10

Transversals and Epsilon Nets

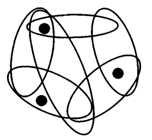
Here we are going to consider problems of the following type: We have a family $\mathcal{F}$ of geometric shapes satisfying certain conditions, and we would like to conclude that $\mathcal{F}$ can be "pierced" by not too many points, meaning that we can choose a bounded number of points such that each set of $\mathcal{F}$ contains at least one of them. Such questions are sometimes called *Gallai-type problems*, because of the following nice problem raised by Gallai: Let $\mathcal{F}$ be a finite family of closed disks in the plane such that every two disks in $\mathcal{F}$ intersect. What is the smallest number of points needed to pierce $\mathcal{F}$? For this problem, the exact answer is known: 4 points always suffice and are sometimes necessary.

We will not cover this particular (quite difficult) result; rather, we consider general methods for proving that the number of piercing points can be bounded. These methods yield numerous results where no other proofs are available. On the other hand, the resulting estimates are usually quite large, and in some simpler cases (such as Gallai's problem mentioned above), specialized geometric arguments provide much better bounds.

Some of the tools introduced in this chapter are widely applicable and sometimes more significant than the particular geometric results. Such important tools include the transversal and matching numbers of set systems, their fractional versions (connected via the duality of linear programming), the Vapnik–Chervonenkis dimension and ways of estimating it, and epsilon nets.

10.1 General Preliminaries: Transversals and Matchings

Let $\mathcal{F}$ be a system of sets on a ground set X; both $\mathcal{F}$ and X may generally be infinite. A subset $T \subseteq X$ is called a *transversal* of $\mathcal{F}$ if it intersects all the sets of $\mathcal{F}$.

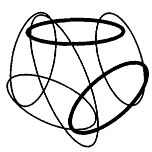


The *transversal number* of $\mathcal{F}$, denoted by $\tau(\mathcal{F})$, is the smallest possible cardinality of a transversal of $\mathcal{F}$.

Many combinatorial and geometric problems, some them considered in this chapter, can be rephrased as questions about the transversal number of suitable set systems.

Another important parameter of a set system $\mathcal{F}$ is the *packing number* (or *matching number*) of $\mathcal{F}$, usually denoted by $\nu(\mathcal{F})$. This is the maximum cardinality of a system of pairwise disjoint sets in $\mathcal{F}$:

$$\nu(\mathcal{F}) = \sup\{|\mathcal{M}|: \mathcal{M} \subseteq \mathcal{F},\ M_1 \cap M_2 = \emptyset \text{ for all } M_1, M_2 \in \mathcal{M},\ M_1 \neq M_2\}.$$



A subsystem $\mathcal{M} \subseteq \mathcal{F}$ of pairwise disjoint sets is called a *packing* (or a *matching*; this refers to graph-theoretic matching, which is a system of pairwise disjoint edges).

Any transversal is at least as large as any packing, and so always

$$\nu(\mathcal{F}) \leq \tau(\mathcal{F}).$$

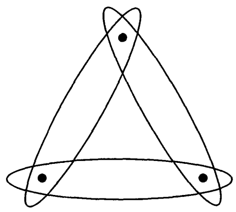
In the reverse direction, very little can be said in general, since $\tau(\mathcal{F})$ can be arbitrarily large even if $\nu(\mathcal{F}) = 1$. As a simple geometric example, we can take the plane as the ground set X and let the sets of $\mathcal{F}$ be n lines in general position. Then $\nu(\mathcal{F}) = 1$, 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.

Fractional packing and transversal numbers. Now we introduce another parameter of a set system, which always lies between ν and τ and which has proved extremely useful in arguments estimating τ or ν. First we restrict ourselves to set systems on *finite* ground sets.

Let $\mathcal{F}$ be a system of subsets of a finite set X. A *fractional transversal* for $\mathcal{F}$ is a function $\varphi: X \to [0,1]$ such that for each $S \in \mathcal{F}$, we have $\sum_{x \in S} \varphi(x) \geq 1$. The size of a fractional transversal φ is $\sum_{x \in X} \varphi(x)$, and the *fractional transversal number* $\tau^*(\mathcal{F})$ is the infimum of the sizes of fractional transversals. So in a fractional transversal, we can take one-third of one point, one-fifth of another, etc., but we must put total weight of at least one full point into every set.

