MODEL THEORY AND COMBINATORICS: CHAPTER 2 (DRAFT)

ARTEM CHERNIKOV

Last updated: 02/28/2017.

Contents

1. Vapnik–Chervonenkis dimension

1.1. **Basic properties and examples.** Let X be a set (finite or infinite), and let $\mathcal{F}$ be a family of subsets of X. A pair $(X, \mathcal{F})$ is called a *set system.*

Given $A \subseteq X$, we say that the family $\mathcal{F}$ *shatters* A if for every $A' \subseteq A$, there is a set $S \in \mathcal{F}$ such that $S \cap A = A'$.

The family $\mathcal{F}$ has *VC-dimension* at most n (written as $\mathrm{VC}(\mathcal{F}) \leq n$), if there is no $A \subseteq X$ of cardinality $n+1$ such that $\mathcal{F}$ shatters A. We say that $\mathcal{F}$ is of VC-dimension n if it is of VC-dimension at most n and shatters some subset of size n.

If for every $n \in \mathbb{N}$ we can find a subset of X of cardinality n shattered by $\mathcal{F}$, then we say that $\mathcal{F}$ has infinite VC-dimension ($\mathrm{VC}(\mathcal{F}) = \infty$). If $\mathrm{VC}(\mathcal{F})$ is finite, we say that $\mathcal{F}$ is a *VC-family.* Note that if $\mathcal{F}' \subseteq \mathcal{F}$ then $\mathrm{VC}(\mathcal{F}') \leq \mathrm{VC}(\mathcal{F})$.

Example 1.1. Let $X = \mathbb{R}$ and let $\mathcal{F}$ be the family of all unbounded intervals. Then $\mathcal{F}$ has VC-dimension 2. Clearly any two-element set can be shattered by $\mathcal{F}$. However, if we take any $a < b < c$, then $\{a, b, c\}$ cannot be shattered by $\mathcal{F}$.

Exercise 1.2. Let $X = \mathbb{R}^2$, and let $\mathcal{F}$ be the set of all half-spaces. Show that $\mathrm{VC}(\mathcal{F}) = 3$.

Exercise 1.3. Let $X = \mathbb{R}^2$ and let $\mathcal{F}$ be the set of all convex polygons. Show that $\mathrm{VC}(\mathcal{F}) = \infty$.

Exercise 1.4. Show that for every $d \in \mathbb{N}$, there is a convex set on the real plane such that the family of all of its isometric copies has VC-dimension at least d.

For a finer analysis, we define the *shatter function* $\pi_{\mathcal{F}} : \mathbb{N} \to \mathbb{N}$ associated to the family $\mathcal{F}$ as follows. For a set $A \subseteq X$ we let $\mathcal{F} \cap A := \{S \cap A : S \in \mathcal{F}\}$. Then we define $\pi_{\mathcal{F}}(n) := \max\{|\mathcal{F} \cap A| : A \subseteq X, |A| = n\}$.

Note that $\pi_{\mathcal{F}}(n) \leq 2^n$, and that $\mathrm{VC}(\mathcal{F}) < n \iff \pi_{\mathcal{F}}(m) < 2^m$ for all $m \geq n$. The following fundamental lemma states that either $\pi_{\mathcal{F}}(n) = 2^n$ for all $n \in \mathbb{N}$, or $\pi_{\mathcal{F}}(n)$ has polynomial growth.

Lemma 1.5. *(Sauer-Shelah lemma) Let $(X, \mathcal{F})$ be a set system of VC-dimension at most k. Then, for all $n \geq k$, we have $\pi_{\mathcal{F}}(n) \leq \sum_{i=0}^{k} \binom{n}{i}$.*

In particular, $\pi_{\mathcal{F}}(n) \leq \left(\frac{en}{k}\right)^k = O\left(n^k\right)$.

There are numerous proofs and generalizations, we give a proof using the so-called “shifting” technique. Notice that the bound is tight: take $\mathcal{F}$ to be the family of all subsets of X of cardinality $\leq k$, then $\mathcal{F}$ has VC-dimension exactly k and its shatter function is equal to the bound in the statement of the lemma. The idea of the proof is to reduce the general situation to this case by modifying the elements of $\mathcal{F}$ making them as small as possible without changing the cardinality or VC-dimension of $\mathcal{F}$.

Proof. Fix an integer $n \geq k$. If $\mathcal{F}$ contradicts the bound, then this is also true for some finite subfamily of $\mathcal{F}$, so for the proof we may assume that $\mathcal{F}$ is finite. Similarly we may assume that X is finite, say, $X = \{x_1, \ldots, x_n\}$, and $\pi_{\mathcal{F}}(n) = |\mathcal{F}|$.

We define recursively families $\mathcal{F}_0, \ldots, \mathcal{F}_n$ of subsets of X. Set $\mathcal{F}_0 = \mathcal{F}$.

Let $l < n$ and assume that $\mathcal{F}_l$ has been defined. Go through the sets in $\mathcal{F}_l$ one by one. For each $S \in \mathcal{F}_l$, if $x_{l+1} \in S$ and $S \setminus \{x_{l+1}\}$ is not in $\mathcal{F}_l$, replace S by $S \setminus \{x_{l+1}\}$. If not, leave S as it is. Let $\mathcal{F}_{l+1}$ be the resulting family.

The following is straightforward by construction:

(1) for each l, $|\mathcal{F}_{l+1}| = |\mathcal{F}_l|$;
(2) let $S \in \mathcal{F}_l$ and $A = S \cap \{x_1, \ldots, x_l\}$, then for every $A_0 \subseteq A$, the set $A_0 \cup (S \setminus A)$ is in $\mathcal{F}_l$;
(3) any $A \subseteq X$ shattered by $\mathcal{F}_{l+1}$ is also shattered by $\mathcal{F}_l$.

It follows from (2) that if $S \in \mathcal{F}_n$, then S is shattered by $\mathcal{F}_n$. It follows from (3) that the $\mathrm{VC}(\mathcal{F}_n) \leq \mathrm{VC}(\mathcal{F})$. Therefore no set in $\mathcal{F}_n$ can have cardinality greater than k. Hence, by (3) we have $\sum_{i=0}^{k} \binom{n}{i} \geq |\mathcal{F}_n| = |\mathcal{F}| = \pi_{\mathcal{F}}(n)$. □

Exercise 1.6. Prove a more general fact, due to Pajor: every finite set system $(X, \mathcal{F})$ shatters at least $|\mathcal{F}|$ subsets of X. Deduce Sauer-Shelah lemma from it.

We consider some general ways of producing VC-families.

Exercise 1.7. (Boolean operations preserve finite VC-dimension) Let $\mathcal{F}_1, \mathcal{F}_2$ be two families of subsets of X of finite VC-dimension. Show that all of the following families have finite VC-dimension:

(1) $\mathcal{F} := \mathcal{F}_1 \cup \mathcal{F}_2$,
(2) $\mathcal{F}_\cap := \{S_1 \cap S_2 : S_i \in \mathcal{F}_i, i = 1, 2\}$,
(3) $\mathcal{F}_\cup := \{S_1 \cup S_2 : S_i \in \mathcal{F}_i, i = 1, 2\}, \mathcal{F}_1^c := \{X \setminus S_1 : S_1 \in \mathcal{F}_1\}$,
(4) $\mathcal{F}_1 \times \mathcal{F}_2 := \{S_1 \times S_2 : S_1 \in \mathcal{F}_1, S_2 \in \mathcal{F}_2\}$ — a family of subsets of $X \times X$.
(5) Besides, if X' is an infinite set and $f : X' \to X$ is a map, let $f^{-1}(\mathcal{F}_1) := \left\{f^{-1}(S) : S \in \mathcal{F}_1\right\}$. Then $\mathrm{VC}\left(f^{-1}(\mathcal{F}_1)\right) \leq \mathrm{VC}(\mathcal{F}_1)$.

(Hint: bound the shattering functions of the corresponding families in terms of $\pi_{\mathcal{F}_1}, \pi_{\mathcal{F}_2}$ and use Lemma 1.5.)

Definition 1.8. Given a set system $(X, \mathcal{F})$, we define the *dual set system* $(X^*, \mathcal{F}^*)$, where $X^* = \mathcal{F}$ and $\mathcal{F}^* = \{\mathcal{F}_a : a \in X\}$ with $\mathcal{F}_a = \{S \in \mathcal{F} : a \in S\}$. We then define the *dual VC-dimension* of $\mathcal{F}$ (written as $\mathrm{VC}^*(\mathcal{F})$) as the VC-dimension of $\mathcal{F}^*$, and the *dual shatter function* $\pi^*_{\mathcal{F}}$ as the shatter function of $\mathcal{F}^*$.

Given a set system $(X, \mathcal{F})$, we can consider its *incidence matrix* M defined as an $|X| \times |\mathcal{F}|$-matrix such that rows are identified with the elements of x, columns with the elements of $\mathcal{F}$ and for any $x \in X, S \in \mathcal{F}$ the corresponding entry $M_{x,S}$ is 1 if $x \in S$ and 0 otherwise. Note that if M is the incidence matrix for $(X, \mathcal{F})$ then the transposed matrix M^{T} is the incidence matrix of the dual system $(X^*, \mathcal{F}^*)$.

Exercise 1.9. Let $X = \mathbb{R}^2$ and let $\mathcal{F}$ be the family of all open half-planes. Then $\pi^*_{\mathcal{F}}(n)$ is the maximal number of regions into which n lines can partition the plane. Show by induction that this number is equal to $\frac{n(n+1)}{2} + 1$.

Lemma 1.10. *(VC duality) We have* $\mathrm{VC}^*(\mathcal{F}) < 2^{\mathrm{VC}(\mathcal{F})+1}$ *and* $\mathrm{VC}(\mathcal{F}) < 2^{\mathrm{VC}^*(\mathcal{F})+1}$.

Proof. Assume that $\mathrm{VC}(\mathcal{F}) \geq 2^n$. Then there is some subset $B \subseteq X$ of size 2^n shattered by $\mathcal{F}$. Write $B = \{b_J : J \subseteq n\}$, and we have $\{S_I : I \subseteq \mathcal{P}(n)\} \subseteq \mathcal{F}$ such that $b_J \in S_I \iff J \in I$. For each $k < n$, let $I_k = \{J \subseteq n : k \in J\}$. Then $J \in I_k \iff k \in J$, hence we have $b_J \in S_{I_k} \iff k \in J$. This shows that the set $\{S_{I_k} : k < n\} \subseteq X^*$ is shattered by $\{\mathcal{F}_{b_J} : J \subseteq n\} \subseteq \mathcal{F}^*$, so $\mathrm{VC}^*(\mathcal{F}) \geq n$.

This proves the second inequality, the first one is proved similarly. □

Exercise 1.11. Show that this bound is optimal.

1.2. **Definable families and Shelah's reduction to one variable.** We describe a large source of examples of VC-families coming from model theory.

A *(first-order) structure* $\mathcal{M} = (M, R_1, R_2, \ldots, f_1, f_2, \ldots, c_1, c_2, \ldots)$ consists of an underlying set M, together with some distinguished relations R_i (subsets of M^{n_i}, $n_i \in \mathbb{N}$), functions $f_i : M^{n_i} \to M$, and constants c_i (distinguished elements of M). We refer to the collection of all these relations, function symbols and constants as *the signature* of $\mathcal{M}$. For example, a group is naturally viewed as a structure $\left(G, \cdot, ^{-1}, 1\right)$, as well as a ring $(R, +, \cdot, 0, 1)$, ordered set $(X, <)$, graph (X, E), etc. A *formula* is an expression of the form $\psi(y_1, \ldots, y_m) = Q_1 x_1 \ldots Q_n x_n \phi(x_1, \ldots, x_n; y_1, \ldots, y_n)$, where $Q_i \in \{\forall, \exists\}$ and ϕ is given by a boolean combination of (superpositions of) the basic relations and functions (and $y_{1,} \ldots, y_n$ are the *free variables* of ψ). We denote the set of all formulas by L. By a partitioned formula $\phi(\bar{x}, \bar{y})$ we mean a formula with its free variables partitioned into two groups $\bar{x}$ (object variables) and $\bar{y}$ (parameter variables). Given a partitioned formula $\phi(\bar{x}, \bar{y})$ and $\bar{b} \in M^{|\bar{y}|}$, we let $\phi\left(M^{|\bar{x}|}, \bar{b}\right)$ be the set of all $\bar{a} \in M^{|\bar{x}|}$ such that $\mathcal{M} \models \phi\left(\bar{a}, \bar{b}\right)$. Sets of this form are called *definable* (or *ϕ-definable*, in this case). We consider the family $\mathcal{F}_{\phi(\bar{x}, \bar{y})}$ of subsets of $M^{|\bar{x}|}$ defined by $\mathcal{F}_{\phi(\bar{x}, \bar{y})} = \left\{\phi\left(M^{|\bar{x}|}, \bar{b}\right) : \bar{b} \in M^{|\bar{y}|}\right\}$.

Example 1.12. Let $G = (V, E)$ be a graph. Then we can consider the formula $E(x, y)$, and for every $v \in V$, then set $E(V, v)$ is the set of all elements connected to v, and the family $\mathcal{F}_{E(x,y)}$ is the family of all neighborhoods of vertices in G. Similarly, for any $k \in \mathbb{N}$ let

$$d_k(x, y) := \exists z_0 \ldots \exists z_{k-1} \left(z_0 = x \wedge z_{k-1} = y \wedge \bigwedge_{1 \leq i \leq k} E(z_{i-1}, z_i) \right).$$

Let $D_k(x, y) := \bigvee_{l \leq k} d_l(x, y)$. Then for every $v \in V$, $D_k(V, v)$ is the set of all vertices at distance at most k from v. But there is no first-order formula expressing that x is in the connected component of y (Exercise).

Theorem 1.13. *(Reduction to formulas in a single variable, Shelah) Let M be a first-order structure. Assume that for every partitioned formula $\phi(x, \bar{y})$ with x a **singleton**, the family $\mathcal{F}_\phi$ has finite VC dimension. Then for any $\phi(\bar{x}, \bar{y}) \in L$, the corresponding family $\mathcal{F}_\phi$ has finite VC dimension.*

Before proving it, we point out some examples of VC-families that it easily applies to.

Example 1.14. (Semialgebraic sets of bounded complexity) Recall that a set $X \subseteq \mathbb{R}^n$ is *semialgebraic* if it is given by a Boolean combination of polynomial equalities and inequalities.

We say that the *description complexity* of a semialgebraic set $X \subseteq \mathbb{R}^d$ is bounded by $t \in \mathbb{N}$ if $d \leq t$ and X can be defined as a Boolean combination of at most t polynomial equalities and inequalities, such that all of the polynomials involved have degree at most t. For example, consider the family of all spheres in $\mathbb{R}^n$, or all cubes in $\mathbb{R}^n$, etc.

We claim that for any t, the family $\mathcal{F}_t$ of all semialgebraic sets of complexity $\leq t$ has finite VC-dimension. To see this, consider the field of real numbers as a first-order structure $\mathcal{M} = (\mathbb{R}, +, \times, 0, 1, <)$. Note that $\mathcal{F}_t$ is contained in the union of finitely many families of the form $\left\{\mathcal{F}_{\phi_i(\bar{x}, \bar{y})} : i < t'\right\}$ where t' only depends on t (since there are only finitely many different polynomials of degree $\leq t$, up to varying coefficients, and only finitely many different Boolean combinations of size $\leq t$). So it is enough to show that every such family has finite VC-dimension (by Exercise 1.7).

By the classical result of Tarski, this structure $\mathcal{M}$ eliminates quantifiers, and so definable sets are precisely the semialgebraic ones. In particular, if we are given a formula of the form $\phi(x, \bar{y})$, for every $b \in M^{|\bar{y}|}$ the set $\phi(M, b)$ is just a union of at most n_ϕ intervals and points, where n_ϕ only depends on ϕ. As the collection of all intervals has finite VC-dimension, in view of Exercise 1.7 we have that for all formulas $\phi(x, \bar{y})$ with $|x| = 1$, $\mathcal{F}_\phi$ has finite VC-dimension. By Theorem 1.13 this implies that the same is true for all formulas.

Remark 1.15. In particular, let $X = \mathbb{R}$ and let $\mathcal{F}$ be the family of all convex n-gons. Then $\mathcal{F}$ has finite VC-dimension (one can verify that $\mathrm{VC}(\mathcal{F}) = 2n+1$) — compare this to Exercise 1.3.

Example 1.16. More generally, definable families in arbitrary o-minimal structures have finite VC-dimension.

A structure $\mathcal{M} = (M, <, \ldots)$ is o-minimal if every definable subset of M is a finite union of singletons and intervals (with endpoints in $M \cup \{\pm\infty\}$). From this assumption one obtains cell decomposition for definable subsets of M^n, for all (see [43] for a detailed treatment of o-minimality, or [41, Section 3] and references there for a quick introduction). Examples of o-minimal structures include (in each of these cases it is a highly non-trivial theorem): $(\mathbb{R}, +, \times, 0, 1, <)$, $\mathbb{R}_{\exp} = (\mathbb{R}, +, \times, e^x)$, $\mathbb{R}_{\mathrm{an}} = \left(R, +, \times, f \restriction_{[0,1]^k}\right)$ for f ranging over all functions real-analytic on some neighborhood of $[0,1]^k$, or the combination of both $\mathbb{R}_{\mathrm{an,exp}}$.

The same argument as in the previous example shows that all definable families in o-minimal structures have finite VC-dimension.

Exercise 1.17. Show that the family $\mathcal{F} = \{X_\lambda : \lambda \in \mathbb{R}\}$ where

$$X_\lambda := \left\{x \in \mathbb{R} : x \geq 0, 0 \leq \frac{x^\lambda - 1}{\lambda}\right\}$$

has finite VC-dimension.

Example 1.18. Definable families in stable structures.

The class of stable structures is well studied in model theory, originating from Morley's theorem and Shelah's work on classification theory. See e.g. [16] for more details. Examples of stable structures:

- $(\mathbb{C}, \times, +, 0, 1)$ (definable sets correspond to the constructible sets, i.e. Boolean combinations of algebraic sets),
- separably closed and differentially closed fields,
- arbitrary planar graphs $G = (V, E)$,
- abelian groups (viewed as structures in the pure group language $(G, \cdot, 1)$),
- [Z. Sela] non-commutative free groups (in the pure group language).

Example 1.19. [21] Let $(G, \cdot, <)$ be an arbitrary ordered abelian group. Then definable families of sets have finite VC-dimension. In particular, all definable families in Presburger arithmetic $(\mathbb{Z}, +, <)$ have finite VC-dimension.

Example 1.20. Let $(\mathbb{Q}_p, \times, +, 0, 1)$ be the field of p-adics. Using quantifier elimination results of Macintyre in this setting, one can show that again all definable families have finite VC-dimension.

Now we work towards a proof of Theorem 1.13. First recall a classical theorem of Ramsey generalizing the pigeon-hole principle.

Theorem 1.21. *(Ramsey) For any $l, k, n \in \mathbb{N}$ there is some $N \in \mathbb{N}$ such that $N \to (n)_l^k$ holds, i.e. for any set A with $|A| \geq N$ and any coloring of k-subsets of A in l colors, $f : \binom{A}{k} \to l$, there is a homogeneous subset $B \subseteq A$ of size $\geq n$, where* homogeneous *means that the value of f is constant on all k-subsets of B.*

Let us fix a first-order structure $\mathcal{M}$. Given a formula $\phi(\bar{x}_1, \ldots, \bar{x}_k)$ with $|\bar{x}_i| = m$ and a sequence $(\bar{a}_i : i \in I)$ of elements of M^m, where I is an arbitrary linearly ordered set, we say that this sequence is *ϕ-indiscernible* if for any $i_1 < \ldots < i_k$ and $j_0 < \ldots < j_k$ from I we have that $\phi(\bar{a}_{i_1}, \ldots, \bar{a}_{i_k}) \iff \phi(\bar{a}_{j_1}, \ldots, \bar{a}_{j_k})$ holds in $\mathcal{M}$. Given a set of such formulas Δ, we say that $(\bar{a}_i : i \in I)$ is *Δ-indiscernible* if it is ϕ-indiscernible for every $\phi \in \Delta$ such that the arities of its variables are compatible with the arities of the tuples in the sequence.

Note that if Δ is a finite set of formulas, then its closure under arbitrary repartitions of the variables of formulas in it is also finite. Thus, in all the arguments below we may assume that our finite sets of formulas are closed under choosing a different partition of the variables.

Lemma 1.22. *For every **finite** set of formulas Δ and every $n \in \mathbb{N}$ there is some $N \in \mathbb{N}$ such that every sequence $(\bar{a}_i : i < N)$ of m-tuples from M of length $\geq N$ contains a Δ-indiscernible subsequence of length $\geq n$.*

Proof. Follows by an iterated application of Ramsey theorem. Let n be fixed. Let's say $\Delta = \{\phi_1, \ldots, \phi_r\}$ with $\phi_i(\bar{x}_1, \ldots, \bar{x}_{k_i})$ for $i = 1, \ldots, r$. Let $k = \max\{k_i : 1 \leq i \leq r\}$. We find an N as wanted by induction on r.

Assume first that $\Delta = \{\phi_1\}$. Let N_0 be as given by Theorem 1.21 such that $N_0 \to (n)_2^k$ holds. Let $(\bar{a}_i : i < N_0)$ from M^m be arbitrary. Consider a coloring of all increasing k-tuples from this sequence in two colors, such that the color of the tuple $(\bar{a}_{i_1}, \ldots, \bar{a}_{i_k})$ corresponds to the truth value of $\phi(\bar{a}_{i_1}, \ldots, \bar{a}_{i_k})$. Then by Theorem 1.21 there is a homogeneous subset of size $\geq n$, which corresponds to a ϕ-indiscernible subsequence of length $\geq n$.

Assume now that we have found a number N_r that works for $\{\phi_1, \ldots, \phi_r\}$, and let $\Delta = \{\phi_1, \ldots, \phi_r, \phi_{r+1}\}$ be given. Let N_{r+1} be as given by Theorem 1.21 such that $N_{r+1} \to (N_r)_2^k$ holds. Again, it follows that an arbitrary sequence $(\bar{a}_i : i < N_{t+1})$ contains a ϕ_{r+1}-indiscernible subsequence $(\bar{a}_i : i \in I)$ for some $I \subseteq N, |I| \geq N_r$. Applying the inductive assumption to the sequence $(\bar{a}_i : i \in I)$ and $\{\phi_1, \ldots \phi_r\}$, we find a Δ-indiscernible subsequence of length $\geq n$. □

Remark 1.23. The dual set system for $\left(M^{|x|}, \mathcal{F}_{\phi(\bar{x},\bar{y})}\right)$ is given by $\left(M^{|y|}, \mathcal{F}_{\phi^*(\bar{y},\bar{x})}\right)$ and $\phi^*(\bar{y}, \bar{x}) = \phi(\bar{x}, \bar{y})$, i.e. we exchange the roles of the object variables and the parameter variables in the partitioned formula (see Definition 1.8).

From now on, by the *VC-dimension of a formula* $\mathrm{VC}(\phi)$ we mean $\mathrm{VC}(\mathcal{F}_\phi)$ (and the same for the dual VC-dimension, shatter function, etc.).

Lemma 1.24. *Let $\phi(\bar{x}, \bar{y})$ be a partitioned formula, and assume that* $\mathrm{VC}(\phi) = \infty$. *Then for any finite set of formulas Δ and every $n \in \mathbb{N}$ there is a Δ-indiscernible sequence $(\bar{a}_i : i < n)$ from $M^{|\bar{y}|}$ and a tuple $\bar{b} \in M^{|x|}$ so that $M \models \phi\left(\bar{b}, \bar{a}_i\right)$ if and only if i is even.*

Proof. Note that if $\mathrm{VC}(\mathcal{F}_\phi) = \infty$, then also $\mathrm{VC}(\mathcal{F}_{\phi^*}) = \infty$ (see Lemma 1.10 and Remark 1.23). This means that for any $N \in \mathbb{N}$ we can find some set $A = (\bar{a}_i : i < N)$ in $M^{|\bar{y}|}$ of size $\geq N$ which is shattered by the family $\mathcal{F}_{\phi^*} = \left\{\phi\left(\bar{b}, M^{|\bar{y}|}\right) : \bar{b} \in M^{|\bar{x}|}\right\}$. Then taking N sufficiently large, Lemma 1.22 ensures that A contains a Δ-indiscernible subsequence $(\bar{a}_i : i < n)$ of length n, and this sequence is still shattered by $\mathcal{F}_{\phi^*}$, in particular there is some $\bar{b}$ such that $\models \phi\left(\bar{b}, \bar{a}_i\right)$ iff i is even. □

Now we give a converse to Lemma 1.24. The point is that a formula of finite VC-dimension cannot cut out the set of even members from a *sufficiently* long and indiscernible sequence. What does "sufficiently" mean here?

Given $\phi(\bar{x}, \bar{y})$, $n \in \mathbb{N}$ and $w \subseteq n$, let

$$\theta_w^\phi(\bar{x}, \bar{y}_0, \ldots, \bar{y}_{n-1}) := \bigwedge_{i \in w} \phi(\bar{x}, \bar{y}_i) \wedge \bigwedge_{i \in n \setminus w} \neg\phi(\bar{x}, \bar{y}_i),$$

$$\rho_w^\phi(\bar{y}_0, \ldots, \bar{y}_{n-1}) := \exists \bar{x} \theta_w^\phi(\bar{x}, \bar{y}_0, \ldots, \bar{y}_{n-1}),$$

$$\Delta_n^\phi := \left\{\rho_w^\phi : w \subseteq n\right\}.$$

Lemma 1.25. *Let $\phi(\bar{x}, \bar{y})$ be a formula with $\mathrm{VC}^*(\phi) \leq d$. Then if $(\bar{a}_i : i < N)$ is a Δ_d^ϕ-indiscernible sequence from $M^{|\bar{y}|}$ and $\bar{b} \in M^{|\bar{x}|}$, there do not exist $i_0 < \ldots < i_{2d-1} < N$ so that $\phi\left(\bar{b}, \bar{a}_{i_j}\right)$ holds if and only if j is even.*

Proof. Assume that $(\bar{a}_i : i < N)$ is a Δ_d^ϕ-indiscernible sequence from $M^{|\bar{y}|}$ and there are $i_0 < \ldots < i_{2d-1}$ and $\bar{b} \in M^{|\bar{x}|}$ such that $M \models \phi\left(\bar{b}, \bar{a}_{i_j}\right)$ iff j is even. We claim that the sequence $(\bar{a}_i : i < d)$ is shattered by the family $\left\{\phi\left(\bar{b}, M^k\right) : \bar{b} \in M^{|\bar{x}|}\right\}$, which would give a contradiction.

