JMLR: Workshop and Conference Proceedings vol 35:1–23, 2014

Sample Compression for Multi-label Concept Classes

Rahim Samei SAMEI20R@CS.UREGINA.CA
Department of Computer Science, University of Regina, Canada

Pavel Semukhin PAVEL.SEMUKHIN@UNIVIE.AC.AT
Kurt Gödel Research Centre, Vienna, Austria

Boting Yang BOTING@CS.UREGINA.CA
Department of Computer Science, University of Regina, Canada

Sandra Zilles ZILLES@CS.UREGINA.CA
Department of Computer Science, University of Regina, Canada

Abstract

This paper studies labeled sample compression for multi-label concept classes. For a specific extension of the notion of VC-dimension to multi-label classes, we prove that every maximum multi-label class of dimension d has a sample compression scheme in which every sample is compressed to a subset of size at most d. We further show that every multi-label class of 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 former, since there are multi-label concept classes of dimension 1 that are not contained in maximum classes of dimension 1.

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

1. Introduction

The combinatorial structure of a concept class is crucial for the complexity of learning concepts from the class. An important combinatorial parameter in this context is the Vapnik-Chervonenkis Dimension (VC-dimension), which provides bounds on the sample complexity in PAC-learning (Blumer et al., 1989). Another parameter that provides sample bounds is the smallest possible size of a *sample compression scheme (SCS)* of the class (Littlestone and Warmuth, 1986; Floyd and Warmuth, 1995). An SCS for a concept class C over the instance space X consists of two mappings, namely, (i) a compression mapping f, which, given a set S of examples over X that are labeled consistently with some concept in C, returns a subset $f(S) = S'$ of S (the compression set for S), and (ii) a decompression mapping g, which, given any compression set $S' = f(S)$ (for some S consistent with at least one concept in C), predicts labels for each instance in X, such that all examples from S are labeled correctly by $g(S')$.[1] The size of such an SCS is the largest cardinality of any compression set that the compression mapping produces on input of sets labeled according to concepts in C.

Almost thirty years ago, Littlestone and Warmuth (1986) asked whether there is any interesting function of the VC-dimension that upper-bounds the size of a smallest SCS. To date, there is no general answer to this question. Partial answers in the literature concern mostly the case of maximum concept classes and the case of concept classes of VC-dimension 1. A finite concept class

1. In the past, various forms of sample compression have been studied, but we focus only on the one described here.

is called maximum of VC-dimension d, if it has the largest possible size among all classes of VC-dimension d over the same instance space, i.e., if its size meets the well-known Sauer bound (Sauer, 1972; Welzl, 1987).[2] Floyd and Warmuth (1995) show that every maximum concept class of VC-dimension d has an SCS of size d. 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).

To the best of our knowledge, the notion of SCS has been studied exclusively for binary concept classes, i.e., subsets of the power set of $\{0,1\}^{|X|}$. This paper extends the study of sample compression to the multi-label case, where a concept may have a number of different labels in each instance. A concept class is then a subset 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 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, 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. 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 be also 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).

The results on which our study builds are those presented by Gurvits (1997), who studies a whole family of analogues of the VC-dimension. As proven by Ben-David et al. (1995), any notion of VC-dimension in this family that satisfies a very natural and simple condition to describe, provides a characterization of PAC-learnability of the class if and only if the VC-dimension is finite. Below, we define a particular notion of dimension, called the VCD_{Ψ^*}-dimension, which belongs to the family studied by Gurvits and Ben-David et al., and thus inherits the properties proven for all members of that family. Given a multi-label class C over $X = \{X_1,\ldots,X_m\}$, consider all m-tuples of mappings ψ_i that map the label set of X_i to $\{0,1\}$. Each such tuple of mappings, when applied to C, yields a binary concept class. The VCD_{Ψ^*}-dimension of C is then the maximum VC-dimension over the binary classes obtained from all such tuples of mappings. The VCD_{Ψ^*}-dimension of C provides an upper bound on all other VC-dimension notions studied in the literature and is hence the most promising version of VC-dimension for proving upper bounds on the size of SCSs in terms of VC-dimension. Gurvits uses linear algebraic methods for generalizing Sauer's bound to his VC-dimension analogues, and thus for the VCD_{Ψ^*}-dimension, so that the notion of

2. In this paper, we consider only finite concept classes, and so we always assume $|X|$ to be finite. However, our results on sample compression apply to the infinite case as well.

maximum class naturally extends to the VCD_{Ψ^*} case (we then use the term "VCD_{Ψ^*}-maximum class.")

The main contributions of our work are the following.

(A) We prove that every VCD_{Ψ^*}-maximum class has an SCS whose size equals its VCD_{Ψ^*}. The scheme and also parts of the proof follow the work on the so-called VC Compression Scheme for binary maximum classes, as introduced by Floyd and Warmuth (1995). However, there are some technical difficulties that need to be overcome in order to adapt Floyd and Warmuth's technique. The latter relies on the fact that every maximum class C of VC-dimension $d < |X|$, in the binary case, induces certain maximum classes over an instance space $X \setminus \{X_i\}$ for any $i \in \{1, \ldots, m\}$: (i) when projecting C onto $X \setminus \{X_i\}$, one obtains a maximum class of VC-dimension d, called the *restriction* of C to $X \setminus \{X_i\}$ and (ii) the set of all concepts c in this restriction for which both the concepts $c \cup \{(X_i, 0)\}$ and $c \cup \{(X_i, 1)\}$ are contained in C, called the *reduction* of C w.r.t. X_i, is maximum of VC-dimension $d-1$. For the multi-label case, Gurvits (1997) proves that the restriction of a VCD_{Ψ^*}-maximum class is again VCD_{Ψ^*}-maximum, but the literature offers no corresponding result for the reduction. In fact, it is not at all obvious how the reduction should even be defined in the multi-label case: should a concept c in the reduction w.r.t. X_i have all N_i+1 possible extensions contained in C (i.e., $c \cup \{(X_i, \ell)\} \in C$ for all $\ell \in \{0, \ldots, N_i\}$) or should we only require there to be at least two different extensions of c in C? A core result of our paper is that, for VCD_{Ψ^*}-maximum classes, we do not have to decide which definition of reduction to chose, since in such classes we obtain that c has either a unique extension to C or all $N_i + 1$ possible extensions to C—no other cases are possible. A large part of Section 4 is devoted to the proof of this technical result, which then allows us to use Floyd and Warmuth's technique.

(B) We show that every class of VCD_{Ψ^*} 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 VCD_{Ψ^*} 1 that is not contained in a VCD_{Ψ^*}-maximum class of VCD_{Ψ^*} 1 over the same instance space. Any such class cannot trivially inherit an SCS of size 1 from a VCD_{Ψ^*}-maximum class of VCD_{Ψ^*} 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 VCD_{Ψ^*} equals 1. This second major contribution of our work is presented in Section 5.

2. Preliminaries and notation

Let $\mathbb{N}^+$ denote the set of all positive natural numbers. For $m \in \mathbb{N}^+$, let $[m] = \{1, \ldots, m\}$ and $[0] = \emptyset$. Let $m \in \mathbb{N}^+$ and $N_i \in \mathbb{N}^+$ for $1 \leq i \leq m$. The finite set $X = \{X_1, \ldots, X_m\}$ is called an *instance space*, where each instance X_i is associated with the value set $X_i = \{0, \ldots, N_i\}$ for all $i \in [m]$. We call $c \in \prod_{i=1}^m X_i$ a *(multi-label) concept* on X, and a *(multi-label) concept class* C is then a set of concepts on X, that is, $C \subseteq \prod_{i=1}^m X_i$. For $c \in C$, let $c(X_i)$ denote the ith coordinate of c. In the rest of the paper, we will always implicitly assume that a given concept class C is a subset of $\prod_{i=1}^m X_i$ for some $m \in \mathbb{N}^+$, where each $X_i = \{0, \ldots, N_i\}$ for some $N_i \in \mathbb{N}^+$. When $N_i = 1$ for all $i \in [m]$, C is a *binary* concept class. In fact, a *binary* concept $c \in \{0,1\}^m$ is an m-dimensional binary vector, and hence a binary concept class is a subset of the m-dimensional vector space over the field $\mathbb{F}_2 = \{\mathbf{0}, \mathbf{1}\}$.

