JMLR: Workshop and Conference Proceedings vol (2013) 1–16

Honest Compressions and Their Application to Compression Schemes

Roi Livni ROI.LIVNI@MAIL.HUJI.COM
Edmond and Lily Safra Center for Brain Sciences and Rachel and Selim Benin School of Computer Science and Engineering, The Hebrew University of Jerusalem Givat Ram, Jerusalem 91904, Israel

Pierre Simon PIERRE.SIMON@NORMALESUP.ORG
Einstein Institute of Mathematics, The Hebrew University of Jerusalem Givat Ram, Jerusalem 91904, Israel

Abstract

The existence of a compression scheme for every concept class with bounded VC-dimension is one of the oldest open problems in statistical learning theory. Here we demonstrate the existence of such compression schemes under stronger assumptions than finite VC-dimension. Specifically, for each concept class we associate a family of concept classes that we call *the alternating concept classes*. Under the assumption that these concept classes have bounded VC-dimension, we prove existence of a compression scheme. This result is motivated by recent progress in the field of model theory with respect to an analogues problem. In fact, our proof can be considered as a constructive proof of these advancements. This means that we describe the reconstruction function explicitly. Not less important, the theorems and proofs we present are in purely combinatorial terms and are available to the reader who is unfamiliar with model theory. Also, using tools from model theory, we apply our results and prove existence of compression schemes in interesting cases such as concept classes defined by hyperplanes, polynomials, exponentials, restricted analytic functions and compositions, additions and multiplications of all of the above.

Keywords: Compression Conjecture, Compression Scheme, NIP Structures.

1. Introduction

A concept class $(X, \mathcal{C})$ consists of a set of points X which we refer to as the domain, and a family of subsets of X (i.e., $\mathcal{C} \subset 2^X$) which we refer to as concepts. Given a concept class $(X, \mathcal{C})$, a k-sample compression scheme for $\mathcal{C}$ is a mapping from sequences of elements of X labeled according to some concept $C \in \mathcal{C}$ to sub-sequences of length k. It is required that the labeling of the input sequence can be reconstructed from the subsequence.

Compression schemes were first introduced by Littlestone and Warmuth (1986), where their close relation to learnability was established. Namely, the existence of a k-compression scheme implies an order k bound on VC-dimension. The reverse implication: whether a bounded VC-dimension implies a $k-$sample compression scheme is an unresolved question in statistical learning theory. A stronger version of this question is the compression conjecture, that every concept class with VC-dimension d has an order $O(d)$ compression scheme.

1.1. Meanwhile in Model Theory

Interestingly the problem has also caught the attention of model theorists. Laskowski (1992) was the first to observe that the property of having finite VC-dimension had been investigated simultaneously and independently in both fields. In model theory, it appeared under the name "NIP" (Negation of the Independence Property). Later Johnson and Laskowski (2010) observed that the compression scheme conjecture had a very natural model-theoretic analogue. Some special cases of the latter conjecture are obtained by Johnson and Laskowski (2010) and Guingona (2012). More recently, Chernikov and Simon (2010, 2012) managed to prove a weak version of this conjecture. Specifically, this latter result assumes finite VC-dimension not only for the concept class $\mathcal{C}$, but also for certain other concept classes built from it. Model-theoretically, this is a very natural assumption which is satisfied by many classes that are known to be of finite VC-dimension.

1.2. Our Main Result

For every concept class $(X, \mathcal{C})$ we associate a family of concept classes that we call *the alternating concept classes*. Under the assumption that these concept classes have finite VC-dimension, we prove existence of a compression scheme. Furthermore, we show that it is enough to verify this assumption for one specific member of the family.

This assumption might seem strange at first. However, we know from model theory that many interesting cases satisfy it. For example, as a special case, we cover concept classes defined by compositions of polynomials, exponentials and restricted analytic functions. An alternative, model theoretic, proof of this case was given by Johnson and Laskowski (2010).

More generally, our assumption is fulfilled by any concept class definable in an NIP structure; this case was proved by Chernikov and Simon (2012). Our proof can be considered as a constructive proof of their result. Our statement, though, demands a simpler and weaker condition. The advantage of our constructive proof is twofold: First, we describe the reconstruction function explicitly. Second, our main result and proof are accessible to the reader who is not familiar with model theory.

1.3. Related Work

The compression conjecture was first introduced by Littlestone and Warmuth (1986). Later, compression schemes were studied by Floyd (1989); Floyd and Warmuth (1995). These papers establish the existence of compression schemes for an important family of concept classes, specifically *maximum concept classes*.

Ben-David and Litman (1998) showed a wide family of classes for which compression schemes of size VC-dim exist. They also proved the existence of compression schemes with additional information for geometric concept classes (specifically *k-size array compression scheme*). The bounds for the compression scheme may be computed in terms of the complexity of the defining formula. Their proof relied on Tarski's theorem stating that any first order formula in the structure $\langle \mathbb{R}; +, \cdot, \leq, 0, 1 \rangle$ is equivalent to one without quantifiers (which is the main step in proving that this structure is NIP). In that respect, Chernikov and Simon (2012) extends this result from geometric concept classes to any concept class in an NIP structure.

More recently Kuzmin and Warmuth (2005) showed the existence of unlabeled compression schemes for maximum concept classes. They conjectured that such compression schemes have a geometric interpretation. Rubinstein et al. (2009), and later Rubinstein and Rubinstein (2012) gave a positive result in that direction by using representation of concept classes in piecewise linear hyperplane arrangement. They further study compressions that arise through operations on geometric representations. Many of the representations they give define geometric concept classes. We discuss this further in Section 6.

Two other works that are relevant to our paper, yet not so much to the compression conjecture, are the works of Macintyre and Sontag (1993); Karpinski and Macintyre (1995). They too applied model theoretic results. Together, these works show polynomial bounds on the VC-dimension of concept classes defined by certain neural networks. The bounds are derived by exploiting the fact that these neural networks are definable in $o-$minimal structures, a property which implies NIP. We elaborate on this in Section 5.

The main focus of this paper is the conceptual connection between learnability and compressibility. The bounds we produce are far from $O(d)$. It is worth comparing our result to results on data size dependent compression schemes. Compression schemes that compress a sample of size m to a subset of size $O(d \log m)$ exist (see for example Freund (1995)). The compression schemes presented here compress a sample set of size m to a subset of size c, where c is a constant independent of m. However, since our bounding techniques involve tools like Ramsey theory, it is questionable if for reasonable size m, the c we present is smaller than $O(d \log m)$. (And in case one has so many examples that $d \log m$ happens to be comparable to c, considering a larger concept class would make sense).

