|\in \{2, \ldots, N_m\}$. This contradicts Lemma 47. ■

Proof of Theorem 44. 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 17 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 48 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, too. By Theorem 15, $C|_{\{X_1,\ldots,X_{d+1}\}}$ 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 48, $c_i|_{\{X_2,\ldots,X_{d+1}\}}$ 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, $c_{\mathrm{new}}|_{\{X_1,\ldots,X_{d+1}\}} \in C|_{\{X_1,\ldots,X_{d+1}\}}$, and thus $C|_{\{X_1,\ldots,X_{d+1}\}} = C_{\mathrm{new}}|_{\{X_1,\ldots,X_{d+1}\}}$. Hence, $\overline{\psi}(C|_{\{X_1,\ldots,X_{d+1}\}}) = \overline{\psi}(C_{\mathrm{new}}|_{\{X_1,\ldots,X_{d+1}\}}) = \{0,1\}^{d+1}$ and C shatters $\{X_1, \ldots, X_{d+1}\}$ in contradiction to $\mathrm{VCD}_{\Psi_{\mathrm{G}}}(C) = d$. ■

Hence, for a $\mathrm{VCD}_{\Psi_{\mathrm{G}}}$-maximum class C, $[C]_{\geq 2}^{X_t} = C^{X_t}$, for all $t \in [m]$. More precisely, for a $\mathrm{VCD}_{\Psi_{\mathrm{G}}}$-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 18 and Theorem 44.

Corollary 49 *Let C be a $\mathrm{VCD}_{\Psi_{\mathrm{G}}}$-maximum class with $\mathrm{VCD}_{\Psi_{\mathrm{G}}}(C) = d$. Then C^{X_t} is $\mathrm{VCD}_{\Psi_{\mathrm{G}}}$-maximum with $\mathrm{VCD}_{\Psi_{\mathrm{G}}}(C^{X_t}) = d-1$, for any $t \in [m]$.*

We remind the reader that $\Psi^* \supseteq \Psi_{\mathrm{G}}$ and also, any VCD_{Ψ^*}-maximum class C is also $\mathrm{VCD}_{\Psi_{\mathrm{G}}}$-maximum with $\mathrm{VCD}_{\Psi_{\mathrm{G}}}(C) = \mathrm{VCD}_{\Psi^*}(C)$. So, all the statements and proofs in this section can be applied to VCD_{Ψ^*} as well.

6.2 Pollard's pseudo-dimension

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

Proposition 50 *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 5. 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 6. As shown in Table 6, 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 VCD$_{\Psi_{\mathrm{P}}}$-maximum of dimension 2. As shown in bold in Table 5, there are two different choices for $\bar{c}$. For example, for $\bar{c} = (0,0)$, $|\{c \in C \mid c - X_3 = \bar{c}\}|= |\{c_1, c_2\}|= 2$, and for $\bar{c} = (1,2)$, $|\{c \in C \mid c - X_3 = \bar{c}\}|= |\{c_{18}, c_{19}\}|= 2$. ■

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: Maximum class C of VCD$_{\Psi_{\mathrm{P}}}$ 2 used in the proof of Proposition 50.

$c \in C'$	$\psi_{P,k_1}(X_1)$	$\psi_{P,k_2}(X_2)$	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 6: Mappings of the concept class C from Table 5.

The class in Table 5 does not stay $\mathrm{VCD}_{\Psi_{\mathrm{P}}}$-maximum when applying either definition of reduction w.r.t. X_3.

Corollary 51 *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 5. As shown in Table 7, $[C]^{X_3}_{\geq 2}$ 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. ■

Remark 52 *The class* C *discussed in the proof of Proposition 50 does have a tight compression scheme, as shown in Table 5. 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.*

$c \in [C]^{X_3}_{\geqslant 2}$	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 7: Both reductions of C where C is the $\mathrm{VCD}_{\Psi_{\mathrm{P}}}$-maximum class from Table 5.

6.3 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 53 *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* $\bar{c} \in C - X_t$*,* $|\{c \in C \mid c - X_t = \bar{c}\}| = 2 \leq N_t$*.*

Proof Consider the concept class C in Table 8. 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 $a, b \in \{0, 1, 2\}$. As shown in bold in Table 8, there are two choices for $\bar{c}$. for $\bar{c} = (0,0)$, $|\{c \in C \mid c - X_3 = \bar{c}\}| = |\{c_1, c_2\}| = 2$, and for $\bar{c} = (2,0)$, $|\{c \in C \mid c - X_3 = \bar{c}\}| = |\{c_9, c_{10}\}| = 2$. ■

The reduction of the class C in Table 8 is not $\mathrm{VCD}_{\Psi_{\mathrm{N}}}$-maximum under either definition of reduction.