Table 1 (left) shows a concept class over $X = \{X_1, X_2, X_3\}$. In this class, $X_1 = X_2 = X_3 = \{0, 1, 2\}$, that is, $N_i = 2$ for all $1 \leq i \leq 3$. We will often use this matrix form to represent a concept class, i.e., a row corresponds to a concept, and a column corresponds to an instance in X.

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

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

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

Table 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 *.

A *labeled example* is a pair (X_t, ℓ), where $X_t \in X$ and $\ell \in \{0, \ldots, N_t\}$. A set of labeled examples is called a *sample*. For a sample S, we define $X(S) = \{X_i \in X \mid (X_i, \ell) \in S \text{ for some } \ell\}$.

For $Y = \{X_{i_1}, \ldots, X_{i_k}\} \subseteq X$ with $i_1 < i_2 < \cdots < i_k$, we denote the *restriction* of a concept c to Y by $c|_Y$ and define it as $c|_Y = (c(X_{i_1}), \ldots, c(X_{i_k}))$. Similarly, $C|_Y = \{c|_Y \mid c \in C\}$ denotes the restriction of C to Y. We 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., $C^{X_t} = \{c \in C - X_t \mid c \times \{0,1\} \subseteq C\}$. Note that C^{X_t} is a subset of $C - X_t$ that has the same reduced domain $X \setminus \{X_t\}$. For a class C over binary instances, Welzl (1987) show that for $X_i, X_j \in X$ with $X_i \neq X_j$, $(C^{X_i})^{X_j} = (C^{X_j})^{X_i}$. Consequently, for $Y = \{X_{i_1}, \ldots, X_{i_k}\} \subseteq X$, C^Y is defined as $((C^{X_{i_1}}) \cdots)^{X_{i_k}}$ (Welzl, 1987). It is not obvious how the definition of reduction should be extended to the multi-label case. One could consider concepts in $C - X_t$ that have at least two distinct extensions, or concepts in $C - X_t$ that have all $N_t + 1$ extensions to concepts in C. We will address this problem again in Section 4.

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

In the literature, a variety of ways to extend the notion of VC-dimension to the non-binary case have been studied (Alon, 1983; Natarajan, 1989; Vapnik, 1989; Pollard, 1990; Gurvits, 1997). We follow the framework proposed by Gurvits (1997), which generalizes over many of these notions.

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 $\mathrm{VCD}_\Psi(C)$ and define it by $\mathrm{VCD}_\Psi(C) = \max_{\overline{\psi} \in \Psi} \mathrm{VCD}(\overline{\psi}(C))$.*

By choosing special families of mappings from $\{0, \ldots, N_i\}$ to $\{0,1\}$, for all $i \in [m]$, we obtain different notions of dimension. For example, the Graph-dimension (Natarajan, 1989) equals $\mathrm{VCD}_{\Psi_{G_1} \times \cdots \times \Psi_{G_m}}$, where, for all i, $\Psi_{G_i} = \{\psi_{G,k} : 0 < k \leq n\}$ and $\psi_{G,k}(x) = 1$ if $x = k$, $\psi_{G,k}(x) = 0$ if $x \neq k$. That means, one considers all ways of mapping the values in each column to 1, if they equal some value k and to 0, if they differ from k. For each column, a different value of k may be used. The largest possible VC-dimension over the resulting binary classes is the Graph-dimension. For instance, 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. 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 when the value of k is set to 2 for both X_1 and X_3, while it is 0 for X_2.

We will focus mostly on the following version of VC-dimension for multi-label concept classes.

Definition 2 *Ψ^* denotes the family of all m-tuples $(\psi_1, \ldots, \psi_m)$ of mappings $\psi_i : X_i \to \{0,1\}$, and $\mathrm{VCD}_{\Psi^*}(C) = \max_{\overline{\psi} \in \Psi^*} \mathrm{VCD}(\overline{\psi}(C))$. C shatters a set $Y \subseteq X$ if there is some $\overline{\psi} \in \Psi^*$ such that the binary class $\overline{\psi}(C)$ shatters Y.*

For example, the concept class C on the left of Table 1 shatters $\{X_1, X_3\}$ and $\mathrm{VCD}_{\Psi^*}(C) = 2$.

For binary classes, the smallest possible *size of a sample compression scheme* yields sample bounds for PAC-learning (Littlestone and Warmuth, 1986; Floyd and Warmuth, 1995) and a big open question is whether this parameter can be upper-bounded by a function linear in the VC-dimension. The notion of sample compression generalizes to the multi-label case in a straightforward way.

Definition 3 *(Littlestone and Warmuth, 1986) A sample compression scheme for C is a pair (f, g) of mappings with the following properties. Given any sample S that is labeled consistently with some concept in C, 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 samples S consistent with some concept in C.*

Floyd and Warmuth (1995) prove that every *binary* maximum class C with $\mathrm{VCD}(C) = d$ (i.e., a class C with $|C| = \sum_{i=0}^{d} \binom{m}{i}$, which is the largest possible size according to Sauer (1972)) has a sample compression scheme of size $\mathrm{VCD}(C)$. Since every binary class of VCD 1 is contained in a maximum class of VCD 1 (Welzl and Woeginger, 1987), all classes C with $\mathrm{VCD}(C) = 1$ have a sample compression scheme of size 1.

3. A generalization of Sauer's bound

Let Ψ_i be a family of mappings $\psi_i : X_i \to \{0,1\}$. The statement "Ψ_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 $\psi_j = (\psi_j(0), \ldots, \psi_j(N_i))$ for $j \in [m]$. 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 $p = (p_1, \ldots, p_n) \in \mathbb{R}^{|C|}$ 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 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 $\mathbb{R}^{|C|}$.

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 $\mathbb{R}^{|C|}$.*

To make the proofs in the paper easier to follow, we make use of the following notation. We define $P^r(N_1, \ldots, N_m)$ to be the following collection of monomials with variables in X:

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

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

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

Theorem 6 *(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 $\mathbb{R}^{|C|}$.*

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

Corollary 7 (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^* is a spanning family, this bound applies also to VCD_{Ψ^*}, and the following general definition of maximum classes in particular defines the notion of VCD_{Ψ^*}-maximum class.

Definition 8 *Let Ψ_i, $1 \le i \le m$, be a family of mappings $\psi_i : X_i \to \{0,1\}$. Let $\Psi = \Psi_1 \times \cdots \times \Psi_m$. 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 VCD_{Ψ^*}-maximum multi-label class, namely the class of concepts that have at most $\mathrm{VCD}_{\Psi^*}(C)$ 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. The instance space X is a set of m lines in general position on the plane, i.e., no two lines are parallel and no three lines share a common point. 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). Then (i) the number of regions (which will be concepts with instance values -1 or $+1$) is $1+m+m(m-1)/2$; (ii) the number of segments and rays (which will be concepts with value 0 in one particular instance and values -1 or $+1$ in all other instances) is m^2; (iii) the number of intersection points (which will be concepts with value 0 on exactly two instances) is $m(m-1)/2$. The sum of these numbers is $1 + 2m^2 = \Phi_2(2,\ldots,2)$. A VCD_{Ψ^*}-maximum class of dimension 2 contains as concepts all regions, segments, rays and intersection points. One can verify that no set of three instances is shattered using any label mapping to a binary class.

In the next section, we will show that every VCD_{Ψ^*}-maximum class of dimension d has a sample compression scheme of size d.