1.4. Structure of the Paper

The structure of this paper is as follows. In Section 2 we elaborate on the results in model theory and further explain them. The rest of the paper is independent of this section. We will however refer to the fact that certain structures are NIP when we will wish to apply our results. In Section 3 we give the necessary definitions, some preliminary facts from Ramsey theory and other combinatorial results. We proceed to our main result in Section 4 which is Theorem 21. We then continue in Section 5 to some applications and to a counter example where our assumptions in Theorem 21 do not apply, thus concluding that the problem of compression scheme is not yet resolved. We conclude in Section 6 where we discuss some notions that arise from our analysis and potential directions for further advancement.

2. NIP Theories and the UDTFS Conjecture

Let us introduce the model-theoretic framework in a simplified version tailored to our needs. An *R-structure* is a (usually infinite) set M equipped with a distinguished binary relation $R(x, y)$ (which is just a subset of M^2). An *R-formula* is a syntactically correct formula written using any number of variable symbols $x_1, x_2, \ldots, y_1, y_2, \ldots$, quantifiers $\exists, \forall$, logical connectives $\wedge, \vee, \neg, \ldots$ and the binary symbols R and $=$. For example

$$\phi(x_1; y_1) := R(y_1, x_1) \wedge (\forall x_2)(R(y_1, x_2) \rightarrow (x_2 = y_1 \vee R(x_1, x_2)))$$

is a formula which, in the case where R happens to be an ordering, says that x_1 is the successor of y_1. We see that the variables x_1, y_1 are not under the scope of any quantifier.

They can thus be replaced by parameters from M: if a_1, b_1 are in M it makes sense to ask whether the formula $\phi(a_1; b_1)$ is true. If this is the case, we write $M \models \phi(a_1; b_1)$.

Assume that $\phi(\boldsymbol{x}; \boldsymbol{y})$ is a formula, where $\boldsymbol{x}, \boldsymbol{y}$ are two sequences of variables of respective lengths $|\boldsymbol{x}|$ and $|\boldsymbol{y}|$. Given a sequence $\boldsymbol{c} \in M^{|\boldsymbol{y}|}$, we can consider the set $\phi(M; \boldsymbol{c}) \subseteq M^{|\boldsymbol{x}|}$ defined as $\phi(M; \boldsymbol{c}) = \{\boldsymbol{a} \in M^{|\boldsymbol{x}|} : M \models \phi(\boldsymbol{a}, \boldsymbol{c})\}$. Such a set is called a *definable set* of M. We associate to the formula $\phi(\boldsymbol{x}; \boldsymbol{y})$ the family $\mathcal{F}_\phi = \{\phi(M; \boldsymbol{c}) : \boldsymbol{c} \in M^{|\boldsymbol{y}|}\} \subseteq \mathcal{P}(M^{|\boldsymbol{x}|})$.

We say that the formula $\phi(\boldsymbol{x}; \boldsymbol{y})$ is *NIP* (No Independence Property) if the family $\mathcal{F}_\phi$ has finite VC-dimension. We say that the structure M is NIP if all R-formulas are NIP. Note that if M is NIP, then in particular the family $\{R(M; c) : c \in M\}$ of subsets of M has finite VC-dimension, but the converse need not hold. The class of NIP structures has been introduced in Shelah (1971) and has been investigated in a number of papers since. See Simon (2012) for a survey.

If $A \subseteq M$ is an arbitrary subset of M, we similarly define $\phi(A; \boldsymbol{b}) = \{\boldsymbol{a} \in A^{|\boldsymbol{x}|} : M \models \phi(\boldsymbol{a}; \boldsymbol{b})\}$. It is the trace of the formula $\phi(x; \boldsymbol{b})$ on the set A. Understanding such traces of definable sets on small subsets is a recurring theme in model theory. A nice situation for example is when this trace coincides with some other trace $\psi(A; \boldsymbol{d})$ where the parameters $\boldsymbol{d}$ come from the set A itself. As we will see, existence of compression schemes is strongly related to this property.

How does this framework relate to concept classes? We have explained how, given a structure M, one associates various concept classes $\mathcal{F}_\phi$ on cartesian powers of M. Conversely, starting with a set X and a concept class $\mathcal{C}$ of subsets of X, we can construct an R-structure M as follows. Let the set M consist of the disjoint union of the sets X and $\mathcal{C}$. Let $R(a, b)$ hold for $a, b \in M$ if and only if a is in X, b is in $\mathcal{C}$ and a belongs to b. Then the formula $R(x, y)$ is NIP if and only if the class $\mathcal{C}$ has finite VC-dimension. Note that saying that the structure M is NIP is a much stronger statement.

In general, model theory considers structures M which are equipped with any—finite or infinite—number of relations R_i, functions $f_j : M^{k_j} \to M$ and constants $c_k \in M$. We write such a structure as $\langle M; (R_i)_i, (f_j)_j, (c_k)_k \rangle$. The definition of a formula and a definable set extend naturally. For example the structure $\langle \mathbb{R}; +, \cdot, \leq, 0, 1 \rangle$ of real numbers, equipped with two ternary relations $+, \cdot$ for addition and multiplication, a binary relation $\leq$ for the order and two constant symbols for 0 and 1, is an NIP structure. Any semi-algebraic set is a definable set in that structure (and by Tarski's theorem, they are the only ones). If we add a symbol for the exponential function, we obtain the structure $\mathbb{R}_{\exp} = \langle \mathbb{R}; +, \cdot, \leq, \exp, 0, 1 \rangle$ which is also NIP. Furthermore, we may add symbols for an arbitrary number of *restricted analytic* functions $(f_i)_i$, that is restrictions to $[0, 1]^n$ of some analytic function defined on a neighborhood of $[0, 1]^n$. (The arity n may vary from function to function.) See Speissegger (2012).

It turns out that many concept classes which are known to be of finite VC-dimension can be written in the form $\mathcal{F}_\phi$ for some formula $\phi(\boldsymbol{x}, \boldsymbol{y})$ from one of those structures: it is the case for example of any concept class defined by polynomial—or exponential polynomial—inequalities. The set of half spaces (in any dimension n) is the simplest example of such. A more complicated example is that of neural networks which we discuss in Section 5. We thus see that the condition that M is NIP as mentioned in the previous paragraph is not as restrictive as it might seem at first.

Johnson and Laskowski (2010) observed that the problem of compression schemes has a natural model-theoretic analogue which they called the UDTFS-conjecture (uniform definability of types over finite sets).

Conjecture 1 (UDTFS) *Let $\phi(\boldsymbol{x};\boldsymbol{y})$ be an NIP-formula. Then there exists a formula $\psi(\boldsymbol{x};\boldsymbol{z})$ such that for any finite subset $A \subset M$ and tuple $\boldsymbol{c} \in M$, there is some $\boldsymbol{d} \in A^{|\boldsymbol{z}|}$ such that $\phi(A;\boldsymbol{c}) = \psi(A;\boldsymbol{d})$.*

This conjecture should be thought of as saying that there is a definable compression scheme for the family $\mathcal{F}_\phi$. The compression function takes as input a finite subset A and the trace $\phi(A;\boldsymbol{c})$ and outputs the tuple $\boldsymbol{d}$. The reconstruction function takes the tuple $\boldsymbol{d}$ as input and outputs the definable set $\psi(M;\boldsymbol{d})$.