Indeed, for any $w \subseteq d$, define a function $f_w : \{0, \ldots, d-1\} \to \{i_0, \ldots, i_{2d-1}\}$ by $f_w(j) := i_{2j}$ if $j \in w$ and $f_w(j) = i_{2j+1}$ if $j \notin w$. Now $\theta_w^\phi\left(\bar{b}, \bar{a}_{f_w(0)}, \ldots, \bar{a}_{f_w(d-1)}\right)$ holds in $\mathcal{M}$, so $\rho_w^\phi\left(\bar{a}_{f_w(0)}, \ldots, \bar{a}_{f_w(d-1)}\right)$ holds in $\mathcal{M}$. However, as $f_w(0) < \ldots < f_w(d-1)$ and $(\bar{a}_i : i < N)$ is Δ_n^ϕ-indiscernible, so in particular ρ_w^ϕ-indiscernible, this implies that $\rho_w^\phi(\bar{a}_0, \ldots, \bar{a}_{d-1})$ holds as well, i.e. there is some $\bar{b}' \in M^{|\bar{x}|}$ such that for all $i < d$ we have $\phi\left(\bar{b}, \bar{a}_i\right)$ holds in $\mathcal{M}$ iff $i \in w$. Since this works for any $w \subseteq d$, we conclude. □

Now we amplify this by finding a large chunk of our sequence that is indiscernible "over $\bar{b}$".

Lemma 1.26. *Let Δ be a finite set of formulas such that for every $\phi(\bar{x}, \bar{y}) \in \Delta$ with $|\bar{x}| \leq l$ we have $\mathrm{VC}^*(\phi) \leq d$. Then for any $n \in \mathbb{N}$ there is a finite set of formulas Δ' and $N \in \mathbb{N}$ such that if $(\bar{a}_i : i < N)$ is a Δ'-indiscernible sequence and $\bar{b} \in M^{|x|}$ is an arbitrary tuple of length $\leq l$, then for some **interval** $I \subseteq N$ with $|I| \geq n$, the sequence of $(|x| + |y|)$-tuples $\left(\bar{a}_i\bar{b} : i \in I\right)$ is Δ-indiscernible.*

Proof. Let us take $\Delta' = \bigcup_{\phi \in \Delta} \Delta_d^\phi$, let $N' = |\Delta|(2d-1)$ and $N = N'n$.

Let $(\bar{a}_i : i < N)$ and $\bar{b} \in M^{|x|}$ be arbitrary.

Let us partition N into N'-many consecutive intervals $(I_j : j < N')$, each of length n. Assume that for each of those intervals, the conclusion of the lemma does not hold. This means that for each $j < N'$ there is some $\phi_j \in \Delta$ and some

$i_0^j < \ldots < i_{m_j-1}^j, h_0^j < \ldots < h_{m_j-1}^j \in I_j$ such that $M \models \phi_j\left(\bar{b}, \bar{a}_{i_0^j}, \ldots, \bar{a}_{i_{m_j-1}^j}\right) \wedge \neg\phi_j\left(\bar{b}, \bar{a}_{h_0^j}, \ldots, \bar{a}_{h_{m_j-1}^j}\right)$.

By the choice of N', throwing away some intervals, we may assume that ϕ_j is constant for all $j < 2d-1$, say $\phi_j = \phi'$ and $m_j = m'$.

For $j < 2d-1$, we consider the sequence of $(m' \, |\bar{y}|)$-tuples $\bar{a}'_j$ defined by taking $\bar{a}'_j$ to be $\bar{a}_{i_0^j} \ldots \bar{a}_{i_{m'-1}^j}$ if j is even, and to be $\bar{a}_{h_0^j} \ldots \bar{a}_{h_{m'-1}^j}$ if j is odd.

Observe that the sequence $\left(\bar{a}'_j : j < 2d-1\right)$ is still $\Delta_d^{\phi'}$-indiscernible since the original sequence was, and $\phi'\left(\bar{b}, \bar{a}'_j\right)$ holds iff j is even. By Lemma 1.25 this contradicts the assumption that $\mathrm{VC}^*(\phi') \leq d$. □

Iterating Lemma 1.26, we obtain:

Corollary 1.27. *Assume that all formulas in M in a single object variable have finite VC-dimension. Then for any Δ a finite set of formulas and $n, l \in \mathbb{N}$ there is a finite set of formulas Δ' and $N \in \mathbb{N}$ such that if $(\bar{a}_i : i < N)$ is a Δ'-indiscernible sequence and $\bar{b} \in M^l$ is an arbitrary l-tuple, then for some interval $I \subseteq N$ with $|I| \geq n$, the sequence of $(l\,|\bar{a}_i|)$-tuples $\left(\bar{a}_i \bar{b} : i \in I\right)$ is Δ-indiscernible.*

Proof. Define recursively $\Delta_0 = \Delta, N_0 = n$ and take N_{j+1}, Δ_{j+1} to be the N and Δ' given by Lemma 1.26 for $n = N_j, \Delta = \Delta_j$ and $d = \max\left\{\mathrm{VC}^*(\phi(x, \bar{y})) : \phi(x, \bar{y}) \in \Delta_j\right\}$. Let $\tilde{N} := N_{l-1}$ and $\tilde{\Delta} = \Delta_{l-1}$.

Let now $\left(\bar{a}_i : i < \tilde{N}\right)$ be $\tilde{\Delta}$-indiscernible and let $\bar{b} \in M^l$ be arbitrary, say $\bar{b} = b_0 \ldots b_{l-1}$. Applying Lemma 1.26 to $\left(\bar{a}_i : i < \tilde{N}\right)$ and b_0 we find some interval $I_{l-1} \subseteq \tilde{N}$ with $|I_{l-1}| \geq N_{l-1}$ such that the sequence $(\bar{a}_i b_0 : i \in I_{l-1})$ is Δ_{l-1}-indiscernible. Again applying Lemma 1.26 to this sequence of thicker tuples $(\bar{a}_i b_0 : i \in I_{l-1})$ and b_1, we find some interval $I_{l-2} \subseteq I_{l-1}$ with $|I_{l-2}| \geq N_{l-2}$ such that the sequence $(\bar{a}_i b_0 b_1 : i \in I_{l-2})$ is Δ_{l-2}-indiscernible. Continuing in this fashion we finally find an interval $I_0 \subseteq N$ with $|I_0| \geq n$ such that the sequence $(\bar{a}_i b_0 \ldots b_{l-1} : i \in I_0)$ is Δ-indiscernible, as wanted. □

Finally, to prove Theorem 1.13, assume that some formula $\phi(\bar{x}, \bar{y})$ has infinite VC-dimension. Let Δ be the finite collection of formulas given by taking arbitrary partitions of the variables in ϕ. Let Δ' and N be as given by Corollary 1.27 for Δ, $l = |\bar{x}|$ and $n = 2$.

Now on the one hand, by Lemma 1.24 we can find a Δ-indiscernible sequence $(\bar{a}_i : i < N)$ and $\bar{b}$ such that $\phi\left(\bar{b}, \bar{a}_i\right)$ holds iff i is even. On the other hand, by Corollary 1.27 we can find an interval $I \subseteq N$ of length ≥ 2 such that the sequence $\left\{\bar{a}_i \bar{b} : i \in I\right\}$ is Δ-indiscernible, so for all $i, i' \in I$ we have $\phi\left(\bar{b}, a_i\right) \iff \phi\left(\bar{b}, a_{i'}\right)$. As I must contain both an even and an odd indices, we get a contradiction.

Remark 1.28. The bounds on the VC-dimension given by this proof are astronomical as we have used Ramsey's theorem iteratively. In most specific cases it is possible to obtain much stronger bounds.

E.g., let $\mathcal{F}_{k,m,n}$ be the family of all semialgebraic subsets of $\mathbb{R}^n$ that can be represented as a Boolean combination of at most k sets of the form $\{\bar{x} \in \mathbb{R}^n : f_j(\bar{x}) > 0\}$ where the functions f_j are real polynomials of maximum degree m. Then $\mathrm{VC}(\mathcal{F}_{k,m,n}) \leq$

$2k\binom{m+n}{m}\log\left(k\left(k+1\right)\binom{m+n}{m}\right)$, and this differs only by a logarithmic factor from the known lower bound (see [10] for the details).

1.3. **Historic remarks.** In model theory, a partitioned formula $\phi\left(\bar{x},\bar{y}\right)$ is called *NIP* (No Independence Property) if the family $\mathcal{F}_\phi$ has finite VC-dimension. A structure $\mathcal{M}$ is NIP if all definable families in it are NIP. Such structures were defined by Shelah around the same time as Vapnik and Chervonenkis have defined their dimension for entirely different purposes, and are currently being actively studied in model theory (see [42] for a survey). The original proof of Lemma 1.13 by Shelah used forcing and absoluteness (see [16] for some more details). It was first finitized by Laskowski [29], and further simplified by Poizat, Adler and others [2]. We avoid the use of compactness and give a purely combinatorial proof which in principle gives explicit bounds, using compactness we could have avoided micromanaging all the numerical parameters involved.

2. VC-density

2.1. **Basic properties and fractional examples.** Let $(X,\mathcal{F})$ be a family of finite VC-dimension. By Lemma 1.5 we know that $\pi_{\mathcal{F}}\left(n\right)=O\left(n^d\right)$, but how exactly can $\pi_{\mathcal{F}}\left(n\right)$ grow?

Definition 2.1. We define the *VC-density* of $\mathcal{F}$ to be $\mathrm{vc}\left(\mathcal{F}\right)=\limsup_{n\to\infty}\frac{\log(\pi_{\mathcal{F}}(n))}{\log n}$. In other words, $\mathrm{vc}\left(\mathcal{F}\right)$ is the infimum over all real numbers $r\geq 0$ for which we have $\pi_{\mathcal{F}}\left(n\right)=O\left(n^r\right)$. Similarly, we define the *VC-codensity* as $\mathrm{vc}^*\left(\mathcal{F}\right)=\mathrm{vc}\left(\mathcal{F}^*\right)$.

We have $\mathrm{vc}\left(\mathcal{F}\right)<\infty \iff \mathrm{VC}\left(\mathcal{F}\right)<\infty$ and by Lemma 1.5 we have $\mathrm{vc}\left(\mathcal{F}\right)\leq \mathrm{VC}\left(\mathcal{F}\right)$. Often they coincide.

Exercise 2.2. (1) Let $\mathcal{F}=\binom{X}{\leq d}$. Show that $\mathrm{VC}\left(\mathcal{F}\right)=\mathrm{vc}\left(\mathcal{F}\right)=d$.
(2) Let $X=\mathbb{R}$, $k\geq 1$ and let $\mathcal{F}$ be the collection whose members are the unions of k disjoint open intervals in $\mathbb{R}$. Show that $\mathrm{VC}\left(\mathcal{F}\right)=\mathrm{vc}\left(\mathcal{F}\right)=2k$, and in fact $\pi_{\mathcal{F}}\left(n\right)=\binom{n}{\leq 2k}$ for each n.

In some sense, the VC-density is a more rigid version of the VC-dimension ignoring small noise.

Example 2.3. Let $(X,\mathcal{F})$ be a set system with $\mathrm{vc}\left(\mathcal{F}\right)<k$. Let $X'=X\cup Y$ where Y is a set of size k disjoint from X. Let $\mathcal{F}'=\mathcal{F}\cup\mathcal{P}\left(Y\right)$. Then $\mathrm{VC}\left(\mathcal{F}'\right)=k$, but $\mathrm{vc}\left(\mathcal{F}'\right)=\mathrm{vc}\left(\mathcal{F}\right)<k$.

Exercise 2.4. (Some properties of VC-density)

(1) Let $(X,\mathcal{F})$ be a set system, let X' be an infinite set and $f:X'\to X$ be a map. Let $f^{-1}\left(\mathcal{F}\right):=\left\{f^{-1}\left(S\right):S\in\mathcal{F}\right\}$. Then $\pi_{f^{-1}(\mathcal{F})}\leq\pi_F$ for all n, with equality if f is surjective. In particular, $\mathrm{vc}\left(f^{-1}\left(\mathcal{F}\right)\right)\leq\mathrm{vc}\left(\mathcal{F}\right)$.
(2) If $\mathcal{F}=\mathcal{F}_1\cup\mathcal{F}_2$ then $\mathrm{vc}\left(\mathcal{F}\right)=\max\left\{\mathrm{vc}\left(\mathcal{F}_1\right),\mathrm{vc}\left(\mathcal{F}_2\right)\right\}$, and $\mathrm{vc}\left(X\setminus\mathcal{F}\right)=\mathrm{vc}\left(\mathcal{F}\right)$ where $X\setminus\mathcal{F}:=\left\{X\setminus S:S\in\mathcal{F}\right\}$.
(3) Given two set systems $\left(X_1,\mathcal{F}_1\right),\left(X_2,\mathcal{F}_2\right)$, consider the set system $\left(X_1\times X_2,\mathcal{F}\right)$ with $\mathcal{F}:=\left(S_1\times S_2:S_i\in\mathcal{F}_i\right)$. Then $\mathrm{vc}\left(\mathcal{F}\right)\leq\mathrm{vc}\left(\mathcal{F}_1\right)+\mathrm{vc}\left(\mathcal{F}_2\right)$.
(4) $\mathcal{F}$ is finite if and only if $\mathrm{vc}\left(\mathcal{F}\right)=0$.

In fact, the last claim can be strengthened.

Proposition 2.5. *If* $\mathrm{vc}\left(\mathcal{F}\right)<1$*, then* $\mathcal{F}$ *is finite (and so VC-density never takes values in the interval* $(0,1)$*).*

Proof. I have originally presented a proof due to Assouad [8, Proposition 2.19], but the following much simpler argument was pointed out to me by Pietro Kreitlon Carolino.

It is enough to show that if $\mathcal{F}$ is infinite, then $\pi_{\mathcal{F}}(n) \geq n$ for all $n \in \mathbb{N}$. This is immediate by the following claim.

Claim. Let $F_1, \ldots, F_m$ be pairwise distinct subsets of a set X. Then there exists an m-subset A of X such that $F_1 \cap A, \ldots, F_m \cap A$ are pairwise distinct.

Proof by induction on m. For $m = 1$ this is obvious. Let $F_1, \ldots, F_{m+1} \subseteq X$ be given, by induction hypothesis there is an m-element set $A \subseteq X$ such that $F_1 \cap A, \ldots, F_m \cap A$ are pairwise distinct. Then $F_{m+1} \cap A$ can only be equal to at most one of the other $F_i \cap A$. If necessary, add an element of X to A that distinguishes that F_i from F_{m+1}.

□

Are there any other restrictions on the possible values of the VC-density? It is clear that for any natural $r \in \mathbb{N}$, the family $\binom{\mathbb{N}}{r}$ has VC-density r. First we observe that there are some natural examples of families with non-integer rational VC-density.

Let (P, Q, E) be a bipartite graph on a set X, i.e. $X = P \cup Q$ is a partition, and $E \subseteq P \times Q$. For $m, n \in \mathbb{N}$, by $K_{m,n}$ we denote the complete bipartite graph with $|P| = m$ and $|Q| = n$ and $E = P \times Q$.

Fact 2.6. *(Kővári-Sós-Turán, [27]) There is some constant $c \in \mathbb{N}$ such that any bipartite graph on n vertices (i.e. $|P| + |Q| = n$) omitting $K_{2,2}$ has at most $cn^{\frac{3}{2}}$ edges.*

Example 2.7. (VC-density $\frac{3}{2}$) Let F_q be a finite field on q elements, where q is a power of a prime p. Let P_q be the set of points on the affine plane over F_q, i.e. $P_q = (F_q)^2$. Let L_q be the set of lines (i.e. subsets of F_q^2 given by $y = ax + b$, $a, b \in F_q$). Finally let E_q be the incidence relation, i.e. $E_q = \{(p, q) : p \in l \in L_q\}$. Consider the bipartite graph $G_q = (P_q, L_q, E_q)$, note that it is $K_{2,2}$-free (there is only one line passing through a pair of points). We have $|P_q| = q^2$, $|L_q| = q^2$ and $|E_q| = q\,|L_q|$, so $|E_q| \geq q^3$ and $|X_q| \leq 3q^2$.

Let $G = (P, Q, E)$ be a bipartite graph given by the disjoint union of G_q for all q. Let $\mathcal{F} := \{\{p, l\} : (p, l) \in E\}$ be a family of subsets of $X = P \cup Q$. We claim that $\operatorname{vc}(\mathcal{F}) = \frac{3}{2}$.

Let A be a subset of X, and consider the bipartite graph

$$G|_A := (P \cap A, Q \cap A, E \cap A \times A)$$

induced on A. As $G|_A$ omits $K_{2,2}$, by Fact 2.6 it has at most $c\,|A|^{\frac{3}{2}}$ edges. Note that for any $S \in \mathcal{F}$, $S \cap A$ is either empty, has one element or appears as an edge in $G|_A$. We thus have $|A \cap \mathcal{F}| \leq 1 + |A| + c\,|A|^{\frac{3}{2}} \leq (c + 2)\,|A|^{\frac{3}{2}}$, so $\operatorname{vc}(\mathcal{F}) \leq \frac{3}{2}$.

On the other hand, $|X_q \cap \mathcal{F}| \geq |E_q| > \frac{1}{8}\,|X_q|^{\frac{3}{2}}$ for all q powers of p, i.e. $\operatorname{vc}(\mathcal{F}) \geq \frac{3}{2}$.

Example 2.8. (VC-density $\frac{4}{3}$) Now we consider the incidence graph on a real plane, i.e. let P be the set of points in $\mathbb{R}^2$, let L be the set of lines in $\mathbb{R}$, and let E be the incidence relation $E = \{(p, l) : p \in l \in L\}$. We consider the bipartite graph (E, P, L), and as above we consider the set system $X = P \cup Q, \mathcal{F} = \{\{p, l\} : (p, l) \in E\}$.

By the famous Szemeredi-Trotter bound, there is some constant c such that for any $P_0 \subseteq P$ with $|P_0| = n$ and $L_0 \subseteq L$ with $L_0 = m$ we have $|E \cap (P_0 \times L_0)| \leq c\left((mn)^{\frac{2}{3}} + m + n\right)$, which implies that for any finite $A \subseteq X$ we have $|A \cap \mathcal{F}| \leq 1 + |A| + c\left(|A|^{\frac{4}{3}} + 2\,|A|\right) \leq (c+2)\,|A|^{\frac{4}{3}}$, so $\mathrm{vc}\,(\mathcal{F}) \leq \frac{4}{3}$.

On the other hand, Szemeredi-Trotter bound is known to be optimal, as witnessed by the following example due to Elekes [18]:

Let k be a positive integer, $t = 4k^3$, and consider the subsets

$$P_0 := \left\{(\eta, \xi) : \eta = 0, 1, \ldots, k-1, \xi = 0, 1, \ldots, 4k^2 - 1\right\},$$

$$L_0 := \left\{(a, b) : a = 0, 1, \ldots, 2k-1, b = 0, 1, \ldots, 2k^2 - 1\right\}$$

of $\mathbb{Z}^2$. Let $A := P_0 \cup L_0 \subseteq X$, note that $|A| \leq t$. For each $i = 0, 1, \ldots, k-1$, each line $\xi = a\eta + b$ with $(a, b) \in L_0$ contains a point $(\eta, \xi) \in P_0$ with $\eta = i$, so $|\mathcal{F} \cap A| \geq |E \cap A| \geq k\,|L_0| \geq 4k^4 = \frac{1}{4^{\frac{1}{3}}} t^{\frac{4}{3}} \geq \frac{1}{4}\,|V|^{\frac{4}{3}}$, hence $\mathrm{vc}\,(\mathcal{F}) \geq \frac{4}{3}$.

It turns out that for any real number $r \in [1, \infty)$, one can find a family $\mathcal{F}_r$ with $\mathrm{vc}\,(\mathcal{F}_r) = r$.

Definition 2.9. Fix $j \in \mathbb{N}$ and let $E \subseteq \mathbb{N}^j$. Let

$$\psi_E\,(n) := \max\left\{|E \cap (A_1 \times \ldots \times A_j)| : A_i \subseteq \mathbb{N}, |A_1| = \ldots = |A_j| = n\right\}.$$

We define the *Blei density* of E to be the infimum of all real numbers $\alpha \geq 0$ for which there is some $c \in \mathbb{N}$ with $\psi_E\,(n) \leq cn^\alpha$ for all $n \in \mathbb{N}$. We denote it by $\mathrm{dens}\,(E)$.

Theorem 2.10. *[Blei, Körner] For every $\alpha \in (1, 2)$, there is some $E \subseteq \mathbb{N}^2$ with $\mathrm{dens}\,(E) = \alpha$.*

Corollary 2.11. *For every $s \in [1, \infty)$ there is some set family $(X, \mathcal{F})$ with $\mathrm{vc}\,(\mathcal{F}) = s$.*

Proof. If s is an integer, we can take $\binom{\mathbb{N}}{s}$. If not, we can write $s = k + \alpha$ with $k \in \mathbb{N}$ and $\alpha \in (1, 2)$.

By Theorem 2.10, let $E \subseteq \mathbb{N}^2$ be a bipartite graph of Blei density α. Let us consider $\mathcal{F}_0 = \{\{a, b\} : (a, b) \in E\} \subseteq \binom{\mathbb{N}}{\leq 2}$. Then the set system $(\mathbb{N}, \mathcal{F}_0)$ has VC-density α (same argument as in Example 2.7).

Let now

$$\mathcal{F} = \left\{S \in \binom{\mathbb{Z}}{k+2} : S \cap \mathbb{N} \in \mathcal{F}_0 \text{ and } |S \setminus \mathbb{N}| = k\right\}.$$

It's easy to see that $\mathrm{vc}\,(\mathcal{F}) = k + \alpha = s$. □

Remark 2.12. Later work [12, 11] generalizes Theorem 2.10 demonstrating the following.

Let $2 \leq d \in \mathbb{N}$ be arbitrary. Then for any $\alpha \in (1, d)$ and $\beta \in [1, \alpha]$ there exists some $E \subseteq \mathbb{N}^d$ such that $\limsup_{s \to \infty} \frac{\log \psi_E(s)}{\log s} = \alpha$ and $\liminf_{s \to \infty} \frac{\log \psi_F(s)}{\log s} = \beta$.

Such sets are not exhibited explicitly, but rather their existence is proved using the probabilistic method. We recall a couple of facts from probability theory that we will need.

2.2. **Interlude on probability theory.** A *probability space* $(\Omega, \mathcal{B}, \mu)$ is a set Ω equipped with a σ-algebra $\mathcal{B}$ and a σ-additive measure μ on $\mathcal{B}$ such that $\mu(\Omega) = 1$.

Example 2.13.
(1) Let Ω be a finite set $\{\omega_1, \ldots, \omega_n\}$, and for each point ω_i we assign a weight $r_i \in \mathbb{R}$ such that $\sum_{1 \leq i \leq n} r_i = 1$. Then let $\mathcal{B} = 2^{\Omega}$, and for $A \subseteq \Omega$, let $\mu(A) = \sum_{\omega_i \in A} r_i$.
(2) Let μ be the Lebesgue measure on $\mathbb{R}^n$ and let $X \subseteq \mathbb{R}^n$ be of finite positive μ-measure. Then we obtain a probability measure μ' on the algebra of Borel subsets by restricting to X, i.e. we define $\mu'(Y) := \frac{\mu(X \cap Y)}{\mu(X)}$.

For each $k \in \mathbb{N}$, the cartesian power Ω^k is equipped with the product σ-algebra $\mathcal{B}^{\otimes k}$, i.e. the σ-algebra generated by the sets of the form $B_1 \times \ldots \times B_k$ with $B_1, \ldots, B_k \in \mathcal{B}$. The *product measure* μ^k is defined as the unique probability measure on $\left(\Omega^k, \mathcal{B}^{\otimes k}\right)$ such that $\mu^k(B_1 \times \ldots \times B_k) = \mu(B_1) \ldots \mu(B_k)$.

A μ-measurable subset $A \subseteq \Omega$ is called an *event*. If A is an event, we let $\mathbf{1}_A$ be its characteristic function, and we write $\mathbb{P}(A) = \mu(A)$. If $f, g : \Omega \to \mathbb{R}$ are measurable functions, we write $\mathbb{P}(f \geq g)$ for the probability of the even $\{\omega \in \Omega : f(\omega) \geq g(\omega)\}$.

A measurable function $X : \Omega \to \mathbb{R}$ is called a (real-valued) *random variable*. The *probability distribution* of X is the probability measure on $\mathbb{R}$ obtained by taking the push-forward of μ by X. It is often convenient to define a random variable just by specifying its distribution and assuming that there is some underlying probability space Ω on which X is defined.

Definition 2.14. Let X be a random variable on $(\Omega, \mathbb{P})$.
(1) The *expected value* of X is defined as $\mathbb{E}(X) = \int_{\Omega} X(\omega)\, d\mathbb{P}$.
(2) The *variance* of X is defined as $\operatorname{Var}(X) = \mathbb{E}\left((X - \mathbb{E}(X))^2\right) = \mathbb{E}\left(X^2\right) - (\mathbb{E}(X))^2$.