All non-trivial claims made in this paper are proven either in the main body or in the Appendix. The Appendix also contains proofs of Theorems 5 and 6, translated into our notation.

4. Sample compression for VCD_{Ψ^*}-maximum classes

Let id_i denote the identity mapping on X_i. We will now show that for a VCD_Ψ-maximum class over a spanning family Ψ, if we only map one column to a set smaller than the original set of labels and keep the other columns unchanged, the resulting class is still maximum of the same dimension.

Lemma 9 *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 $\overline{\varphi_t} = (\mathrm{id}_1, \ldots, \mathrm{id}_{t-1}, \varphi_t, \mathrm{id}_{t+1}, \ldots, \mathrm{id}_m)$. Let $\Psi'_t = \{\mathrm{id}_t, 1 - \mathrm{id}_t\}$ [3] and $\Psi' = \Psi_1 \times \cdots \times \Psi_{t-1} \times \Psi'_t \times \Psi_{t+1} \cdots \times \Psi_m$. Then $\overline{\varphi_t}(C)$ is $\mathrm{VCD}_{\Psi'}$-maximum of dimension $\mathrm{VCD}_\Psi(C)$.*

Proof Let $d = \mathrm{VCD}_\Psi(C)$. W.l.o.g., let $t = 1$, i.e, $\varphi_1 : X_1 \to \{0, 1\}$ and $\overline{\varphi_1} = (\varphi_1, \mathrm{id}_2, \ldots, \mathrm{id}_m)$. Let $X'_1 = \varphi_1(X_1) = \{0, 1\}$ and $C' = \overline{\varphi_1}(C)$. Then, $\mathrm{VCD}_{\Psi'}(C') = \max_{\overline{\psi} \in \Psi'} \mathrm{VCD}(\overline{\psi}(C')) \leq \max_{\overline{\psi} \in \Psi} \mathrm{VCD}(\overline{\psi}(C)) = d$. Since $\mathrm{VCD}_{\Psi'}(C') \leq d$, the monomials from $P^d(1, N_2, \ldots, N_m)$ with variables in $\{X'_1, X_2, \ldots, X_m\}$ span $\mathbb{R}^{|C'|}$, by Theorem 6. If C' is not $\mathrm{VCD}_{\Psi'}$-maximum, then the monomials in $P^d(1, N_2 \ldots, N_m)$ must be linearly dependent. We will show that a linear dependency between the monomials in $P^d(1, N_2, \ldots, N_m)$ with variables in $\{X'_1, X_2, \ldots, X_m\}$ implies a linear dependency between the monomials in $P^d(N_1, \ldots, N_m)$ with variables in $\{X_1, \ldots, X_m\}$. This will contradict the assumption that C is VCD_Ψ-maximum because if $|C| = \Phi_d(N_1, \ldots, N_m)$ then the monomials from $P^d(N_1, \ldots, N_m)$ must be linearly independent.

Assume that 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'.

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, $a_{n_1} X_1^{n_1} X_{i_1}^{n_{i_1}} \cdots X_{i_t}^{n_{i_t}}$ 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. ■

Lemma 9 may be of interest beyond the study of VCD_{Ψ^*}, as it applies to a broad class of notions of VC-dimension. As an immediate corollary we obtain

Lemma 10 *Let C be VCD_{Ψ^*}-maximum and let $\varphi_t : X_t \to \{0, 1\}$ be a non-constant mapping, for some $t \in [m]$, and $\overline{\varphi_t} = (\mathrm{id}_1, \ldots, \mathrm{id}_{t-1}, \varphi_t, \mathrm{id}_{t+1}, \ldots, \mathrm{id}_m)$. Then $\overline{\varphi_t}(C)$ is VCD_{Ψ^*}-maximum of dimension $\mathrm{VCD}_{\Psi^*}(C)$.*

3. The mapping $1 - \mathrm{id}_t$ is only needed to make Ψ'_t a spanning family.

The proof of the following lemma is not quite as obvious and thus can be found in the Appendix.

Lemma 11 *Let C be a VCD_{Ψ^*}-maximum class and let $\overline{\varphi} = (\varphi_1, \ldots, \varphi_m)$ be a tuple of non-constant mappings such that each φ_i is either the identity mapping on X_i or $\varphi_i : X_i \to \{0,1\}$. Then $\overline{\varphi}(C)$ is also a VCD_{Ψ^*}-maximum class of dimension $\mathrm{VCD}_{\Psi^*}(C)$.*

It is obvious that if one of the φ_i's is a constant mapping, then $\overline{\varphi}(C)$ is not maximum because it contains a constant column of 0s or 1s. Thus we obtain the following corollary.

Corollary 12 *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 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 13 *(Gurvits, 1997) Let Ψ_i, $1 \leq i \leq m$, be a spanning family of mappings $\psi_i : X_i \to \{0,1\}$, and $\Psi = \Psi_1 \times \cdots \times \Psi_m$. Let C be VCD_{Ψ}-maximum with $\mathrm{VCD}_{\Psi}(C) = d$, and $Y \subseteq X$ with $|Y| \geq d$. Then $C|_Y$ is VCD_{Ψ}-maximum with $\mathrm{VCD}_{\Psi}(C|_Y) = d$.*

One of our core results is that a reduction of a VCD_{Ψ^*}-maximum class is also VCD_{Ψ^*}-maximum. Before proving this claim, we show that for any VCD_{Ψ^*}-maximum class C, each concept $c \in C - X_t$, for all $t \in [m]$, either has a unique extension in C or has all possible extensions in C.

Lemma 14 *Let C be a VCD_{Ψ^*}-maximum class with $\mathrm{VCD}_{\Psi^*}(C) = m - 1$. Let $t \in [m]$ and $\overline{c} \in C - X_t$. Then $|\{c \in C \mid c - X_t = \overline{c}\}| \in \{1, N_t + 1\}$.*

Proof It suffices to prove the claim for $t = m$. For $N_m = 1$ there is nothing to show. Thus let $N_m \geq 2$. We show that if there is a concept $\overline{c} \in C - X_m$ that has more than one but fewer than $N_m + 1$ extensions to concepts in C, then C is not VCD_{Ψ^*}-maximum. The idea is to show that there is a tuple of non-constant mappings that transform C to a class that is not VCD_{Ψ^*}-maximum of dimension $m - 1$. Then, by Lemma 11, C is not VCD_{Ψ^*}-maximum.

Let $1 \leq k < N_m$. Suppose there is some $\overline{c} \in C - X_m$ such that $|\{c \in C \mid c - X_m = \overline{c}\}| = k+1$. Let $c_0, \ldots, c_k \in C$ such that $c_i \neq c_j$ and $c_i - X_m = c_j - X_m = \overline{c}$, for all $i, j \in \{0, \ldots, k\}$ with $i \neq j$. W.l.o.g., let $c_i(X_m) = i$ for $i \in \{0, \ldots, k\}$. Further, for $i \in [m-1]$, let $\overline{c}(X_i) = l_i$.

Since $k < N_m$, we have $k + 1 \in X_m \setminus \{0, \ldots, k\}$. Let $c_{\mathrm{new}} = \overline{c} \cup \{(X_m, k+1)\}$ and $C_{\mathrm{new}} = C \cup \{c_{\mathrm{new}}\}$. C is VCD_{Ψ^*}-maximum of dimension $m - 1$, so C_{new} shatters X. Thus, there is a tuple $\overline{\psi} = (\psi_1, \ldots, \psi_m)$ of mappings, where $\psi_i : X_i \to \{0,1\}$ for all $i \in [m]$, and $\overline{\psi}(C_{\mathrm{new}}) = \{0,1\}^m$.