Corollary 54 *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.*

$c \in C$	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**

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

Table 8: Maximum class C of $\mathrm{VCD}_{\Psi_\mathrm{N}}$ 1 used in the proof of Proposition 53 and its reduction.

Proof Consider the $\mathrm{VCD}_{\Psi_\mathrm{N}}$-maximum class C in Table 8 and Natarajan family of mappings Ψ_N. Clearly, C^{X_3} is the empty set and also as shown in Table 8 (right), $[C]^{X_3}_{\geq 2}$ is of $\mathrm{VCD}_{\Psi_\mathrm{N}}$ 1 with $\sum_{i=0}^{0}\binom{3}{i}\binom{3}{2}^i < \text{size}([C]^{X_3}_{\geq 2}) < \sum_{i=0}^{1}\binom{3}{i}\binom{3}{2}^i$. So, in either case, the reduction of C w.r.t. X_3 is not $\mathrm{VCD}_{\Psi_\mathrm{N}}$-maximum. ■

Remark 55 *The class C in Table 8 has* no *tight compression scheme. Note that the Natarajan-dimension violates both premises of Theorems 22 and 30—it violates the reduction property, and it is not based on a spanning family.*

7. Tight compression schemes and recursive teaching

In this section we connect a recently introduced teaching notion, namely the recursive teaching dimension (RTD) Zilles et al. (2011), to tight compression schemes. We first generalize the algebraic characterization of teaching sets by Samei et al. (2014a) to the multi-label case.

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 $\mathrm{TD}(C) = \max_{c \in C} \mathrm{TD}(c, C)$ (Goldman and Kearns, 1995; Shinohara and Miyano, 1991).

The next two lemmas generalize the core idea of Samei et al. (2014a) in algebraic characterization of teaching sets in the binary case to multi-label concept classes. The proofs are analogous to those in the binary case (Theorem 1 and Lemma 1) by Samei et al. (2014a). The tricky point in the multi-label case is to use the following type of polynomials.

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:

$$p_{i,k}(X_i) = \begin{cases} 1 & \text{if } X_i = k \\ 0 & \text{if } X_i \in \{0, \ldots, N_i\} \setminus \{k\}. \end{cases} \tag{5}$$

We can find such a polynomial using interpolation. In the binary case, as discussed by Samei et al. (2012),

$$p_{i,0}(X_i) = 1 - X_i \ \text{ and } \ p_{i,1}(X_i) = X_i. \tag{6}$$

Lemma 56 *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 (sketch) Let $\{(X_{i_1}, n_{i_1}), \ldots, (X_{i_k}, n_{i_k})\}$ be a teaching set for a concept c in C. Let $p(X_{i_1}, \ldots, X_{i_k}) = p_{i_1,n_{i_1}}(X_{i_1}) \times \cdots \times p_{i_k,n_{i_k}}(X_{i_k})$. Since 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 $P^k(N_{i_1}, \ldots, N_{i_k})$. ■

The next lemma is a stronger result where the VCD_Ψ of the class comes into account. The proof is analogous to the proof of Lemma 1 in (Samei et al., 2014a).

Lemma 57 *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 $P^d(N_{i_1}, \ldots, N_{i_k})$.*

The following definitions are based on previous literature on recursive teaching (Doliwa et al., 2010; Zilles et al., 2011). 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 $S_i \in \text{TS}(c_i, \{c_i, \ldots, c_n\})$ for all $i = 1, \ldots, n$. The *order* of the teaching plan P is $\text{ord}(P) = \max_{i=1,\ldots,n}|S_i|$. The *recursive teaching dimension* of C is

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

A teaching plan of C whose order equals $\text{RTD}(C)$ 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 $\text{RTD}(c_i, C)$.

We first present a Sauer-type bound on the size of a concept class with a given RTD. The following theorem is a generalization of the same result in the binary case (Samei et al., 2014a) that is proved with the same technique.

Theorem 58 *Let $C \subseteq \prod_{i=1}^{m}\{0, \ldots, N_i\}$. If $\text{RTD}(C) = r$ then the monomials from $P^r(N_1, \ldots, N_m)$ span the vector space $\mathbb{R}^{|C|}$.*

As a corollary from Theorem 58, we obtain a generalized Sauer-type bound for RTD.