Example 2.15. (Bernoulli random variable) By a Bernoulli random variable on $(\Omega, \mathbb{P})$ we mean a $\{0,1\}$-valued random variable X such that $\mathbb{P}(X = 1) = p$ and $\mathbb{P}(X = 0) = 1 - p$ for some real $p \in [0,1]$. We have:
(1) $\mathbb{E}(X) = \mathbb{P}(X = 1) \cdot 1 + \mathbb{P}(X = 0) \cdot 0 = p$.
(2) $\mathbb{E}\left(X^2\right) = \mathbb{P}(X = 1) \cdot 1^2 + \mathbb{P}(X = 0) \cdot 0^2 = p$.
(3) $\operatorname{Var}(X) = \mathbb{E}\left(X^2\right) - (\mathbb{E}(X))^2 = p - p^2 = p(1 - p)$.

Random variables $X_1, \ldots, X_n$ on Ω are *mutually independent* if for any Borel sets $B_1, \ldots, B_n \subseteq \mathbb{R}$ we have $\mathbb{P}\left(\bigcap_{k \leq n} \{X_k \in B_k\}\right) = \prod_{k \leq n} \mathbb{P}(X_k \in B_k)$.

Example 2.16. Let X_i be a random variable on the space $(\Omega_i, \mathcal{B}_i, \mu_i)$. Consider the product probability space $(\Omega, \mathcal{B}, \mu)$, i.e. $\Omega = \Omega_1 \times \ldots, \times \Omega_n$, $\mathcal{B}$ is the sigma algebra generated by all sets of the form $B_1 \times \ldots \times B_n$ with $B_i \in \mathcal{B}_i$, and μ is the product measure $\mu_1 \times \ldots \times \mu_n$. Let $\pi_i : \Omega \to \Omega_i$ be the projection map onto the i'th coordinate. For each i, we can define a copy of X_i living on Ω by taking $X'_i(\omega_1, \ldots, \omega_i, \ldots, \omega_n) := X_i(\omega_i)$. Then for all i, X_i and X'_i have the same probability distribution, and $X'_1, \ldots, X'_n$ are mutually independent random variables on Ω.

Fact 2.17. *Let $X_1, \ldots, X_k$ be some random variables on Ω.*
(1) *Expectation is linear: for any $r_1, \ldots, r_k \in \mathbb{R}$ we have $\mathbb{E}(r_1 X_1 + \ldots + r_k X_k) = r_1 \mathbb{E}(X_1) + \ldots + r_k \mathbb{E}(X_k)$.*

(2) *If $X_1, \ldots, X_k$ are mutually independent, then* $\mathbb{E}(X_1 \cdot \ldots \cdot X_k) = \mathbb{E}(X_1) \cdot \ldots \cdot \mathbb{E}(X_k)$ *and* $\operatorname{Var}(X_1 + \ldots + X_k) = \operatorname{Var}(X_1) + \ldots + \operatorname{Var}(X_k)$.

We recall some useful inequalities for estimating probabilities.

Fact 2.18. *(Markov's inequality) Let X be a random variable taking non-negative values. Then for all $r > 0$ we have $\mathbb{P}(X \geq r) \leq \frac{\mathbb{E}(X)}{r}$.*

Proof. Write $\mathbb{E}(X) \geq \mathbb{E}\left(X \cdot \mathbf{1}_{\{f \geq r\}}\right) \geq r\mathbb{P}(X \geq r)$. □

Fact 2.19. *(Chebyshev's inequality) Let X be an $\mathbb{R}$-valued random variable on $(\Omega, \mathbb{P})$. Then for any $k > 0$ we have*

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq k) \leq \frac{\operatorname{Var}(X)}{k^2}.$$

Proof. Consider the random variable $(X - \mathbb{E}(X))^2$, by Markov's inequality we have $\mathbb{P}(|X - \mathbb{E}(X)| \geq k) = \mathbb{P}\left((X - \mathbb{E}(X))^2 \geq k^2\right) \leq \frac{\mathbb{E}\left((X-\mathbb{E}(X))^2\right)}{k^2} \leq \frac{\operatorname{Var}(X)}{k^2}$. □

2.3. Blei-Körner example with irrational VC-density.

Now we go back to the proof of Theorem 2.10. The idea is basically to take a random bipartite graph with the edge density α. But we will need a couple of auxiliary lemmas to show that it actually works.

Lemma 2.20. *Let $\alpha \in (1, 2)$ and $0 < M \in \mathbb{N}$ be arbitrary. Then there exists $n = n(M) \geq M$ and $F \subseteq [n]^2$ so that:*

(1) $\psi_F(n) \geq \frac{1}{2} n^\alpha$,
(2) $\psi_F(s) \leq s^\alpha$ *for all $s \geq L(\alpha)$, where $L(\alpha) = \min\{s : (2 - \alpha)s^\alpha - 2s \geq 1\}$. Note that since $\alpha \in (1, 2)$, $L(\alpha)$ is well-defined.*

Proof. Let $k \geq M$ be an arbitrary integer, and let $\left(X_{ij}^{(k)} : i, j \in \mathbb{N}\right)$ be a collection of independent Bernoulli $\{0, 1\}$-valued random variables on some probability space $(\Omega, \mathbb{P})$ with $\mathbb{P}\left(X_{ij}^{(k)} = 1\right) = k^{\alpha-2}$ (so $\mathbb{P}\left(X_{ij}^{(k)} = 0\right) = 1 - k^{\alpha-2}$).

Suppose that $L(\alpha) \leq s \leq M$ and $A, B \in \binom{\mathbb{N}}{s}$. Clearly we have

$$\mathbb{P}\left(\sum_{i \in A, j \in B} X_{ij}^{(k)} \geq s^\alpha\right) \leq \sum_{m = s^\alpha}^{s^2} \binom{s^2}{m} k^{(\alpha-2)m} \left(1 - k^{\alpha-2}\right)^{s^2 - m} \leq k^{(\alpha-2)s^\alpha} 2^{s^2}.$$

Summing over all $A, B \in \binom{[k]}{s}$ we deduce from it (as $L(\alpha) = \min\{s : 2s - s^\alpha(2 - \alpha) \leq -1\}$) that

$$(a)\, \mathbb{P}\left(\sum_{i \in A, j \in B} X_{ij}^{(k)} \geq s \text{ for some } A, B \in \binom{[k]}{s}\right) \leq k^{2s} k^{(\alpha-2)s^\alpha} 2^{s^2} \leq \frac{2^{M^2}}{k}.$$

By the mutual independence of $\left(X_{ij}^{(k)} : i, j \in [k]\right)$, Example 2.15 and Fact 2.17 we have:

- $\mathbb{E}\left(\sum_{i,j \in [k]} X_{ij}^{(k)}\right) = \sum_{i,j \in [k]} \mathbb{E}\left(X_{ij}^{(k)}\right) = k^2 k^{\alpha-2} = k^\alpha$,
- $\operatorname{Var}\left(\sum_{i,j \in [k]} X_{ij}^{(k)}\right) = \sum_{i,j \in [k]} \operatorname{Var}\left(X_{ij}^{(k)}\right) = \sum_{i,j \in [k]} k^{\alpha-2}\left(1 - k^{\alpha-2}\right) = k^\alpha\left(1 - k^{\alpha-2}\right) \leq k^\alpha$.

By Chebyshev's inequality (Fact 2.19) we have

$$(b)\,\mathbb{P}\left(\left|\sum_{i,j\in[k]} X_{ij}^{(k)} - k^{\alpha}\right| \geq k\right) \leq \frac{\operatorname{Var}\left(\sum_{i,j\in[k]} X_{ij}^{(k)}\right)}{k^2} \leq \frac{k^{\alpha}}{k^2} = k^{\alpha-2}.$$

Note that the probabilities of the events in (a) and $(b) \to 0$ as $k \to \infty$ (and all the other parameters are fixed). Thus fixing $k \in \mathbb{N}$ sufficiently large we obtain with high probability an element $\omega \in \Omega$ such that

$$(c) \sum_{i,j\in[k]} X_{ij}^{(k)}(\omega) \geq \frac{1}{2}k^{\alpha} \text{ and } \sum_{i\in A, j\in B} X_{ij}^{(k)}(\omega) \leq s^{\alpha} \text{ for all } A, B \in \binom{[k]}{s}, L(\alpha) \leq s \leq M.$$

Let $F_{\omega} := \left\{(i,j) \in [k]^2 : X_{ij}^{(k)}(\omega) = 1\right\}$. By (c) we have:

$$(d)\,\psi_{F_{\omega}}(k) \geq \frac{1}{2}k^{\alpha},$$

$$(e)\,\psi_{F_{\omega}}(s) \leq s^{\alpha} \text{ for all } L(\alpha) \leq s \leq M.$$

Let $n = \min\left\{j \geq M : \psi_{F_{\omega}}(j) \geq \frac{1}{2}j^{\alpha}\right\}$. By (d) we have $M \leq n \leq k$. Now we have $\psi_{F_{\omega}}(s) \leq s^{\alpha}$ for all $L(\alpha) \leq s \leq M$ by (c) and $\psi_{F_{\omega}}(s) \leq s^{\alpha}$ for all $M < s < n$ by the choice of n. Taking $F = F_{\omega}|_{[n]^2}$, possibly with some $\left(\psi_{F_{\omega}}(n) - \frac{1}{2}n^{\alpha} - 1\right)$ edges removed to ensure that the upper bound holds for $\psi_F(n)$ as well (while still keeping the lower bound), satisfies all the assumptions. □

Two sets $F_1, F_2 \subseteq \mathbb{N}^2$ are *bi-disjoint* if $\pi_1(F_1) \cap \pi_1(F_2) = \pi_2(F_1) \cap \pi_2(F_2) = \emptyset$, where $\pi_1(n,m) = n$ and $\pi_2(n,m) = m$ are the canonical projections from $\mathbb{N}^2$ onto $\mathbb{N}$.

Lemma 2.21. *Let $\{F_j : j \in \mathbb{N}\}$ be a collection of mutually bi-disjoint sets so that for each j we have $\psi_{F_j}(s) \leq Ks^{\alpha}$ for all $s \in \mathbb{N}$. Let $F = \bigcup_{j\in\mathbb{N}} F_j$. Then $\psi_F(s) \leq 2Ks^{\alpha}$ for all $s \in \mathbb{N}$.*

Proof. By assumption we can choose $I_j^{(1)}, I_j^{(2)}$ for $j \in \mathbb{N}$ such that $F_j \subseteq I_j^{(1)} \times I_j^{(2)}$ and $I_j^{(l)} \cap I_j^{(l)} = \emptyset$ for all $j \neq k \in \mathbb{N}$ and $l \in \{1,2\}$. Let $A, B \in \binom{\mathbb{N}}{s}$ be arbitrary, and let $A_j := A \cap I_j^{(1)}, B_j := B \cap I_j^{(2)}$. Then we have

$$|F \cap (A \times B)| = \sum_{j\in\mathbb{N}} |F \cap (A_j \times B_j)| \leq K \sum_{j\in\mathbb{N}} (\max\{|A_j|, |B_j|\})^{\alpha} \leq$$

$$\leq K\left(\left(\sum_{j\in\mathbb{N}} |A_j|\right)^{\alpha} + \left(\sum_{j\in\mathbb{N}} |B_j|\right)^{\alpha}\right) \leq 2Ks^{\alpha}.$$

□

Finally we can establish Theorem 2.10.

Proof. Fix $\alpha \in (1,2)$. For each $1 \leq j \in \mathbb{N}$, let $n(j)$ be as in Lemma 2.20. Let $\{I_j : j \in \mathbb{N}\}$ be a collection of mutually disjoint subsets of $\mathbb{N}$ with $|I_j| = n(j)$. By Lemma 2.20, we find $F_i \subseteq I_j \times I_j$ such that $\psi_{F_j}(s) \leq Ks^{\alpha}$ for all $s \geq 1$, where $K := (L(\alpha))^2$, and $\psi_{F_j}(n(j)) \geq \frac{1}{2}n(j)^{\alpha}$. By Lemma 2.21, $F = \bigcup_j F_j$ has Blei density α. □

So far we were only evaluating the exponent of the growth, the VC-density — this is the important parameter that we will need in the future.

But let us point out that $\pi_{\mathcal{F}}(n)$ need not grow as a power function in general. Again, an example comes from a rare case of a known tight bound in incidence geometry.

Fact 2.22. *(Pach, Sharir* [38]*) Let $\alpha \in (0, \pi)$ be a real number. The maximum number of times that α occurs as an angle among the ordered triples of t points in the plane is $O\left(t^2 \log t\right)$.*

Furthermore, suppose $\tan \alpha \in \mathbb{Q}\sqrt{d}$ *where $d \in \mathbb{N}$ is not a square. Then there exists a constant $C = C(\alpha) > 0$ and, for every $t > 3$, a t-element set $S_t \subseteq \mathbb{R}^2$ such that at least $Ct^2 \log t$ ordered triples of points from S_t determine the angle α.*

Thus we can consider the family

$$\mathcal{F} = \left\{ \{a, b, c\} \in \binom{\mathbb{R}^2}{3} : \text{the vectors } b - a, c - a \text{ are non-zero and } \angle(b, a, c) = \frac{\pi}{3} \right\}$$

of subsets of $\mathbb{R}^2$. As $\tan \frac{\pi}{3} = \sqrt{3}$, both parts of the theorem apply. Then for any $A \subseteq \mathbb{R}^2$ we have $\mathcal{F} \cap A = \binom{A}{\leq 2} \cup \left\{ \{a, b, c\} \in \binom{A}{3} : \text{ the condition above holds} \right\}$, and so there are some constants $C_1, C_2 > 0$ such that $C_1 t^2 \log t \leq \pi_{\mathcal{F}}(t) \leq C_2 t^2 \log t$ for every $t > 0$. That is, $\pi_{\mathcal{F}}(t) = \Theta\left(t^2 \log t\right)$ as $t \to \infty$.

2.4. **Growth of VC-density in first-order structures.** VC-density for definable families of sets is studied in [32, 7]. That is, we are back in the context of Theorem 1.13, i.e. we fix a structure $\mathcal{M}$ with some distinguished functions and relations, and we consider families of the form $\mathcal{F}_\phi$ where $\phi(\bar{x}, \bar{y})$ is a partitioned formula.

For $n \in \mathbb{N}$, define $\mathrm{vc}_{\mathcal{M}}(n) := \max\{\mathrm{vc}(\mathcal{F}_\phi) : \phi(\bar{x}, \bar{y}) \text{ is a formula with } |\bar{y}| \leq n\}$. Note that we are bounding the size of the parameter variables $\bar{y}$, but not of the object variables.

Note that even if all definable families $\mathcal{F}_\phi$ in $\mathcal{M}$ have finite VC-dimension, it is still possible that $\mathrm{vc}_{\mathcal{M}}(1) = \infty$ (due to the allowed growth of $|\bar{x}|$).

Exercise 2.23. Construct an example with this property.

However, it is not known if the analog of Shelah's theorem 1.13 holds for VC-density instead of VC-dimension.

Problem 2.24. Assume that $\mathrm{vc}_{\mathcal{M}}(1) < \infty$. Does it imply that $\mathrm{vc}_{\mathcal{M}}(n) < \infty$ for all $n \in \mathbb{N}$?

In all known examples this is true, and moreover $\mathrm{vc}_{\mathcal{M}}(n)$ grows linearly with n.

Example 2.25. [32, Section 6.2] Let $\mathcal{M}$ be quasi-o-minimal (i.e. every definable subset of M is a finite Boolean combination of singletons, intervals in M and $\emptyset$-definable sets). Then $\mathrm{vc}_{\mathcal{M}}(n) = n$ for all n.

This applies in particular to all o-minimal structures (including semialgebraic families) and to Presburger arithmetic $(\mathbb{Z}, +, 0, 1, <)$, as well as variations such as $(\mathbb{Z}^n, <, +)$ with $<$ the lexicographic ordering on $\mathbb{Z}^n$.

Problem 2.26. Assume that all definable families $\mathcal{F}_\phi$ in $\mathcal{M}$ *have finite VC-dimension. Is it true then that* $\mathrm{vc}(\mathcal{F}_\phi)$ *is rational for all formulas ϕ?*

2.5. **Historic remarks.** The notion of VC density was studied by Assouad and others, Lemma 2.5 and Example 2.7 are from [8]. This article contains many other examples of calculations of VC-density, in particular in the metric setting, and the connection between Blei's density and VC-density is also due to him. In the context of model theory, VC-density was studied in [32, 7].

3. The VC theorem

3.1. **VC theorem in finite probability spaces.** Recall the classical fact from probability theory.

Fact 3.1. *(Weak law of large numbers) Let $(\Omega, \mathcal{B}, \mathbb{P})$ be a probability space. Let $A \subseteq \Omega$ be an event and let $\varepsilon > 0$ be fixed. Then for any $n \in \mathbb{N}$ we have:*

$$\mathbb{P}^n\left((\omega_1, \ldots, \omega_n) \in \Omega^n : \left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_A(\omega_i) - \mathbb{P}(A)\right| \geq \varepsilon\right) \leq \frac{1}{4n\varepsilon^2}.$$

Note that this probability $\to 0$ as $n \to \infty$. In particular this means that fixing an arbitrary error ε, we can take n large enough so that with high probability the measure of A can be determined up to ε by picking n points at random and counting the proportion of them in A.

Proof. Fix $n \in \mathbb{N}$. For $i \leq n$, the Bernoulli random variable $\mathbf{1}_A(\omega_i) : \Omega^n \to \mathbb{R}$ has expectation $\mathbb{P}(A)$ and variance $\mathbb{P}(A)(1-\mathbb{P}(A)) \leq \frac{1}{4}$. Also the variables $\mathbf{1}_A(\omega_i)$, $i = 1, \ldots, n$ are mutually independent. Hence $\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_A(\omega_i)\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(\mathbf{1}_A(\omega_i)) = \frac{1}{n}n\mathbb{P}(A) = \mathbb{P}(A)$ and $\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_A(\omega_i)\right) = \sum_{i=1}^{n}\mathrm{Var}\left(\frac{\mathbf{1}_A(\omega_i)}{n}\right) = \sum_{i=1}^{n}\left(\frac{\mathbb{P}(A)(1-\mathbb{P}(A))}{n^2}\right) \leq \frac{1}{4n}$. We can then conclude by Chebyshev's inequality (Fact 2.19). □

The key result in VC-theory is the theorem of Vapnik and Chervonenkis [44] demonstrating that a *uniform* version of the weak law of large numbers holds for *families of events* of finite VC-dimension. That is, with high probability sampling on a sufficiently long random tuple gives a good estimate for the measure of all sets in the family $\mathcal{F}$ simultaneously. To prove it, we will need a finer (exponential decay) estimate for the tail distribution of a sum of independent random variables than the one provided by Markov's or Chebyshev's inequalities.

Fact 3.2. *(Chernoff's bound, special case, see e.g.* [6, Appendix A]*) Let $X_1, \ldots, X_n$ be independent $\{-1, 1\}$-valued random variables such that $\mathbb{P}(X_k = 1) = \mathbb{P}(X_k = -1) = \frac{1}{2}$ for all k. Then for any $\varepsilon > 0$ we have*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{k=1}^{n}X_k\right| \geq \varepsilon\right) \leq 2\exp\left(-\frac{n\varepsilon^2}{2}\right).$$

Remark 3.3. Chernoff's bound also applies to a slightly more general situation when we have random variables $X_1, \ldots, X_n$ such that $X_1, \ldots, X_m$ satisfy the hypothesis of Fact 3.2 and $X_{m+1}, \ldots X_n$ are equal to 0. Let $Y = \frac{1}{n}\sum_{k=1}^{n}X_k$ and let $Y' = \frac{1}{m}\sum_{k=1}^{m}X_k$. Indeed, we have

$$\mathbb{P}(|Y| \geq \varepsilon) = \mathbb{P}\left(|Y'| \geq \frac{n}{m}\varepsilon\right) \leq 2\exp\left(-\frac{m\left(\frac{n\varepsilon}{m}\right)^2}{2}\right) \leq 2\exp\left(-\frac{n\varepsilon^2}{2}\right).$$

Let us fix some notation. For $S \in \mathcal{F}$ and $(x_1, \ldots, x_n) \in X^n$ we define

$$\mathrm{Av}\left(x_1, \ldots, x_n; S\right) := \frac{1}{n}\left|\{1 \leq i \leq n : x_i \in S\}\right|.$$

Theorem 3.4. *(VC-theorem) Let (X, μ) be a* ***finite*** *probability space, and $\mathcal{F} \subseteq \mathcal{P}(X)$ a family of subsets of X. Then for every $\varepsilon > 0$ we have*

$$\mu^n\left(\sup_{S \in \mathcal{F}}\left|\mathrm{Av}\left(x_1, \ldots, x_n; S\right) - \mu(S)\right| > \varepsilon\right) \leq 8\pi_{\mathcal{F}}(n)\exp\left(-\frac{n\varepsilon^2}{32}\right).$$

Remark 3.5. Note that if $\mathrm{VC}(\mathcal{F}) = d$, then $\pi_{\mathcal{F}}(n) = O\left(n^d\right)$ and so the right part converges to 0 as n grows. Thus, as long as the VC-dimension of $\mathcal{F}$ is bounded, starting with $\mathcal{F}$ of arbitrary large finite size and an arbitrary measure, we still get an approximation up to an error ε for all sets in $\mathcal{F}$ by sampling on a random tuple of length depending just on d, ε.

Proof. Fix some integer n. For $\bar{x} = (x_1, \ldots, x_n), \bar{x}' = (x'_1, \ldots, x'_n)$ and $S \in \mathcal{F}$, let

$$f\left(\bar{x}, \bar{x}'; S\right) := \left|\mathrm{Av}\left(x_1, \ldots, x_n; S\right) - \mathrm{Av}\left(x'_1, \ldots, x'_n; S\right)\right|.$$

Let $x_1, \ldots, x_n, x'_1, \ldots, x'_n$ be mutually independent random elements from X, each with distribution μ (i.e. we have some probability space $(\Omega, \mathcal{B}, \mathbb{P})$ and measurable functions $x_i, x'_i : \Omega \to X$ such that μ is the push-forward of the probability measure $\mathbb{P}$ according to x_i). Let also $\sigma_1, \ldots, \sigma_n$ be random variables independent from each other and from the previous ones such that $\mathbb{P}(\sigma_i = 1) = \mathbb{P}(\sigma_i = -1) = \frac{1}{2}$. They will play an auxiliary role allowing us to apply Chernoff's bound.

Claim 1. We have

$$\mathbb{P}\left(\sup_{S \in \mathcal{F}} f\left(\bar{x}, \bar{x}'; S\right) > \frac{\varepsilon}{2}\right) \leq 2\mathbb{P}\left(\sup_{S \in \mathcal{F}}\frac{1}{n}\left|\sum_{i=1}^{n}\sigma_i \mathbf{1}_S(x_i)\right| > \frac{\varepsilon}{4}\right).$$

Note that for a fixed i and S, the random variable $\mathbf{1}_S(x_i) - \mathbf{1}_S(x'_i)$ has expectation 0 and a symmetric distribution around 0 (i.e. it takes the values 1 and -1 with the same probability). Therefore its distribution does not change if we multiply it by σ_i (check!). We then have:

$$\begin{aligned}
\mathbb{P}\left(\sup_{S \in \mathcal{F}} f\left(\bar{x}, \bar{x}'; S\right) > \frac{\varepsilon}{2}\right) &= \\
\mathbb{P}\left(\sup_{S \in \mathcal{F}}\frac{1}{n}\left|\sum_{i=1}^{n}\left(\mathbf{1}_S(x_i) - \mathbf{1}_S(x'_i)\right)\right| > \frac{\varepsilon}{2}\right) &= \\
\mathbb{P}\left(\sup_{S \in \mathcal{F}}\frac{1}{n}\left|\sum_{i=1}^{n}\sigma_i\left(\mathbf{1}_S(x_i) - \mathbf{1}_S(x'_i)\right)\right| > \frac{\varepsilon}{2}\right) &\leq \\
\mathbb{P}\left(\sup_{S \in \mathcal{F}}\frac{1}{n}\left|\sum_{i=1}^{n}\sigma_i \mathbf{1}_S(x_i)\right| > \frac{\varepsilon}{4} \text{ or } \sup_{S \in \mathcal{F}}\frac{1}{n}\left|\sum_{i=1}^{n}\sigma_i \mathbf{1}_S(x'_i)\right| > \frac{\varepsilon}{4}\right) &\leq \\
2\mathbb{P}\left(\sup_{S \in \mathcal{F}}\frac{1}{n}\left|\sum \sigma_i \mathbf{1}_S(x_i)\right| > \frac{\varepsilon}{4}\right),&
\end{aligned}$$

where the last inequality comes from the obvious union bound and the fact that $\bar{x}$ and $\bar{x}'$ have the same distribution.