Note that $c_{\mathrm{new}} - X_m = \overline{c}$ and thus $\overline{\psi}(c_{\mathrm{new}})|_{\{X_1, \ldots, X_{m-1}\}} = \overline{\psi}(c_i)|_{\{X_1, \ldots, X_{m-1}\}}$ for all $i \in \{0, \ldots, k\}$. Let $\overline{\psi}(c_{\mathrm{new}})|_{\{X_1, \ldots, X_{m-1}\}} = (p_1, \ldots, p_{m-1})$, where $p_i \in \{0,1\}$ for all $i \in [m-1]$. If $\psi_m(k+1) = \psi_m(i)$ for some $i \in \{0, \ldots, k\}$, then $\overline{\psi}(c_{\mathrm{new}}) = \overline{\psi}(c_i)$ and consequently, $\overline{\psi}(C_{\mathrm{new}}) = \overline{\psi}(C)$. This means that X is shattered by C which is not possible, because $\mathrm{VCD}_{\Psi^*}(C) = m - 1$. So, $\psi_m(k+1) \neq \psi_m(i)$, for all $i \in \{0, \ldots, k\}$. W.l.o.g., assume that $\psi_m(k+1) = 1$ and $\psi_m(i) = 0$ for all $i \in \{0, \ldots, k\}$. So,

$$\psi_m(x) = \begin{cases} 1 & \text{if } x = k+1 \\ 0 & \text{if } x \in \{0, \ldots, k\} \\ 0 \text{ or } 1 & \text{if } x \in X_m \setminus \{0, \ldots, k, k+1\}. \end{cases}$$

and $\overline{\psi}(c_{\text{new}}) = (p_1, \ldots, p_{m-1}, 1)$. Consequently, $\overline{\psi}(C) = \{0,1\}^m \setminus \{(p_1, \ldots, p_{m-1}, 1)\}$.

We show that changing ψ_m to ψ'_m as follows will not affect $\overline{\psi}(C_{\text{new}})$ or $\overline{\psi}(C)$. Let

$$\psi'_m(x) = \begin{cases} 1 & \text{if } x = k+1 \\ 0 & \text{if } x \in X_m \setminus \{k+1\}. \end{cases}$$

Let $\overline{\psi'} = (\psi_1, \ldots, \psi_{m-1}, \psi'_m)$. We claim that $\overline{\psi'}(C) = \overline{\psi}(C) = \{0,1\}^m \setminus \{(p_1, \ldots, p_{m-1}, 1)\}$. It is obvious that, for any $q \in X_m$, $\psi'_m(q) = 1$ implies $\psi_m(q) = 1$. Thus $(p_1, \ldots, p_{m-1}, 1) \notin \overline{\psi'}(C)$ follows from $(p_1, \ldots, p_{m-1}, 1) \notin \overline{\psi}(C)$. By Corollary 12, $\overline{\psi'}(C)$ is a VCD-maximum class with $\text{VCD}(\overline{\psi'}(C)) = m-1$. So, $|\overline{\psi'}(C)| = 2^m - 1$ and consequently, $\overline{\psi'}(C) = \overline{\psi}(C) = \{0,1\}^m \setminus \{(p_1, \ldots, p_{m-1}, 1)\}$. Thus, we fix the tuple of mappings $\overline{\psi} = (\psi_1, \ldots, \psi_m)$ for the rest of the proof and we can choose ψ_m to be

$$\psi_m(x) = \begin{cases} 1 & \text{if } x = k+1 \\ 0 & \text{if } x \in X_m \setminus \{k+1\}. \end{cases}$$

and we still have

$$\overline{\psi}(C) = \{0,1\}^m \setminus \{(p_1, \ldots, p_{m-1}, 1)\}. \tag{1}$$

In particular, for any concept $c \in C$, we obtain

$$\text{if } \overline{\psi}(c)|_{\{X_1,\ldots,X_{m-1}\}} = (p_1, \ldots, p_{m-1}) \text{ then } c(X_m) \neq k+1\,. \tag{2}$$

We define $\overline{\psi^0} = (\psi_1, \ldots, \psi_{m-1}, \psi^0_m)$ and $\overline{\psi^1} = (\psi_1, \ldots, \psi_{m-1}, \psi^1_m)$ with

$$\psi^0_m(x) = \begin{cases} 1 & \text{if } x \in \{0, k+1\} \\ 0 & \text{if } x \in X_m \setminus \{0, k+1\} \end{cases} \quad \text{and } \psi^1_m(x) = \begin{cases} 1 & \text{if } x \in \{1, k+1\} \\ 0 & \text{if } x \in X_m \setminus \{1, k+1\}. \end{cases}$$

Claim. (See Appendix for proof of Claim.)

1. $\overline{\psi^0}(C) = \{0,1\}^m \setminus \{(r_1, \ldots, r_{m-1}, 0)\}$, for some $(r_1, \ldots, r_{m-1}) \in \{0,1\}^{m-1}$ satisfying $(p_1, \ldots, p_{m-1}) \neq (r_1, \ldots, r_{m-1})$.
2. $\overline{\psi^1}(C) = \{0,1\}^m \setminus \{(s_1, \ldots, s_{m-1}, 0)\}$, for some $(s_1, \ldots, s_{m-1}) \in \{0,1\}^{m-1}$ satisfying $(p_1, \ldots, p_{m-1}) \neq (s_1, \ldots, s_{m-1})$.
3. $(r_1, \ldots, r_{m-1}) \neq (s_1, \ldots, s_{m-1})$ for the $(r_1, \ldots, r_{m-1})$, $(s_1, \ldots, s_{m-1})$ as in the above two statements.

To finish the proof we need to show that there is a tuple of non-constant mappings that maps C to a class that is not VCD_{Ψ^*}-maximum of dimension $m-1$. For this purpose, we change ψ_m in $\overline{\psi}$ to be the identity mapping on X_m and define a new tuple of mappings $\overline{\psi''} = (\psi_1, \ldots, \psi_{m-1}, \text{id}_m)$. Note that

$$\overline{\psi''}(c)|_{\{X_1,\ldots,X_{m-1}\}} = \overline{\psi}(c)|_{\{X_1,\ldots,X_{m-1}\}} \text{ for all } c \in C. \tag{3}$$

Let $C'' = \overline{\psi''}(C)$. We show that C'' is not VCD_{Ψ^*}-maximum of dimension $m-1$. Assume that C'' is VCD_{Ψ^*}-maximum with $\text{VCD}_{\Psi^*}(C'') = m-1$. By Theorem 13, $C''|_{\{X_1,\ldots,X_{m-1}\}}$ is also VCD_{Ψ^*}-maximum with $\text{VCD}_{\Psi^*}(C''|_{\{X_1,\ldots,X_{m-1}\}}) = m-1$. Thus $C''|_{\{X_1,\ldots,X_{m-1}\}} = \{0,1\}^{m-1}$ and, in particular, $\{(p_1, \ldots, p_{m-1}), (r_1, \ldots, r_{m-1}), (s_1, \ldots, s_{m-1})\} \subseteq C''|_{\{X_1,\ldots,X_{m-1}\}}$.

Since C'' is maximum of VCD_{Ψ^*} $m-1$, by Corollary 7, $|C''| = \Phi_{m-1}(\overbrace{1,\ldots,1}^{m-1}, N_m)$.

Now, we count the maximum number of concepts that can exist in C''. First, note that (2) and (3) imply that for any concept $c \in C''$ with $c|_{\{X_1,\ldots,X_{m-1}\}} = (p_1,\ldots,p_{m-1})$, $c(X_m) \neq k+1$. Thus $(p_1,\ldots,p_{m-1})$ has at most N_m extensions in C''. Second, note that $(r_1,\ldots,r_{m-1})$ and $(s_1,\ldots,s_{m-1})$ each have at most 2 extensions in C'' (namely extensions with 0 and $k+1$, and extensions with 1 and $k+1$, respectively), because otherwise $(r_1,\ldots,r_m,0) \in \overline{\psi^0}(C)$ or $(s_1,\ldots,s_m,0) \in \overline{\psi^1}(C)$, which contradicts Claims 1 or 2, respectively. Third, by Claims 1, 2, and 3, there are exactly $2^{m-1}-3$ binary vectors remaining in $\{0,1\}^{m-1}$. Each one of these has at most N_m+1 extensions in C''. In total, $|C''| \leq (2^{m-1}-3)\times(N_m+1)+N_m+2+2 < \Phi_{m-1}(1,\ldots,1,N_m)$ (see Appendix for derivation).