Corollary 59 *Let $C \subseteq \prod_{i=1}^{m}\{0, \ldots, N_i\}$. If $\text{RTD}(C) = r$ then*

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

Definition 60 *Let $C \subseteq \prod_{i=1}^{m}\{0, \ldots, N_i\}$ with $\text{RTD}(C) = r$. C is RTD-maximum if $|C| = \Phi_r(N_1, \ldots, N_m)$. C is called RTD-maximal if $\text{RTD}(C \cup \{c\}) > r$ for any concept $c \notin C$.*

In the binary case, Doliwa et al. (2010, 2014) proved that for every binary VCD-maximum class, RTD and VCD are equal. 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 $\mathrm{RTD}(C) \leq d$. On the other hand, C is VCD_Ψ-maximum and $|C|= \Phi_d(N_1, \ldots, N_m)$, so by Definition 60, 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 $\mathrm{tail}_{X_t}(C)$. Recall that when C is VCD_Ψ-maximum with $\mathrm{VCD}_\Psi(C) = d$, then C^{X_t} is VCD_Ψ-maximum of VCD_Ψ $d-1$ and $\mathrm{Forb}(C^{X_t})$ denotes the set of forbidden labelings of size d for C^{X_t}. We have already shown that there is a bipartite graph between $\mathrm{tail}_{X_t}(C)$ and $\mathrm{Forb}(C^{X_t})$ with a unique perfect matching such that there is an edge between $c \in \mathrm{tail}_{X_t}(C)$ and $S \in \mathrm{Forb}(C^{X_t})$ iff S is consistent with c, i.e., $S \subseteq c$ (see Theorem 41). 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 $\mathrm{Forb}(C^{X_t})$ that is consistent with only one concept in $\mathrm{tail}_{X_t}(C)$.

Lemma 61 (Lovász and Plummer, 1986; Zhongyuana and Zhibob, 2013) *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. (2010) revealed a strong relationship between sample compression schemes and recursive teaching sets. In particular, they showed that for a VCD-maximum class, there exists a teaching plan for 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. Here we generalize that result to the multi-label case.

Theorem 62 *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 each $r(c_i)$, $i \in \{1, \ldots, |C|\}$, is the compression set for c_i resulting from Algorithm 2.*

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. 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, $C^{X_s} \times X_s = \{c_i^k \mid 1 \le i \le l \text{ and } 0 \le k \le N_s\}$. Let the teaching plan $\mathcal{P}^2$ for $C^{X_s} \times X_s$ be as follows:

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

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

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

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 $\mathrm{tail}_{X_s}(C)$ are taught by their corresponding compression sets before the concepts in $C^{X_s} \times X_s$. For simplicity, let $C' = \mathrm{tail}_{X_s}(C)$ 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 $\overline{f} \in \mathrm{Forb}(C^{X_s})$, $\overline{f}$ 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 $c' \in C'$, $r(c') \in \mathrm{TS}(c', \{c'\} \cup C^{X_s} \times X_s)$. 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 $r(c_i') \in \mathrm{TS}(c_i', \{c_i', \ldots, c_{l'}'\})$, for all $i \in \{1, \ldots, l'\}$. Such an ordering along with $\mathcal{P}^2$ yields the teaching plan

$$\begin{aligned}\mathcal{P} = \{&(c_1', r(c_1')), \ldots, (c_{l'}', r(c_{l'}')), (c_1^{N_s}, r(c_1^{N_s})), \ldots, (c_1^0, r(c_1^0)), \ldots,\\ &(c_l^{N_s}, r(c_l^{N_s})), \ldots, c_l^0, r(c_l^0))\}\end{aligned}$$

of order d for C.

By Theorem 41, there is a bipartite graph $G = (C' \cup \mathrm{Forb}(C^{X_s}), E')$ with a unique perfect matching between C' and $\mathrm{Forb}(C^{X_s})$, where there is an edge between a concept in C' and a forbidden labeling in $\mathrm{Forb}(C^{X_s})$ if this forbidden labeling is contained in the concept. By Lemma 61, there is a forbidden labeling $\overline{f_1} \in \mathrm{Forb}(C^{X_s})$ 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, $r(c'_1) \in \mathrm{TS}(c'_1, \{c'_1, \ldots, c'_{l'}\})$. The subgraph G_1 induced by $C' \setminus \{c'_1\} \cup \mathrm{Forb}(C^{X_s}) \setminus \{\overline{f_1}\}$ also has a unique perfect matching, because otherwise $G = (C', \mathrm{Forb}(C^{X_s}))$ cannot have a unique perfect matching. Similarly, by using Lemma 61, we conclude that there is a concept $c'_2 \in C' \setminus \{c'_1\}$ such that $r(c'_2) \in \mathrm{TS}(c'_2, \{c'_2, \ldots, c'_{l'}\})$. 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 (Doliwa et al., 2010; Samei et al., 2012, 2014a), Theorem 62 establishes this result with a completely different approach from the one in the literature.