Claim 2. We have

$$\mathbb{P}\left(\sup_{S\in\mathcal{F}} f\left(\bar{x},\bar{x}';S\right)>\frac{\varepsilon}{2}\right)\leq 4\pi_{\mathcal{F}}\left(n\right)\exp\left(-\frac{n\varepsilon^{2}}{32}\right).$$

First let us fix a tuple $\bar{a}=\left(a_{1},\ldots,a_{n}\right)\in X^{n}$ and $S\in\mathcal{F}$. Let $A_{S}\left(\bar{a}\right)$ be the event "$\frac{1}{n}\left|\sum_{i=1}^{n}\sigma_{i}\mathbf{1}_{S}\left(a_{i}\right)\right|>\frac{\varepsilon}{4}$" (so the only randomness left is in the σ_{i}'s). By Remark 3.3 we can apply Chernoff's bound to this situation, obtaining

$$\mathbb{P}\left(A_{S}\left(\bar{a}\right)\right)\leq 2\exp\left(-\frac{n\varepsilon^{2}}{32}\right).$$

Note that the event $A_{S}\left(\bar{a}\right)$ depends only on the set of those elements in the tuple $\bar{a}$ that are in S. As S varies in $\mathcal{F}$, there are at most $\pi_{\mathcal{F}}\left(n\right)$ different possible values for that set (if there are repetitions among $a_{1},\ldots,a_{n}$, this can only reduce the number of possibilities). Hence also at most $\pi_{\mathcal{F}}\left(n\right)$ events A_{S} to consider. Thus the union bound shows that the disjunction $\bigcup_{S\in\mathcal{F}}A_{S}\left(\bar{a}\right)$ has probability at most $2\pi_{\mathcal{F}}\left(n\right)\exp\left(-\frac{n\varepsilon^{2}}{32}\right)$. By Claim 1 we have

$$\begin{aligned}\mathbb{P}\left(\sup_{S\in\mathcal{F}} f\left(\bar{x},\bar{x}';S\right)>\frac{\varepsilon}{2}\right) & \leq 2\mathbb{P}\left(\sup_{S\in\mathcal{F}}\frac{1}{n}\left|\sum_{i=1}^{n}\sigma_{i}\mathbf{1}_{S}\left(x_{i}\right)\right|>\frac{\varepsilon}{4}\right)\\ & = 2\mathbb{P}\left(\sup_{S\in\mathcal{F}}A_{S}\left(\bar{x}\right)\right)\\ & \leq 2\mathbb{P}\left(\bigcup_{S\in\mathcal{F}}A_{S}\left(\bar{x}\right)\right)\leq 4\pi_{\mathcal{F}}\left(n\right)\exp\left(-\frac{n\varepsilon^{2}}{32}\right).\end{aligned}$$

To conclude the proof of the theorem, we may assume that $n>\frac{2}{\varepsilon^{2}}$ (since otherwise the right hand side is larger than 1, and so theorem obviously holds).

Let $X_{0}\subseteq X^{n}$ be the set of all $\bar{b}\in X^{n}$ such that $\mathbb{P}\left(\sup_{S\in\mathcal{F}}f\left(\bar{x},\bar{b};S\right)>\frac{\varepsilon}{2}\right)\geq\frac{1}{2}$. By Claim 2 we have $\mu^{n}\left(X_{0}\right)\leq 8\pi_{\mathcal{F}}\left(n\right)\exp\left(-\frac{n\varepsilon^{2}}{32}\right)$ (recalling the assumptions on the random variables $\bar{x},\bar{x}'$).

Fix $\bar{a}\in X^{n}\setminus X_{0}$ and $S\in\mathcal{F}$. By the weak law of large numbers (Fact 3.1) we have

$$\mathbb{P}\left(\left|\mathrm{Av}\left(x_{1},\ldots,x_{n};S\right)-\mu\left(S\right)\right|>\frac{\varepsilon}{2}\right)\leq\frac{1}{n\varepsilon^{2}}<\frac{1}{2}$$

by the assumption on n. It follows that there is some $\bar{x}\in X^{n}$ satisfying simultaneously

- $f\left(\bar{x},\bar{a};S\right)\leq\frac{\varepsilon}{2}$,
- $\left|\mathrm{Av}\left(\bar{x};S\right)-\mu\left(S\right)\right|\leq\frac{\varepsilon}{2}$.

Unwinding the definition of f, together this implies that $\left|\mathrm{Av}\left(\bar{a};S\right)-\mu\left(S\right)\right|\leq\varepsilon$. As $S\in\mathcal{F}$ was arbitrary, we conclude that for any $\bar{a}\in X^{n}\setminus X_{0}$ we have $\sup_{S\in\mathcal{F}}\left|\mathrm{Av}\left(\bar{x};S\right)-\mu\left(S\right)\right|\leq\varepsilon$, and the theorem follows. □

Corollary 3.6. *Let $d\in\mathbb{N}$ and $\varepsilon>0$ be arbitrary. Then there is some $N=N\left(d,\varepsilon\right)\in\mathbb{N}$ such that any set system $\left(X,\mathcal{F}\right)$ on a finite probability space $\left(X,\mu\right)$ with $\mathrm{VC}\left(\mathcal{F}\right)\leq d$ admits an* ε-approximation *of size at most N.*

That is, there is a multi-set $\left\{x_{1},\ldots,x_{N}\right\}$ of elements from X (repetitions are allowed) such that for all $S\in\mathcal{F}$ we have

$$\left|\mathrm{Av}\left(x_{1},\ldots,x_{N};S\right)-\mu\left(S\right)\right|\leq\varepsilon.$$

Proof. By Remark 3.5, it follows from Theorem 3.4 that for N large enough (with respect to d and ε), with high probability any N-tuple from X works as a ε-approximation (so in particular that is at least one N -tuple with this property). □

Remark 3.7. Note that repetitions among the points $x_1, \ldots, x_n$ are necessary — think of a measure on a finite set, giving certain *different* weights to different points.

Proposition 3.8. *One can take $N = C\frac{d}{\varepsilon^2} \ln \frac{d}{\varepsilon}$, where C is an absolute constant independent of d or ε.*

Proof. Fix some $\delta \in [0, 1]$, we want to find an n such that

$$\mu^n \left(\sup_{S \in \mathcal{F}} |\mathrm{Av}\,(x_1, \ldots, x_n; S) - \mu\,(S)| > \varepsilon \right) \leq \delta.$$

By Theorem 3.4, enough to show that $8\pi_{\mathcal{F}}\,(n) \exp\left(-\frac{n\varepsilon^2}{32}\right) \leq \delta$. As $\mathrm{VC}\,(\mathcal{F}) \leq d$ by assumption, by Lemma 1.5 there is some $C' = C'\,(d)$ so that $\pi_{\mathcal{F}}\,(n) \leq C'n^d$ for all n. Hence enough to show that $8C'n^d \exp\left(-\frac{n\varepsilon^2}{32}\right) \leq \delta$. Taking ln and rearranging, this is equivalent to $\frac{n\varepsilon^2}{32} \geq d \ln n + \ln \frac{8C'}{\delta}$.

Assume $n \geq \max\left\{\frac{64}{\varepsilon^2} \ln \frac{8C'}{\delta}, \frac{128d}{\varepsilon^2} \ln \frac{128d}{\varepsilon^2}\right\}$. Then $\frac{n\varepsilon^2}{64} \geq \ln \frac{8C'}{\delta}$, hence we only need to show that $\frac{n\varepsilon^2}{64} \geq d \ln n$.

Substituting $n = \frac{128d}{\varepsilon^2} \ln \frac{128d}{\varepsilon^2}$ and calculating, this is equivalent to $2d \ln \frac{128d}{\varepsilon^2} \geq d \ln \left(\frac{128d}{\varepsilon^2} \ln \frac{128d}{\varepsilon^2}\right)$, which is equivalent to $\left(\frac{128d}{\varepsilon^2}\right)^2 \geq \frac{128d}{\varepsilon^2} \ln \frac{128d}{\varepsilon^2}$, which is equivalent to $\frac{128d}{\varepsilon^2} \geq \ln \frac{128d}{\varepsilon^2}$. But this is true for all $\varepsilon \in [0, 1]$ and $d \geq 1$. Clearly this argument works for all $n \geq \frac{128d}{\varepsilon^2} \ln \frac{128d}{\varepsilon^2}$ as well.

Hence we can take N to be any n which works some for $\delta < 1$, say for $\delta = \frac{1}{2}$.

Recall that $C'\,(d) \leq \left(\frac{e}{d}\right)^d \leq e^d$. Hence $\frac{64}{\varepsilon^2} \ln \frac{8C'}{\delta} \leq \frac{64}{\varepsilon^2} (\ln 16 + d) \leq 384 \frac{1}{\varepsilon^2} d$ for any $\varepsilon \in [0, 1]$ and $d \geq 1$.

On the other hand, $\frac{128d}{\varepsilon^2} \ln \frac{128d}{\varepsilon^2} = 128d \frac{1}{\varepsilon^2} \left(\ln 128 + \ln d + 2 \ln \frac{1}{\varepsilon}\right) \leq 256d \frac{1}{\varepsilon^2} \left(\ln 128 + \ln \frac{d}{\varepsilon}\right) \leq 128^3 \frac{d}{\varepsilon^2} \ln \frac{d}{\varepsilon}$ for any $\varepsilon \in [0, 1]$ and $d \geq 1$. Hence taking $N := 128^3 \frac{d}{\varepsilon^2} \ln \frac{d}{\varepsilon}$ works. □

Remark 3.9.
1. Proposition 3.8 immediately implies that for every d there is some $C\,(d)$ so that $N = C\,(d) \left(\frac{1}{\varepsilon}\right)^2 \ln \frac{1}{\varepsilon}$ works for any $\varepsilon > 0$.
2. Using different methods from discrepancy theory, a slightly better bound is proved in [36, Theorem 1.3]: for every d there is some $C\,(d)$ so that for every finite family of VC-dimension d and $\varepsilon > 0$, there is an ε-approximation of size at most $C\,(d) \left(\frac{1}{\varepsilon}\right)^{2 - \frac{2}{d+1}} \left(\ln \frac{1}{\varepsilon}\right)^{2 - \frac{1}{d+1}}$ if $d > 1$ and $C\,(d) \frac{1}{\varepsilon} \left(\ln \frac{1}{\varepsilon}\right)^{\frac{5}{2}}$ if $d = 1$, for all $\varepsilon > 0$. This bound is known to be optimal, see [36, Section 3] and [3].

3.2. **Generalizations to arbitrary probability spaces.** The assumption of finiteness of the space (X, μ) can be relaxed. In fact it is easy to see that the proof goes through verbatim for an arbitrary probability space (X, μ) and an arbitrary family $\mathcal{F}$ of subsets of X as long as the following assumptions are satisfied:

1. Every set $S \in \mathcal{F}$ is measurable;
2. For each n, the function

$$(x_1, \ldots, x_n) \mapsto \sup_{S \in \mathcal{F}} |\mathrm{Av}\,(x_1, \ldots, x_n; S) - \mu\,(S)|$$

from X^n to $\mathbb{R}$ is measurable;

(3) For each n, the function

$$(x_1, \ldots, x_n, x'_1, \ldots, x'_n) \mapsto \sup_{S \in \mathcal{F}} |\mathrm{Av}(x_1, \ldots, x_n; S) - \mathrm{Av}(x'_1, \ldots, x'_n; S)|$$

from X^{2n} to $\mathbb{R}$ is measurable.

By basic measure theory, the first condition implies the other two when the family $\mathcal{F}$ is countable (and of course all the conditions hold when X is finite). However, this conditions are necessary in general.

Exercise 3.10. Let $X = \omega_1$ (i.e X is of uncountable size, and on it we have a total linear order without infinite decreasing chains). Let $\mathcal{B}$ be the σ-algebra generated by the intervals. Let μ be defined on $\mathcal{B}$ by $\mu(A) = 1$ if A contains an end-segment of X and $\mu(A) = 0$ otherwise. This defines a σ-additive measure on $(X, \mathcal{B})$. Let $\mathcal{F}$ be the family of all intervals in X. Check that it has VC-dimension 2, but the conclusion of the VC-theorem does not hold for $\mathcal{F}$. Namely, one checks that there are no finite ε-approximations for $\varepsilon < 1$ with respect to $\mathcal{F}$ and μ.

Under an additional set-theoretic assumption (continuum hypothesis) we can turn this into an example showing that assumption (1) alone is not enough even on a *standard* probability space.

Example 3.11. [Durst, Dudley] Let $(X, \mathcal{B})$ be an uncountable standard Borel space, e.g. a Borel space associated to the unit interval $X = [0, 1]$. Continuum Hypothesis is equivalent to the existence of a total order $\prec$ on X with the property that every half-open initial segment $I_y = \{x \in X : x \prec y\}$, $y \in X$, is countable and $\prec$ is a well-ordering (i.e. every non-empty subset of X has the $\prec$-smallest element).

Let $\mathcal{F}$ consist of all half-open initial segments $I_y, y \in X$ as above. Clearly the VC-dimension of $\mathcal{F}$ is one.

Now let μ be the Lebesgue measure on $[0, 1]$. As described above, under the Continuum Hypothesis every element of $\mathcal{F}$ is a countable set, therefore Borel measurable of measure 0. At the same time, for every finite set $x_1, \ldots, x_n$ of elements from X, there is a countable initial segment $I_y \in \mathcal{F}$ containing all of $x_1, \ldots, x_n$, so $\mathrm{Av}(x_1, \ldots, x_n; I_y) = 1$. Thus, no finite set of points works as an ε-approximation for $\mu, \mathcal{F}$ with $\varepsilon < 1$.

However, there are also natural uncountable families satisfying all the necessary measurability assumptions.

Exercise 3.12. Let G be a (locally) compact Polish group. Let S be a Borel subset of G, and consider the family $\mathcal{F} = \{gS : g \in G\}$. In general this is an uncountable family of subsets of G, however all of the assumptions (1)–(3) above are satisfied with respect to the (left) G-invariant Haar measure on G. (Hint: use absolute measurability of analytic sets in Polish spaces.)

Definition 3.13. Let $(X, \mathcal{F})$ be a set system, let μ be a probability measure on X and let $\varepsilon > 0$ be given. We say that a set $A \subseteq X$ is an ε-net with respect to μ if for every $S \in \mathcal{F}$ with $\mu(S) \geq \varepsilon$ we have $S \cap A \neq \emptyset$.

Proposition 3.14. *For any $\varepsilon > 0$ and $d \in \mathbb{N}$ there is some $N = N(\varepsilon, d)$ such that:*

if $(X, \mathcal{F})$ is a set-system with $\mathrm{VC}(\mathcal{F}) \leq d$ and μ is a probability measure on X, then there is an ε-net $A \subseteq X$ with respect to μ and $\mathcal{F}$ of size $\leq N$.

Proof. Note that every ε-approximation is an ε-net (every set of measure $\geq \varepsilon$ must contain at least one point from an ε-approximation) and apply Corollary 3.6 to find an ε-approximation of size $C\frac{1}{\varepsilon}\log\frac{1}{\varepsilon}$, with C depending just on d. □

Remark 3.15. One can achieve a better bound on the size of ε-nets than on the size of ε-approximations (morally because we can get rid of all the possible repeated elements required in an ε-approximation in general). In fact one has upper bound $C\frac{1}{\varepsilon}\log\frac{1}{\varepsilon}$ [23], and even $(1+o(1))\left(\frac{d}{\varepsilon}\log\frac{1}{\varepsilon}\right)$ [26]. It is also demonstrated in [39] that there is a matching lower bound already for $d=2$, even in natural geometric families.

Exercise 3.16. Note that if $(X,\mathcal{F})$ is a set system with $\mathrm{VC}(\mathcal{F}) \leq d$ and $Y \subseteq X$ is arbitrary, then for the set system $(Y, \mathcal{F}\cap Y)$ with $\mathcal{F}\cap Y = \{S\cap Y : S\in\mathcal{F}\}$ we again have $\mathrm{VC}(\mathcal{F}\cap Y)\leq d$. Thus for every such subsystem $(Y,\mathcal{F}\cap Y)$ we can find an ε-net $A\subseteq Y$ of size $\leq N = N(\varepsilon, d)$.

Show that this property characterizes set families of finite VC-dimension. Namely, if the VC-dimension of $\mathcal{F}$ is infinite, then for every $n\in\mathbb{N}$ we can find a finite set $Y\subseteq X$ of size n such that the smallest ε-net for $(Y,\mathcal{F}\cap Y)$ with respect to the uniform counting measure $\mu(S) = \frac{|S\cap Y|}{|Y|}$ is of size at least $(1-\varepsilon)n$.

3.3. Historic remarks.

Theorem 3.4 was established in [44] and started the whole area of VC-theory, there are many versions and generalizations of this result (for example, a version of the VC-inequality holds not only for the case of i.i.d. random processes as presented here, but for arbitrary ergodic processes [1]). Our presentation of the VC-theorem is from [42].

4. Finding small transversals (ε-nets and Helly-type theorems)

4.1. Transversals and packing numbers.

Let $(X,\mathcal{F})$ be a set system, with $\mathcal{F}$ and X possibly infinite.

Definition 4.1.

1. A subset $T\subseteq X$ is a *transversal* of $\mathcal{F}$ if $T\cap S\neq\emptyset$ for all $S\in\mathcal{F}$.
   The *transversal number* of $\mathcal{F}$, denoted $\tau(\mathcal{F})$, is the smallest possible cardinality of a transversal of $\mathcal{F}$.
2. A subsystem $\mathcal{G}\subseteq\mathcal{F}$ of pairwise-disjoint sets is called a *packing*. The *packing number* of $\mathcal{F}$, denoted $\nu(\mathcal{F})$, is the maximum cardinality of a packing $\mathcal{G}\subseteq\mathcal{F}$.

Remark 4.2. When $\mathcal{F}$ is the system of edges of a graph, this corresponds to the vertex cover and the matching number.

Any transversal of $\mathcal{F}$ is at least as large as any packing, so we always have $\nu(\mathcal{F})\leq\tau(\mathcal{F})$. Very little can be said in the reverse direction in general.

Example 4.3. Let X be the real plane, and let $\mathcal{F}_n$ be the set of n lines on the plane in general position. Then $\nu(\mathcal{F})=1$ since every two lines intersect, but $\tau(\mathcal{F})\geq\frac{1}{2}n$ because no point is contained in more than two of the lines.

Exercise 4.4. Let $\mathcal{F}$ be a system of finitely many closed intervals on the real line. Prove that $\nu(\mathcal{F})=\tau(\mathcal{F})$.

Remark 4.5. If $\mathcal{F}$ is the family of hyper-edges in a uniform r-hypergraph then $\tau\leq r\nu$ (the union of the edges of a maximal matching gives a transversal). For a bipartite graph we have $\nu(\mathcal{F}) = \tau(\mathcal{F})$ (König's theorem), and in general Ryser's conjecture says that $\tau \leq (r-1)\nu$ (only known for $r = 2$).

Exercise 4.6. Recall Hall's matching theorem: if G is a bipartite graph with parts A and B such that every subset $S \subseteq A$ has at least $|S|$ neighbors in B, then there is a matching in G containing all vertices of A. Derive König's theorem from Hall's theorem (and reversely)

Now we introduce another parameter of a set system that always lies between ν and τ, and which is useful in estimating their values. For this we first restrict to set systems with a *finite* underlying set X.

Definition 4.7. Let $(X, \mathcal{F})$ be a set system with X finite. A *fractional transversal* for $\mathcal{F}$ is a function $\phi : X \to [0,1]$ such that for each $S \in \mathcal{F}$ we have $\sum_{x \in S} \phi(x) \geq 1$.

The *size* of a fractional transversal ϕ is $\sum_{x \in X} \phi(x)$, and the *fractional transversal number* $\tau^*(\mathcal{F})$ is the infimum of the sizes of fractional transversals for $\mathcal{F}$.

So, intuitively, in a fractional transversal we can take one-third of one point, one-fifth of another, etc., but we must put a total weight of at least one full point into every set in the family.

Similarly:

Definition 4.8. A *fractional packing* for $\mathcal{F}$ is a function $\psi : \mathcal{F} \to [0,1]$ such that for each $x \in X$ we have $\sum_{\{S \in \mathcal{F} : x \in S\}} \psi(S) \leq 1$.

The *size* of a fractional packing ψ is $\sum_{S \in \mathcal{F}} \psi(S)$, and the fractional packing number $\nu^*(\mathcal{F})$ is the supremum of the sizes of all fractional packings for $\mathcal{F}$.

So in a fractional packing, each set in the family is assigned a weight, and the total weight of all sets in $\mathcal{F}$ containing any given point in X must not exceed 1.

Example 4.9. Consider the "triangle" set system with $X = \{a_1, a_2, a_3\}$ and $\mathcal{F} = \{\{a_1, a_2\}, \{a_2, a_3\}, \{a_3, a_1\}\}$. Check that $\nu = 1$, $\tau = 2$ and $\nu^* = \tau^* = \frac{3}{2}$.

Clearly $\tau^*(\mathcal{F}) \leq \tau(\mathcal{F})$ and $\nu \leq \nu^*$ (by assigning weight 1 to every point in the transversal, resp. to every set in a packing). In general the gap between τ^* and τ can be arbitrarily large:

Exercise 4.10.
(1) Let $X = [m]$ and let $\mathcal{F} = \binom{[m]}{n}$. Then $\tau^* = \frac{m}{n}$ while $\tau = m - n + 1$. Thus, when $m = 2n$ we get $\tau^* = 2$ and $\tau = n + 1$.
(2) Similarly, find a set system with ν bounded by a constant and ν^* arbitrarily large.
(3) Show that $\tau(\mathcal{F}) \leq \tau^*(\mathcal{F}) \ln(|\mathcal{F}| + 1)$ for all finite set systems $\mathcal{F}$ (hint: choose a transversal as a random sample).

Theorem 4.11. *For every set system $(X, \mathcal{F})$ with X finite we have $\nu^*(\mathcal{F}) = \tau^*(\mathcal{F})$.*
Moreover, the common value is a rational number, and there exists an optimal fractional transversal and an optimal fractional packing attaining only rational values.

The proof relies on the duality of linear programming (see e.g. [33, Proposition 10.1.2]).

Fact 4.12. *("Strong duality of linear programming") Let A be an $m \times n$ real matrix, $b \in \mathbb{R}^m$ a (column) vector and $c \in \mathbb{R}^n$ a (column) vector. Let $P = \{x \in \mathbb{R}^n : x \geq 0, Ax \geq b\}$ and $D = \left\{y \in \mathbb{R}^m : y \geq 0, y^T A \leq c^T\right\}$ (the inequalities* *should hold in every component). If both $P \neq \emptyset$ and $D \neq \emptyset$ then* $\min \left\{c^T x : x \in P\right\} = \max \left\{y^T b : y \in D\right\}$. *In particular, both the minimum and the maximum are well-defined and attained.*

Proof. (of Theorem 4.11) Set $n = |X|, m = |\mathcal{F}|$ and let A be the $m \times n$ incidence matrix of the set system $\mathcal{F}$: rows correspond to sets, columns correspond to points, and the entry corresponding to a point p and a set S is $\mathbf{1}_S(p)$. It is easy to check that:

$$\tau^* (\mathcal{F}) = \min \left\{\mathbf{1}_n^T x : x \geq 0, Ax \geq \mathbf{1}_m\right\},$$

$$\nu^* (\mathcal{F}) = \max \left\{y^T \mathbf{1}_m : y \geq 0, y^T A \leq \mathbf{1}_n^T\right\},$$

where $\mathbf{1_n} \in \mathbb{R}^\mathbf{n}$ denotes the (column) vector of all 1's of length n. Indeed, the vectors $x \in \mathbb{R}^n$ satisfying $x \geq 0$ and $Ax \geq \mathbf{1}_m$ correspond precisely to the fractional transversals of $\mathcal{F}$, and similarly, the $y \in \mathbb{R}^n$ with $y \geq 0$ and $y^\mathrm{T} A \leq \mathbf{1}_n^\mathrm{T}$ correspond to the fractional packings. There is at least one fractional transversal (e.g. $x = \mathbf{1}_n$), and at least one fractional packing ($y = 0$), so Fact 4.12 applies and shows that $\nu^* (\mathcal{F}) = \tau^* (\mathcal{F})$.

At the same time, $\tau^* (\mathcal{F})$ is the minimum of the linear function $x \mapsto \mathbf{1}_n^\mathrm{T} x$ over a polyhedron, and such a minimum, since it is finite, is attained at a vertex. The inequalities describing the polyhedron have rational coefficients, and so all vertices have rational coordinates as well. □

So the moral is that it is easy to calculate or bound the fractional transversal number of a set system (linear programming is in **P**), while it is difficult to calculate the actual transversal number (integer programming in **NP**-hard). Thus it is often very useful when one can obtain some bounds for τ in terms of τ^* (and independent of the size of $\mathcal{F}$).

Such a bound can easily be obtained using ε-nets.

Corollary 4.13. *Let $\mathcal{F}$ be a finite set system on a (possible infinite) set X with* $\mathrm{VC}(\mathcal{F}) \leq d$. *Then we have* $\tau(\mathcal{F}) \leq Cd\tau^*(\mathcal{F}) \ln \tau^*(\mathcal{F})$.

Proof. Let $r := \tau^* (\mathcal{F})$. Since $\mathcal{F}$ is finite, we may assume that an optimal fractional transversal $\phi : X \to [0, 1]$ is concentrated on a finite set Y (we can pick a single point from each class in the Venn diagram in our family and work with this finite set instead of X). This ϕ, after rescaling, defines a probability measure μ on X, by taking $\mu(\{y\}) = \frac{1}{r}\phi(y)$ for all $y \in Y$. Then each $S \in \mathcal{F}$ has $\mu(S) \geq \frac{1}{r}$ by the definition of a fractional transversal (Definition 4.7), and so a $\frac{1}{r}$-net for $\mathcal{F}$ with respect to μ is a transversal. By Remark 3.14 it follows that there is a transversal of size $O(dr \ln r)$. □

4.2. **"Weak ε-nets" and convex sets.** By Exercise 3.16 we know that existence of ε-nets for all subsystems $(Y, \mathcal{F} \cap Y)$ of $(X, \mathcal{F})$ of size depending just on εimplies finite VC dimension. However, there are some very important examples of families admitting ε-nets, yet without all of their subsystems admitting ε-nets. Somewhat confusingly, this is called "admitting weak ε-nets" in the literature.

Fact 4.14. *("Weak ε-net theorem for convex sets") For every $d \geq 1$, $\varepsilon > 0$ there is some $N = N(d, \varepsilon)$ such that:*

for every finite $X \subseteq \mathbb{R}^d$, if μ is a probability measure concentrated on X, then there exists an ε-net of size $\leq N$ for the family of all convex subsets of $\mathbb{R}^d$, with respect to μ.