Hence $C'' = \overline{\psi''}(C)$ is not VCD_{Ψ^*}-maximum of dimension $m-1$. Therefore, by Lemma 11, the class C is not VCD_{Ψ^*}-maximum either. ■

We now generalize Lemma 14 as follows.

Theorem 15 *Let C be a VCD_{Ψ^*}-maximum class with $\text{VCD}_{\Psi^*}(C) = d$. Let $t \in [m]$ and $\overline{c} \in C - X_t$. Then $|\{c \in C \mid c - X_t = \overline{c}\}| \in \{1, N_t+1\}$.*

Proof Note that, by definition, $m \geq d$.

For $m = d$, we obtain $\text{VCD}_{\Psi^*}(C) = m$ 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 14 and is thus proven. So suppose $m > d+1$.

Consider a VCD_{Ψ^*}-maximum class $C \subseteq \prod_{i=1}^{m} X_i$ with $\text{VCD}_{\Psi^*}(C) = d$. 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 $c_{\text{new}} = \overline{c} \cup \{(X_1, k+1)\}$ and $C_{\text{new}} = C \cup \{c_{\text{new}}\}$. C is VCD_{Ψ^*}-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 $\overline{\psi}(C_{\text{new}})|_{\{X_1,\ldots,X_{d+1}\}} = \{0,1\}^{d+1}$.

We show that $\{X_1,\ldots,X_{d+1}\}$ is shattered by C, too. By Theorem 13, $C|_{\{X_1,\ldots,X_{d+1}\}}$ is VCD_{Ψ^*}-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 14, $c_i|_{\{X_2,\ldots,X_{d+1}\}}$ has either a unique or all extensions to concepts in $C|_{\{X_1,\ldots,X_{d+1}\}}$. Since $\overline{c}$ has more than one extension to concepts in C, we obtain that $\overline{c}|_{\{X_2,\ldots,X_{d+1}\}}$ has more than one extension—and thus all possible extensions—to concepts in $C|_{\{X_1,\ldots,X_{d+1}\}}$. In particular, there is a concept $c' \in C|_{\{X_1,\ldots,X_{d+1}\}}$, such that $c'|_{\{X_2,\ldots,X_{d+1}\}} = \overline{c}|_{\{X_2,\ldots,X_{d+1}\}}$, and $c'(X_1) = k+1$. Equivalently, $c_{\text{new}}|_{\{X_1,\ldots,X_{d+1}\}} \in C|_{\{X_1,\ldots,X_{d+1}\}}$, and thus $C|_{\{X_1,\ldots,X_{d+1}\}} = C_{\text{new}}|_{\{X_1,\ldots,X_{d+1}\}}$. Hence, $\overline{\psi}(C|_{\{X_1,\ldots,X_{d+1}\}}) = \overline{\psi}(C_{\text{new}}|_{\{X_1,\ldots,X_{d+1}\}}) = \{0,1\}^{d+1}$ and C shatters $\{X_1,\ldots,X_{d+1}\}$ in contradiction to $\text{VCD}_{\Psi^*}(C) = d$. ■

Hence, for a VCD_{Ψ^*}-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. Formally, we define $C^{X_t} = \{c \in C - X_t \mid c \times X_t \subseteq C\}$. Then the reduction of a VCD_{Ψ^*}-maximum class is also VCD_{Ψ^*}-maximum:

Theorem 16 *Let C be a VCD_{Ψ^*}-maximum class with $\mathrm{VCD}_{\Psi^*}(C) = d$. Then C^{X_t} is VCD_{Ψ^*}-maximum with $\mathrm{VCD}_{\Psi^*}(C^{X_t}) = d - 1$, for any $t \in [m]$.*

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, as $(C^{X_i})^{X_j} = (C^{X_j})^{X_i}$ for all $i, j \in [m]$, as in the binary case.

Proposition 17 *For any X_i, X_j with $i \neq j$, $(C^{X_i})^{X_j} = (C^{X_j})^{X_i}$.*

We now closely follow the technique that Floyd and Warmuth (1995) use to show that any binary VCD-maximum class has a sample compression scheme of the size of its VC-dimension.

Corollary 18 *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.*

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. In particular, for any tuple $(n_{i_1}, \ldots, n_{i_d}) \in \prod_{j=1}^{d} X_{i_j}$, $c_{Y,C} \cup (n_{i_1}, \ldots, n_{i_d}) \in C$. 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 19 *Let C be a VCD_{Ψ^*}-maximum class with $\mathrm{VCD}_{\Psi^*}(C) = d < m$. Let S with $|S| = d$ be a sample consistent with some concepts in C 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.*

In order to have a compression scheme of size d, any sample of size at least $d+1$ consistent with some concepts in C 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, $C|_Y$ is VCD_{Ψ^*}-maximum and Definition 19 applies to $C|_Y$, too.

Theorem 20 *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 $c = c_{S,C}$.*

Corollary 21 *Let C be a VCD_{Ψ^*}-maximum class with $\mathrm{VCD}_{\Psi^*}(C) = d$. Then C has a sample compression scheme of size d.*

An inspection of the proof will show that Corollary 21 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.

5. Sample compression for classes of VCD_{Ψ^*} 1

For binary concept classes, compression schemes of size d for maximum classes of VC-dimension d, like the VC Scheme proposed by Floyd and Warmuth (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.

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 Corollary 21. However, this approach is fruitless, since there is a VCD_{Ψ^*}-maximal class C such that $\mathrm{VCD}_{\Psi^*}(C) = 1$ and C is not VCD_{Ψ^*}-maximum. As an example, consider the class $\hat{C} \subset \{0,1,2\} \times \{0,1,2\}$ given by $\hat{C} = \{(0,0),(1,1),(2,2)\}$. Clearly, $\mathrm{VCD}_{\Psi^*}(\hat{C}) = 1$ and $\hat{C}$ is too small to be VCD_{Ψ^*}-maximum. However, it is VCD_{Ψ^*}-maximal.

It is easy to see that, considering the family of mappings $\Psi_{G_1} \times \ldots \times \Psi_{G_m}$ used for computing the Graph-dimension (see Section 2), the class $\hat{C}$ is still $\mathrm{VCD}_{\Psi_{G_1} \times \ldots \times \Psi_{G_m}}$-maximal. So, even when restricting ourselves to some special families of mappings studied in the literature previously, classes of dimension 1 do not necessarily maintain the same structural properties as in the binary case. We will prove that, despite the changes in structural properties when compared to the binary case, every multi-label class C with $\mathrm{VCD}_{\Psi^*}(C) = 1$ has a sample compression scheme of size 1.

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).

Lemma 22 *Let* $\mathrm{VCD}_{\Psi^*}(C) = 1$. *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.* $C|_{\{X_i, X_j\}}$.

This result does not generalize to the case when $\mathrm{VCD}_{\Psi^*}(C) = 2$, not even for binary classes. For example, the VCD-maximum class of VC-dimension 2 over 3 instances that corresponds to the class of all sets of size at most 2 has 4 concepts in $C|_{\{X_1, X_2, X_3\}} = C$ of teaching dimension 3, namely the empty concept $(0,0,0)$ and the singletons $(1,0,0)$, $(0,1,0)$, and $(0,0,1)$. For $\mathrm{VCD}_{\Psi^*}(C) = 1$, Lemma 22 will help us to compress a set of two examples to one example.

Definition 23 *Let* C *be a concept class and let* S *be a sample consistent with some concept in* C. *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).