Corollary 63 *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 59, $\mathrm{RTD}(C) \geq d$. On the other hand, by Theorem 62, there is a teaching plan for C of order d. Hence, $\mathrm{RTD}(C) = d$. ■

As shown by Samei et al. (2014a), the other direction of the above corollary is not always true. In fact, there is a binary RTD-maximum class that is not VCD-maximum.

8. One-inclusion Hypergraph

This section studies the one-inclusion hypergraph of multi-label concept classes. 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 64 (Rubinstein et al., 2009) *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 65 *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.*

8.1 Shortest-path Closedness

Kuzmin and Warmuth (2007) proved that when a class is VCD-maximum, then in 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. Samei et al. (2014a) provided an alternative proof by using the algebraic characterization of teaching sets. Here, we apply the same approach and show that the one-inclusion hypergraph for a VCD_Ψ-maximum class is also shortest-path closed.

We first present a lemma that is a generalization of Lemma 17 in Kuzmin and Warmuth (2007) in which they proved that when C is VCD-maximum, then for any $c \in C$, the set of instances corresponding to the incident edges for c in the one-inclusion graph of C, is a teaching set for c. While Kuzmin and Warmuth (2007) used a combinatorial argument in their proof, we apply Linear Algebra here. The proof is omitted, as it is established analogously to the first part of the proof of Theorem 3 by Samei et al. (2014a).

Lemma 66 (Kuzmin and Warmuth, 2007) *Let Ψ_i, for all $i \in [m]$, be a spanning family of mappings on X_i, $\Psi = \Psi_1 \times \cdots \times \Psi_m$ and C be a VCD_Ψ-maximum class. Then for every $c \in C$, $I_C(c)$ is a teaching set for c.*

By using Lemma 66, the proof of the following theorem can be established analogously to the proof of Theorem 1 in Samei et al. (2014a).

Theorem 67 *Let Ψ_i, for all $i \in [m]$, be a spanning family of mappings on X_i and $\Psi = \Psi_1 \times \cdots \times \Psi_m$. If C is a VCD_Ψ-maximum class, then C is shortest-path closed.*

8.2 Connection to the tight compression scheme

Assume that VCD_Ψ fulfills the reduction property. As in the binary case (Kuzmin and Warmuth, 2007), 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 (2007), 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 $r(c')$ (Kuzmin and Warmuth, 2007). 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 68 *Let C be VCD_Ψ-maximum of dimension d and r be a representation mapping between C and some $\mathrm{LRep}_{\leq d}(X)$. Let $G(C) = (V, E)$ be the one-inclusion hypergraph for C. Then for any hyperedge $e = \{c_0, c_1, \dots, c_{N_t}\}$ labeled with X_t in $E(G)$, $t \in [m]$, e charges exactly N_t incident concepts to e.*

Proof See the appendix. ■

Corollary 69 *Let r be a representation mapping for the VCD_Ψ-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 68 along with (10) show 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. ■

9. Sample compression for classes of VCD_Ψ 1

In the binary case, compression schemes of size d for maximum classes of VC-dimension d, like the VC Scheme proposed by Floyd and Warmuth (1995), 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 (Welzl and Woeginger, 1987). 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 9: 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 22 or Theorem 30. 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 9. 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 9 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.

9.1 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 70 *Our approach for* $\mathrm{VCD}_{\Psi_{\mathrm{G}}}$ 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 71 *Let* $\mathrm{VCD}_{\Psi_{\mathrm{G}}}(C) = 1$*. 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 10: Illustration of the proof of Lemma 71.

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 $\mathrm{VCD}_{\Psi_{\mathrm{G}}}(C) = 1$. 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 $\mathrm{VCD}_{\Psi_{\mathrm{G}}}(C) = 2$, not even for binary classes.

Definition 72 *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 11: Concept class $C \subseteq \{0, 1, 2\}^4$ of $\mathrm{VCD}_{\Psi_\mathrm{G}}$ 1 and the representatives.

Using Definition 72, we obtain a simple lemma.

Lemma 73 *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 $\mathrm{VCD}_{\Psi_\mathrm{G}}(C) = 1$ 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 71, $\{e_i, e_j\}$ 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 74 *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 $\mathrm{VCD}_{\Psi_\mathrm{G}}(C) = 1$ 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 73 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 75 *Let $\mathrm{VCD}_{\Psi_\mathrm{G}}(C) = 1$, 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 $\mathrm{VCD}_{\Psi_\mathrm{G}}(C) = 1$, e_1 uniquely implies e_3 by Lemma 73. ■

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

Proof For $|S|= 1$, there is nothing to show, and for $|S|= 2$, Corollary 74 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 74, 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 75, e_i explicitly and thus uniquely implies e_j. If r implicitly implies e_j, then by Lemma 75, 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 11 shows the representatives for concepts in a $\mathrm{VCD}_{\Psi_{\mathrm{G}}}$ 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 76 now allows us to define a compression scheme of size 1 for any $\mathrm{VCD}_{\Psi_{\mathrm{G}}}$ 1 class.