Similarly, a *fractional packing* for $\mathcal{F}$ is a function $\psi: \mathcal{F} \to [0,1]$ such that for each $x \in X$, we have $\sum_{S\in\mathcal{F}:\, x\in S} \psi(S) \leq 1$. So sets receive weights and the total weight of sets containing any given point must not exceed 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 $\mathcal{F}$.

It is instructive to consider the "triangle" system of 3 sets on 3 points,



and check that $\nu = 1$, $\tau = 2$, and $\nu^* = \tau^* = \frac{3}{2}$.

Any packing $\mathcal{M}$ yields a fractional packing (by assigning weight 1 to the sets in $\mathcal{M}$ and 0 to others), and so $\nu \leq \nu^*$. Similarly, $\tau^* \leq \tau$.

We promised one parameter but introduced two: τ^* and ν^*. But they happen to be the same.

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

This is not a trivial result; the proof is a nice application of the duality of linear programming. Here is the version of the linear programming duality we need.

10.1.2 Proposition (Duality of linear programming). *Let A be an $m \times n$ real matrix, $b \in \mathbf{R}^m$ a (column) vector, and $c \in \mathbf{R}^n$ a (column) vector. Let*

$$P = \{x \in \mathbf{R}^n\colon x \geq 0,\ Ax \geq b\}$$

and

$$D = \left\{y \in \mathbf{R}^m\colon y \geq 0,\ y^T A \leq c^T\right\}$$

(the inequalities between vectors should hold in every component). If both $P \neq \emptyset$ and $D \neq \emptyset$, then

$$\min\left\{c^T x\colon x \in P\right\} = \max\left\{y^T b\colon y \in D\right\};$$

in particular, both the minimum and the maximum are well-defined and attained.

This result can be quickly proved by piecing together a larger matrix from A, b, and c and applying a suitable version of the Farkas lemma (Lemma 1.2.5) to it (Exercise 6). It can also be derived directly from the separation theorem.

Let us remark that there are several versions of the linear programming duality (differing, for example, in including or omitting the requirement $x \geq 0$, or replacing $Ax \geq b$ by $Ax = b$, or exchanging minima and maxima), and they are easy to mix up.

Proof of Theorem 10.1.1. Set $n = |X|$ and $m = |\mathcal{F}|$, and let A be the $m \times n$ incidence matrix of the set system $\mathcal{F}$: Rows correspond to sets, columns to points, and the entry corresponding to a point p and a set S is 1 if $p \in S$ and 0 if $p \notin S$. It is easy to check that $\nu^*(\mathcal{F})$ and $\tau^*(\mathcal{F})$ are solutions to the following optimization problems:

$$\begin{aligned} \tau^*(\mathcal{F}) &= \min\left\{\mathbf{1}_n^T x\colon x \geq 0,\ Ax \geq \mathbf{1}_m\right\}, \\ \nu^*(\mathcal{F}) &= \max\left\{y^T \mathbf{1}_m\colon y \geq 0,\ y^T A \leq \mathbf{1}_n^T\right\}, \end{aligned}$$

where $\mathbf{1}_n \in \mathbf{R}^n$ denotes the (column) vector of all 1's of length n. Indeed, the vectors $x \in \mathbf{R}^n$ satisfying $x \geq 0$ and $Ax \geq \mathbf{1}_m$ correspond precisely to the fractional transversals of $\mathcal{F}$, and similarly, the $y \in \mathbf{R}^n$ with $y \geq 0$ and $y^T A \leq \mathbf{1}_n^T$ correspond to the fractional packings. There is at least one fractional transversal, e.g., $x = \mathbf{1}_n$, and at least one fractional packing, namely, $y = 0$, and so Proposition 10.1.2 applies and shows that $\nu^*(\mathcal{F}) = \tau^*(\mathcal{F})$.

At the same time, $\tau^*(F)$ is the minimum of the linear function $x \mapsto \mathbf{1}_n^T x$ 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 are rational points. □

Remark about infinite set systems. Set systems encountered in geometry are usually infinite. In almost all the considerations concerning transversals, the problem can be reduced to a problem about finite sets, usually by a simple ad hoc argument. Nevertheless, we include here a few remarks that can aid a simple consistent treatment of the infinite case. However, they will not be used in the sequel in any essential way.