Using the above definition, we obtain a simple lemma.

Lemma 24 *Let* S *be a sample consistent with some concept in* C *and* $(X_i, l_i), (X_j, l_j) \in S$, *such that* (X_i, l_i) *implies* (X_j, l_j). *If* $\mathrm{VCD}_{\Psi^*}(C) = 1$ *then* (X_i, l_i) *uniquely implies* (X_j, l_j).

Corollary 25 *Let S be a sample consistent with some concept in C and $(X_i, l_i), (X_j, l_j) \in S$. If $\mathrm{VCD}_{\Psi^*}(C) = 1$ then at least one of the following statements is true: (i) (X_i, l_i) explicitly implies (X_j, l_j), (ii) (X_j, l_j) explicitly implies (X_i, l_i), (iii) (X_i, l_i) implicitly implies (X_j, l_j) and (X_j, l_j) implicitly implies (X_i, l_i).*

Example 1 *Consider the class in Table 2. For $S = \{(X_1, 0), (X_2, 0), (X_3, 0)\}$, $(X_2, 0)$ explicitly implies $(X_1, 0)$ and implicitly implies $(X_3, 0)$. For $S' = \{(X_1, 1), (X_2, 1), (X_3, 0)\}$, $(X_1, 1)$ explicitly implies $(X_2, 1)$; $(X_2, 1)$ explicitly implies $(X_3, 0)$; $(X_1, 1)$ explicitly implies $(X_3, 0)$.*

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".

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

Table 2: Concept class used in Example 1.

Lemma 26 *Let $\mathrm{VCD}_{\Psi^*}(C) = 1$, and let S be a sample consistent with some concept in C 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.*

Theorem 27 *Let $\mathrm{VCD}_{\Psi^*}(C) = 1$. Then any sample S consistent with some concept in C has a representative.*

Proof For $|S| = 1$, there is nothing to show, and for $|S| = 2$, Corollary 25 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 25, 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 26, e_i explicitly and thus uniquely implies e_j. If r implicitly implies e_j, then by Lemma 26, 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 $\{e_1, \ldots, e_i\}$. ■

Corollary 28 *Let $\mathrm{VCD}_{\Psi^*}(C) = 1$. Then C has a sample compression scheme of size 1.*

For example, consider the class in Table 2. Decompression of the set $\{(X_2, 1)\}$ here would yield c_3, since $(X_2, 1)$ explicitly implies $(X_3, 0)$ and implicitly implies $(X_1, 0)$.

The assumption that X is finite is not used in the proof of Corollary 28, so that the latter applies also to infinite concept classes of VCD_{Ψ^*}-dimension 1. Further, all label mappings used to verify Corollary 28 are of the form used for defining the Graph-dimension, that is, for each instance exactly one value is mapped to 1. Hence all results in Section 5 apply also to classes of Graph-dimension 1. In particular, every class of Graph-dimension 1 has a compression scheme of size 1.

Acknowledgments

We thank the anonymous reviewers for insightful questions and comments. We also acknowledge financial support from the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

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

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. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36(4):929–965, 1989.

A. Daniely, S. Sabato, S. Ben-David, and S. Shalev-Shwartz. Multiclass learnability and the ERM principle. In *COLT*, volume 19 of *JMLR Proceedings*, pages 207–232. JMLR.org, 2011.

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. *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.

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

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.

N. Sauer. On the density of families of sets. *J. Comb. Theory, Ser. 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, San Francisco, CA, USA, 1989. Morgan Kaufmann Publishers Inc.

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

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

Appendix A. Proofs of Results Proven by Gurvits (1997)

For convenience, we include proofs of some of Gurvits's results, translated into our notation.

To prove Theorem 6, Gurvits first observes the following.

Lemma 29 *(Gurvits, 1997) Let Ψ_i be a spanning family of mappings on X_i for all $i \in [m]$ and $\Psi = \Psi_1 \times \cdots \times \Psi_m$. Then Ψ is spanning on $X_1 \times \cdots \times X_m$, where $(\psi_1, \psi_2, \ldots, \psi_m)(x_1, x_2, \ldots, x_m) = \psi_1(x_1) \cdot \psi_2(x_2) \cdot \ldots \cdot \psi_m(x_m)$ for $(\psi_1, \psi_2, \ldots, \psi_m) \in \Psi$ and $(x_1, x_2, \ldots, x_m) \in X_1 \times \cdots \times X_m$.*

Theorem 6 *(Gurvits, 1997) Let Ψ_i, $1 \leq i \leq m$, be a spanning family of mappings $\psi_i : X_i \to \{0, 1\}$, and $\Psi = \Psi_1 \times \cdots \times \Psi_m$. If $\mathrm{VCD}_\Psi(C) = d$ then the monomials from $P^d(N_1, \ldots, N_m)$ span the vector space $\mathbb{R}^{|C|}$.*

Proof We show that any vector from $\mathbb{R}^{|C|}$ can be expressed as a linear combination of monomials from $P^d(N_1, \ldots, N_m)$.

By Lemma 29, we know that if Ψ_i is spanning on X_i, then Π_Ψ is spanning on $X_1 \times \cdots \times X_m$. In particular, any vector from $\mathbb{R}^{|C|}$ 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, $\mathrm{VCD}(C') \leq \mathrm{VCD}_\Psi(C) = d$. By Theorem 5, the monomial $X'_1 \cdot \ldots \cdot X'_m$ can be expressed as a linear combination of short products

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

It follows that $\psi_1(X_1) \cdot \ldots \cdot \psi_m(X_m)$ can be expressed as a linear combination of short products $\{\psi_{i_1}(X_{i_1}) \cdot \ldots \cdot \psi_{i_k}(X_{i_k}) : k \leq d\}$.

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

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

So, any 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 $\mathbb{R}^{|C|}$. ■

Theorem 13 *(Gurvits, 1997) Let Ψ_i, $1 \leq i \leq m$, be a spanning family of mappings $\psi_i : X_i \to \{0,1\}$, and $\Psi = \Psi_1 \times \cdots \times \Psi_m$. Let C be VCD_Ψ-maximum with $\mathrm{VCD}_\Psi(C) = d$, and $Y \subseteq X$ with $|Y| \geq d$. Then $C|_Y$ is VCD_Ψ-maximum with $\mathrm{VCD}_\Psi(C|_Y) = d$.*

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

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

On the one hand, $\mathrm{VCD}_\Psi(C|_Y) \leq d$, and by Theorem 6, the monomials from $P^d(N_{i_1}, \ldots, N_{i_k})$ span the vector space $\mathbb{R}^{\mathrm{size}(C|_Y)}$. So, $\mathrm{size}(C|_Y) \leq |P^d(N_{i_1}, \ldots, N_{i_k})| = \Phi_d(N_{i_1}, \ldots, N_{i_k})$. Hence, $\mathrm{size}(C|_Y) = \Phi_d(N_{i_1}, \ldots, N_{i_k})$. Considering the size of $C|_Y$, $\mathrm{VCD}_\Psi(C|_Y)$ cannot be smaller than d. Hence $C|_Y$ is a VCD_Ψ-maximum class of dimension $d = \mathrm{VCD}_\Psi(C)$. ■

Appendix B. Proofs Omitted From Section 4

B.1. Proof of Lemma 11

Lemma 11 *Let C be a VCD_{Ψ^*}-maximum class and let $\overline{\varphi} = (\varphi_1, \ldots, \varphi_m)$ be a tuple of non-constant mappings such that each φ_i is either the identity mapping on X_i or $\varphi_i : X_i \to \{0,1\}$. Then $\overline{\varphi}(C)$ is also a VCD_{Ψ^*}-maximum class of dimension $\mathrm{VCD}_{\Psi^*}(C)$.*

Proof W.l.o.g., let $\varphi_i : X_i \to \{0,1\}$, for all $1 \leq i \leq k$ 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. ■

B.2. Missing Parts From the Proof of Lemma 14

Claim.