In Chernikov and Simon (2012), the following theorem is proved.

Theorem 2 *Let M be an NIP structure, then the UDTFS conjecture holds for all formulas over M.*

Note that this statement is weaker than the original conjecture in that it assumes NIP for all formulas and not just for one. From a model-theoretic perspective, this is a very natural assumption. As noted above, it is satisfied in particular by all concept classes defined in the field of real numbers with exponential; this includes many interesting examples.

The proof of this theorem requires two ingredients. The first one is the notion of 'honest definitions' introduced in Chernikov and Simon (2010).

Theorem 3 *Assume that M is an NIP structure and let $\phi(x;\boldsymbol{y})$ be a formula. Fix some parameters $\boldsymbol{c} \in M^{|\boldsymbol{y}|}$ and subset $A \subseteq M$. Then there is a formula $\psi(x;\boldsymbol{z})$ with the following property: For any finite $A_0 \subseteq A$, there is some $\boldsymbol{d}_{A_0} \in A^{|\boldsymbol{z}|}$ such that $\phi(A_0;\boldsymbol{c}) = \psi(A_0;\boldsymbol{d}_{A_0})$ and $\psi(A;\boldsymbol{d}_{A_0}) \subseteq \phi(A;\boldsymbol{c})$.*

The condition $\psi(A;\boldsymbol{d}_{A_0}) \subseteq \phi(A;\boldsymbol{c})$ is called the honesty condition because it says that the formula $\psi(x;\boldsymbol{d}_{A_0})$ does not lie on A: whenever it holds on some $a \in A$, then the initial formula $\phi(x;\boldsymbol{c})$ also holds on a. This theorem can be seen as saying that the trace of a definable set on an arbitrary set A is approximable from the inside by instances of some formula $\psi(x;\boldsymbol{z})$ taking the parameters in A. Note that as we assumed the structure M to be NIP, the family $\{\psi(A;\boldsymbol{d}_{A_0}) : A_0 \subset A\}$ of approximations has finite VC-dimension.

The other ingredient in the proof of Theorem 2 is Matoušek's (p,k)-theorem which we discuss later in this paper. It is used to show that in the special case where A is finite, one can find a small number of approximations whose union is equal to the full set $\phi(A;\boldsymbol{b})$.

3. Preliminaries and Notations

3.1. VC-Dimension and Compression Schemes

We start with the definition of a concept class with a finite VC-dimension.

Definition 4 (Concept class) *A concept class $(X,\mathcal{C})$ consists of a pair: a domain X, and a family $\mathcal{C}$ of subsets of X. The elements of X are called points and the elements of $\mathcal{C}$ are called concepts. We will call a concept class non-degenerate if for every $x \in X$ there is a concept $C \in \mathcal{C}$ such that $x \in C$.*

With abuse of notation we will identify a concept with its indicator function. In other words, a concept C will also denote a function from X to $\{0,1\}$. Given a sequence $\boldsymbol{x}$ of length m, we will notate by $C(\boldsymbol{x})$ the sequence: $(C(x_1), C(x_2), \ldots, C(x_m)) \in \{0,1\}^m$.

Definition 5 (VC dimension) *Given a concept class $(X, \mathcal{C})$, a finite set $A \subseteq X$ is said to be shattered if $\{A \cap C : C \in \mathcal{C}\} = 2^A$. The VC dimension of a concept class, VC-dim($\mathcal{C}$), is defined as the maximal cardinality of a finite set A that is shattered. If such a maximum doesn't exists, we write VC-dim($\mathcal{C}$) $= \infty$.*

The relation between finite VC-dimension and the learnability of a concept class was established in Vapnik and Chervonenkis (1971), and is arguably the most important result in statistical learning theory.

Definition 6 (Sample) *The set $S(X^m)$ of samples of size m contains all the sequences in $(X \times \{0,1\})^m$. Given a concept C and a sequence $\boldsymbol{x} = (x_1, \ldots, x_m)$ of size m, we will denote by $S_C(\boldsymbol{x})$ the sample S whose coordinates are $(x_i, C(x_i))$.*

We continue to the definitions of compression schemes. We begin with the basic compression scheme, first introduced in Littlestone and Warmuth (1986).

Definition 7 (Compression scheme) *Given a concept class $\mathcal{C}$, a k-compression scheme consists of a pair of mappings*

$$\kappa : \cup_{m=k}^{\infty} S(X^m) \to S(X^k) \quad \text{and} \quad \rho : S(X^k) \times X \to \{0,1\}$$

such that:

1. *For any $C \in \mathcal{C}$, $m \geq k$ and $\boldsymbol{x} \in X^m$, $\kappa(S_C(\boldsymbol{x}))$ is a subsequence of $S_C(\boldsymbol{x})$.*
2. *For any $C \in \mathcal{C}$, $m \geq k$, $\boldsymbol{x} \in X^m$ and a point x_i in $\boldsymbol{x}$, $\rho(\kappa(S_C(\boldsymbol{x})), x_i) = C(x_i)$.*

We call the function κ, the compression function and the function ρ the reconstruction function. In this paper we we will focus on compression schemes with additional information also introduced in the same paper.

Definition 8 (Compression scheme with additional information) *Given a finite set Q, a k-compression scheme with additional information Q consists of a pair of functions ρ and κ, as in Definition 7, with the modification that the image of κ is $Q \times S(X^k)$ and the domain of ρ is $\left(Q \times S(X^k)\right) \times X$.*

It is often convenient to encode the additional information via reordering of the subsequence or using a bounded amount of repetitions of certain elements. In our case we will use the following type of a compression scheme, which can be considered a special case of a compression scheme with additional information.

Definition 9 (Word compression scheme with an extra bit- b) *A k-word compression scheme with an extra bit b consists of a pair of functions ρ and κ, as in Definition 7, with the modification that the image of κ is $(b \cup X)^k$ and the domain of ρ is $\left((b \cup X)^k\right) \times X$. We replace requirement 1 with the requirement that all elements in the sequence $\kappa(S_C(\boldsymbol{x}))$ are either the extra bit b or elements in the sequence $\boldsymbol{x}$.*

To see that this is a special case of a compression scheme with additional information, note that given a k-word compression function κ, we can define a compression function $\tilde{\kappa}$ whose image consists of a subsequence of $S_C(\boldsymbol{x})$ that contains all elements that appear in the word $\kappa(S_C(\boldsymbol{x}))$ (there are at most k such elements) and as additional information we need $(k+1)^k$ elements in order to encode all the possible arrangements of $k+1$ symbols in a sequence of length k.