There is no problem with the definitions of ν and τ in the infinite case, but one has to be a little careful with the definition of ν^* and τ^* to preserve the equality $\nu^* = \tau^*$. Everything is still fine if we have finitely many sets on an infinite ground set: The infinite ground set can be factored into finitely many equivalence classes, where two points are equivalent if they belong to the same subcollection of the sets. One can choose one point in each equivalence class and work with a finite system.

For infinitely many sets, some sort of compactness condition is certainly needed. For example, the system of intervals $\{[i, \infty)\colon i = 1, 2, \ldots\}$ has, according to any reasonable definition, $\nu^* = 1$ but $\tau^* = \infty$.

If we let $\mathcal{F}$ be a family of closed sets in a compact metric space X (compact Hausdorff space actually suffices), we can define $\nu^*(\mathcal{F})$ as $\sup_\psi \sum_{S \in \mathcal{F}} \psi(S)$, where the supremum is over all $\psi\colon \mathcal{F} \to [0,1]$ attaining only finitely many nonzero values and such that $\sum_{S \in \mathcal{F}\colon x \in S} \psi(S) \leq 1$ for each $x \in X$.

For the definition of τ^*, the first attempt might be to consider all functions $\varphi: X \to [0,1]$ attaining only finitely many nonzero values and summing up to at least 1 over every set. But this does not work very well: For example, if we let $\mathcal{F}$ be the system of all compact subsets of $[0,1]$ of Lebesgue measure $\frac{1}{2}$, say, then $\nu^* \le 2$ but τ^* would be infinite, since any finite subset is avoided by some member of $\mathcal{F}$. It is better to define a *fractional transversal* of $\mathcal{F}$ as a Borel measure μ on X such that $\mu(S) \ge 1$ for all $S \in \mathcal{F}$, and $\tau^*(\mathcal{F})$ as the infimum of $\mu(X)$ over all such μ. With this definition, the validity of the first part Theorem 10.1.1 is preserved; i.e., $\nu^*(\mathcal{F}) = \tau^*(\mathcal{F})$ for all systems $\mathcal{F}$ of closed sets in a compact X. The proof uses a little of functional analysis, and we omit it; it can be found in [KM97a]. The rationality of ν^* and τ^* no longer holds in the infinite case.

Bibliography and remarks. Gallai's problem about pairwise intersecting disks mentioned at the beginning of this chapter was first solved by Danzer in 1956, but he hasn't published the solution. For another solution and a historical account see Danzer [Dan86].

Attempting to summarize the contemporary knowledge about the transversal number and the packing number in combinatorics would mean taking a much larger bite than can be swallowed, so we restrict ourselves to a few sketchy remarks. An excellent source for many combinatorial results is Lovász's problem collection [Lov93].

A quite old result relating ν and τ is the famous König's edge-covering theorem from 1912, asserting that $\nu(\mathcal{F}) = \tau(\mathcal{F})$ if $\mathcal{F}$ is the system of edges of a bipartite graph (this is also easily seen to be equivalent to Hall's marriage theorem, proved by Frobenius in 1917; see Lovász and Plummer [LP86] for the history). On the other hand, an appropriate generalization to systems of triples, namely, $\tau \le 2\nu$ for any tripartite 3-uniform hypergraph, is a celebrated recent result of Aharoni [Aha01] (based on Aharoni and Haxell [AH00]), while the generalization $\tau \le (k-1)\nu$ for k-partite k-uniform hypergraphs, known as Ryser's conjecture, remains unproved for $k \ge 4$.

While computing ν or τ for a given $\mathcal{F}$ is well known to be NP-hard, τ^* can be computed in time polynomial in $|X| + |\mathcal{F}|$ by linear programming (this is another reason for the usefulness of the fractional parameter). The problem of approximating τ is practically very important and has received considerable attention. More often it is considered in the dual form, as the *set cover problem*: Given $\mathcal{F}$ with $\bigcup \mathcal{F} = X$, find the smallest subcollection $\mathcal{F}' \subseteq \mathcal{F}$ that still covers X. The size of such $\mathcal{F}'$ is the transversal number of the set system *dual* to $(X, \mathcal{F})$, where each set $S \in \mathcal{F}$ is assigned a point y_S and each point $x \in X$ gives rise to the set $\{y_S: x \in S\}$.