Corollary 77 *Let* $\mathrm{VCD}_{\Psi_{\mathrm{G}}}(C) = 1$. *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 76.

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 3 *Consider the class in Table 11. 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 77, so that the latter applies also to infinite concept classes of $\mathrm{VCD}_{\Psi_{\mathrm{G}}}$-dimension 1.

9.2 Pollard's Pseudo-dimension

Although we could not prove that $\mathrm{VCD}_{\Psi_{\mathrm{P}}}$ 1 classes have compression schemes of size 1, we show that the approach that we used for classes of $\mathrm{VCD}_{\Psi_{\mathrm{G}}}$ 1 does not work here. In particular, we illustrate that Lemma 71 does not hold for $\mathrm{VCD}_{\Psi_{\mathrm{P}}}$ 1 classes.

Proposition 78 *There is a multi-label class* C *of* $\mathrm{VCD}_{\Psi_{\mathrm{P}}}$ 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 12. 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. ■

Note that the concept class in Table 9 is not $\mathrm{VCD}_{\Psi_{\mathrm{P}}}$-maximal. In fact, we could neither find a proper $\mathrm{VCD}_{\Psi_{\mathrm{P}}}$-maximal class of $\mathrm{VCD}_{\Psi_{\mathrm{P}}}$ 1 nor prove that every $\mathrm{VCD}_{\Psi_{\mathrm{P}}}$ 1 class can be embedded in a $\mathrm{VCD}_{\Psi_{\mathrm{P}}}$-maximum class of $\mathrm{VCD}_{\Psi_{\mathrm{P}}}$ 1.

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

Table 12: Class of $\mathrm{VCD}_{\Psi_\mathrm{P}}$ 1 with 2 concepts of teaching dimension 2.

9.3 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 71 does not hold for $\mathrm{VCD}_{\Psi_\mathrm{N}}$ 1 classes though.

Proposition 79 *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 13. 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 13: Class of $\mathrm{VCD}_{\Psi_\mathrm{N}}$ 1 with 3 concepts of teaching dimension 2.

Note that the concept class in Table 9 is not $\mathrm{VCD}_{\Psi_\mathrm{N}}$-maximal. In fact, we could neither find a proper $\mathrm{VCD}_{\Psi_\mathrm{N}}$-maximal class of $\mathrm{VCD}_{\Psi_\mathrm{N}}$ 1 nor prove that every $\mathrm{VCD}_{\Psi_\mathrm{N}}$ 1 class can be embedded in a $\mathrm{VCD}_{\Psi_\mathrm{N}}$-maximum class of $\mathrm{VCD}_{\Psi_\mathrm{N}}$ 1.

Acknowledgments

Boting Yang and Sandra Zilles were supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

Bibliography

N. Alon. On the density of sets of vectors. *Discrete Mathematics*, 46(2):199–202, 1983.

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.

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.

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.

M. Darnstädt, T. Doliwa, Simon H.U., and S. Zilles. Order compression schemes. In *Proceedings of the 24th International Conference on Algorithmic Learning Theory (ALT)*, pages 173–187, 2013.

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.

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.

S. Floyd and M. K. Warmuth. Sample compression, learnability, and the Vapnik-Chervonenkis dimension. *Machine Learning*, 21(3):269–304, 1995.

S. A. Goldman and M. J. Kearns. On the complexity of teaching. In *Proceedings of the fourth Annual Workshop on Computational Learning Theory*, COLT '91, pages 303–314, 1991.

S. A. Goldman and M. J. Kearns. On the complexity of teaching. *Journal of Computer and System Sciences*, 50:20–31, 1995.

L. Gurvits. Linear algebraic proofs of VC-dimension based inequalities. In *Proceedings of the Third European Conference on Computational Learning Theory*, EuroCOLT '97, pages 238–250, London, UK, 1997. Springer-Verlag.

D. Haussler and P. M. Long. A generalization of Sauer's lemma. *Journal of Combinatorial Theory, Series A*, 71(2):219–240, 1995.

D. Haussler, N. Littlestone, and M.K. Warmuth. Predicting $\{0, 1\}$-functions on randomly drawn points. *Information and Computation*, 115(2):248–292, 1994.