1. $\overline{\psi^0}(C) = \{0,1\}^m \setminus \{(r_1, \ldots, r_{m-1}, 0)\}$, for some $(r_1, \ldots, r_{m-1}) \in \{0,1\}^{m-1}$ satisfying $(p_1, \ldots, p_{m-1}) \neq (r_1, \ldots, r_{m-1})$.
2. $\overline{\psi^1}(C) = \{0,1\}^m \setminus \{(s_1, \ldots, s_{m-1}, 0)\}$, for some $(s_1, \ldots, s_{m-1}) \in \{0,1\}^{m-1}$ satisfying $(p_1, \ldots, p_{m-1}) \neq (s_1, \ldots, s_{m-1})$.
3. $(r_1, \ldots, r_{m-1}) \neq (s_1, \ldots, s_{m-1})$ for the $(r_1, \ldots, r_{m-1})$, $(s_1, \ldots, s_{m-1})$ as in the above two statements.

Proof of Claim 1. It is clear that ψ^0_m is a non-constant mapping. Note that $\overline{\psi}$ and $\overline{\psi^0}$ differ only in the mth mapping. Thus,

$$\overline{\psi^0}(c)|_{\{X_1, \ldots, X_{m-1}\}} = \overline{\psi}(c)|_{\{X_1, \ldots, X_{m-1}\}} \text{ for all } c \in C. \tag{4}$$

By Corollary 12, $\overline{\psi^0}(C)$ is VCD-maximum with $\mathrm{VCD}(\overline{\psi^0}(C)) = m-1$. So, $|\overline{\psi^0}(C)| = 2^m - 1$.

Assume $\overline{\psi^0}(C) = \{0,1\}^m \setminus \{(r_1,\ldots,r_{m-1},1)\}$, where $r_i \in \{0,1\}$ for all $i \in [m-1]$. We show that $(r_1,\ldots,r_{m-1},1) \neq (p_1,\ldots,p_{m-1},1)$ and also $(r_1,\ldots,r_{m-1},1) \notin \overline{\psi}(C)$. That is, $|\overline{\psi}(C)| \leq 2^m - 2$ which is a contradiction.

First, it is obvious that $\overline{\psi^0}(c_0) = \overline{\psi^0}(c_{\mathrm{new}}) = (p_1,\ldots,p_{m-1},1)$, and thus $(p_1,\ldots,p_{m-1},1) \in \overline{\psi^0}(C)$. So,

$$(r_1,\ldots,r_{m-1},1) \neq (p_1,\ldots,p_{m-1},1). \tag{5}$$

Second, $\psi_m(x) = 1$ implies $x = k+1$ and thus also $\psi_m^0(x) = 1$. Having this and (4), we conclude that $(r_1,\ldots,r_{m-1},1) \in \overline{\psi}(C)$ implies $(r_1,\ldots,r_{m-1},1) \in \overline{\psi^0}(C)$. So,

$$(r_1,\ldots,r_{m-1},1) \notin \overline{\psi}(C). \tag{6}$$

From (5) and (6) we conclude that $\overline{\psi}(C) \subseteq \{0,1\}^m \setminus \{(p_1,\ldots,p_{m-1},1),(r_1,\ldots,r_{m-1},1)\}$ which contradicts (1). Hence, our initial assumption is false, so that we obtain $\overline{\psi^0}(C) = \{0,1\}^m \setminus \{(r_1,\ldots,r_{m-1},0)\}$, where $r_i \in \{0,1\}$ for all $i \in [m-1]$. So, for any concept $c \in C$, $\overline{\psi^0}(c) \neq (r_1,\ldots,r_{m-1},0)$.

To establish Claim 1, it remains to show that $(p_1,\ldots,p_{m-1}) \neq (r_1,\ldots,r_{m-1})$. By the definition of ψ_m^0, for any concept $c \in C$ with $\overline{\psi^0}(c)|_{\{X_1,\ldots,X_{m-1}\}} = (r_1,\ldots,r_{m-1})$, either $c(X_m) = 0$ or $c(X_m) = k+1$. From (4) we then have, for any concept $c \in C$ with $\overline{\psi}(c)|_{\{X_1,\ldots,X_{m-1}\}} = (r_1,\ldots,r_{m-1})$, either $c(X_m) = 0$ or $c(X_m) = k+1$. Since $c_1(X_m) = 1$, we conclude that $\overline{\psi}(c_1)|_{\{X_1,\ldots,X_{m-1}\}} \neq (r_1,\ldots,r_{m-1})$. Hence, $(p_1,\ldots,p_{m-1}) \neq (r_1,\ldots,r_{m-1})$. ■ (Claim 1.)

Proof of Claim 2. By the same arguments as used for establishing Claim 1. ■ (Claim 2.)

Proof of Claim 3. Equation (1) implies $(r_1,\ldots,r_{m-1},0) \in \overline{\psi}(C)$. So, there is some $c' \in C$ with $\overline{\psi}(c') = (r_1,\ldots,r_{m-1},0)$, i.e., $\overline{\psi}(c')|_{\{X_1,\ldots,X_{m-1}\}} = (r_1,\ldots,r_{m-1})$ and $\psi_m(c'(X_m)) = 0$. Also, as shown in the proof of Claim 1, for any $c \in C$ with $\overline{\psi}(c)|_{\{X_1,\ldots,X_{m-1}\}} = (r_1,\ldots,r_{m-1})$, either $c(X_m) = 0$ or $c(X_m) = k+1$. So, $c'(X_m) = 0$ since $\psi_m(k+1) = 1$. Thus, $\psi_m^1(c'(X_m)) = \psi_m^1(0) = 0$. Also, from the analogue of (4) for $\overline{\psi^1}$, $\overline{\psi}(c')|_{\{X_1,\ldots,X_{m-1}\}} = \overline{\psi^1}(c')|_{\{X_1,\ldots,X_{m-1}\}}$. Hence $(r_1,\ldots,r_{m-1},0) \in \overline{\psi^1}(C)$. Therefore, $(r_1,\ldots,r_{m-1}) = (s_1,\ldots,s_{m-1})$ would imply $(s_1,\ldots,s_{m-1},0) \in \overline{\psi^1}(C)$ which contradicts Claim 2. Consequently, we obtain $(s_1,\ldots,s_{m-1}) \neq (r_1,\ldots,r_{m-1})$. ■ (Claim 3.)

Details for the derivation of $|C''| < \Phi_{m-1}(1,\ldots,1,N_m)$ near the end of the proof of Lemma 14:

$$\begin{aligned}
|C''| &\leq (2^{m-1}-3)\times(N_m+1) + N_m + 2 + 2 \\
&= 2^{m-1}N_m - 3N_m - 3 + 2^{m-1} + N_m + 4 = 2^{m-1} + 2^{m-1}N_m - 2N_m + 1 \\
&< 2^{m-1} + 2^{m-1}N_m - N_m \ \ (\text{since } N_m \geq 2) \\
&= \Phi_{m-1}(1,\ldots,1,N_m).
\end{aligned}$$

B.3. Proof of Results Concerning Reductions of Maximum Classes

Proposition 17 *For any X_i, X_j with $i \neq j$, $(C^{X_i})^{X_j} = (C^{X_j})^{X_i}$.*

Proof

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

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

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

■

Theorem 16 *Let C be a VCD_{Ψ^*}-maximum class with $\mathrm{VCD}_{\Psi^*}(C) = d$. 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 a mapping such that

$$\psi_m(x) = \begin{cases} 0 & \text{if } x = 0 \\ 1 & \text{if } x \in X_m \setminus \{0\} \end{cases}$$

and $\overline{\psi^{1,m}} = (\psi_1, \ldots, \psi_{m-1}, \psi_m)$. By the definition of reduction, 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 Theorem 15, 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 13, $C - X_m$ 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 $\mathrm{VCD}_{\Psi^*}(C^{X_t}) \leq d-1$ and $|C^{X_t}| = \Phi_{d-1}(N_1, \ldots, N_{m-1})$, the reduction C^{X_m} is VCD_{Ψ^*}-maximum with $\mathrm{VCD}_{\Psi^*}(C^{X_m}) = d - 1$. ■