For the set cover problem, it was shown by Chvátal and independently by Lovász that the greedy algorithm (always take a set covering the maximum possible number of yet uncovered points) achieves a solution whose size is no more than $(1 + \ln |X|)$ times larger than the optimal one.[1] Lovász actually observed that the proof implies, for any finite set system $\mathcal{F}$,

$$\tau(\mathcal{F}) \leq \tau^*(\mathcal{F}) \cdot (1 + \ln \Delta(\mathcal{F})),$$

where $\Delta(\mathcal{F})$ is the maximum degree of $\mathcal{F}$, i.e., the maximum number of sets with a common point (Exercise 4). The weaker bound with $\Delta(\mathcal{F})$ replaced by $|\mathcal{F}|$ is easy to prove by probabilistic argument (Exercise 3). It shows that in order to have a large gap between τ^* and τ, the set system must have very many sets.

Exercises

1. (a) Find examples of set systems with τ^* bounded by a constant and τ arbitrarily large. $\boxed{1}$
   (b) Find examples of set systems with ν bounded by a constant and ν^* arbitrarily large. $\boxed{1}$
2. Let $\mathcal{F}$ be a system of finitely many closed intervals on the real line. Prove that $\nu(\mathcal{F}) = \tau(\mathcal{F})$. $\boxed{3}$
3. Prove that
   $$\tau(\mathcal{F}) \leq \tau^*(\mathcal{F}) \cdot \ln(|\mathcal{F}|+1)$$
   for all (finite) set systems $\mathcal{F}$. Choose a transversal as a random sample. $\boxed{3}$
4. (Analysis of the greedy algorithm for transversal) Let $\mathcal{F}$ be a finite set system. We choose points $x_1, x_2, \ldots, x_t$ of a transversal one by one: x_i is taken as a point contained in the maximum possible number of uncovered sets (i.e., sets of $\mathcal{F}$ containing none of $x_1, \ldots, x_{i-1}$).
   (a) Prove that the size t of the resulting transversal satisfies
   $$t \leq \frac{1}{1 \cdot 2}\,\nu_1(\mathcal{F}) + \frac{1}{2 \cdot 3}\,\nu_2(\mathcal{F}) + \cdots + \frac{1}{(d-1)d}\,\nu_{d-1}(\mathcal{F}) + \frac{1}{d}\,\nu_d(\mathcal{F}),$$
   where $d = \Delta(\mathcal{F})$ is the maximum degree of $\mathcal{F}$ and $\nu_k(\mathcal{F})$ is the maximum size of a *simple k-packing* in $\mathcal{F}$. A subsystem $\mathcal{M} \subseteq \mathcal{F}$ is a simple k-packing if $\Delta(\mathcal{M}) \leq k$ (so $\nu_1(\mathcal{F}) = \nu(\mathcal{F})$). $\boxed{5}$
   (b) Conclude that $\tau(\mathcal{F}) \leq t \leq \tau^*(\mathcal{F}) \cdot \sum_{k=1}^{d} \frac{1}{k}$. $\boxed{2}$
5. König's edge-covering theorem asserts that if E is the set of edges of a bipartite graph, then $\nu(E) = \tau(E)$. Hall's marriage theorem states that if G is a bipartite graph with color classes 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.

[1] As a part of a very exciting development in complexity theory, it was recently proved that no polynomial-time algorithm can do better in general unless P = NP; see, e.g., [Hoc96] for proofs and references.

(a) Derive König's edge-covering theorem from Hall's marriage theorem. $\boxed{2}$

(b) Derive Hall's marriage theorem from König's edge-covering theorem. $\boxed{2}$

6. Let A, b, c, P, and D be as in Proposition 10.1.2.
(a) Check that $c^T x \geq y^T b$ for all $x \in P$ and all $y \in D$. $\boxed{1}$
(b) Prove that if $P \neq \emptyset$ and $D \neq \emptyset$, then the system $Ax \leq b$, $y^T A \geq c$, $c^T x \geq y^T b$ has a nonnegative solution x, y (which implies Proposition 10.1.2). Apply the version of the Farkas lemma as in Exercise 1.2.7(b). $\boxed{4}$

10.2 Epsilon Nets and VC-Dimension

Large sets should be easier to hit by a transversal than small ones. The notion of ε-net and the related theory elaborate on this intuition. We begin with a special case, where the ground set is finite and the size of a set is simply measured as the cardinality.