D. Kuzmin and M. K. Warmuth. Unlabeled compression schemes for maximum classes. *Journal of Machine Learning Research*, 8:2047–2081, 2007.

N. Littlestone and M. Warmuth. Relating data compression and learnability. Unpublished notes, 1986.

L. Lovász and M. D. Plummer. *Matching Theory*, volume 121, page 139. North-Holland Mathematics Studies, North-Holland Publishing, Amsterdam, 1986.

S. Moran and M.K. Warmuth. Labeled compression schemes for extremal classes. *CoRR*, abs/1506.00165, 2015. URL `http://arxiv.org/abs/1506.00165`.

S. Moran and A. Yehudayoff. Proper PAC learning is compressing. *CoRR*, abs/1503.06960, 2015. URL `http://arxiv.org/abs/1503.06960`.

B. K. Natarajan. On learning sets and functions. *Machine Learning*, 4(1):67–97, 1989.

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.

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.

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.

R. Samei, P. Semukhin, B. Yang, and S. Zilles. Algebraic methods proving Sauer's bound for teaching complexity. *Theoretical Computer Science*, 558:35–50, 2014a.

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, 2014b.

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, 2014c.

N. Sauer. On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13(1):145–147, 1972.

A. Shinohara and S. Miyano. Teachability in computational learning. *New Generation Computing*, 8(4):337–347, 1991.

H. U. Simon and B. Szörényi. One-inclusion hypergraph density revisited. *Information Processing Letters*, 110(8-9):341–344, 2010.

R. Smolensky. Well-known bound for the VC-dimension made easy. *Computational Complexity*, 6(4):299–300, 1997.

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 '89, pages 3–21, 1989.

E. Welzl. Complete range spaces. Unpublished notes, 1987.

E. Welzl and G. Woeginger. On Vapnik-Chervonenkis dimension one. Unpublished notes, 1987.

C. Zhongyuana and C. Zhibob. Conjugated circuits and forcing edges. *Communications in Mathematical and in Computer Chemistry / MATCH*, 69(3):721–732, 2013.

S. Zilles, S. Lange, R. Holte, and M. Zinkevich. Models of cooperative teaching and learning. *Journal of Machine Learning Research*, 12:349–384, 2011.

Appendix A. Proofs Omitted From Section 4

Theorem 18 *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 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 15, $C - X_m$ is VCD_Ψ-maximum of dimension d. So,

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

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

Proposition 19 *For any X_i, X_j with $i \neq j$, $(C^{X_i})^{X_j} = (C^{X_j})^{X_i}$.*

Proof

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

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

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

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

■

Appendix B. Proofs Omitted From Section 5

B.1. Proof of Results Concerning the Generalization of Floyd and Warmuth's Compression Scheme

Lemma 26 *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 (Analogous to the proof of Lemma 2 in (Floyd and Warmuth, 1995)) 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 15, $C|_{\{X_1,\ldots,X_d\}}$ is VCD_Ψ-maximum of dimension d and by Corollary 12, $\overline{\psi^{1,d}}(C|_{\{X_1,\ldots,X_d\}}) = \{0,1\}^d$. Let $\psi_{d+1} : X_{d+1} \to \{0,1\}$ be $\psi_{0\neq 1}$, as discussed in Remark 3 and $\overline{\psi^{1,d+1}} = (\psi_1, \ldots, \psi_d, \psi_{d+1})$. Lemma 25 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. ■

Lemma 27 *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 $c_{S_i,C} - X_t = c_{S,C^{X_t}}$.*

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. From Theorem 15, we obtain that $C|_{\{X_1,\ldots,X_d\}}$ is VCD_Ψ-maximum of dimension d and by Corollary 12, $\overline{\psi^{1,d}}(C|_{\{X_1,\ldots,X_d\}}) = \{0, 1\}^d$.

On the one hand, $\{\{0, 1\}^d \times \{0\}\} \subseteq \overline{\psi^{1,d+1}}(C|_{\{X_1,\ldots,X_{d+1}\}})$, from Lemma 25.

On the other hand, because $c_{S,C^{X_d}}(X_{d+1}) = 1$, 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 17, 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 some $c \in C^{X_d}$, 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. ■

Theorem 28 *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 (Theorem 10 in (Floyd and Warmuth, 1995)) 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 15, $C - X_m$ 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 23, 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 26, $c_{S,C} - X_m = c_{S,C-X_m}$, 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 18, 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 23, 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 27, $c_{S_i,C} - X_m = c_{S,C^{X_m}}$, 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 $c_{S_i,C} = c \cup \{(X_m, i)\}$.