Remark 4.15. Best known upper bounds are $N\left(2, \frac{1}{r}\right) = O\left(r^2\right)$ in the plane and $O\left(r^d (\log r)^{c(d)}\right)$ for all d, with a fixed constant $c = c(d) > 0$. Slightly super-linear lower bounds are known, there are examples with $\Omega\left(r \log^{d-1} r\right)$ [15].

Example 4.16. Let $X = \mathbb{R}^2$ and let Y be the set of points on the unit circle. Let $\mathcal{F}$ be the family of convex polygons. Now considering the sets system $(Y, \mathcal{F} \cap Y)$, it seems conceivable that to choose an ε-net of points on Y we should put sufficiently many points equidistantly on the circle. Since any finite set of points on the unit circle can be shattered by convex polygons inscribed into the circle, the smallest size of an ε-net is forced to depend on the choice of a probability measure μ (Exercise 3.16). However, for the original system $(X, \mathcal{F})$ this is not the case since we can choose an ε-net of bounded size using points *inside* the unit disk.

There are some striking and mutually enriching similarities between the convex and the finite VC-dimension worlds when one is concerned with packings, transversals, etc. In our proofs we will try to work at the maximal level of generality, concentrating on the finite VC-dimension setting (and commenting on the convex counterparts of the results).

4.3. **Fractional Helly property.** Now we investigate further methods of bounding $\tau^*(\mathcal{F})$. Recall classical theorem of Helly about convex sets.

Fact 4.17. *(Helly's theorem) Let $\mathcal{F}$ be a finite family of convex sets in $\mathbb{R}^d$. Assume that any $d+1$ sets from $\mathcal{F}$ have a point in common. Then the whole family $\mathcal{F}$ has a non-empty intersection.*

Exercise 4.18. Helly's theorem does not hold for families of finite VC dimension. Namely, for every $d \in \mathbb{N}$, find an example of a finite family $\mathcal{F}$ of VC-dimension 2 such that it does not satisfy the conclusion of Fact 4.17 with respect to d.

What if in the setting of Helly's theorem, not all $(d+1)$-tuples of sets from $\mathcal{F}$ have non-empty intersections, but only a large fraction of them?

Fact 4.19. *(Fractional Helly theorem for convex sets) For every dimension $d \geq 1$, for every $\alpha > 0$ there exists $\beta = \beta(d, \alpha) > 0$ with the following property.*

If $S_1, \ldots, S_n$ are convex sets in $\mathbb{R}^d$, $n \geq d+1$, and for at least $\alpha\binom{n}{d+1}$ of the $(d+1)$-subsets I of $[n]$ we have $\cap_{i \in I} S_i \neq \emptyset$, then there is a point contained in at least βn sets among the F_i's.

Remark 4.20. Fact 4.19 was established by Katchalski and Liu in [25], the optimal bound is known to be $\beta = 1 - (1-\alpha)^{\frac{1}{d+1}}$ [24].

Definition 4.21. Let $(X, \mathcal{F})$ be a set system. We say that $\mathcal{F}$ has *fractional Helly number k* if for every $\alpha > 0$ there exists a $\beta > 0$ such that if $S_1, \ldots, S_n \in \mathcal{F}$ are such that $\bigcap_{i \in I} S_i \neq \emptyset$ for at least $\alpha\binom{n}{k}$ sets $I \in \binom{[n]}{k}$, then there is some $J \subseteq [n]$ such that $|J| \geq \beta n$ and $\bigcap_{i \in J} S_i \neq \emptyset$.

By the *fractional Helly number* of $\mathcal{F}$ we mean the smallest k with this property, and we say that $\mathcal{F}$ satisfies *fractional Helly property* if it has a finite fractional Helly number.

Remark 4.22. Again, turns out that this fractional Helly property is more robust and better behaved than the original Helly property.

1. For convex lattice sets in $\mathbb{Z}^d$ (i.e., intersections of convex sets in $\mathbb{R}^d$ with the d-dimensional integer lattice), the Helly number is 2^d, while the fractional Helly number is only $d+1$ [9].
2. If $\mathcal{F}$ has fractional Helly number k then the family $\{S_1 \cup S_2 : S_1, S_2 \in \mathcal{F}\}$ also has fractional Helly number k. This fails badly for Helly numbers.

Besides, it turns out that fractional Helly property holds for families of finite VC-dimension.

Theorem 4.23. *(Matousek, [34]) Let $(X, \mathcal{F})$ be a set system with $\pi^*_{\mathcal{F}}(n) = o\left(n^k\right)$ as $n \to \infty$ (in particular this holds if $\mathrm{vc}^*(\mathcal{F}) < k$, e.g. when $\mathrm{VC}^*(\mathcal{F}) \leq k-1$). Then $\mathcal{F}$ has fractional Helly number k.*

Proof. Let $\mathcal{F}$ and k satisfy the assumption, and let $\alpha > 0$ be arbitrary.

Let $S_1, \ldots, S_n \in \mathcal{F}$ be arbitrary. Given $I \subseteq [n]$ we write S_I for $\bigcap_{i \in I} S_i$.

So assume now that $S_I \neq \emptyset$ for at least $\alpha\binom{n}{k}$-many $I \in \binom{[n]}{k}$. It is enough to prove the conclusion of the theorem for all n large enough (as otherwise for β sufficiently small it is enough to have a point in a single S_i).

Using the assumption $\pi^*(m) = o\left(m^k\right)$, we may thus fix m so that $\pi^*_{\mathcal{F}}(m) < \frac{1}{4}\alpha\binom{m}{k}$ and set $\beta = \frac{1}{2m}$. Finally, by the previous paragraph we may assume that n is so large that $\beta n \geq m$.

For contradiction, suppose that no point in X is common to βn of the S_i's. Let us fix $J \in \binom{[n]}{m}$ and $I \in \binom{J}{k}$. We call the pair (J, I) *good* if there is a point $x \in X$ with $x \in S_i$ for all $i \in I$ and $x \notin S_j$ for all $j \in J \setminus I$. We bound from below the probability that a pair (J, I) chosen uniformly at random is good.

We first choose a random $I \in \binom{[n]}{k}$, and then we choose $m-k$ elements of $J \setminus I$ at random from $[n] \setminus I$. By assumption the probability that $S_I \neq \emptyset$ is at least α. If $S_I \neq \emptyset$, we fix one point $x \in S_I$. By the assumption x is contained in fewer than βn of the S_i's, and so the probability that none of the sets S_j with $j \in J \setminus I$ contains x is at least

$$\frac{\binom{\lceil (1-\beta)n \rceil}{m-k}}{\binom{n-k}{m-k}} \geq \prod_{i=0}^{m-k-1} \frac{(1-\beta)n - i}{n-i} \geq \prod_{i=0}^{m-1} \frac{(1-\beta)n - m}{n-m} \geq \left(\frac{(1-\beta)n - m}{n-m}\right)^m .$$

Since we assumed that $m \leq \beta n$ and $\beta = \frac{1}{2m}$, the above expression is at least $\left(\frac{n - \beta n - m}{n-m}\right)^m = \left(1 - \frac{n}{n-m}\beta\right)^m \geq (1-2\beta)^m = \left(1 - \frac{1}{m}\right)^m \geq \frac{1}{4}$. Therefore the probability of a random pair (J, I) being good is at least $\frac{1}{4}\alpha$.

If we choose a random $J \in \binom{[n]}{m}$, the expected number of $I \in \binom{J}{k}$ with (J, I) good is at least $N = \frac{1}{4}\alpha\binom{m}{k}$, and so there exists a J with at least this many I's.

But this violates the assumption $\pi^*_{\mathcal{F}}(m) < N$ since the sets indexed by J have at least N non-empty fields in their Venn diagram. □

Remark 4.24. I don't know if the assumption $\pi_{\mathcal{F}}(n) = \mathrm{o}\left(n^k\right)$ as $n \to \infty$ is strictly weaker than the assumption $\mathrm{vc}(\mathcal{F}) < k$. It is conceivable that there are set systems $(X, \mathcal{F})$ with $\pi_{\mathcal{F}}(n) \sim \frac{n^2}{\log n}$, which would give a counterexample, however I don't know any such example (Fact 2.22 gives an example with $\pi_{\mathcal{F}}(n) = n^2 \log n$).

4.4. (p, q)-theorems.

Definition 4.25. For integers $p \geq q$ we say that a set system $(X, \mathcal{F})$ satisfies the *(p, q)-property* if out of any p sets from $\mathcal{F}$, there are at least q of sets among them with a non-empty intersection.

Example 4.26. Let μ be a probability measure on $\mathbb{R}^d$ and consider all convex sets with measure at least δ. If $\delta > \frac{q}{p}$ then this family satisfies the (p, q)-property. We'll see that in some sense all families with the (p, q)-property of this form (with possibly a much smaller δ).

With this terminology, Helly's theorem (Fact 4.17) says that any finite family of convex sets in $\mathbb{R}^d$ satisfying the $(d+1, d+1)$-property has a non-empty intersection, i.e. admits a transversal of size 1.

A generalization of this was conjectured by Hadwiger and Debrunner, and many years later proved by Alon and Kleitman.

Fact 4.27. *(Alon, Kleitman* [5]*) Let p, q, d be integers with $p \geq q \geq d + 1$. Then there exists a number $N = N(d, p, q)$ such that: if $\mathcal{F}$ is a finite family of convex sets in $\mathbb{R}^d$ satisfying the (p, q)-property then $\tau(\mathcal{F}) \leq N$.*

The proof has a modular structure combining all of the ideas considered so far in this section and admits certain generalizations. We prove that an analog holds for families of finite VC-dimension (it does not formally imply the Alon-Kleitman theorem for convex sets).

Theorem 4.28. *[Alon, Kleitman + Matousek] Let $p \geq q \geq d + 1$ be arbitrary natural numbers. Then there is some $N = N(d, p, q)$ such that if $(X, \mathcal{F})$ is a finite set system of* VC*-codensity $\leq d$, then $\tau(\mathcal{F}) \leq N$.*

Proof. Let $(X, \mathcal{F})$ be a finite set system with $n = |\mathcal{F}|$. Since we are not trying to optimize N, it is enough to prove the theorem for $q = d + 1$.

(1) By Corollary 4.13 we know that $\tau(\mathcal{F})$ is bounded by a function of $\tau^*(\mathcal{F})$ (due to the existence of ε-nets), so it is enough to bound $\tau^*(\mathcal{F})$. By Theorem 4.11 we know that $\tau^*(\mathcal{F}) = \nu^*(\mathcal{F})$, so it is enough to bound $\nu^*(\mathcal{F})$.

(2) The first observation is that if $\mathcal{F}$ satisfies the $(p, d+1)$ condition, then many $(d+1)$-tuples from $\mathcal{F}$ have a non-empty intersection. This can be seen by double counting. Every p-tuple of sets from $\mathcal{F}$ contains at least one $(d+1)$-tuple with a non-empty intersection, and a single $(d+1)$-tuple is contained in $\binom{n-d+1}{p-d+1}$ p-tuples. Therefore there are at least

$$\frac{\binom{n}{p}}{\binom{n-(d+1)}{p-(d+1)}} \geq \alpha \binom{n}{d+1}$$

intersecting $(d+1)$-tuples, with $\alpha = \alpha(p, d) > 0$. The fractional Helly theorem (Theorem 4.23) then implies that there is some $\beta = \beta(d, \alpha) > 0$ such that at least βn sets from $\mathcal{F}$ have a point in common.

By removing this βn sets from $\mathcal{F}$ and iterating, we would get that $\tau(\mathcal{F}) \leq \mathrm{O}(\log n)$. However, to get rid of this $\log n$ factor needs some more work.

(3) How is this related to the fractional packing number $\nu^*(\mathcal{F})$? This shows that a factional packing $\psi : \mathcal{F} \to [0, 1]$ that has the same value on all sets of $\mathcal{F}$ cannot have size larger than $\frac{1}{\beta}$, as otherwise a point lying in βn sets would receive weight greater than 1, contradicting the definition of fractional packing. The trick

for handling other fractional packings is to consider the sets in $\mathcal{F}$ with appropriate multiplicities.

(4) Let $\psi : \mathcal{F} \to [0,1]$ be an optimal fractional packing ($\sum_{S\in\mathcal{F},x\in S}\psi(S) \leq 1$ for all $x \in X$). As noted in Theorem 4.11 we may assume that the values of ψ are rational numbers. Write $\psi(S) = \frac{m(S)}{D}$, where D and $m(S)$ are integers (D is a common denominator). Let us form a new collection $\mathcal{F}_m$ of sets, by putting $m(S)$ copies of each $S \in \mathcal{F}$ into $\mathcal{F}_m$ — so $\mathcal{F}_m$ is a multiset of sets.

Let $N = |\mathcal{F}_m| = \sum_{S\in\mathcal{F}} m(S) = D\nu^*(\mathcal{F})$. Suppose that we could conclude the existence of a point a lying in at least βN sets in $\mathcal{F}_m$ (counted with multiplicity). Then

$$1 \geq \sum_{S\in\mathcal{F}:a\in S} \psi(S) = \sum_{S\in\mathcal{F}:a\in S} \frac{m(S)}{D} = \frac{1}{D}\beta N = \beta\nu^*(\mathcal{F}),$$

and so $\nu^*(\mathcal{F}) \leq \frac{1}{\beta}$.

(5) So we would like to apply Helly's theorem to $\mathcal{F}_m$. If $\mathcal{F}$ is a family of finite VC dimension, then the fractional Helly theorem holds for finite multisets of sets from $\mathcal{F}$ (Exercise).

The new family $\mathcal{F}_m$ does not have to satisfy the $(p, d+1)$-condition, since the $(p, d+1)$-condition for $\mathcal{F}$ speak only of p-tuples of distinct sets from $\mathcal{F}$, while a p-tuple of sets from $\mathcal{F}_m$ may contain multiple copies of the same set.

Fortunately, $\mathcal{F}_m$ does satisfy the $(p', d+1)$-condition with $p' = d(p-1)+1$. Indeed, a p'-tuple of sets of $\mathcal{F}_m$ contains at least $d+1$ copies of the same set, or it contains p distinct sets, in the first case those $d+1$ copies clearly have a non-empty intersection, and in the second case the $(p, d+1)$-condition for $\mathcal{F}$ applies.

Using the fractional Helly theorem as before, we find a point a in common to at least βN sets of $\mathcal{F}_m$ for some $\beta = \beta(p,d)$. By (4) this gives an upper bound on ν^*, and by (1) we can conclude. □

4.5. **Another sufficient condition for bounded transversals.** So the two main assumptions on the set system used in the proof of the Alon-Kleitman theorem are the fractional Helly property and the existence of ε-nets. We show that the second assumption can be omitted at the price of strengthening the first one (we are following [4]).

We generalize the notion of the packing number first.

Definition 4.29. Let $\nu_d(\mathcal{F})$ denote the largest size of $\mathcal{M} \subseteq \mathcal{F}$ such that any point in $x \in X$ belongs at most to d sets from $\mathcal{M}$.

Remark 4.30. (1) $\nu_1(\mathcal{F}) = \nu(\mathcal{F})$,

(2) $\mathcal{F}$ satisfies the (p,q)-property iff $\nu_{q-1}(\mathcal{F}) < p$,

(3) $\frac{\nu_d(\mathcal{F})}{d} \leq \nu^*(\mathcal{F})$.

Definition 4.31. For $(X, \mathcal{F})$ a set system, the *set-cover number* $\rho(\mathcal{F})$ is the minimal number of sets from $\mathcal{F}$ required to cover all the points in X. Note that $\rho(\mathcal{F}) = \tau(\mathcal{F}^*)$, where $\mathcal{F}^*$ is the dual set system (see Definition 1.8). Similarly, the *fractional set-cover number* is defined by $\rho^*(\mathcal{F}) = \tau^*(\mathcal{F}^*)$.

Exercise 4.32. The following are equivalent for $\mathcal{F}$ (with $g(x) = f\left(\frac{1}{x}\right)$):

(1) There is a function g such that $\tau(\mathcal{G}) \leq g(\tau^*(\mathcal{G}))$ for every $\mathcal{G} \subseteq \mathcal{F}$.

(2) There is a function f such that for every ε and every multiset $Y \subseteq X$ there is an ε-net of size at most $f(\varepsilon)$ with respect to the uniform counting

measure (i.e. there is $Z \subseteq X$ with $|Z| \leq f(\varepsilon)$ such that $Z \cap S \neq \emptyset$ for every $S \in \mathcal{F}$ with $|S \cap Y| \geq \varepsilon |Y|$).

Definition 4.33. We write that $\mathcal{F}$ satisfies $\mathrm{FH}(k, \alpha, \beta)$ if for every $S_1, \ldots, S_n \in \mathcal{F}$ (possibly with repetitions) such that the number of k-subsets $I \subseteq [n]$ with $\bigcap_{i \in I} S_i \neq \emptyset$ is at least $\alpha \binom{n}{k}$, then there exists $J \subseteq [n]$ such that $|J| \geq \beta n$ and $\bigcap_{j \in J} S_j \neq \emptyset$.

So $k \in \mathbb{N}$ is a fractional Helly number for $\mathcal{F}$ if $\forall \alpha > 0 \exists \beta = \beta(\alpha) > 0$ such that $\mathrm{FH}(k, \alpha, \beta)$ holds. It may happen that we cannot find a $\beta > 0$ for all $\alpha > 0$ but there are some α and $\beta > 0$ with $\mathrm{FH}(k, \alpha, \beta)$. Then we speak of the weak fractional Helly property.

First we generalize the first part of the proof of Alon and Kleitman.

Theorem 4.34. (1) *For every d and p there exists an $\alpha > 0$ such that the following holds.*
For any finite family $\mathcal{F}$ satisfying $\mathrm{FH}(d+1, \alpha, \beta)$ with some $\beta > 0$ and with $\nu_d(\mathcal{F}) < p$ we have $\tau^*(\mathcal{F}) \leq T$ for some $T = T(p, d, \beta)$.

(2) *For every d and $p, k \geq d+1$, and $\beta_0 > 0$ there exists $\alpha > 0$ such that the following holds.*
For any finite family $\mathcal{F}$ satisfying $\mathrm{FH}(d+1, 1, \beta_0)$, $\mathrm{FH}(k, \alpha, \beta)$ for some $\beta > 0$ and $\nu_d(\mathcal{F}) < p$ we have $\tau^*(\mathcal{F}) \leq T$ for some $T = T(p, d, k, \beta_0, \beta)$.

Proof. (1) The proof of the first part is the same as in Theorem 4.28, let us recall it briefly.

Namely, to bound τ^* by T we need to bound ν^* by T. For that we take a fractional packing $\psi : \mathcal{F} \to [0,1]$ with rational coefficients attaining the maximum 4.11, say $\psi(S) = \frac{n_S}{D}$. Then we take a new family $\{S_1, \ldots, S_n\}$ containing n_S copies of S for each $S \in \mathcal{F}$. This family satisfies the $(p', d+1)$-property with $p' = (p-1)d+1$. By double counting we get that there are at least $\binom{n}{p'} / \binom{n-d-1}{p'-d-1} \geq \alpha \binom{n}{d+1}$ index sets $I \in \binom{[n]}{p'}$ with $\bigcap_{i \in I} S_i \neq \emptyset$, for a suitable $\alpha = \alpha(p, d)$. By $\mathrm{FH}(d+1, \alpha, \beta)$ there is a point x in at least βn of the S_i's. On the other hand, since the multiset $\{S_1, \ldots, S_n\}$ was defined using a fractional matching, no point is in more than $n / \nu^*(\mathcal{F})$ of the sets S_i, and we conclude that $\tau * (\mathcal{F}) = \nu^*(\mathcal{F}) \leq 1/\beta$.

(2) We assume that $\mathcal{F}$ satisfies $\mathrm{FH}(d+1, 1, \beta_0)$ and $\mathrm{FH}(k, \alpha, \beta)$ with a suitable $\alpha > 0$ and some $\beta > 0$, and has the $(p, d+1)$-property. We define $S_1, \ldots, S_n$ using an optimal fractional packing as above, and again it suffices to show that there is a point common to at least βn of the S_i's.

We want to show that there are at least $\alpha \binom{n}{k}$ good index sets $K \in \binom{[n]}{k}$ with $\alpha = \alpha(p, d, k, \beta_0) > 0$, as then we can use $\mathrm{FH}(k, \alpha, \beta)$ (where K is *good* if $\bigcap_{i \in K} S_i \neq \emptyset$).

To this end, let $m = m(p, d, k, \beta_0)$ be a sufficiently large integer (independent of n). It suffices to prove that each index set $M \in \binom{[n]}{m}$ contains at least one good k-element subindex K, since then the total number of good k-tuples is at least $\binom{n}{m} / \binom{n-k}{m-k} \geq \alpha \binom{n}{k}$. To exhibit a good k-tuple in a given m-tuple M we use Ramsey's theorem.

For each $I \in \binom{M}{p'}$, we choose a good $(d+1)$-element $J = J(I) \subset I$ (we use the $(p', d+1)$-property where p' is as in the proof of (1)). This $J(I)$ has one of $\binom{p'}{d+1}$ types, where the type is given by the relative positions of the elements of $J(I)$ among the elements of I (in the natural ordering of I). By Ramsey's theorem, if m is sufficiently large, there exists an r-element $N \subseteq M$, with r still large, such that all $I \in \binom{N}{p'}$ have the same type of $J(I)$. Let $i_1 < i_2 < \ldots < i_r$ be the elements of

N in the increasing order, let $s = \left\lfloor \frac{r}{p'} \right\rfloor$ and let $L = \{i_{p'}, i_{2p'}, \ldots, i_{sp'}\}$. Now all the $J \in \binom{L}{d+1}$ are good, since for each of them we can find an $I \in \binom{N}{p'}$ with $J(I) = J$ (as we have enough space between the elements of J, we can choose the remaining elements of I such that J has the order type of $J(I)$, and so is good).

By $\mathrm{FH}(d+1, 1, \beta_0)$ applied to $\{S_i : i \in L\}$, there are at least $\beta_0 s$ sets among $\{S_i : i \in L\}$ sharing a common point. If $\beta_0 s \geq k$, which can be guaranteed by setting m sufficiently large, we have obtained a good k-tuple contained in M. This allows to conclude as explained above. □

Then the second part is an abstraction of the proof of the existence of weak ε-nets for convex sets.

Theorem 4.35. *For every integer $d \geq 1$ there exists $\alpha > 0$ such that the following holds.*

Let $\mathcal{F}$ be a finite family of sets and let $\mathcal{F}^{\cap} = \{\bigcap \mathcal{H} : \mathcal{H} \subseteq \mathcal{F}\}$ be the family of all intersections of the sets in $\mathcal{F}$. If $\mathcal{F}^{\cap}$ satisfies $\mathrm{FH}(d+1, \alpha, \beta)$ *with some $\beta > 0$ then we have $\tau(\mathcal{F}) \leq c_1 \tau^*(\mathcal{F})^{c_2}$, where c_1 and c_2 depend only on d and β.*

Remark 4.36. The proof gives c_2 exponential in d, in the known examples τ is only slightly super-linear in τ^*.

The proof is an abstraction of one of the proofs in the convex case. Even though we are guided by the geometric example, somewhat amusingly the proof proceeds just by counting.

Recall Tverberg's theorem.

Fact 4.37. *(Tverberg's theorem) Let d and r be given natural numbers. Then for any set $A \subseteq \mathbb{R}^d$ of at least $(d+1)(r-1)+1$ points there exist r pairwise disjoint subsets $A_1, \ldots, A_r \subseteq A$ such that $\bigcap_{i=1}^{r} \mathrm{Conv}(A_i) \neq \emptyset$.*

Let $c : 2^X \to 2^X$ denote the closure operation induced by $\mathcal{F}$ given by $c(A) = \bigcap \{S \in \mathcal{F} : A \subseteq S\}$ and $c(A) = X$ if no $S \in \mathcal{F}$ contains A. This is an abstract analogue of the convex hull. For a multiset $\{x_1, \ldots, x_m\} \subseteq X$ and $I \subseteq [m]$ put $G_I = c(\{x_i : i \in I\})$.

Proposition 4.38. *(A Tverberg-type theorem) Let $\mathcal{F}$ be a finite family and suppose that $\mathcal{F}^{\cap}$ satisfies* $\mathrm{FH}\left(d+1, \frac{1}{4}, \beta\right)$ *for some $\beta > 0$. Then there exist integers $a = a(d, \beta)$ and $b = b(d, \beta)$ such that:*

for every multiset $\{x_1, \ldots, x_{ab}\} \subseteq X$ there are $d+1$ pairwise disjoint subsets $I_1, \ldots, I_{d+1} \in \binom{[ab]}{a}$ with

$$\bigcap_{i=1}^{d+1} G_{I_i} \neq \emptyset.$$

That is, a sufficiently large (multi-)set can be partitioned into $d+1$ parts whose closures have a common point.

Remark 4.39. Here $\alpha = \frac{1}{4}$ can be replaced by any other constant < 1, if a and b are chosen suitably.

Proof. Let $b = \left\lceil \frac{d}{\beta} \right\rceil + 1$ and $a = b^d$. Let $m = \binom{ab}{a}$ and consider the multiset $\mathcal{S} = \left\{ G_I : I \in \binom{[ab]}{a} \right\}$, its sets are members of $\mathcal{F}^{\cap}$. We want to apply fractional

Helly to $\mathcal{S}$, so we need to show that at least $\frac{1}{4}$ of the $(d+1)$-tuples of sets in $\mathcal{S}$ intersect.