Corollary 18 *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 Let $Y = \{X_{i_1}, \ldots, X_{i_d}\}$. By applying Theorem 16 to $C^Y = ((C^{X_{i_1}}) \cdots)^{X_{i_d}}$ repeatedly, C^Y is a VCD_{Ψ^*}-maximum class of dimension 0. So, $|C^Y| = 1$. ■

B.4. Proof of Theorem 20 and Corollary 21

To prove Theorem 20, we need two lemmas. We first have to 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 30 *Let C be a VCD_{Ψ^*}-maximum class with $\mathrm{VCD}_{\Psi^*}(C) = d < m$. Let S be a sample consistent with some concepts in C, with $X(S) \subseteq Y \subseteq X$, and $|X(S)| = d$. Then $(c_{S,C})|_Y = c_{S,C|_Y}$.*

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

Let $\overline{\psi^{1,d}} = (\psi_1, \ldots \psi_d)$ be a tuple of non-constant mappings $\psi_i : X_i \to \{0, 1\}$. From Theorem 13, $C|_{\{X_1,\ldots,X_d\}}$ is VCD_{Ψ^*}-maximum of dimension d and by Lemma 11, $\overline{\psi^{1,d}}(C|_{\{X_1,\ldots,X_d\}}) = \{0, 1\}^d$. Let $\psi_{d+1} : X_{d+1} \to \{0, 1\}$ such that

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

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

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

On the other hand, because $c_{S,C|_Y}(X_{d+1}) = 1$, 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. ■

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 $X \setminus X_t$.

Lemma 31 *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 $0 \leq i \leq N_t$. Then $c_{S_i,C} - X_t = c_{S,C^{X_t}}$.*

Proof W.l.o.g, let $t = d$ and $X(S) = \{X_1, \ldots, X_{d-1}\}$. Clearly, $c_{S_i,C}$ and $c_{S,C^{X_d}}$ agree on $X(S)$. Assume that $c_{S,C}$ and $c_{S,C|_Y}$ differ on some $X_t \in X \setminus \{X_1, \ldots, X_d\}$. 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 $\psi_i : X_i \to \{0,1\}$, such that

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

From Theorem 13, we obtain that $C|_{\{X_1,\ldots,X_d\}}$ is VCD_{Ψ^*}-maximum of dimension d and by Lemma 11, $\overline{\psi^{1,d}}(C|_{\{X_1,\ldots,X_d\}}) = \{0,1\}^d$.

On the one hand, for all $0 \leq i \leq N_d$, the fact that $c_{S_i,C}(X_{d+1}) = 0$ implies, for each labeling $((X_1, n_1), \ldots, (X_d, n_d))$ of $X(S_i)$, the existence of a concept $c \in C$ that is consistent with that labeling and fulfills $c(X_{d+1}) = 0$. That is, for each tuple $(n_1, \ldots, n_d) \in C|_{\{X_1,\ldots,X_d\}}$, there is a concept $c \in C$, such that $c|_{\{X_1,\ldots,X_d\}} = (n_1, \ldots, n_d)$ and $c(X_{d+1}) = 0$. Consequently, 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 $c \in C$, such that $\overline{\psi^{1,d}}(c|_{\{X_1,\ldots,X_d\}}) = (\psi_1(n_1), \ldots, \psi_d(n_d))$ and $c(X_{d+1}) = 0$. So, $\{\{0,1\}^d \times \{0\}\} \subseteq \overline{\psi^{1,d+1}}(C|_{\{X_1,\ldots,X_{d+1}\}})$.

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 the definition of reduction, for each $c \in C^{X_d}$, $c \cup (i) \in C$, 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. ■

Now, we are ready to show that for each concept in a VCD_{Ψ^*}-maximum class, there exists a compression set of size VCD_{Ψ^*}-dimension of the class.

Theorem 20 *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 As Floyd and Warmuth (1995), we do 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 13, $C - X_m$ 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 Corollary 18, 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 30, $c_{S,C} - X_m = c_{S,C-X_m}$, 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 assumption for 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 onto the concepts in C. Clearly, $c \in C^{X_m}$.

By Theorem 16, 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 Corollary 18, 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 31, $c_{S_i,C} - X_m = c_{S,C^{X_m}}$, 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 $c_{S_i,C} = c \cup \{(X_m, i)\}$.

■

Corollary 21 *Let C be a VCD_{Ψ^*}-maximum class with $\text{VCD}_{\Psi^*}(C) = d$. Then C has a sample compression scheme of size d.*

Proof The compression function, on the input of a sample S of size at least d, where S agrees with at least one concept in C, works as follows: S is a concept $c \in C|_{X(S)}$. Since $C|_{X(S)}$ is VCD_{Ψ^*}-maximum with $\text{VCD}_{\Psi^*}(C) = d$, Theorem 20 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.

The decompression function, given a compression set S' of size d and an $X_i \in X$, works as follows. If $X_i \in X(S')$, then the output is the label l for which $(X_i, l) \in S'$. If $X_i \notin X(S')$, then the output label l is the same as the one predicted by the decompression set of S' with respect to $C|_{X(S)\cup\{X_i\}}$, which exists because $C|_{X(S)\cup\{X_i\}}$ is a VCD_{Ψ^*}-maximum class with $\text{VCD}_{\Psi^*}(C|_{X(S)\cup\{X_i\}}) = d$. In fact, the decompression function returns as a hypothesis the concept $c_{S',C}$ on X from the class C. ■

For an infinite instance space and for a sample S consistent with some concept in C with $X(S) \subseteq X' \subset X$, such that X' is finite and $|S| = d$, we define $c_{X(S),C}$ on the instances in $X' \setminus X(S)$ as $c_{X(S),C|_{X'}}$. Consequently, $c_{S,C}$ is defined as $c_{S,C|_{X'}}$. Note that X' can contain finitely many instances from X and since C is maximum, $C|_{X'}$ is also maximum. So, by Lemma 30 , $c_{X(S),C}$

assigns a unique label to each instance $X_i \in X \setminus X(S)$. That is, the concept $c_{S',C}$ on X is consistent with the original sample set $c_{S',C|_{X(S)}}$. So, the Corollary holds also for infinite X.

Appendix C. Proofs Omitted From Section 5

Lemma 22 *Let* $\mathrm{VCD}_{\Psi^*}(C) = 1$. *Then for any* $X_i, X_j \in X$ *with* $X_i \neq X_j$, *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
$\vdots$		

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 $\mathrm{VCD}_{\Psi^*}(C) = 1$. 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. ■

Lemma 24 *Let* C *be a concept class and let* S *be a sample consistent with some concept in* C *and* $(X_i, l_i), (X_j, l_j) \in S$, *such that* (X_i, l_i) *implies* (X_j, l_j). *If* $\mathrm{VCD}_{\Psi^*}(C) = 1$ *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 22, $\{e_i, e_j\}$ 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 25 *Let C be a concept class and let S be a sample consistent with some concept in C and $(X_i, l_i), (X_j, l_j) \in S$. If $\mathrm{VCD}_{\Psi^*}(C) = 1$ 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 24 e_i uniquely implies e_j and e_j uniquely implies e_i. ■

Lemma 26 *Let $\mathrm{VCD}_{\Psi^*}(C) = 1$, and let S be a sample consistent with some concept in C 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 $\mathrm{VCD}_{\Psi^*}(C) = 1$, e_1 uniquely implies e_3 by Lemma 24. ■

Corollary 28 *Let $\mathrm{VCD}_{\Psi^*}(C) = 1$. 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 27.

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. ■