■

B.2. Proof of Results Concerning the Tight Compression Scheme

Lemma 32 *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 By contradiction (analogous to the proof of Lemma 1 in (Kuzmin and Warmuth, 2007)).

$2 \Rightarrow 1$: 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

$$\begin{aligned}\mathrm{size}(C|_{X(S)}) &= \Phi_d(N_{i_1}, \ldots, N_{i_k}) = |\mathrm{LRep}_{\leq d}(X(S))| \\ &= |\{c \in C \mid r(c) \in \mathrm{LRep}_{\leq d}(X(S))\}| \qquad (7)\end{aligned}$$

and thus by the pigeon hole 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. ■

Corollary 33 *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 (Kuzmin and Warmuth, 2007)) 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 32, 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 bijective property of r_Y. By replacing $X(S)$ with Y in (7), and applying the same counting argument as in the second part of the proof of Lemma 32, we conclude that r_Y is bijective. ■

Lemma 34 *Let $s, t \in [m]$ with $s \neq t$. Then the following statements are true.*

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

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

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

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

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

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

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

We need to show that if $0\bar{c} \in \mathrm{tail}_{X_s}(C^{X_t}) \setminus \mathrm{tail}_{X_s}(C - X_t)$, there is an $l \in X_t$ for which $0\bar{c}l \notin \mathrm{tail}_{X_s}(C)$. Assume that for all $j \in X_t$, $0\bar{c}j \in \mathrm{tail}_{X_s}(C)$, 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 $0\bar{c} \in \mathrm{tail}_{X_s}(C - X_t)$.

2. Let $0\bar{c} \in \mathrm{tail}_{X_s}(C - X_t)$. We show that for each $j \in X_t$ with $0\bar{c}j \in C$, we have $0\bar{c}j \in \mathrm{tail}_{X_s}(C)$. W.l.o.g., assume that $0\bar{c}0 \in C$, but $0\bar{c}0 \notin \mathrm{tail}_{X_s}(C)$. 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, $0\bar{c} \notin \mathrm{tail}_{X_s}(C - X_t)$—a contradiction.

For a concept $0\bar{c} \in \mathrm{tail}_{X_s}(C - X_t) \cap \mathrm{tail}_{X_s}(C^{X_t})$, 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 $0\bar{c}j \in \mathrm{tail}_{X_s}(C)$, for all $j \in X_t$.

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

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

Second, from Statements 1 and 2, any concept in $\mathrm{tail}_{X_s}(C^{X_t})$ can be mapped to N_t concepts in $\mathrm{tail}_{X_s}(C)$, and any concept in $\mathrm{tail}_{X_s}(C - X_t)$ can be mapped to a single concept in

$\mathrm{tail}_{X_s}(C)$. Hence, each concept in $\mathrm{tail}_{X_s}(C)$ is an extension of either a concept in $\mathrm{tail}_{X_s}(C^{X_t})$ or a concept in $\mathrm{tail}_{X_s}(C - X_t)$. ■

Lemma 35 *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 (Kuzmin and Warmuth, 2007)) 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, $C^{X_s} - X_t = (C - X_t)^{X_s}$. ■

Lemma 37 *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 15 and Theorem 18, 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 $|\mathrm{Forb}(C^{X_s})|= N_t|\mathrm{Forb}((C^{X_s})^{X_t})|+|\mathrm{Forb}((C-X_t)^{X_s})|$ as follows.

$$\begin{aligned}
|\mathrm{Forb}(C^{X_s})| &= \sum_{\substack{1\le i_1<\cdots<i_d\le m\\ i_j\neq s}} N_{i_1}\cdots N_{i_d}\\
&= N_t \sum_{\substack{1\le i_1<\cdots<i_{d-1}\le m\\ i_j\neq s\\ i_j\neq t}} N_{i_1}\cdots N_{i_{d-1}} + \sum_{\substack{1\le i_1<\cdots<i_d\le m\\ i_j\neq s\\ i_j\neq t}} N_{i_1}\cdots N_{i_d}\\
&= N_t|\mathrm{Forb}((C^{X_s})^{X_t})|+|\mathrm{Forb}((C-X_t)^{X_s})|. \qquad (8)
\end{aligned}$$

So,

$$|\mathrm{Forb}((C^{X_s})^{X_t})|= \frac{1}{N_t}|\mathrm{Forb}(C^{X_s},Y)| \qquad (9)$$

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

Therefore, any set of N_t forbidden labelings $S_i = S\cup\{(X_t,l_i)\}$, $1\le i\le N_t$ for C^{X_s} 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 (9)), we conclude that any forbidden labeling for $(C^{X_s})^{X_t}$ can be extended to N_t forbidden labelings for C^{X_s}. ■