3.2. Homogeneous and Alternating Sequences

In this section we discuss some relevant results that are corollaries of Ramsey's theorem.

Definition 10 (Homogeneous sequence) *Let $(X, \mathcal{C})$ be a concept class. A sequence $\boldsymbol{x}$ of elements in X is called d-homogeneous if for any two subsequences of size d, $\boldsymbol{x}^{(1)}$ and $\boldsymbol{x}^{(2)}$, for any concept $C_1 \in \mathcal{C}$, there is a concept $C_2 \in \mathcal{C}$ such that*

$$C_1(\boldsymbol{x}^{(1)}) = C_2(\boldsymbol{x}^{(2)}).$$

The following result is an immediate corollary of Ramsey's Theorem for hypergraphs, applied to the coloring of all d subsequences using the 2^{2^d} possible concept classes over sets of size d as our colors (see for example Graham et al. (1990)).

Theorem 11 *For any k and d, there exists a large enough p such that: Given a concept class $(X, \mathcal{C})$, any sequence $\boldsymbol{x}$ of size p has a d-homogeneous subsequence of size k.*

Definition 12 (Alternation number) *Given a concept class $(X, \mathcal{C})$, a sequence $\boldsymbol{y}$ is said to be alternating in $\mathcal{C}$ if there is a concept $C \in \mathcal{C}$ such that $C(y_i) \neq C(y_{i-1})$ for every $1 < i \leq |\boldsymbol{y}|$. The alternation number of a sequence $\boldsymbol{x}$ in $\mathcal{C}$, denoted $\mathrm{alt}(\boldsymbol{x}; \mathcal{C})$, is the size of the largest subsequence of $\boldsymbol{x}$ that is alternating in $\mathcal{C}$. Formally*

$$\mathrm{alt}(\boldsymbol{x}; \mathcal{C}) := \max_{C \in \mathcal{C}} |\{i : C(x_i) \neq C(x_{i-1})\}| + 1.$$

Theorem 13 *Let $(X, \mathcal{C})$ be a concept class of VC dimension d and let $\boldsymbol{x}$ be a $(d+1)$-homogeneous sequence, then $\mathrm{alt}(\boldsymbol{x}; \mathcal{C}) \leq 2d + 1$.*

A proof of Theorem 13 can be found for example in Adler (2008) (Proposition 3). For those who wish to avoid model theoretic terminology we repeat the proof in Appendix A.1.

As a corollary of Theorem 13 and Theorem 11 we have the following result:

Corollary 14 *For any k and d, there exists a large enough p such that, given a concept class $(X, \mathcal{C})$ with VC-dim$(\mathcal{C}) = d$, any sequence $\boldsymbol{x}$ of length p has a subsequence $\boldsymbol{y}$ of length k such that*

$$\mathrm{alt}(\boldsymbol{y}; \mathcal{C}) \leq 2d + 1.$$

3.3. Alternating Concept Classes and Honest Concepts

The next definition is the last definition needed to formulate our main result.

Definition 15 (Alternating concept class) *Given a concept class $(X,\mathcal{C})$ and $\boldsymbol{x}$ such that $|\boldsymbol{x}| = m$, for every $i \leq m$, we consider the subset of X:*

$$\mathrm{ALT}(\boldsymbol{x};\mathcal{C}) := \{y : \mathrm{alt}((y, x_1, \ldots, x_2, \ldots, x_m);\mathcal{C}) \leq 2\,VC\text{-}dim(\mathcal{C}) + 1\}\,.$$

Given m and a subset (possibly infinite) A, we define the following family of subsets of X:

$$\mathcal{ALT}_m(A;\mathcal{C}) = \{\mathrm{ALT}(\boldsymbol{x};\mathcal{C}) : \boldsymbol{x} \in A^m\}\,.$$

The concept class $(X, \mathcal{ALT}_m(A;\mathcal{C}))$ is called "The alternating concept class of $(X,\mathcal{C})$ with respect to sequences of length m from A".

We leave the proof of the following lemma to Appendix A.2.

Lemma 16 *If $\mathcal{ALT}_{2\,VC\text{-}dim(\mathcal{C})+1}(X;\mathcal{C})$ has finite VC-dimension then for every m $\mathcal{ALT}_m(X;\mathcal{C})$ has finite VC-dimesnion.*

Definition 17 (Honest concept) *Given a concept C and a finite set A, we say that the set $\tilde{C}$ (not necessarily a concept in $\mathcal{C}$) is honest for A with respect to C if $\tilde{C} \cap A \subseteq C \cap A$.*

Definition 18 (Honest alternating concept class) *Let $(X,\mathcal{C})$ be a concept class, $A \subseteq X$ a finite set and $C \in \mathcal{C}$ a concept. The concept class $(X, \mathcal{ALT}_m(A, C;\mathcal{C}))$ consists of all concepts in $(X, \mathcal{ALT}_m(A;\mathcal{C}))$ that are honest for A with respect to C.*

3.4. The $(\mathbf{p},\mathbf{k})$ Theorem

We now proceed to a result by Matoušek (2004) that is the backbone of our proof.

Definition 19 ($(\mathbf{p},\mathbf{k})$ property) *Given a concept class $(X,\mathcal{C})$, a finite set $A \subseteq X$ is said to have the (p,k) property with respect to $(X,\mathcal{C})$, if for every subset A_0 of cardinality p, $A_0 \cap C$ is of cardinality at least k for some $C \in \mathcal{C}$.*

Theorem 20 ($(\mathbf{p},\mathbf{k})$ Theorem for bounded VC dimension concept class) *For every k and p there exists an N such that: Given a non degenerate concept class $(X,\mathcal{C})$ with VC-dim$(\mathcal{C}) < k$, if a finite subset $A \subseteq X$ has the (p,k) property with respect to $(X,\mathcal{C})$, then there are N concepts $C_1, \ldots, C_N$ such that $A \subseteq \cup_{i=1}^{N} C_i$.*

Matoušek's original result is formalized in a dual form and applies to a wider class of concepts for which N is not uniform. We go through the full derivation in Appendix A.3.

4. Main Result

The following theorem is the main result of this paper.

Theorem 21 *Let $(X, \mathcal{C})$ be a concept class with VC-dimension d and assume that $\mathcal{ALT}_{2d+1}(X; \mathcal{C})$ has finite VC dimension. Then $(X, \mathcal{C})$ has a bounded compression scheme with additional information.*

The proof is a corollary of Lemma 24 and Theorem 23 which is a generalization of Matoušek's theorem. We first present the idea of proof, which illustrates how the tools we developed can be used to construct compression schemes.