We check that at least $\frac{1}{4}$ of all $(d+1)$-tuples $(I_1, \ldots, I_{d+1})$ of pairwise distinct a-element index sets $I_i \subset [ab]$ satisfy $\bigcap_{i=1}^{d+1} I_i \neq \emptyset$ (this is more than necessary, and definitely implies that $\bigcap_{i=1}^{d+1} G_{I_i} \neq \emptyset$). Intuitively this is because $d+1$ independent random a-element subsets of $[ab]$ are very likely to be all distinct and to have a point in common, since a is very large compared to b.

The relative fraction of intersecting $(d+1)$-tuples of distinct a-element subsets of $[ab]$ is

$$\frac{\left|\left\{(I_1, \ldots, I_{d+1}) \in \binom{[ab]}{a}^{d+1} : I_i \neq I_j \text{ for } i \neq j \text{ and } \bigcap_{i=1}^{d+1} I_i \neq \emptyset\right\}\right|}{m\,(m-1)\ldots(m-d)} \geq$$

$$\frac{\left|\left\{(I_1, \ldots, I_{d+1}) \in \binom{[ab]}{a}^{d+1} : \bigcap_{i=1}^{d+1} I_i \neq \emptyset\right\}\right|}{m\,(m-1)\ldots(m-d)} - \frac{m^{d+1} - m\,(m-1)\ldots(m-d)}{m\,(m-1)\ldots(m-d)} \geq$$

$$\frac{ab\binom{ab-1}{a-1}^{d+1} - \binom{ab}{2}\binom{ab-2}{a-2}^{d+1}}{m^{d+1}} - \frac{1}{4} \geq$$

$$\frac{a}{b^d} - \frac{a^2}{2b^{2d}} - \frac{1}{4} = 1 - \frac{1}{2} - \frac{1}{4} = \frac{1}{4}$$

Here the second line is obtained by considering "(intersecting $(d+1)$-tuples) - ($(d+1)$-tuples - $(d+1)$-sets, i.e. tuples with repetitions)", and the third line is obtained by "pick an element in the intersection, pick the rest of the tuples) -(tuples with 2 points in common)".

By FH $\left(d+1, \frac{1}{4}, \beta\right)$ applied to $\mathcal{S}$, there exists an $\mathcal{H} \subseteq \binom{[ab]}{a}$ such that $\bigcap_{I \in \mathcal{H}} G_I \neq \emptyset$ and

$$|\mathcal{H}| \geq \lfloor \beta m \rfloor > \frac{d}{b}\binom{ab}{a}.$$

Thus $\mathcal{H}$ contains a significant fraction of all possible a-tuples of indices, and such a large system has to contain $d+1$ disjoint a-tuples. With our parameters, we can use the following.

Fact 4.40. *(Frankl,* [20, Theorem 10.3]*) Suppose that $\mathcal{F} \subseteq \binom{X}{k}$, $|X| = n \geq ks$ and $\mathcal{F}$ contains no s pairwise disjoint sets. Then $|\mathcal{F}| \leq (s-1)\binom{n-1}{k-1}$ holds. (For $s = 2$ this is the Erdős-Ko-Rado theorem).*

It implies that there are pairwise disjoint $I_1, \ldots, I_{d+1} \in \mathcal{H}$. This proves the proposition. □

Barany proved the following selection lemma:

Fact 4.41. *If $P \subset \mathbb{R}^d$ is an n-point multiset, then there exists a point x contained in the convex hulls of at least $c\binom{n}{d+1}$ subsets of P of cardinality $d+1$, where $c = c\,(d) > 0$.*

The following is an abstract analogue.

Proposition 4.42. *(Selection lemma) Let $\mathcal{F}$ be a finite family such that $\mathcal{F}^{\cap}$ satisfies* FH $(d+1, \alpha, \beta)$ *with a suitable $\alpha = \alpha\,(d) > 0$ and some $\beta > 0$. Then for any* *multiset* $\{x_1, \ldots, x_n\} \subseteq X$ *there exists a family* $\mathcal{H} \subseteq \binom{[n]}{a}$ *such that* $|\mathcal{H}| \geq \lambda \binom{n}{a}$ *and* $\bigcap_{I \in \mathcal{H}} G_I \neq \emptyset$*, where* $a = a(d, \beta)$ *is as in Proposition 4.38 and* $\lambda = \lambda(d, \beta) > 0$*.*

Proof. Let $\mathcal{S} = \left\{ G_I : I \in \binom{[n]}{a} \right\}$, we want to show that a significant fraction of the $(d+1)$-tuples from $\mathcal{S}$ intersect, in order to apply fractional Helly. Let

$$T = \left\{ \{I_1, \ldots, I_{d+1}\} : I_i \in \binom{[n]}{a}, I_i \cap I_j = \emptyset \text{ for } i \neq j \text{ and } \bigcap_{i=1}^{d+1} G_{I_i} \neq \emptyset \right\}.$$

Proposition 4.38 implies that for each subset $J \in \binom{[n]}{ab}$, there exist pairwise disjoint $I_1, \ldots, I_{d+1} \in \binom{J}{a}$ such that $\bigcap_{i=1}^{d+1} G_{I_i} \neq \emptyset$, and so each J contributes a $(d+1)$-tuple in T. On the other hand, for any given $\{I_1, \ldots, I_{d+1}\} \in T$, the $a(d+1)$ indices in $I_1 \cup \ldots \cup I_{d+1}$ are contained in $\binom{n-a(d+1)}{ab-a(d+1)}$ of the (ab)-tuples J. Therefore

$$|T| \geq \frac{\binom{n}{ab}}{\binom{n-a(d+1)}{ab-a(d+1)}} \geq \left(\frac{n}{ab}\right)^{a(d+1)} \geq \frac{1}{(ab)^{a(d+1)}} \binom{\binom{n}{a}}{d+1}$$

and we can conclude by $\mathrm{FH}(d+1, \alpha, \beta)$ applied to $\mathcal{S}$ with $\alpha = \frac{1}{(ab)^{a(d+1)}}$. □

Proof. (of Theorem 4.35) Again by Theorem 4.11, we know that the value of $\tau^*(\mathcal{F})$ is attained on some rational-valued $f : X \to [0, 1]$ which is non-zero only on finitely many points, say $x_1, \ldots, x_r$. We write $f(x_i) = \frac{n_i}{D}$ with integers n_i and D, and we let $Y = \{y_1, \ldots, y_n\}$ be the multiset obtained from X by taking each x_i with multiplicity n_i. We have $|Y| = n = \sum_{i=1}^{r} n_i = \tau^*(\mathcal{F}) D$ and $|Y \cap S| \geq D = \frac{n}{\tau^*(\mathcal{F})}$ for all $S \in \mathcal{F}$.

(We are following an argument of the existence of "weak ε-nets" in the convex case).

We choose a transversal Z for $\mathcal{F}$ by the following greedy algorithm.

Initially $Z = \emptyset$. Having already put $z_1, \ldots, z_k$ into Z, we check if there is a D-element subset $J \subseteq [n]$ such that $G_J = c(\{y_i : i \in J\})$ contains none of the $z_1, \ldots, z_k$.

- If there is no such J then the current Z intersects the closures of all D-element subsets of Y and, in particular, it is a transversal for $\mathcal{F}$ (as by above $|Y \cap S| \geq D$ for all $S \in \mathcal{F}$).
- If such a J exists, we apply selection lemma (Proposition 4.42) to the set $\{y_i : i \in J\}$. This yields some point z_{k+1} that is contained in G_I for at least $\lambda \binom{D}{a}$ a-tuples $I \subseteq J$. (We may assume $D \geq a$ and thus $\lambda \binom{D}{a} > 0$, for otherwise Y will work as a small transversal).

Call an a-tuple $I \subseteq [n]$ *alive* if $G_I \cap \{z_1, \ldots, z_k\} = \emptyset$ and *dead* otherwise.

Initially, all the $\binom{n}{a}$ of a-tuples are alive, and on each step adding z_{k+1} to Z kills at least $\lambda \binom{D}{a}$ of the a-tuples currently alive. As $\mathcal{F}$ is finite, after finitely many steps all tuples will be dead. So the size of the transversal found by the algorithm is at most

$$\frac{\binom{n}{a}}{\lambda \binom{D}{a}} \leq \frac{1}{\lambda} \left(\frac{en}{D}\right)^a \leq \frac{e^a}{\lambda} \left(\tau^*(\mathcal{F})\right)^a,$$

as wanted. □

Combining, we obtain an abstract (p, q)-theorem.

Theorem 4.43. *For every $p \geq d \geq 1$ there is some $\alpha > 0$ such that the following holds.*

if $\mathcal{F}$ is a finite set system such that $\mathcal{F}^{\cap}$ satisfies $\mathrm{FH}(d+1, \alpha, \beta)$ and $\mathcal{F}$ satisfies the $(d+1, p)$-property, then $\tau(\mathcal{F}) \leq T$ for some $T = T(d, p, \beta)$.

4.6. **Complementary examples.** First we discuss a general construction.

Let $G = (V, E)$ be a graph, and let Ξ denote the system of all non-empty anti-cliques (independent sets) in G. We define a family $\mathcal{F}$ with Ξ as the ground set and with the sets $S_v = \{A \in \Xi : v \in A\}$ for $v \in V$. The following properties are easy to check:

- $\mathcal{F}$, as well as $\mathcal{F}^{\cap}$, have Helly number 2 (i.e. satisfy $\mathrm{FH}(2, 1, 1)$).
- If G contains no clique of size p as a subgraph, then $\mathcal{F}$ has the $(p, 2)$-property.
- $\tau(\mathcal{F}) = \chi(G)$ (the chromatic number of G, i.e. the smallest number of colors needed to color the vertices of G so that no two adjacent vertices have the same color).
- $\tau^*(\mathcal{F}) = \chi_f(G)$ (the fractional chromatic number; say that a graph has an a/b-coloring if to each vertex of G one can assign a b-element subset of $\{1, 2, \ldots, a\}$ in such a way that adjacent vertices are assigned disjoint subsets; define $\chi_f(G) = \inf\left\{\frac{a}{b} : G \text{ can be } a/b\text{-colored}\right\}$).

Example 4.44. There exists a family $\mathcal{F}$ such that $\mathcal{F}^{\cap}$ has Helly number 2 and with $\nu(\mathcal{F}) \leq 2$ (i.e. satisfying the $(3, 2)$-property) for which $\tau^*(\mathcal{F})$ is arbitrarily large.

In the above construction, it suffices to choose a triangle-free graph G with arbitrarily large fractional chromatic number. It is known that $\chi_f(G) \geq |V(G)| / \alpha(G)$, where $\alpha(G)$ is the independence number (i.e. the cardinality of the largest independent set). So it suffices that $|V(G)| / \alpha(G)$ is arbitrarily large. Many constructions of such graphs are known, for example the well-known probabilistic construction of Erdős of graphs with large girth (i.e. the length of the shortest cycle in the graph) and large chromatic number works here.

Example 4.45. There exists a family $\mathcal{F}$ satisfying the $(3, 2)$-property and $\mathrm{FH}\left(2, 0, \frac{1}{3}\right)$ (i.e. among any n-sets from $\mathcal{F}$, at least $\frac{n}{3}$ have a common point), such that $\mathcal{F}^{\cap}$ has Helly number 2, with $\tau^*(\mathcal{F}) \leq 3$ and with $\tau(\mathcal{F})$ arbitrarily large.

We start with G a Kneser graph with the vertex set $\binom{[m]}{k}$ and with two k-tuples connected by an edge if and only if they are disjoint. It is well-known that the chromatic number is $m - 2k + 2$ [31]. If we set $m = 3k - 1$, it is easy to see that G is triangle-free and $\chi_f < 3$.

To verify $\mathrm{FH}\left(2, 0, \frac{1}{3}\right)$ we need to check that for every multiset $\{S_1, \ldots, S_n\}, S_i \in \binom{[3k-1]}{k}$, there is a subsystem of at least $\frac{n}{3}$ k-tuples with a common intersection. This is because the sum of the sizes of the S_i's is $nk > \frac{nm}{3}$ and so some point is contained in at least $\frac{n}{3}$ of the S_i's.

Problem 4.46. Is Fractional Helly for distinct sets sufficient? (as opposed to sets with repetitions).

5. Zarankiewicz's problem

5.1. **Zarankiewicz's problem in finite VC-dimension.** Let $G = (P, Q, E)$ be a bipartite graph, i.e. $E \subseteq P \times Q$ and P, Q are disjoint.

The following is a more general version of Fact 2.6.

Fact 5.1. *(Kővári-Sós-Turán,* [27]*) Let $G = (P, Q, E)$ be a bipartite graph with $|P| = m, |Q| = n$. Then if G is $K_{k,k}$-free, we have*

$$|E(G)| \leq c_k \left(mn^{1-\frac{1}{k}} + n \right),$$

where c_k only depends on k.

This bound is tight for $k = 2, 3$ and for $k \geq 5$ and $m = n$ the lower bound $\Omega \left(n^{2-\frac{2}{k}} (\log k)^{1/(k^2-1)} \right)$ is known [14].

For any vertex $q \in Q$, let $N_G(q)$ denote the neighborhood of q in G, i.e. the set of vertices in P connected to q.

Consider the set system $(P, \mathcal{F})$, where $\mathcal{F} = \{N_G(q) \subseteq P : q \in Q\}$. The dual of $(P, \mathcal{F})$ is the set system $(\mathcal{F}, \mathcal{F}^*)$, where $\mathcal{F}^* = \{\{A \in \mathcal{F} : p \in A\} : p \in P\}$.

Theorem 5.2. *(Fox, Pach, Sheffer, Suk, Zahl* [19]*) Let $G = (P, Q, E)$ be a bipartite graph with $|P| = m, |Q| = n$ and such that the set system $\mathcal{F}_1 = \{N(q) : q \in Q\}$ satisfies $\pi_{\mathcal{F}_1}(z) = O(z^d)$. Then if G is $K_{k,k}$-free, we have*

$$|E(G)| \leq c \left(mn^{1-\frac{1}{d}} + n \right),$$

where $c = c(c, d, k)$.

So the exponent only depends on the VC-dimension of the graph, and k only affects the constant.

We need some preparatory lemmas.

Let $(P, \mathcal{F})$ be a set system on a ground set P. The *distance* between two sets $A_1, A_2 \in \mathcal{F}$ is $|A_1 \Delta A_2|$. The *unit distance* graph $\mathrm{UD}(\mathcal{F})$ is the graph with vertex set $\mathcal{F}$, and its edges are pairs of sets (A_1, A_2) from $\mathcal{F}$ that have distance 1. (Equivalently, this corresponds to the edges of the Boolean cube).

Lemma 5.3. *(Haussler* [22, Lemma 2]*) If $\mathcal{F}$ is a set system of VC-dimension d_0 on a ground set P, then the unit distance graph $\mathrm{UD}(\mathcal{F})$ has at most $d_0 |\mathcal{F}|$ edges.*

Proof. This is a slight elaboration on the proof of the Sauer-Shelah lemma using the shifting technique (Lemma 1.5), and we use it as an opportunity to recall the details.

Let us identify P with $[n]$, and let us view $\mathcal{F}$ as a subset of $\{0, 1\}^n$, and $\mathrm{UD}(\mathcal{F}) = (V, E)$.

For each index i, $1 \leq i \leq n$, and each $v = (v_1, \ldots, v_n) \in V$, if $v_i = 1$ and the vector $v' = (v_1, \ldots, v_{i-1}, 0, v_{i+1}, \ldots, v_n)$ is not in V, then let $S_{i,V}(v) = v'$ (we say that v is *shifted* to v'), otherwise let $S_{i,V}(v) = v$.

We define the shift of V on index i, denoted $S_i(V)$, by $S_i(V) = \{S_{i,V}(v) : v \in V\}$.

Let $S_i(E)$ denote the set of edges in the subgraph of the unit graph induced by $S_i(V)$. We claim that:

(1) $|S_i(V)| = |V|$,
(2) $|S_i(E)| \geq |E|$,
(3) for any index set I, if I is shattered by $S_i(V)$ then I is shattered by V. Hence $\mathrm{VC}(S_i(V)) \leq \mathrm{VC}(V)$.

(1) is obvious.

To verify (2), we map the edges of E in a 1-1 manner into the edges of $S_i(E)$.

Assume $(u,v) \in E$. If neither u nor v are shifted then this edge is unaffected by the shift, so map it to itself.

If both u and v are shifted then this edge is mapped to the edge $(S_{i,V}(u), S_{i,V}(v))$.

Finally, assume that v is shifted, but u is not. Then u and v must differ on some index $j \neq i$, and we must have $u_i = v_i = 1$. Since u is not shifted, $u' = (u_1, \ldots, u_{i-1}, 0, u_{i+1}, \ldots, u_n) \in V$. It follows that $(u', S_{i,V}(v)) \in S_i(E)$. Hence we can map (u,v) to $(u', S_{i,V}(v))$. The resulting map is easily 1-1.

(3) Suppose that a sequence I of k indices is shattered by $S_i(V)$. If i is not in I, then clearly I is also shattered by V since $V|_I = S_i(V)|_i$ in this case. So let us assume that $i \in I$, without loss of generality $i = 1$ and $I = (1, \ldots, k)$. Since I is shattered by $S_i(V)$, for every $u \in \{0,1\}^k$ there is a $v \in S_i(V)$ with $v_j = u_j$, $1 \leq j \leq k$. However, if $u_1 = 1$ then we must have v and $v' = (0, v_2, \ldots, v_n)$ both in V, otherwise v would have been shifted, and hence not be in $S_i(V)$. This implies that I is shattered by V.

Now beginning with V, shift V repeatedly on any sequence of (not necessarily distinct) indices until no more non-trivial shifts are possible, i.e. until we obtain a set W such that $S_i(W) = W$ for all $1 \leq i \leq n$. This must happen eventually since each non-trivial shift reduces the total number of ones in the vectors of V. Let F be the set of edges in the unit distance graph induced by W. By the above result we have $|W| = |V|$, $|F| \geq |E|$ and $\mathrm{VC}(W) \leq d_0$.

Let us write $u \leq v$ if $u_i \leq v_i$ for all i, $1 \leq i \leq n$. We claim that W is closed downward under $\leq$, i.e. if $v \in W$ and $u \leq v$, then $u \in W$. It is clear that if $u \leq v \in W$ and u differs from v on only one index i, then $u \in W$ (otherwise one more non-trivial shift of W would be possible). The claim follows by induction.

It follows that if $v \in W$, then the set of indices i for which $v_i = 1$ is shattered by W. Since $\mathrm{VC}(W) \leq d_0$, this implies that no vector in W contains more than d_0 ones. Therefore

$$|V| = |W| \leq \sum_{i=0}^{d_0} \binom{n}{i}$$

and

$$\frac{|E|}{|V|} \leq \frac{|F|}{|W|} \leq d_0.$$

The last inequality is since a vector in $\{0,1\}^n$ with at most d ones can have unit edges to at most d vectors with fewer ones. □

We say that a set system $\mathcal{F}$ is (k,δ)-separated if for any k sets $A_1, \ldots, A_k \in \mathcal{F}$ we have $|(A_1 \cup \ldots \cup A_k) \setminus (A_1 \cap \ldots \cap A_k)| \geq \delta$.

So e.g. if $k = 2$, then $\mathcal{F}$ is $(2,\delta)$-separated if $|A_1 \Delta A_2| \geq \delta$ for all $A_1, A_2 \in \mathcal{F}$.

We have the following upper bound on the number of sets in a separated VC-family.

Lemma 5.4. *(Packing lemma) Let $(P, \mathcal{F})$ be a set system, $|P| = m$ and $\pi_{\mathcal{F}}(z) = O(z^d)$. If $\mathcal{F}$ is (k,δ)-separated, then $|\mathcal{F}| \leq c' \left(\frac{m}{\delta}\right)^d$, where $c' = c'(c,d,k)$.*

Before proving this lemma, we show how it can be used to prove Theorem 5.2.

Proof. (of Theorem 5.2). Let $\mathcal{F}_1 = \{N(q) : q \in Q\}$ and $\mathcal{F}_2 = \{N(p) : p \in P\}$. Notice that the dual set system of $\mathcal{F}_2$ is isomorphic to the set system $\mathcal{F}_1$.

Given a set of k points $\{q_1, \ldots, q_k\} \subseteq Q$, we say that a set $B \in \mathcal{F}_2$ crosses $\{q_1, \ldots, q_k\}$ if $\{q_1, \ldots, q_k\} \cap B \neq \emptyset$ and $\{q_1, \ldots, q_k\} \cap (Q \setminus B) \neq \emptyset$.

Claim. There exist k points $q_1, \ldots, q_k \in Q$ such that at most $2c' \frac{m}{n^{\frac{1}{d}}}$ sets from $\mathcal{F}_2$ cross $\{q_1, \ldots, q_k\}$, where c' is as defined in Lemma 5.4.

For the sake of contradiction, suppose that every set of k points has at least $2c'm/n^{1/d}$ sets from $\mathcal{F}_2$ crossing it. Then the dual set system $\mathcal{F}_2^*$ is (k, δ)-separated with $\delta = 2c'm/n^{1/d}$, and has the property that $\pi_{\mathcal{F}_2^*}(z) = \pi_{\mathcal{F}_1}(z) \leq cz^d$ for all z. Then by Lemma 5.4 we have

$$n = |\mathcal{F}_2^*| \leq c' \left(\frac{m}{\delta}\right)^d.$$

Hence $\delta \leq (c')^{1/d} m/n^{1/d}$, a contradiction — and the claim is proved.

So let now $q_1, \ldots, q_k$ be the set of k points such that at most $2c'm/n^{1/d}$ sets in $\mathcal{F}_2$ cross it. Since G is $K_{k,k}$-free, there are at most $(k-1)$ points $p_1, \ldots, p_{k-1} \in P$ with the property that the neighborhood $N_G(p_i)$ contains $\{q_1, \ldots, q_k\}$ for $1 \leq i \leq k-1$. Therefore, the neighborhood of q_1 contains at most $2c'm/n^{1/d} + (k-1)$ points (i.e. the sets that cross it, upper bound given by the lemma + the sets that don't cross it, of which there are $k-1$). We remove q_1, and repeat this argument until there are less than k vertices remaining in Q and see that

$$|E(G)| \leq (k-1)m + \sum_{i=k}^{n} \left(2c' \frac{m}{i^{1/d}} + (k-1)\right) \leq c_1 \left(mn^{1-\frac{1}{d}} + n\right)$$

for sufficiently large $c_1 = c_1(c, d, k)$. □

Exercise 5.5. Prove a weaker conclusion that $|\mathcal{F}| \leq O\left(\left(\frac{m}{\delta}\right)^d \log^d\left(\frac{m}{\delta}\right)\right)$ using ε-nets to obtain a small set with few sets cutting it instead of Lemma 5.4.

Now we prove Lemma 5.4.

Proof. Assume for contradiction that $|\mathcal{F}| > c' \left(\frac{m}{\delta}\right)^d$ (where the constant c' depends on c, d, k and is determined below).

Since, say, $\pi_{\mathcal{F}}(z) \leq cz^d$ for all z, we have $2^{\mathrm{VC}(\mathcal{F})} \leq c\,\mathrm{VC}(\mathcal{F})^d$, which implies $\mathrm{VC}(\mathcal{F}) \leq 4d \log(cd) =: d_0$.

If $\delta \leq 4k(k-1)d_0$, then the statement is trivial for sufficiently large c' (by the assumption $|\mathcal{F}| \leq cm^d$). Hence we may assume that $\delta > 4k(k-1)d_0$.

Let $S \subseteq P$ be a random s-element subset, for $s = \lceil 4k(k-1)d_0 m/\delta \rceil$. Let $\mathcal{T} = \{A \cap S : A \in \mathcal{F}\}$, and for each $B \in \mathcal{T}$ we define its weight $w(B)$ as the number of sets $A \in \mathcal{F}$ with $A \cap S = B$. Notice that

$$\sum_{B \in \mathcal{T}} w(B) = |\mathcal{F}|.$$

We let E be the edge set of the unit distance graph $\mathrm{UD}(\mathcal{T})$, and define the weight of an edge $e = (B_1, B_2)$ in E as $\min\{w(B_1), w(B_2)\}$. Finally, let

$$W = \sum_{e \in E} w(e).$$

We estimate the expectation of W in two ways.

1) By Lemma 5.3 we know that $\mathrm{UD}(\mathcal{T})$ has a vertex $B \in \mathcal{T}$ of degree at most $2d_0$. Since the weight of all edges containing B is at most $w(B)$, by removing the

vertex $B \in \mathcal{T}$, the total edge weight drops by at most $2d_0 w(B)$. By repeating this argument until there are no vertices left, we have

$$W \leq 2d_0 \sum_{B \in \mathcal{T}} w(B) = 2d_0 |\mathcal{F}|,$$

so in particular $\mathbb{E}(W) \leq 2d_0 |\mathcal{F}|$.

2) Now we bound $\mathbb{E}(W)$ from below. Suppose we first choose a random $(s-1)$-element subset $S' \subseteq P$, and then choose a single element $p \in P \setminus S'$. Then the set $S = S' \cup \{p\}$ is a random s-element set. Let $E_1 \subseteq E$ be the edges in $\mathrm{UD}(\mathcal{T})$ that differ by the element p, and let

$$W_1 = \sum_{e \in E_1} w(e).$$

We have $\mathbb{E}(W) = s\mathbb{E}(W_1)$. Hence we need to bound $E(W_1)$ from below.

To do so, we will estimate $\mathbb{E}(W_1|S')$ from below (the expected value of W_1 when $S' \subseteq P$ is a fixed $(s-1)$-element subset and we choose p at random from $P \setminus S'$.

Divide $\mathcal{F}$ into equivalence classes $\mathcal{F}_1, \ldots, \mathcal{F}_r$ where $A_1, A_2 \in \mathcal{F}$ are in the same class iff $A_1 \cap S' = A_2 \cap S'$. By assumption $\pi_{\mathcal{F}}(z) \leq cz^d$ for all z, we have $r \leq \pi_{\mathcal{F}}(s-1) \leq c_0 \left(\frac{m}{\delta}\right)^d$, where $c_0 = c_0(c, k, d)$.