Theorem 41 *For any $X_s\in X$, there is a bipartite graph between the set $\mathrm{tail}_{X_s}(C)$ and the set $\mathrm{Forb}(C^{X_s})$, with an edge between a concept and a forbidden labeling 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 (Kuzmin and Warmuth, 2007)) 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 $s\in\{1,\dots,m\}$. 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 $d'\le d$, $m'\le m$ and $m'+d' < m+d$. Pick $X_s, X_t\in X$. First, by Lemma 38, 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 34(3), any concept in $\mathrm{tail}_{X_s}(C)$ is an extension of either a concept in $\mathrm{tail}_{X_s}(C^{X_t})$ or a concept in $\mathrm{tail}_{X_s}(C-X_t)$. Also, $\mathrm{tail}_{X_s}(C^{X_t})$ is a VCD_Ψ-maximum class of dimension $d-1$ and $\mathrm{tail}_{X_s}(C-X_t)$ 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 $\mathrm{tail}_{X_s}(C-X_t)$ and $\mathrm{Forb}((C-X_t)^{X_s})$, and also, between $\mathrm{tail}_{X_s}(C^{X_t})$ and $\mathrm{Forb}((C^{X_s})^{X_t})$. We combine these two matchings to form a matching for $\mathrm{tail}_{X_s}(C)$. This is done in steps 2, 3 and 4 in Algorithm 3, as described in the following paragraphs.

Concepts in $\mathrm{tail}_{X_s}(C^{X_t})\setminus\mathrm{tail}_{X_s}(C-X_t)$ are matched to the forbidden labelings for $(C^{X_s})^{X_t}$ of size $d-1$. Consider a concept $\bar{c}\in\mathrm{tail}_{X_s}(C^{X_t})\setminus\mathrm{tail}_{X_s}(C-X_t)$ which gets matched to a forbidden labeling F for $(C^{X_s})^{X_t}$. On the one hand, by Lemma 34(1), there are N_t concepts $c_i\in\mathrm{tail}_{X_s}(C)$ such that $c_i - X_t = \bar{c}$, for $i\in\{1,\dots,N_t\}$. W.l.o.g., assume that for $i\in\{1,\dots,N_t\}$, $c_i = \bar{c}\cup\{(X_t,i)\}$, that is, $c_0 = \bar{c}\cup\{(X_t,0)\}$ is not in $\mathrm{tail}_{X_s}(C)$ and thus, $c_0\in C^{X_s}$. Since F is contained in $\bar{c}$, it is contained in c_i, $i\in\{0,\dots,N_t\}$, too. On the other hand, by Lemma 37, 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 34(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 36, 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 37, 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 36, 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 35, 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 $\mathrm{tail}_{X_s}(C)$ splits into a matching for $\mathrm{tail}_{X_s}(C^{X_t})$ and a matching for $\mathrm{tail}_{X_s}(C - X_t)$, and consequently, the existence of two matchings for $\mathrm{tail}_{X_s}(C)$ implies the existence of two matchings for $\mathrm{tail}_{X_s}(C^{X_t})$ or two matchings for $\mathrm{tail}_{X_s}(C - X_t)$. ■

Theorem 43 *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 (Kuzmin and Warmuth, 2007)) 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 42, 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, $r(c_i) \nsubseteq c_j$ and $r(c_j) \nsubseteq c_i$. 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, $r(c_i) \nsubseteq c_0$. 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 42, no clashes occur between the concepts in $\mathrm{tail}_{X_s}(C)$. ■

Appendix C. Proof Omitted From Section 8

Proposition 68 *Let* C *be* VCD_Ψ*-maximum of dimension* d *and* r *be a representation mapping between* C *and some* $\mathrm{LRep}_{\leq d}(X)$*. 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 $\mathrm{Chg}(e, X_t)$ 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 $\mathrm{Chg}(e, X_t) \geq N_t$. 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, $\sum\limits_{e \in E} \mathrm{Chg}(e, X_t)$, is lower-bounded by

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

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

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

On the other hand, 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 $\mathrm{LRep}_{\leq d}(X)$. That is,

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

Hence,

$$\begin{aligned}\sum_{1\leq t\leq m}\sum_{e\in E}\operatorname{Chg}(e,X_t) &= \sum_{S\in \mathrm{LRep}_{\leq d}(X)} |S| \qquad (10)\\ &= \sum_{1\leq t\leq m} N_t\Phi_{d-1}(N_1,\ldots,N_{t-1},N_{t+1},\ldots,N_m)\end{aligned}$$

and consequently,

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

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