Idea of proof By the assumptions of Theorem 21 and by Lemma 16, the VC-dimension of $\mathcal{ALT}_m(A, C; \mathcal{C})$ is bounded uniformly for all A and C. Now assume that for some m, $\mathcal{ALT}_m(A, C; \mathcal{C})$ is non degenerate and $A \cap C$ has the (p, k) property with respect to $\mathcal{ALT}_m(A, C; \mathcal{C})$ for all A, $C \in \mathcal{C}$, a fixed p and sufficiently large k.

Matoušek's theorem states that there are N sequences $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ of length m such that $A \cap C \subseteq \cup_{j=1}^{N} \mathrm{ALT}(\boldsymbol{x}_j; \mathcal{C})$. We compress our sample to $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$. We then define $\rho([\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N])(x) = 1$ if for some j we have $x \in \mathrm{ALT}(\boldsymbol{x}_j; \mathcal{C})$. The honesty property of $\mathrm{ALT}(\boldsymbol{x}_j; \mathcal{C})$ ensures we won't be wrong on any $x \in A$ such that $C(x) = 0$. Matoušek's theorem ensures that we are not wrong on any $x \in A$ such that $C(x) = 1$.

In general, $\mathcal{ALT}_m(A, C; \mathcal{C})$ need not have the (p, k) property. That is why we need to relax the assumptions of Theorem 20. We do that by introducing the (p, k, k) property which is weaker then the (p, k) property.

The structure of the rest of this section is as follows: In Definition 22 we define the (p, k, k) property. In Theorem 23 we prove the analogous of Matoušek's theorem to the (p, k, k) property. In Lemma 24 we show that indeed the (p, k, k) property may be assumed for the complex $\mathcal{ALT}(X; \mathcal{C})$. We finish this section with a formal proof of Theorem 21 and Corollary 25.

Definition 22 ($(\mathbf{p}, \mathbf{k_1}, \mathbf{k_2}, \ldots, \mathbf{k_d})$ property) *Let $(X, \boldsymbol{\mathcal{C}}) = (X, \mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_d)$ be d concept classes sharing a common domain X.*

A subset $A_0 \subseteq X$ of size p is said to suggest property (p, k_n) with respect to $\mathcal{C}_n$ if $A_0 \cap C_n$ is of cardinality at least k_n for some $C_n \in \mathcal{C}_n$.

A subset $A \subseteq X$ is said to have the $(p, k_1, k_2, \ldots, k_d)$ property with respect to $(X, \boldsymbol{\mathcal{C}})$ if every subset $A_0 \subseteq A$ of size p suggests property (p, k_n) with respect to $\mathcal{C}_n$ for some $n \leq d$.

Theorem 23 ($(\mathbf{p}, \mathbf{k_1}, \ldots, \mathbf{k_d})$ Theorem for bounded VC dimension concept classes) *For every $p, k_1, \ldots, k_d$ there exists an N such that: Given d non-degenerate concept classes sharing a common domain $(X, \boldsymbol{\mathcal{C}}) = (X, \mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_d)$, with VC-dim$(\mathcal{C}_i) < k_i$ for every i, for any finite subset A with the $(p, k_1, \ldots, k_d)$ property with respect to $(X, \boldsymbol{\mathcal{C}})$, there are N concepts $\{C_1, \ldots, C_N\} \subseteq \cup_{i=1}^{d} \mathcal{C}_i$ such that $A \subseteq \cup_{j=1}^{N} C_j$.*

Proof The concept class $(X, \cup_{i=1}^{d} \mathcal{C}_i)$ is the union of d concept classes with finite VC-dimension hence its VC-dimension is bounded, say by D (a bound D that is dependent

on $k_1, \ldots, k_d$ may be produced). If we can show that the $(p, k_1, \ldots, k_d)$ property implies $(P, D+1)$-property with respect to $(X, \cup_{i=1}^{d} \mathcal{C}_i)$ for some P, we are done by Theorem 20.

Let N be larger then any bound N_i whose existence follows from Theorem 20 applied to a subset with the (p, k_i) property with respect to a concept class with VC-dim$(C_i) < k_i$.

For each subset $A_0 \subseteq A$ of cardinality p color it with some n such that it suggests property (p, k_n) with respect to $\mathcal{C}_n$. We now apply Ramsey's theorem for hypergraph with respect to this coloring. There is a P large enough so that any subset of size P contains a monochromatic subset of cardinality $N \cdot (D+1)$. In other words, a set A of size P will contain a subset A_0 of size $N \cdot (D+1)$ with the (p, k_i) property with respect to $\mathcal{C}_i$ for some i. By Theorem 20, there are $N_i < N$ concepts such that any element in A_0 is in one of them. By the pigeonhole principle one concept contains $D+1$ points. ■

Lemma 24 *For every $\theta_1, \theta_2, \ldots, \theta_{d+1}$ there exists a p such that: Given $(X, \mathcal{C})$, a concept class of VC-dimension d, for any finite set A and concept C, the set $A \cap C$ has the $(p, \theta_1, \theta_2, \ldots, \theta_{d+1})$ property with respect to the concept classes*

$$(X, \boldsymbol{\mathcal{C}}) = \left(X, \mathcal{ALT}_{2d+2}(A, C; \mathcal{C}), \mathcal{ALT}_{\theta_1+2d+2}(A, C; \mathcal{C}), \ldots, \mathcal{ALT}_{\sum_{i=1}^{d} \theta_i + 2d+2}(A, C; \mathcal{C})\right).$$

Proof Choose p such that any sequence $\boldsymbol{x}$ of size p contains a subsequence $\boldsymbol{y}$ of length $m = \sum_{i=1}^{d+1} \theta_i + 2d + 2$ with alt$(\boldsymbol{y}; \mathcal{C}) \leq 2d+1$ (by Corollary 14). We claim that this p will suffice.

Indeed, given a set $A_0 \subseteq A \cap C$ of size p, let t be the maximal number (may be 0), for which there is a sequence

$$\boldsymbol{y}^t = \left\langle y_1, y_2, \ldots y_{\theta_{d+1-t}+1}, z_1, y_{\theta_{d+1-t}+2}, z_2, \ldots, y_{\theta_{d+1-t}+t}, z_t, y_{\theta_{d+1-t}+t+1}, y_{\theta_{d+1-t}+t+2}, \ldots, y_n \right\rangle$$

such that

1. $\{y_1, y_2, \ldots, y_n\} \subseteq A_0$;
2. $\{z_1, \ldots, z_t\}$ are in A but not in C;
3. alt$(\boldsymbol{y}^t; \mathcal{C}) \leq 2d+1$;
4. $|\boldsymbol{y}^t| = \sum_{i=1}^{d+1-t} \theta_i + 2d + 2$ (where we define $\sum_{i=1}^{d+1-t} \theta_i = 0$ for $t > d$).