Let $\mathcal{F}_i$ be one of the equivalence classes, such that $|\mathcal{F}_i| = b$. If an element $p \in P \setminus S'$ is chosen, such that b_1 sets from $\mathcal{F}_i$ contain p and $b_2 = b - b_1$ sets from $\mathcal{F}_i$ do not contain p, then $\mathcal{F}_i$ gives rise to an edge in E_1 of weight $\min\{b_1, b_2\}$. Since $\min\{b_1, b_2\} \geq \frac{b_1 b_2}{b}$, we will estimate $\mathbb{E}(b_1 b_2)$ from below when picking p at random. Notice that $b_1 b_2$ is the number of ordered pairs of sets in $\mathcal{F}_i$ that differ on point p. Hence

$$\mathbb{E}(b_1 b_2) \geq \sum_{(A_1, A_2) \in \mathcal{F}_i \times \mathcal{F}_i} \mathbb{P}(p \in A_1 \Delta A_2) = \sum_{(A_1, A_2) \in \mathcal{F}_i \times \mathcal{F}_i} \frac{|A_1 \Delta A_2|}{m - s + 1} . (*)$$

Now given any k sets $A_1, \ldots, A_k \in \mathcal{F}_i$ we have

$$\bigcup_{2 \leq i \leq k} A_1 \Delta A_j = (A_1 \cup \ldots \cup A_k) \setminus (A_1 \cap \ldots \cap A_k).$$

Since $\mathcal{F}_i$ is (k, δ)-separated, we have

$$\sum_{2 \leq j \leq k} |A_1 \Delta A_j| \geq |(A_1 \cup \ldots \cup A_k) \setminus (A_1 \cap \ldots \cap A_k)| \geq \delta.$$

Therefore every k sets in $\mathcal{F}_i$ contain a pair of sets (A_1, A_j) such that $|A_1 \Delta A_j| \geq \frac{\delta}{k-1}$.

Define an auxiliary graph $G_i = (\mathcal{F}_i, E_i)$ whose vertices are the members in $\mathcal{F}_i$ and two sets $A_1, A_2 \in \mathcal{F}_i$ are adjacent if and only if $|A_1 \Delta A_2| \geq \frac{\delta}{k-1}$.

Recall Turán's theorem: If G is a graph on n vertices which is K_{r+1}-free, then the number of edges in G is at most $\frac{r-1}{r} \frac{n^2}{2}$.

Since G_i does not contain an independent set of size k, by Turán's theorem we have $|E_i| \geq \frac{b(b-k)}{2k}$. $(b^2 - \frac{k-1}{k} \frac{b^2}{2} = \frac{2kb^2 - kb^2 + b^2}{2k} = \frac{kb^2 + b^2}{2k} \ldots)$. Therefore

$$\sum_{(A_1, A_2) \in \mathcal{F}_i \times \mathcal{F}_i} |A_1 \Delta A_2| \geq 2 \frac{b(b-k)}{2k} \frac{\delta}{k-1} = \frac{\delta}{k(k-1)} b(b-k). (**)$$

By combining $(*)$ and $(**)$, we have

$$\mathbb{E}\left(b_1 b_2\right) \geq \frac{\delta}{k\left(k-1\right)m} b\left(b-k\right).$$

Since $\min\left\{b_1, b_2\right\} \geq \frac{b_1 b_2}{b}$, the expected contribution of $\mathcal{F}_i$ to W_1 is at least $\frac{\delta}{k(k-1)m}\left(b-k\right)$. Summing over all classes, we have

$$\begin{aligned}\mathbb{E}\left(W_1\right) &\geq \frac{\delta}{k\left(k-1\right)m}\sum_{i=1}^{r}\left(\left|\mathcal{F}_i\right|-k\right)\\ &= \frac{\delta}{k\left(k-1\right)m}\left(\left|\mathcal{F}\right|-kr\right)\\ &\geq \frac{\delta}{k\left(k-1\right)m}\left(\left|\mathcal{F}\right|-kc_0\left(\frac{m}{\delta}\right)^d\right).\end{aligned}$$

Finally, recall that we are assuming that $\left|\mathcal{F}\right| > c'\left(\frac{m}{\delta}\right)^d$ and that we took $s = \lceil 4k\left(k-1\right)d_0 m/\delta\rceil$. Taking c' sufficiently large with respect to k and c_0, we have:

$2d_0\left|\mathcal{F}\right| \geq \mathbb{E}\left(W\right) = s\mathbb{E}\left(W_1\right) \geq \frac{s\delta}{k(k-1)m}\left(\left|\mathcal{F}\right|-kc_0\left(\frac{m}{\delta}\right)^d\right) \geq 4d_0\left|\mathcal{F}\right|-k4d_0c_0\left(\frac{m}{\delta}\right)^4$,

which implies $\left|\mathcal{F}\right| \leq c'\left(\frac{m}{\delta}\right)^d$, where $c' = c'\left(c,d,k\right)$. □

This result can be used (along with other ideas) to obtain even stronger bounds in the case of semialgebraic graphs (in particular giving a nice generalization of the Szemeredi-Trotter theorem over the reals).

Theorem 5.6. *(Fox, Pach, Sheffer, Suk, Zahl [19]) Let $G = (P, Q, E)$ be a semialgebraic bipartite graph in $(\mathbb{R}^{d_1}, \mathbb{R}^{d_2})$ such that E has description complexity at most t, $|P| = m, |Q| = n$. If G is $K_{k,k}$-free, then*

(1) $|E(G)| \leq c_1\left((mn)^{\frac{2}{3}} + m + n\right)$ *for* $d_1 = d_2 = 2$,

(2) $|E(G)| \leq c_2\left((mn)^{\frac{d}{d+1}+\varepsilon} + m + n\right)$ *for* $d_1 = d_2 = d$,

(3) $|E(G)| \leq c_3\left(m^{\frac{d_2(d_1-1)}{d_1d_2-1}} n^{\frac{d_1(d_2-1)}{d_1d_2-1}} + m + n\right)$ *for all* d_1, d_2.

Here, ε is an arbitrarily small constant and $c_1 = c_1(t,k), c_2 = c_2(d,t,k,\varepsilon), c_3 = c_3(d_1,d_2,t,k,\varepsilon)$.

This theorem, in turn, admits certain model-theoretic generalizations (work in progress with Sergei Starchenko).

5.2. **Historic remarks.** Packing lemma for $(2,\delta)$-separated systems was originally proved by Chazelle, see [35].

6. Compression schemes and PAC learning

6.1. **Compression schemes.** We stick to the situation with a finite underlying set for now, to avoid any measurability issues, etc.

Definition 6.1. (Concept class) Let $(X, \mathcal{F})$ be a set system, in the context of learning theory the term *concept class* is used, and the elements of $\mathcal{F}$ are called *concepts*. We will identify a concept with its indicator function, i.e. we identify $C \in \mathcal{F}$ with the function $\mathbf{1}_S : X \to \{0,1\}$. Given a sequence $\bar{x} = (x_1, \ldots, x_m) \in X^m$, we write $C\left(\bar{x}\right)$ for the sequence $\left(C\left(x_1\right), \ldots, C\left(x_m\right)\right) \in \{0,1\}^m$.

Definition 6.2. (Sample) The set $S_m(X)$ of *samples* of size m contains all the sequences in $(X \times \{0,1\})^m$. Given a concept C and a sequence $\bar{x} = (x_1, \ldots, x_m) \in X^m$, we denote by $S_C(\bar{x})$ the sample $S = (x_i, C(x_i) : i = 1, \ldots m)$. We let $S_m(\mathcal{F}) = \{S_C(\bar{x}) : \bar{x} \in X^n, n \leq m, C \in \mathcal{F}\}$ be the set of $\mathcal{F}$-labeled samples of size m, and let $S(\mathcal{F}) = \bigcup_{m \in \mathbb{N}} S_m(\mathcal{F})$ be the set of all $\mathcal{F}$-labeled samples of finite size.

Definition 6.3. (Compression scheme) Given a concept class $(X, \mathcal{F})$, a *k-compression scheme with side information I* consists of the following:

(1) A finite set I,
(2) The *compression map* $\kappa : S(\mathcal{F}) \to S_k(\mathcal{F}) \times I$ taking $S_C(\bar{x})$ to some $(S_{C'}(\bar{x}'), i)$ such that $S_{C'}(\bar{x}')$ is a subsequence of $S_C(\bar{x})$ and $i \in I$.
(3) The *reconstruction map* $\rho : S_k(\mathcal{F}) \times I \to \{0,1\}^X$ so that for all $S_C(\bar{x}) \in S(\mathcal{F})$ we have $\rho(\kappa(S_C(\bar{x}))) \restriction_{\bar{x}} = C(\bar{x})$, i.e. the reconstruction map is invertible.

First we observe that if $(X, \mathcal{F})$ admits a compression scheme, then it has bounded VC-dimension.

Proposition 6.4. *Assume that $(X, \mathcal{F})$ admits a k-compression scheme. Then* $\mathrm{vc}(\mathcal{F}) \leq k$.

Proof. Let $A \subseteq X$ be an arbitrary finite set, say $A = (a_1, \ldots, a_m)$. For every $C \in \mathcal{F}$, consider the sample $S_C(A)$. We know that $\kappa(S_C(A))$ is given by $S_C(A'), i$ for some $A' \subseteq A$ of size k and $i \in I$, and that $S_C(A)$ can be reconstructed from it. This implies that $|\mathcal{F} \cap A| \leq |A|^k \times |I|$, and so $\pi_{\mathcal{F}}(m) = O(m^k)$, which implies that $\mathcal{F}$ is of finite VC-dimension by Sauer-Shelah. □

The converse was an open problem due to Warmuth which was very recently solved by Moran and Yehudayoff [37]. That is, they show that every family of finite VC-dimension d admits a k-compression scheme with k bounded by a function of d. We present their proof.

Remark 6.5. A model-theoretic version of this problem is concerned with uniform definability of types over finite sets, and for families definable in NIP structures it was resolved in [17].

6.2. **PAC learning.**

Definition 6.6. (1) The concept class $(X, \mathcal{F})$ is *PAC learnable* with d samples, error ε and probability of success $1 - \delta$ if there is a learning map $H : S_d(\mathcal{F}) \to \mathcal{P}(X)$ so that for every $C \in \mathcal{F}$ and for every probability distribution μ on X,

$$\mu^d\left(\left\{\bar{x} \in X^d : \mu\left(H\left(S_C(\bar{x})\right) \Delta C\right) \leq \varepsilon\right\}\right) \geq 1 - \delta.$$

(2) If moreover the image of H is contained in $\mathcal{F}$ then we say that $\mathcal{F}$ is *properly PAC learnable*.
(3) We say that $\mathcal{F}$ is *consistently* PAC learnable with d samples if any map $H : S_d(\mathcal{F}) \to \mathcal{P}(X)$ such that $H(S_C(\bar{x}))|_{\bar{x}} = C|_{\bar{x}}$ for all $\bar{x} \in X^d$ works.

We show that PAC learnability of $\mathcal{F}$ is equivalent to finite VC dimension, but first let us recall ε-nets and ε-approximations.

Fact 6.7. *(VC theorem rephrased, with an improved bound due to Talagrand and Yi, Long, Srinivasan, see* [30]*) Let $d \in \mathbb{N}$ and $\varepsilon, \delta > 0$ be arbitrary. Then there is some $n = n(d, \varepsilon) \in \mathbb{N}$ such that:*

Any set system $(X, \mathcal{F})$ on a finite probability space (X, μ) with $\mathrm{VC}(\mathcal{F}) \leq d$ satisfies
$\mu^n\left(\{(a_1, \ldots, a_n) \in X^n : \sup_{S \in \mathcal{F}} |\mu(S) - \mathrm{Av}(a_1, \ldots, a_n; S)| \geq \varepsilon\}\right) \leq \delta.$
Moreover, n can by bounded by $O\left(\frac{d}{\varepsilon^2 \delta}\right)$.

In particular, *there is* an *ε-approximation* of size at most $O\left(\frac{d}{\varepsilon^2}\right)$.

Theorem 6.8. *(*[13]*, or see* [45] *for a very detailed exposition of the theory) The following are equivalent:*

1. *$\mathcal{F}$ is PAC learnable.*
2. *$\mathrm{VC}(\mathcal{F})$ is finite.*

Proof. (2) implies (1). Let $\mathcal{F}$ have a finite VC-dimension d. Then for any $C \in \mathcal{F}$, the family $\mathcal{F}_C := C \Delta \mathcal{F}$ has the same VC-dimension (Exercise). Let μ be arbitrary. By the ε-net theorem, a randomly chosen tuple $\bar{x}$ of length n that is sufficiently long compared to d is an ε-net for the family $C \Delta \mathcal{F}$, with high probability. That is, $\mu^n\left(\{(a_1, \ldots, a_n) \in X^n : S \in \mathcal{F}, \mu(S \Delta C) > \varepsilon, \bigwedge a_i \notin S \Delta C\}\right) \leq \delta.$

But then any consistent function $H : S_n(\mathcal{F}) \to \mathcal{F}$ such that $H(S_C(\bar{x}))|_{\bar{x}} = C|_{\bar{x}}$ for all $\bar{x} \in X^n$ works.

(1) implies (2). Assume that $\mathrm{VC}(\mathcal{F}) = \infty$, but that $\mathcal{F}$ is PAC learnable via H, with $\varepsilon = \delta = 0.1$ and with samples of length m. That is, for any measure μ, after sampling on a sample of length m, H generates a hypothesis in $\mathcal{F}$ based on it such that $\mu(\mathrm{error}(H) < \varepsilon) > 1 - \delta$.

By assumption there is an $\mathcal{F}$-shattered set S of size $2m$. Let μ be a uniform measure concentrated on S, i.e. probability of every element in S is $\frac{1}{2m}$.

We choose a concept $C \in \mathcal{F}$ that we will try to learn at random on S, i.e. $\mu(C(x_i) = 0) = \frac{1}{2}, \forall x_i \in S$.

Now if the learner selects an iid sample of m instances $\bar{S}$, which by the choice of the measure implies that $\bar{S} \subseteq S$ and outputs some $H = H\left(\bar{S}\right) \in \mathcal{F}$.

The probability of error for each $x_i \notin \bar{S}$ is $\mu(C(x_i) \neq H(x_i)) = \frac{1}{2}$ (as S is shattered, we can select the labels on the $2m - m$ elements of S not seen by $\bar{S}$ arbitrarily). Regardless if H, the probability of the mistake is 0.5.

The expectation on the error of H is $\mathbb{E}(\mathrm{error}(H)) = m \cdot 0 \cdot \frac{1}{2m} + m \cdot \frac{1}{2} \cdot \frac{1}{2m} = \frac{1}{4}$. (this is because we have $2m$ points to sample, from which the error on half of them is 0 as H is consistent on $\bar{S}$ and the error on the remaining half is 0.5).

However, according to the choice of ε, δ and the learnability assumption we have that with probability of at least 0.9 we have $\mathrm{error}(h) \leq 0.1$, and with probability 0.1 then $\mathrm{error}(h) = \beta$ where $0.1 < \beta \leq 1$. Taking the worst case $\beta = 1$ we have $\mathbb{E}(\mathrm{error}(h)) \leq 0.9 \cdot 0.1 + 0.1 \cdot 1 < \frac{1}{4}$. This is a contradiction. □

The following lemma can be thought of as an abstract "approximate" version of the Caratheodory's theorem (if $X \subseteq \mathbb{R}^d$, then each point of $\mathrm{Conv}(X)$ is a convex combination of at most $d + 1$ points of X.)

6.3. Minimax theorem and the construction of compression schemes.

Lemma 6.9. *Let $(X, \mathcal{F})$ be given (as before X and $\mathcal{F}$ are finite), and let $d^* = \mathrm{VC}(\mathcal{F}^*)$. Let μ be a probability measure on $\mathcal{F}$ (i.e. each set in $\mathcal{F}$ is assigned a* *weight) and let $\varepsilon > 0$. Then μ can be ε-approximated in L^∞ by an average of at most $O\left(\frac{d^*}{\varepsilon^2}\right)$ elements from $\mathcal{F}$. That is, there is a multiset $\mathcal{G} \subseteq \mathcal{F}$ of size $|\mathcal{G}| \leq O\left(\frac{d^*}{\varepsilon^2}\right)$ so that for every point $x \in X$ we have*

$$\left| \mu\left(\{S \in \mathcal{F} : x \in S\}\right) - \frac{|\{S \in \mathcal{G} : x \in S\}|}{|\mathcal{G}|} \right| \leq \varepsilon.$$

Proof. Applying the VC-theorem to the dual system $\mathcal{F}^*$. □

The following *minimax theorem of Von Neumann* is a seminal result in game theory. Consider a zero-sum game with 2 players, a row player and a column player. A *pure strategy* of the row player is $r \in [m]$, and a pure strategy of the column player is $j \in [n]$. A *mixed strategy* is a probability measure on pure strategies.

Let M be a binary matrix so that $M(i,j) = 1$ if and only if the row player wins the game when the pure strategies r, j are played.

Then the minimax theorem says:

- if for every mixed strategy q of the column player, there is a mixed strategy p of the row player that guarantees the row player wins with probability at least V, then there is a mixed strategy p^* of the row player so that for all mixed strategies q of the column player, the row player wins with probability at least V.

As a probability measure on $[m]$ is determined by an assignment of weights to each $r \in [m]$, we can view it as a vector $(x_1, \ldots, x_m)$ of length m such that $x_i \geq 0$ and $\sum_{i=1}^m x_i = 1$ (and the same for columns).

Theorem 6.10. *(Minimax) Let $M \in \mathbb{R}^{m \times n}$ be a real matrix. Then*

$$\max_{p \in \Delta^m} \min_{q \in \Delta^n} p^T M q = \min_{q \in \Delta^n} \max_{p \in \Delta^m} p^T M q,$$

where Δ^l is the set of all probability measures on $[l]$.

Proof. Note that the probability of the column player winning with mixed strategies p and q can be computed as $\sum_{i=1}^m \sum_{j=1}^n p_i M(i,j) q_j = p^t M q$.

Now the theorem is a corollary of the strong LP duality theorem 4.12. Consider the following linear programs.

Primal.

Maximize t subject to the following conditions:

$$\begin{aligned} t - \sum_{j=1}^n M(i,j) q_j &\leq 0 \text{ for all } i,\ (x_i) \\ \sum_{j=1}^n q_j &\leq 1,\ (s) \\ q_j &\geq 0 \text{ for all } j. \end{aligned}$$

Dual.

Minimize s subject to the following conditions:

$$\begin{aligned} s-\sum_{i=1}^{m} M\left(i,j\right)p_i &\geq 0 \text{ for all } j,\ \left(y_j\right) \\ \sum_{i=1}^{m} p_i &\geq 1,\ (t) \\ p_i &\geq 0 \text{ for all } i. \end{aligned}$$

Technically we require that $\sum_{i=1}^{m} p_i = \sum_{j=1}^{n} q_j = 1$ as both are probability measures, the relaxation to inequalities are sufficient since $\sum_j q_j < 1$ and $\sum_i p_i$ correspond to suboptimal values for their respective objective functions.

Maximizing t corresponds to the right hand side, and minimizing s corresponds to the left hand side, and by the LP duality these values are equal. □

Finally, we are ready construct the compression scheme.

Theorem 6.11. *(Moran, Yehudayoff* [37]*) If $\mathcal{F} \subseteq \{0,1\}^X$ satisfies $\mathrm{VC}\left(\mathcal{F}\right) = d$ and the dual set system $\mathcal{F}^*$ satisfies $\mathrm{VC}\left(\mathcal{F}^*\right) = d^*$, then $\mathcal{F}$ has a compression scheme of size $O\left(dd^*\right)$.*

In particular (as $\mathrm{VC}\left(\mathcal{F}\right) \leq d$ implies $\mathrm{VC}\left(\mathcal{F}^*\right) < 2^{d+1}$), only assuming that $\mathrm{VC}\left(\mathcal{F}\right) \leq d$ we get a compression scheme of size $2^{O(d)}$.

Proof. First we describe roughly the compression scheme. Recall that by finite VC dimension, $\mathcal{F}$ is PAC learnable.

Given a sample of the form (Y,y) (so Y is a subset of X, and $y = S|_Y$ for some $S \in \mathcal{F}$), the compression identifies $T \leq O\left(d^*\right)$ subsets $Z_1,\ldots,Z_T$ of Y, each of size at most d. It then compresses (Y,y) ti (Z,z) with $Z = \bigcup_{t\in[T]} Z_t$ and $z = y|_Z$. The additional information $i \in I$ allows to recover $Z_1,\ldots,Z_T$ from Z.

The reconstruction process uses the information $i \in I$ to recover $Z_1,\ldots,Z_T$ from Z, and then uses the PAC learning map H to generate T hypothesis $h_1,\ldots,h_T$ defined as $h_T = H\left(Z_t, z|_{Zt}\right)$. The final reconstruction hypothesis $h = \rho\left(\left(Z,z\right),i\right)$ is the majority vote over $h_1,\ldots,h_T$.

Now we give the details.

Since the VC-dimension of $\mathcal{F}$ is d, by Theorem 6.8 there is $s = O\left(d\right)$ and a proper learning map $H : S_s\left(\mathcal{F}\right) \to \mathcal{F}$ so that for every $c \in \mathcal{F}$ and for every probability measure μ on X, there is some $Z \subseteq \mathrm{supp}\left(\mu\right)$ such that $|Z| \leq s$ and $\mu\left(\{x \in X : h_Z\left(x\right) \neq c\left(x\right)\}\right) \leq \frac{1}{3}$, where $h_Z = H\left(Z, c|_Z\right)$.

We will define the compression map. Let $(Y,y) \in S\left(\mathcal{F}\right)$ be arbitrary. Let

$$\mathcal{H} = \mathcal{H}_{Y,y} = \left\{H\left(Z,z\right) : Z \subseteq Y, |Z| \leq s, z = y|_Z\right\} \subseteq \mathcal{F}.$$

Claim. There are $T \leq O\left(d^*\right)$ sets $Z_1,\ldots,Z_T \subseteq Y$, with $|Z_t| \leq s$, so that the following holds.

For $t \in [T]$, let

$$f_t = H\left(Z_t, y|_{z_t}\right). \ \ (*)$$

Then, for every $x \in Y$, we have

$$\left|\left\{t \in [T] : f_t\left(x\right) = y\left(x\right)\right\}\right| > \frac{T}{2}. \ \ (**)$$

Proof of the claim. By the choice of H, for every probability measure μ on Y, there is some $h \in \mathcal{H}$ so that

$$\mu\left(\{x : h(x) = y(x)\}\right) \geq \frac{2}{3}.$$

By the minimax theorem (Theorem 6.10), there is a probability measure ν on $\mathcal{H}$ such that for every $x \in Y$,

$$\nu\left(\{h \in \mathcal{H} : h(x) = y(x)\}\right) \geq \frac{2}{3}.$$

By Lemma 6.9 applied to $\mathcal{H}$ and ν with $\varepsilon = \frac{1}{8}$, there is a multiset $F = \{f_1, \ldots, f_T\} \subseteq \mathcal{H}$ of size $T \leq O(d^*)$ so that for every $x \in Y$,

$$\frac{|t \in T : f_t(x) = y(x)|}{T} \geq \nu\left(\{h \in \mathcal{H} : h(x) = y(x)\}\right) - \frac{1}{8} > \frac{1}{2}.$$

For every $t \in [T]$, by the definition of $\mathcal{H}$ let Z_t be a subset of Y of size $|Z_t| \leq d$ so that $H\left(Z_t, y|_{Z_t}\right) = f_t$.

Assuming the claim, the compression $\kappa(Y, y)$ is defined as $(Z, z), i$ where $Z = \bigcup_{t \in [T]} Z_t, z = y|_Z$, and the additional information $i \in I$ allows to recover the sets $Z_1, \ldots, Z_T$ from the set Z. Thus we can bound the size of I by k^k with $k = O(d^*)\, s \leq O(dd^*)$.

Now we define the reconstruction map.

Given $((Z, z), i)$, i is interpreted as a list of T subsets $Z_1, \ldots, Z_T$ of Z, each of size at most d. For $t \in [T]$, let $h_t = H\left(Z_t, z|_{Z_t}\right)$.

Define $h = \rho((Z, z), i)$ as follows. For every $x \in X$, let $h(x)$ be an element of $\{0, 1\}$ that appears most in the list $\lambda_x((Z, z), i) = (h_1(x), \ldots, h_T(x))$

Finally, we check that this reconstruction map is correct.

Fix $(Y, y) \in S(\mathcal{F})$. Let $((Z, z), i) = \kappa(Y, y)$ and $h = \rho((Z, z), i)$. For $x \in Y$, consider the list

$$\phi_x(Y, y) = (f_1(x), \ldots, f_T(x))$$

defined in the compression process of (Y, y). The list $\phi_x(Y, y)$ is identical to the list $\lambda_x((Z, z), i)$ due to the following reasons:

- Equation $(*)$,
- The information i allows to correctly recover $Z_1, \ldots, Z_T$,
- $y|_{Z_t} = z|_{Z_t}$ for all $t \in [T]$.

Finally, by $(**)$, for every $x \in Y$, the symbol $y(x)$ appears in more than half of the list $\lambda_x((Z, z), i)$, so indeed $h(x) = y(x)$. □

6.4. **Set-theoretic issues around the equivalence of PAC learning and finite VC-dimension.** This part is based on a paper of Pestov [40], and was presented in the Spencinar (`http://www.math.ucla.edu/~sunger/seminars/index.html`).

We consider the more general situation when the underlying set X is infinite.

- Let $(X, \mathcal{B})$ be a standard Borel space, i.e. a complete separable metric space equipped with the σ-algebra of Borel subsets.