We have $t \geq 0$ by the property of p (by simply arranging the elements of A_0 in some sequence and taking the proper subsequence), and $t < d+1$ because otherwise alt$(\boldsymbol{y}^t; \mathcal{C}) > 2d+1$ (by considering C). Consider the sequence $\boldsymbol{y}^{t+1}$ defined by removing the first θ_{d+1-t} elements in $\boldsymbol{y}^t$. The cardinality of $\boldsymbol{y}^{t+1}$ is $\sum_{i=1}^{d-t} \theta_i + 2d + 2$. If ALT$(\boldsymbol{y}^{t+1}; \mathcal{C})$ is not honest for A with respect to C, there must be some $z_0 \in A$ such that $z_0 \notin C$ and alt$((z_0, \boldsymbol{y}^{t+1}); \mathcal{C}) \leq 2d+1$.

Note that the property alt$(\boldsymbol{x}; \mathcal{C}) \leq 2d+1$ is preserved under cyclic permutation of $\boldsymbol{x}$. Indeed, if $x_{i_1}, \ldots x_{i_{2d+2}}$ is alternating in $\mathcal{C}$ then for any cyclic permutation, since the number of elements in the sequence is even, we will get a sequence that is alternating in $\mathcal{C}$.

Apply a cyclic permutation that takes the last $\theta_{d-t}+1$ elements in the sequence $(z_0, \boldsymbol{y}^{t+1})$ and moves them to the beginning of the sequence. Next, remove the last element of the

permuted sequence (recall that we added z_0 to the sequence and increased its size by one). We Finally get the following sequence which contradicts the maximality of t:

$$\left\langle y_{n-\theta_{d-t}}, \ldots, y_n, z_0, y_{\theta_{d+1-t}+1}, z_1, \ldots, y_{\theta_{d+1-t}+t}, z_t, \ldots, y_{n-\theta_{d-t}-2} \right\rangle.$$

Hence $\mathrm{ALT}(\boldsymbol{y}^{t+1}; \mathcal{C})$ is honest. Also we have $\{y_1, \ldots, y_{\theta_{d+1-t}}\} \subseteq A_0 \cap \mathrm{ALT}(\boldsymbol{y}^{t+1}; \mathcal{C})$. ■

Proof of Theorem 21: Let C and $S_C(\boldsymbol{x}) \in S(X^m)$ be given. We will prove the existence of a word compression scheme with extra bit defined in Definition 9. Existence of a compression scheme with additional information then follows as discussed in the last paragraph of Subsection 3.1.

First, let $\tilde{S}$ be the set of all elements $x \in X$ that appear in the sample $S_C(\boldsymbol{x})$.

The VC-dimension of $\mathcal{ALT}_k(\tilde{S}, C; \mathcal{C})$ is bounded by the VC-dimension of $\mathcal{ALT}_k(X; \mathcal{C})$ which is finite by Lemma 16.

Define $\theta_1 := \text{VC-dim}\left(\mathcal{ALT}_{2d+2}(X; \mathcal{C})\right) + 1$. Define inductively for every $n \leq d+1$

$$\theta_n := \text{VC-dim}\left(\mathcal{ALT}_{\sum_{i=1}^{n-1} \theta_i + 2d+2}(X; \mathcal{C})\right) + 1.$$

Lemma 24 states that for an appropriate p, $\tilde{S} \cap C$ has the $(p, \theta_1, \theta_2, \ldots, \theta_{d+1})$ property with respect to the $d+1$ concept classes with common domain X:

$$\left(X, \mathcal{ALT}_{2d+2}(\tilde{S}, C; \mathcal{C}), \ldots, \mathcal{ALT}_{\sum_{i=1}^{d} \theta_i + 2d+2}(\tilde{S}, C; \mathcal{C})\right). \tag{1}$$

First, for simplicity, assume that all concepts $\mathcal{ALT}(\tilde{S}, C; \mathcal{C})$ that appear in Equation 1 are non-degenerate.

By Theorem 23 there are N honest concepts in $\bigcup \mathcal{ALT}(\tilde{S}, C; \mathcal{C})$, say

$$\{\mathrm{ALT}(\boldsymbol{x}_1; \mathcal{C}), \ldots, \mathrm{ALT}(\boldsymbol{x}_N; \mathcal{C})\},$$

such that for every $x \in \tilde{S} \cap C$, we have $x \in \mathrm{ALT}(\boldsymbol{x}_i; \mathcal{C})$ for some i. Each $\boldsymbol{x}_i$ is at most of size $\sum_{i=1}^{d} \theta_i + 2d + 2$. Define:

$$\kappa(S_C(\boldsymbol{x})) = (\boldsymbol{x}_1, b, \boldsymbol{x}_2, b, \ldots, b, \boldsymbol{x}_N).$$

The reconstruction function works as follows: for any x we put $\rho(x) = 1$ if for some sequence $\boldsymbol{x}_i$ we have $x \in \mathrm{ALT}(\boldsymbol{x}_i; \mathcal{C})$; otherwise we set $\rho(x) = 0$. The honesty property ensures that if $C(x) = 0$ then $\rho(x) = 0$. Matoušek's theorem ensures that if $C(x) = 1$, we will have $\rho(x) = 1$.

When some of the concepts in Equation 1 are degenerate, simply add to the degenerate concept classes $\mathcal{ALT}(\tilde{S}, C; \mathcal{C})$ all the singletons. This will increase the VC dimension by only a constant factor. By the same method of proof one will still obtain for some N, N bounded concepts, $\{\{x_1\}, \{x_2\}, \ldots, \{x_M\}, \mathrm{ALT}(\boldsymbol{x}_{M+1}; \mathcal{C}), \ldots, \mathrm{ALT}(\boldsymbol{x}_N; \mathcal{C})\}$. The concepts of the form $\mathrm{ALT}(\boldsymbol{x}; \mathcal{C})$ will be honest. As for the singleton concepts, we may assume $\{x_i\}$ is contained in $\tilde{S} \cap C$ and we add them to the compressed sequence. ■

It is easy to see that in any structure where the original concept class is definable, the alternating concept classes are also definable using a single quantifier over concepts.

Our result covers as a special case the existence of compression schemes in o-minimal structures established in Johnson and Laskowski (2010). An immediate corollary is the following generalization of Ben-David and Litman (1998) result on compression schemes for geometric concept classes.