- We consider Borel probability measures μ on X (and don't distinguish between μ and its Lebesgue completion, i.e. an extension of μ over a larger σ-algebra of Lebesgue-measurable subsets of Ω).
- A set $A \subseteq X$ is *universally measurable* if it is Lebesgue μ-measurable for every probability measure on X.
- Recall that a subset $N \subseteq X$ is *universally null* if for every non-atomic probability measure μ on $(X, \mathcal{B})$ we have $\mu(N') = 0$ for some Borel set $N' \supseteq N$. Universally null Borel sets are just countable sets.
- Let $\mathcal{F}$ be a family of universally measurable subsets of X ($\mathcal{F}$ is called sometimes a *concept class*).

In the learning model, a set $\mathcal{P}$ of probability measures on X is fixed. Usually $\mathcal{P} = \mathcal{P}(X)$ is the set of all probability measures on X (*distribution-free learning*) or $\mathcal{P} = \{\mu\}$ is a single measure (*learning under a fixed distribution*).

A *learning sample* is a pair (A, B) of finite multisets of elements from X, where $B \subseteq A$ is thought of as the set of points belonging to an unknown set $S \in \mathcal{F}$ that we want to learn by sampling.

The set of all samples of size n (i.e. with $|A| = n$) is usually identified with $(X \times \{0,1\})^n$ (we pair each element $b \in B$ with $\mathbf{1}_A(b)$).

Definition 6.12. A *learning rule* (for $\mathcal{F}$) is a mapping $\mathcal{L} : \bigcup_{n\in\mathbb{N}} (X \times \{0,1\})^n \to \mathcal{F}$ which satisfies the following measurability condition: for every $S \in \mathcal{F}$, $n \in \mathbb{N}$ and $\mu \in \mathcal{P}$, the function $A \mapsto \mu(\mathcal{L}(A, S\cap A)\Delta S)$ from X^n to $\mathbb{R}$ is μ-measurable.
A learning rule is *consistent* (with $\mathcal{F}$) if for all $S \in \mathcal{F}, n \in \mathbb{N}$ and $A \in X^n$ we have $\mathcal{L}(A, S \cap A) \cap A = S \cap A$.
A learning rule $\mathcal{L}$ is *Probably Approximately Correct* (PAC) *under* $\mathcal{P}$ if for every $\varepsilon > 0$ we have

$$\mu^n (A \in X^n : \mu(\mathcal{L}(A, S \cap A)\Delta S) > \varepsilon) \to 0$$

as $n \to \infty$, uniformly over all $S \in \mathcal{F}$ and $\mu \in \mathcal{P}$ (where μ^n is the product measure). Rephrasing, a learning rule $\mathcal{L}$ is PAC with sample complexity function $s(\varepsilon, \delta)$, where ε is the error and δ is confidence, such that for each $S \in \mathcal{F}$ and $\mu \in \mathcal{P}$, for any $n \geq s(\varepsilon, \delta)$ we have $\mu^n (A \in X^n : \mu(\mathcal{L}(A, S \cap A)\Delta S) < \varepsilon) \geq 1 - \delta$.

So the idea is that by sampling on sufficiently long tuples, with high probability our learning rule allows us to guess which set from $\mathcal{F}$ we are looking at.

A family $\mathcal{F}$ is *PAC learnable under* $\mathcal{P}$ if there exists a PAC learning rule $\mathcal{L}$ for $\mathcal{F}$ under $\mathcal{P}$. $\mathcal{F}$ is *consistently learnable* (under $\mathcal{P}$) if every learning rule $\mathcal{L}$ consistent with $\mathcal{F}$ is PAC under $\mathcal{P}$. If $\mathcal{P} = \mathcal{P}(X)$ then $\mathcal{F}$ is *distribution-free PAC learnable.*

For now we are not concerned with the (computational) complexity of our learning rule, but only with its existence.

Definition 6.13. $\mathcal{F}$ is *uniform Glivenko-Cantelli, or UGC, with respect to a family of measures* $\mathcal{P}$ if for each $\varepsilon > 0$ we have

$$\sup_{\mu\in\mathcal{P}} \mu^n \left(\left\{ (a_1, \ldots, a_n) \in X^n : \sup_{S\in\mathcal{F}} |\mu(S) - \mathrm{Av}(a_1, \ldots, a_n; S)| \geq \varepsilon \right\} \right) \to 0 \text{ as } n \to \infty.$$

From Theorem 3.4 (and Section 3.2) we have:

Fact 6.14. *Let $\mathcal{F}$ be a countable family and* $\mathrm{VC}(\mathcal{F})$ *is finite. Then if $\mathcal{P}$ is a family of measures such that each $S \in \mathcal{F}$ is measurable with respect to each $\mu \in \mathcal{P}$ then $\mathcal{F}$ is UGC with respect to $\mathcal{P}$.*

Remark 6.15. The countability assumption can be somewhat relaxed, see Section 3.2 for a discussion. In particular, it is enough to assume that the family $\mathcal{F}$ is *image admissible Souslin* (i.e. $\mathcal{F}$ can be parametrized as $\mathcal{F} = \{C_t : t \in [0,1]\}$ so that the set $\{(x,y) : x \in C_t, t \in [0,1]\}$ is an analytic subset of $\Omega \times [0,1]$) or is *universally separable.*

Proposition 6.16. *Every UGC family $\mathcal{F}$ (with respect to $\mathcal{P}$) is consistently PAC learnable (with respect to $\mathcal{P}$).*

Theorem 6.17. *Every distribution-free PAC learnable $\mathcal{F}$ has finite VC-dimension.*

Proof. The same proof as in Theorem 6.8 goes through. □

Under the measurability assumptions, by the VC-theorem we have that finite VC-dimension implies uniform Glivenko-Cantelli, so implies consistent learnability. Still, to ensure PAC learnability from consistent learnability, we need to prove the existence of a consistent learning rule satisfying the measurability assumptions.

So, for

We saw in Example 3.11 of Durst and Dudley that under CH, without any additional assumptions there is an example of a family $\mathcal{F}$ of VC-dimension 1 on a standard probability space which is not uniformly Glivenko-Cantelli. We consider a slight modification of that example.

Example 6.18. (Blumer et. al.) Assume CH. Let $(X, \mathcal{F})$ be as in the Example 3.11, and let $\mathcal{F}' := \mathcal{F} \cup \{X\}$. We still have that $\mathrm{VC}(\mathcal{F}') = 1$. For a finite sample (A, B) (recall $B \subseteq A$) define $\mathcal{L}(A,B) = I_z$ with $z = \min\{y \in X : B \subseteq I_y\}$.

This learning rule $\mathcal{L}$ is consistent with $\mathcal{F}'$. At the same time, $\mathcal{L}$ is not PAC. Indeed, for the set $X \in \mathcal{F}$ the value of the learning rule $\mathcal{L}(A.X \cap A) = \mathcal{L}(A,A)$ always returns a countable concept I_y for some $y \in X$, and if μ is a non-atomic Borel probability measure on X, then $\mu(X \Delta I_y) = 1$. Thus the set $X \in \mathcal{F}$ cannot be learned with accuracy $\varepsilon < 1$ with a non-zero confidence.

It is important to note that, again under CH, the class $\mathcal{F}'$ is distribution-free PAC learnable.

Indeed, redefine a well-ordering on $\mathcal{F} = \{I_x : x \in X\} \cup \{X\}$ by making X the smallest element (instead of the largest one) and keep the order relation between the other elements the same. Denote the new order by $\prec_1$, and define a learning rule $\mathcal{L}_1$ similarly to $\mathcal{L}$ but with respect to $\prec_1$:

$$\mathcal{L}_1(A,B) = \min_{\prec_1} \left\{ S \in \mathcal{F} : S \cap A = \bigcap_{B \subseteq D \in \mathcal{F}} D \right\}.$$

In essence, $\mathcal{L}_1$ examines all the sets in $\mathcal{F}$ following the transfinite order on them, and returns the first encountered set consistent with sample, provided it exists. To see the difference, let μ be again a non-atomic probability measure on Ω. If $S = X$, then for every sample (A,B) consistent with S, the rule $\mathcal{L}_1$ will return S. If $S \neq X$, then S is countable, and for μ-almost all samples A, $A \cap S$ will be empty, and the set $\mathcal{L}_1(A, \emptyset)$ returned by $\mathcal{L}_1$, while possibly different from S, will be countable, meaning that $\mu(S \Delta \mathcal{L}_1(A,\emptyset)) = 0$.

We only used CH to ensure that every I_y, $y \in X$ is a universally measurable set. We'll see that this can be achieved under a much weaker assumption of MA.

Problem 6.19. Under CH (or MA), this gives an example of a PAC learnable class which is not uniformly Glivenko-Cantelli (even if having finite VC dimension). Can the same combination of properties be achieved without additional set-theoretic assumptions?

To achieve PAC, we may relax the assumption of it being Glivenko-Cantelli.

Lemma 6.20. *Let $\mathcal{F}$ be a class, and let $\mathcal{P}$ be a family of probability measures on X. Suppose there exists a function $s(\varepsilon, \delta)$ and a learning rule $\mathcal{L}$ for $\mathcal{F}$ with the property that for every $S \in \mathcal{F}$, the set $\mathcal{L}^S \cup \{S\}$ is Glivenko-Cantelli with respect to $\mathcal{P}$ with the sample complexity $s(\varepsilon, \delta)$, where $\mathcal{L}^S = \{\mathcal{L}(S \cap A) : A \in X^n, n \in \mathbb{N}\}$. Then $\mathcal{L}$ is PAC under $\mathcal{P}$ with sample complexity $s(\varepsilon, \delta)$.*

Proof. The proof of 6.8 works locally with respect to a fixed set S. □

We recall

Definition 6.21. *Martin's Axiom* (MA) says that no compact Hausdorff topological space with the countable chain condition is a union of strictly less than continuum many nowhere dense subsets.

Thus, it is a stronger statement than the Baire Category Theorem. In particular, CH implies MA, but also MA is consistent with the negation of CH.

Fact 6.22. *(Martin-Solovay, see e.g.* [28, Theorem 2.21]*) Let (X, μ) be a standard Lebesgue non-atomic probability space. Under MA, the Lebesgue measure is $2^{\aleph_0}$-additive. That is, if $\kappa < 2^{\aleph_0}$ and A_α, $\alpha < \kappa$ is a family of pairwise-disjoint measurable subsets of X, then $\bigcup_{\alpha<\kappa} A_\alpha$ is Lebesgue-measurable and $\mu\left(\bigcup_{\alpha<\kappa} A_\alpha\right) = \sum_{\alpha<\kappa} \mu(A_\alpha)$.*

In particular, the union of strictly less than continuum many null subsets of X is a null set.

First we observe that under MA, if every countable subcollection of $\mathcal{F}$ is UGC, then every subcollection of size $< 2^{\aleph_0}$ is UGC. Note that $|\mathcal{F}| \leq 2^{\aleph_0}$ as $\mathcal{F}$ consists of Borel subsets of a standard Borel space.

Lemma 6.23. *(MA) Let $\mathcal{F}$ be a set system and $\mathcal{P}$ a family of probability measures on a standard Borel space $(X, \mathcal{B})$. Then the following are equivalent:*

1. *Every countable subclass of $\mathcal{F}$ is uniform Glivenko-Cantelli with regard to $\mathcal{P}$.*
2. *There is a function $s(\varepsilon, \delta)$ so that every subclass $\mathcal{F}' \subseteq \mathcal{F}$ of cardinality $< 2^{\aleph_0}$ is uniform Glivenko-Cantelli with regard to $\mathcal{P}$, with sample complexity $s(\varepsilon, \delta)$.*

Proof. The implication (2) $\Longrightarrow$ (1) is obvious.

Assume (1) holds, and fix arbitrary $\varepsilon, \delta > 0$. For each countable subclass $\mathcal{F}' \subseteq \mathcal{F}$ let $s(\mathcal{F}, \varepsilon, \delta)$ be the smallest value of the sample complexity which works for all measures in $\mathcal{P}$ (exists as all countable subclasses are uniformly GC). If there is no uniform bound on it, we could choose $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \ldots$ an increasing countable sequence of countable classes such that $s(\mathcal{F}_n, \varepsilon, \delta) > n$. But then $\mathcal{F} = \bigcup_{n\in\mathbb{N}} \mathcal{F}_n$ is countable, and for every $n \in \mathbb{N}$ there is some measure $\mu \in \mathcal{P}$ which requires samples of size $\geq n$ with respect to $\mathcal{F}$, i.e. $\mathcal{F}$ is not UGC with respect to $\mathcal{P}$ — contradicting the assumption (1). So we obtain a uniform bound $s(\varepsilon, \delta)$ that works for all countable subclasses of $\mathcal{F}$.

Now assume MA, and we prove that (2) holds for this s by transfinite induction on $\kappa = |\mathcal{F}'|$, for $\kappa < 2^{\aleph_0}$.

For $\kappa = \aleph_0$ this is the assumption, otherwise write $\mathcal{F}' = \bigcup_{\alpha<\kappa} \mathcal{F}_\alpha$ where $\mathcal{F}_\alpha$ is an increasing chain and $|\mathcal{F}_\alpha| < \kappa$, so by inductive assumption (2) holds for each of $\mathcal{F}_\alpha$.

For every ε and $n \in \mathbb{N}$ the set

$$\left\{ \bar{a} \in X^n : \sup_{S \in \mathcal{F}'} |\mathrm{Av}(\bar{a}; S) - \mu(S)| < \varepsilon \right\} = \bigcap_{\alpha<\kappa} \left\{ \bar{a} \in X^n : \sup_{S \in \mathcal{F}_\alpha} |\mathrm{Av}(\bar{a}; S) - \mu(S)| < \varepsilon \right\}$$

is measurable by Fact 6.22. Given $\delta > 0$ and $n \geq s(\varepsilon, \delta)$, again by Fact 6.22 for every $\mu \in \mathcal{P}$ we have

$$\begin{aligned} \mu^n \left(\left\{ \bar{a} \in X^n : \sup_{S \in \mathcal{F}'} |\mathrm{Av}(\bar{a}; S) - \mu(S)| < \varepsilon \right\} \right) &= \\ \mu^n \left(\bigcap_{\alpha<\kappa} \left\{ \bar{a} \in X^n : \sup_{S \in \mathcal{F}_\alpha} |\mathrm{Av}(\bar{a}; S) - \mu(S)| < \varepsilon \right\} \right) &= \\ \inf_{\alpha<\kappa} \mu^n \left(\left\{ \bar{a} \in X^n : \sup_{S \in \mathcal{F}_\alpha} |\mathrm{Av}(\bar{a}; S) - \mu(S)| < \varepsilon \right\} \right) &\geq \\ 1 - \delta, \end{aligned}$$

as required. □

Next we observe that under MA, if in addition we have a learning rule $\mathcal{L}$ for $\mathcal{F}$ such that $\mathcal{L}$ produces families of hypothesis of size smaller than continuum, then $\mathcal{L}$ is PAC.

Lemma 6.24. *(MA) Let $\mathcal{F}$ be a concept class such that all countable subclasses are UGC with respect to $\mathcal{P}$, and let $\mathcal{L}$* be a learning rule for $\mathcal{F}$ such that for every $S \in \mathcal{F}$, the family $\mathcal{L}^{S,n} = \{\mathcal{L}(S \cap A) : A \in X^n\}$ has cardinality strictly less than continuum. Then $\mathcal{L}$ is PAC with respect to $\mathcal{P}$.

Proof. Let $\mathcal{L}^S = \bigcup_{n \in \mathbb{N}} \mathcal{L}^{S,n}$, note that still $\left|\mathcal{L}^S\right| < 2^{\aleph_0}$ (as $2^{\aleph_0}$ is a regular cardinal, i.e. there are no countable cofinal subsets). By Lemma 6.23 $\mathcal{L}^S$ is UGC with sample complexity $s(\varepsilon, \delta)$. By Lemma 6.20 we conclude. □

Next, we show that a learning rule with this property always exists.

Lemma 6.25. *Let $\mathcal{F}$ be a concept class on a Borel space X, let $\kappa = |\mathcal{F}|$. There exists a consistent learning rule $\mathcal{L}$ for $\mathcal{F}$ such that for every $S \in \mathcal{F}$ and each n, the set $\mathcal{L}^{S,n}$ has cardinality $< \kappa$.*

Moreover, under MA this rule satisfies the measurability condition: for every $S \in \mathcal{F}$, $n \in \mathbb{N}$ and $\mu \in \mathcal{P}$, the function $A \mapsto \mu(\mathcal{L}(A, S \cap A) \Delta S)$ from X^n to $\mathbb{R}$ is μ-measurable.

Proof. Choose a minimal well-ordering of elements of $\mathcal{F}$: $\mathcal{F} = \{S_\alpha : \alpha < \kappa\}$, and for every $A \in X^n$ and $B \subseteq A$ set $\mathcal{L}(A, B) = S_\beta$, where $\beta = \min\{\alpha < \kappa : S_\alpha \cap A = B\}$ provided such a β exists.

Clearly for each $\alpha < \kappa$ we have $\mathcal{L}(A, S_\alpha \cap A) \in \{S_\beta : \beta \leq \tau\}$, which ensures that $\left|\mathcal{L}^{S,n}\right| < \kappa$. Besides, the learning rule $\mathcal{L}$ is clearly consistent.

It remains to check the moreover part. Fix $S = S_\alpha \in \mathcal{F}$, $\alpha < \kappa$. For every $\beta \leq \alpha$ define $D_\beta = \{A \in X^n : S \cap A = S_\beta \cap A\}$. The sets D_β are measurable, and the function $A \to \mu(\mathcal{L}(S \cap A) \Delta S)$ from X^n to $\mathbb{R}$ takes a constant value $\mu(S \Delta S_\alpha)$ on each set $D_\beta \setminus \bigcup_{\gamma<\beta} D_\gamma$, $\beta \leq \alpha$. Such sets, as well as their unions, are measurable under MA using Fact 6.22, and their union is X^n. This implies the measurability condition from the moreover part. □

Combining, we obtain the following theorem.

Theorem 6.26. *(Pestov [40]) Assuming MA, let $\mathcal{F}$ be a family consisting of Borel measurable subsets of a standard Borel space X, and let $\mathcal{P}$ be a family of probability measures on X. Assume that every countable subfamily of $\mathcal{F}$ is UGC with respect to $\mathcal{P}$. Then $\mathcal{F}$ is PAC learnable under $\mathcal{P}$.*

Corollary 6.27. *Under MA, the following are equivalent for every family $\mathcal{F}$ consisting of universally measurable subsets of a Borel space X.*

(1) *$\mathcal{F}$ is distribution-free PAC learnable,*
(2) $\mathrm{VC}(\mathcal{F}) < \infty$.

Proof. (1) $\implies$ (2) always holds (Theorem 6.8). For the converse, as $\mathrm{VC}(\mathcal{F}') \leq \mathrm{VC}(\mathcal{F})$ for every $\mathcal{F}' \subseteq \mathcal{F}$, by Fact every countable subfamily of $\mathcal{F}$ is UGC. Then Theorem 6.26 applies with respect to $\mathcal{P} = \mathcal{P}(X)$. □

7. Colorful versions of fractional Helly, (p, q)-theorem, etc.

8. Discrepancy

References

[1] Terrence M Adams, Andrew B Nobel, et al. Uniform convergence of vapnik–chervonenkis classes under ergodic sampling. *The Annals of Probability*, 38(4):1345–1367, 2010.

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

[3] Ralph Alexander. Geometric methods in the study of irregularities of distribution. *Combinatorica*, 10(2):115–136, 1990.

[4] Noga Alon, Gil Kalai, Jiři Matoušek, and Roy Meshulam. Transversal numbers for hypergraphs arising in geometry. *Advances in Applied Mathematics*, 29(1):79–101, 2002.

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

[6] Noga Alon and Joel H Spencer. *The probabilistic method*. John Wiley & Sons, 2004.

[7] Matthias Aschenbrenner, Alf Dolich, Deirdre Haskell, Dugald Macpherson, Sergei Starchenko, et al. Vapnik–chervonenkis density in some theories without the independence property, ii. *Notre Dame Journal of Formal Logic*, 54(3-4):311–363, 2013.

[8] Patrick Assouad. Densité et dimension. In *Annales de l'Institut Fourier*, volume 33, pages 233–282, 1983.

[9] Imre Bárány et al. A fractional helly theorem for convex lattice sets. *Advances in Mathematics*, 174(2):227–235, 2003.

[10] Shai Ben-David and Michael Lindenbaum. Localization vs. identification of semi-algebraic sets. *Machine Learning*, 32(3):207–224, 1998.

[11] Ron Blei. *Analysis in integer and fractional dimensions*, volume 71. Cambridge University Press, 2001.

[12] Ron Blei, Yuval Peres, and James Schmerl. Fractional products of sets. *Random Structures & Algorithms*, 6(1):113–119, 1995.

[13] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth. Learnability and the vapnik-chervonenkis dimension. *Journal of the ACM (JACM)*, 36(4):929–965, 1989.

[14] Tom Bohman and Peter Keevash. The early evolution of the h-free process. *Inventiones Mathematicae*, 181(2):291–336, 2010.

[15] Boris Bukh, Jiří Matoušek, and Gabriel Nivasch. Lower bounds for weak epsilon-nets and stair-convexity. *Israel Journal of Mathematics*, 182(1):199–228, 2011.

[16] Artem Chernikov. Lecture notes on stability theory (math 285D). *http://www.math.ucla.edu/~chernikov/teaching/StabilityTheory285D/StabilityNotes.pdf*.

[17] Artem Chernikov and Pierre Simon. Externally definable sets and dependent pairs II. *Transactions of the American Mathematical Society*, 367(7):5217–5235, 2015.

[18] György Elekes. Sums versus products in number theory, algebra and erdos geometry. *Paul Erdos and his Mathematics II*, 11:241–290, 2001.

[19] Jacob Fox, János Pach, Adam Sheffer, Andrew Suk, and Joshua Zahl. A semi-algebraic version of zarankiewicz's problem. *arXiv preprint arXiv:1407.5705*, 2014.

[20] Peter Frankl. The shifting technique in extremal set theory. *Surveys in combinatorics, http://www.renyi.hu/~pfrankl/1987-4.pdf*, 123:81–110, 1987.

[21] Yuri Gurevich and Peter H Schmitt. The theory of ordered abelian groups does not have the independence property. *Transactions of the American Mathematical Society*, 284(1):171–182, 1984.

[22] David Haussler. Sphere packing numbers for subsets of the boolean n-cube with bounded vapnik-chervonenkis dimension. *Journal of Combinatorial Theory, Series A*, 69(2):217–232, 1995.

[23] David Haussler and Emo Welzl. ε-nets and simplex range queries. *Discrete & Computational Geometry*, 2(1):127–151, 1987.

[24] Gil Kalai. Intersection patterns of convex sets. *Israel Journal of Mathematics*, 48(2-3):161–174, 1984.

[25] M Katchalski and A Liu. A problem of geometry in $\mathbb{R}^n$. *Proceedings of the American Mathematical Society*, 75(2):284–288, 1979.

[26] János Komlós, János Pach, and Gerhard Woeginger. Almost tight bounds for ε-nets. *Discrete & Computational Geometry*, 7(1):163–173, 1992.

[27] Tamás Kovári, Vera Sós, and Pál Turán. On a problem of k. zarankiewicz. In *Colloquium Mathematicae*, volume 3, pages 50–57. Institute of Mathematics Polish Academy of Sciences, 1954.

[28] Kenneth Kunen. Set theory, volume 102 of studies in logic and the foundations of mathematics, 1980.

[29] Michael C Laskowski. Vapnik-chervonenkis classes of definable sets. *Journal of the London Mathematical Society*, 2(2):377–384, 1992.

[30] Yi Li, Philip M Long, and Aravind Srinivasan. Improved bounds on the sample complexity of learning. *Journal of Computer and System Sciences*, 62(3):516–527, 2001.

[31] László Lovász. Kneser's conjecture, chromatic number, and homotopy. *Journal of Combinatorial Theory, Series A*, 25(3):319–324, 1978.

[32] HD Macpherson, M Aschenbrenner, A Dolich, D Haskell, and S Starchenko. Vapnik-chervonenkis density in some theories without the independence property, i. *Transactions of the American Mathematical Society*.

[33] Jiří Matoušek. *Lectures on discrete geometry*, volume 212. Springer New York, 2002.

[34] Jirí Matousek. Bounded vc-dimension implies a fractional helly theorem. *Discrete & Computational Geometry*, 31(2):251–255, 2004.

[35] Jiri Matousek. *Geometric discrepancy: An illustrated guide*, volume 18. Springer Science & Business Media, 2009.

[36] Jiří Matoušek, Emo Welzl, and Lorenz Wernisch. Discrepancy and approximations for bounded vc-dimension. *Combinatorica*, 13(4):455–466, 1993.

[37] S. Moran and A. Yehudayoff. Sample compression schemes for VC classes. *ArXiv e-prints*, March 2015.

[38] János Pach and Micha Sharir. Repeated angles in the plane and related problems. *Journal of Combinatorial Theory, series A*, 59(1):12–22, 1992.

[39] János Pach and Gábor Tardos. Tight lower bounds for the size of epsilon-nets. *Journal of the American Mathematical Society*, 26(3):645–658, 2013.

[40] Vladimir Pestov. Pac learnability versus vc dimension: a footnote to a basic result of statistical learning. In *Neural Networks (IJCNN), The 2011 International Joint Conference on*, pages 1141–1145. IEEE, 2011.

[41] Thomas Scanlon. o-minimality as an approach to the andré-Oort conjecture. *Panoramas et Synthèses, to appear*.

[42] Pierre Simon. *A guide to NIP theories*, volume 44. Cambridge University Press, 2015.

[43] Lou Van den Dries. *Tame topology and o-minimal structures*, volume 248. Cambridge university press, 1998.

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

[45] Mathukumalli Vidyasagar. *Learning and generalisation: with applications to neural networks*. Springer Science & Business Media, 2013.