Corollary 25 *Let $R(\boldsymbol{x}, \boldsymbol{y})$ be a $j + m$-ary relation on the field $\mathbb{R}$ which is definable in a structure of the form $\langle \mathbb{R}; +, \cdot, \leq, \exp, (f_i)_i \rangle$ where the f_i's are restricted analytic functions of any arity, then $R(\boldsymbol{x}, \boldsymbol{y})$ has a compression scheme with additional information.*

5. Applications and Examples

Example 1 (Application to neural networks) *Neural networks with activation function $\sigma = \frac{1}{1+e^{-x}}$ are definable in the NIP structure $\langle \mathbb{R}, +, \cdot, \leq, \exp, 0, 1 \rangle$. It follows from Corollary 25 that any such neural network has a compression scheme with additional information.*

For an elaborate discussion on neural networks see Macintyre and Sontag (1993). There the question whether any such neural network has finite VC-dimension was resolved. Later, using the theory of o-minimal structures, polynomial bounds were attained in Karpinski and Macintyre (1995).

Next we construct a concept class such that VC-dim $(\mathcal{ALT}_m(X; \mathcal{C}))$ is infinite.

Example 2 (Counter example) *Construct a set $X_1 = \{(B, i) : B \subset \mathbb{N} \quad |B| < \infty, i = \{1, \cdots, 5\}\}$.*

We consider the domain $X = \mathbb{N} \cup X_1$. For each $|B| < \infty$ and $n \in B$ let $C(B, n) = \{n, (B, 2), (B, 4)\}$ and finally let $\mathcal{C} = \{C(B, n), |B| < \infty\}$. It is easy to verify that VC-dim$(\mathcal{C}) = 2$.

Consider the sequence: $\boldsymbol{x}_n = n, (B, 1), (B, 2), (B, 3), (B, 4), (B, 5)$.

For every $n \in B$ we have $\text{alt}(\boldsymbol{x}_n; \mathcal{C}) = 6$ and for every $n \notin B$ we have $\text{alt}(\boldsymbol{x}_n; \mathcal{C}) = 5$. Hence for every finite set $A \subseteq \mathbb{N}$ and subset B, we have $B^c = \text{ALT}((B, 1), \ldots, (B, 5); \mathcal{C})$, and A is shattered.

6. Discussion

The compression conjecture is one of the oldest open problems in the field of statistical learning theory. The UDTFS conjecture discussed in Section 2 is a stronger version of the original conjecture (without $O(d)$ bound). In a sense, it states that we can find not just any compression scheme, but one that can be computed via simple queries on the existence of certain elements and certain concepts in our theory. This is done without imposing any additional structure on our data (e.g. ordering, linearity, convexity etc.). Concept classes definable in NIP structures have this type of compression scheme.

The reconstruction function that arises from our construction has some interesting properties. For any point x the labeling depends only on the trace of the concept class on the set $\boldsymbol{x} \cup x$ where $\boldsymbol{x}$ is the compressed sample. In other words, if one can efficiently compute

the trace of the concept class, then one can efficiently reconstruct. It would be surprising if such a simple compression scheme would work for every concept class. Indeed, the compactness theorem of first order logic would imply the existence of a uniform formula defining a reconstruction function for all concept classes of given VC-dimension. One could however imagine such a uniform formula with, for example, quantifiers over the domain X. The computation would then involve queries on the whole underlying concept class $(X, \mathcal{C})$. Its computational complexity would then depend on the concept class.

Our results apply to a wide family of concept classes, most notably, geometric concept classes and concept classes definable in the structure $\langle \mathbb{R}; +, \cdot, \leq, \exp, 0, 1 \rangle$. One important family of concept classes for which we do not know if our theorem applies is the family of maximum concept classes, first studied by Floyd and Warmuth (1995). A concept class $(X, \mathcal{C})$ is called maximum if for every finite set A, the number of concepts on A equals the upper bound attained by Sauer's lemma. As discussed earlier, such concept classes always have a VC-dim compression scheme (in fact an unlabeled compression scheme).

Rubinstein and Rubinstein (2012) suggest studying compression schemes via operating on geometric representations. They study classes that correspond to Hyperbolic arrangements and PL hyperplanes. It is natural to ask what kind of compression schemes arise from general representations in NIP structures. If any maximal concept class can be represented in an NIP structure, this would give a positive solution to the problem of sample compression (without $O(d)$ bound however). On the other hand, a maximal concept class with VC-dim$(\mathcal{ALT}(X; \mathcal{C})) = \infty$ cannot be embedded in an NIP structure without an increase in the VC-dimension.

Finally we wish to discuss further directions toward a complete solution to the problem. Due to our extensive use of Ramsey's theorem, it would be challenging to use our methods to obtain VC-dim bounds on compression schemes. However the methods presented in this paper might be used for the construction of compression schemes in general, in particular definable compression schemes. Theorem 20 due to Matoušek is part of a family of Helly type theorems. An analogue theorem exists, for example, for convex sets whose VC-dimension is unbounded (see, Alon and Kleitman (1992)). Thus, our requirements from $\mathcal{ALT}(X; C)$ may be relaxed or altered. Other concept classes derived from the original concept class might also be exploited. Finally, it would be interesting to know if there are necessary conditions to the existence of compression schemes that relate to NIP structures. This might enable the construction of a counter example to the compression conjecture.

Acknowledgments

The authors would like to thank Amir Globerson for some helpful comments and discussions, and to the first anonymous reviewer for some suggestions and simplifications in the proofs.

References

H. Adler. An introduction to theories without the independence property. *Archive for Mathematical Logic*, 2008.

N. Alon and D.J. Kleitman. Piercing convex sets and the hadwiger-debrunner (p, q)-problem. *Advances in Mathematics*, 96(1):103–112, 1992.

S. Ben-David and A. Litman. Combinatorial variability of vapnik-chervonenkis classes with applications to sample compression schemes. *Discrete Applied Mathematics*, 86(1):3–25, 1998.

A. Chernikov and P. Simon. Externally definable sets and dependent pairs. *Israel Journal of Mathematics*, pages 1–17, 2010.

A. Chernikov and P. Simon. Externally definable sets and dependent pairs ii. *arXiv preprint arXiv:1202.2650*, 2012.

S. Floyd. Space-bounded learning and the vapnik-chervonenkis dimension. In *Proceedings of the second annual workshop on Computational learning theory*, pages 349–364. Morgan Kaufmann Publishers Inc., 1989.

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

Y. Freund. Boosting a weak learning algorithm by majority. *Information and computation*, 121(2):256–285, 1995.

R.L. Graham, B.L. Rothschild, and J.H. Spencer. *Ramsey theory*, volume 2. Wiley, 1990.

Vincent Guingona. On uniform definability of types over finite sets. *Journal of Symbolic Logic*, 77(2):499–514, 2012.

H. R. Johnson and M. C. Laskowski. Compression schemes, stable definable families, and o-minimal structures. *Discrete Comput. Geom.*, 43(4):914–926, 2010. ISSN 0179-5376. doi: 10.1007/s00454-009-9201-3. URL http://dx.doi.org/10.1007/s00454-009-9201-3.

M. Karpinski and A. Macintyre. Polynomial bounds for vc dimension of sigmoidal neural networks. In *Proceedings of the twenty-seventh annual ACM symposium on Theory of computing*, pages 200–208. ACM, 1995.

D. Kuzmin and M. Warmuth. Unlabeled compression schemes for maximum classes. *Learning Theory*, pages 801–814, 2005.

M.C Laskowski. Vapnik-chervonenkis classes of definable sets. *J. London Math. Soc*, 45: 377–384, 1992.

N. Littlestone and M. Warmuth. Relating data compression and learnability. Technical report, Technical report, University of California, Santa Cruz, 1986.

A. Macintyre and E.D. Sontag. Finiteness results for sigmoidal "neural" networks. In *Proceedings of the twenty-fifth annual ACM symposium on Theory of computing*, pages 325–334. ACM, 1993.

J. Matoušek. Bounded vc-dimension implies a fractional helly theorem. *Discrete & Computational Geometry*, 31(2):251–255, 2004.

B.I.P. Rubinstein and J.H. Rubinstein. A geometric approach to sample compression. *The Journal of Machine Learning Research*, 98888:1221–1261, 2012.

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.

S. Shelah. Stability, the f.c.p., and superstability; model theoretic properties of formulas in first order theory. *Ann. Math. Logic*, 3(3):271–362, 1971. ISSN 0168-0072.

P Simon. Lecture notes on nip theories. ArXiv: 1208.3944, 2012.

P. Speissegger. Pfaffian sets and o-minimality. In Chris Miller, Jean-Philippe Rolin, and Patrick Speissegger, editors, *Lecture Notes on O-Minimal Structures and Real Analytic Geometry*, volume 62 of *Fields Institute Communications*, pages 179–218. Springer New York, 2012. ISBN 978-1-4614-4041-3. doi: 10.1007/978-1-4614-4042-0_5. URL http://dx.doi.org/10.1007/978-1-4614-4042-0_5.

V.N. Vapnik and A.Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, 1971.

Appendix A.

A.1. A proof of Theorem 13

Let $\boldsymbol{x}$ be a $(d+1)$ homogeneous sequence with $\text{alt}(\boldsymbol{x};\mathcal{C}) \geq 2d+2$. We will show that $\text{VC-dim}(\mathcal{C}) \geq d+1$. Without loss of generality we may assume that $\boldsymbol{x}$ is a sequence of length $2d+2$, and there is a concept C such that $C(x_i) = 1$ if and only if i is even. Given a boolean sequence $\boldsymbol{b} \in \{0,1\}^{d+1}$ we construct a subsequence $\boldsymbol{y}$ of $\boldsymbol{x}$ such that $C(\boldsymbol{y}) = \boldsymbol{b}$.

Choose $y_1 = x_1$ if $b_1 = 0$ and $y_1 = x_2$ otherwise. We define y_n inductively for every $n \leq d+1$. Assume we chose x_i to be y_{n-1}. Let $y_n = x_{i+2}$ if $b_n = b_{n-1}$, and $y_n = x_{i+1}$ if $b_n \neq b_{n-1}$.

By homogeneity there exists a concept C_2 such that $C_2(x_1,\ldots,x_{d+1}) = C(\boldsymbol{y}) = \boldsymbol{b}$. As $\boldsymbol{b}$ was arbitrary, $x_1,\ldots,x_{d+1}$ is shattered.

A.2. A proof of Lemma 16

Let $(X,\mathcal{C})$ be a concept class with VC-dimension d. Given m and $y \in X$. If $y \notin \text{ALT}(\boldsymbol{x};\mathcal{C}) \neq \emptyset$ for some $\text{ALT}(\boldsymbol{x};\mathcal{C}) \in \mathcal{ALT}_m(X;\mathcal{C})$ then $\boldsymbol{x}$ has a subsequence $\hat{\boldsymbol{x}} = (x_{i_1}, x_{i_2}, \ldots, x_{i_{2d+1}})$ such that

$$\text{alt}((y, x_{i_1}, \ldots, \ldots, x_{2d+1});\mathcal{C}) = 2d+2. \tag{2}$$

In other words $y \notin \text{ALT}(\boldsymbol{x};\mathcal{C})$ if and only if $\boldsymbol{x}$ has a subsequence such that

$$y \notin \text{ALT}((x_{i_1},\ldots,x_{i_2},\ldots,x_{i_{2d+1}});\mathcal{C}).$$

We've shown that we can describe every element in $\mathcal{ALT}_m(X;\mathcal{C})$ as the finite intersection of $\binom{m}{2d+1}$ elements in $\mathcal{ALT}_{2d+1}(X;\mathcal{C})$. The result follows immediately.

A.3. Derivation of Theorem 20

The (p,k) Theorem for bounded VC dimension concept classes is formalized as follows:

Theorem 26 *Let $\mathcal{F}$ be a set system with $\pi^*_{\mathcal{F}}(m) = o(m^k)$ for some integer k, and let $p \geq k$. Then there is a constant N such that the following holds for every finite family $\mathcal{G} \subset \mathcal{F}$: If $\mathcal{G}$ has the (p,k)-property, meaning amongst every p sets of $\mathcal{G}$, some k intersect, then there is an N-point set intersecting all sets of $\mathcal{G}$.*

Recall that the dual shatter function $\pi^*_{\mathcal{F}}(m)$ is defined to be the maximum number of nonempty fields of the Venn diagram of m sets of $\mathcal{F}$.

Given a non degenerate concept class $(X, \mathcal{C})$ with VC dimension d, consider the set system

$$\mathcal{F} = \{\mathcal{F}_x : \mathcal{F}_x = \{C \in \mathcal{C} : x \in C\}\}.$$

By the definition of the dual shatter function and Sauer's Lemma, it is easy to see that $\pi^*(\mathcal{F}) = o(m^k)$ for any $k > d$.

For a finite set $G \subseteq X$ we associate a finite set $\mathcal{G} \subseteq \mathcal{F}$ by

$$\mathcal{G} = \{\mathcal{F}_x : x \in G\}.$$

It is easy to see that if G has the (p,k) property with respect to $(X, \mathcal{C})$ then $\mathcal{G} \subseteq \mathcal{F}$ has the (p,k) property (i.e. every subset of cardinality p of $\mathcal{G}$, some k intersect). By Matousek's theorem there are N concepts $C_i \in \mathcal{C}$ such that every $\mathcal{F}_x \in \mathcal{G}$ contains some C_i. In other words $G \subseteq \cup C_i$.

The N given so far is not uniform for all concept classes of VC dimension smaller then k. To see that it can be chosen uniformly, simply apply the result to a concept class $(X, \mathcal{C})$ that contains all possible finite concept classes with VC dimension smaller then k. One can construct such a universal concept class by simply taking the disjoint countable union of all such concept classes. Since any finite set is embeddable in such a concept class, we obtain a uniform N.