10.2.1 Definition (Epsilon net, a special case). *Let $(X, \mathcal{F})$ be a set system with X finite and let $\varepsilon \in [0, 1]$ be a real number. A set $N \subseteq X$ (not necessarily one of the sets of $\mathcal{F}$) is called an* ε-net *for $(X, \mathcal{F})$ if $N \cap S \neq \emptyset$ for all $S \in \mathcal{F}$ with $|S| \geq \varepsilon|X|$.*

So an ε-net is a transversal for all sets larger than $\varepsilon|X|$. Sometimes it is convenient to write $\frac{1}{r}$ instead of ε, with $r \geq 1$ a real parameter. A beautiful result (Theorem 10.2.4 below) describes a simple combinatorial condition on the structure of $\mathcal{F}$ that guarantees the existence of $\frac{1}{r}$-nets of size only $O(r \log r)$ for all $r \geq 2$.

If we want to deal with infinite sets, measuring the size as the number of points is no longer appropriate. For example, a "large" subset of the unit square could naturally be defined as one with large Lebesgue measure. So in general we consider an arbitrary probability measure μ on the ground set. In concrete situations we will most often encounter μ concentrated on finitely many points. This means that there is a finite set $Y \subseteq X$ and a positive function $w: Y \to (0, 1]$ with $\sum_{y \in Y} w(y) = 1$, and μ is given by $\mu(A) = \sum_{y \in A \cap Y} w(y)$. In particular, if the weights of all points $y \in Y$ are the same, i.e., $\frac{1}{|Y|}$, we speak of the *uniform measure on* Y. Another common example of μ is a suitable multiple of the Lebesgue measure restricted to some geometric figure.

10.2.2 Definition (Epsilon net). *Let X be a set, let μ be a probability measure on X, let $\mathcal{F}$ be a system of μ-measurable subsets of X, and let $\varepsilon \in [0, 1]$ be a real number. A subset $N \subseteq X$ is called an* ε-net for $(X, \mathcal{F})$ with respect to μ *if $N \cap S \neq \emptyset$ for all $S \in \mathcal{F}$ with $\mu(S) \geq \varepsilon$.*

VC-dimension. In order to describe the result promised above, about existence of small ε-nets, we need to introduce a parameter of a set system called the *Vapnik–Chervonenkis dimension*, or *VC-dimension* for short. Its applications are much wider than for the existence of ε-nets.

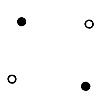
Let $\mathcal{F}$ be a set system on X and let $Y \subseteq X$. We define the *restriction of $\mathcal{F}$ on Y* (also called the *trace* of $\mathcal{F}$ on Y) as

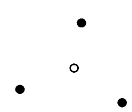
$$\mathcal{F}|_Y = \{S \cap Y \colon S \in \mathcal{F}\}.$$

It may happen that several distinct sets in $\mathcal{F}$ have the same intersection with Y; in such a case, the intersection is still present only once in $\mathcal{F}|_Y$.

10.2.3 Definition (VC-dimension). *Let $\mathcal{F}$ be a set system on a set X. Let us say that a subset $A \subseteq X$ is* shattered *by $\mathcal{F}$ if each of the subsets of A can be obtained as the intersection of some $S \in \mathcal{F}$ with A, i.e., if $\mathcal{F}|_A = 2^A$. We define the* VC-dimension *of $\mathcal{F}$, denoted by* $\dim(\mathcal{F})$, *as the supremum of the sizes of all finite shattered subsets of X. If arbitrarily large subsets can be shattered, the VC-dimension is ∞.*

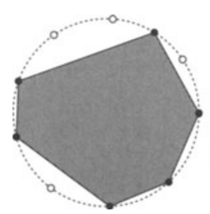
Let us consider two examples. First, let $\mathcal{H}$ be the system of all closed half-planes in the plane. We claim that $\dim(\mathcal{H}) = 3$. If we have 3 points in general position, each of their subsets can be cut off by a half-plane, and so such a 3-point set is shattered. Next, let us check that no 4-point set can be shattered. Up to possible degeneracies, there are only two essentially different positions of 4 points in the plane:





In both these cases, if the black points are contained in a half-plane, then a white point also lies in that half-plane, and so the 4 points are not shattered. This is a rather ad hoc argument, and later we will introduce tools for bounding the VC-dimension in geometric situations. We will see that bounded VC-dimension is rather common for families of simple geometric objects in Euclidean spaces.

A rather different example is the system $\mathcal{K}_2$ of all convex sets in the plane. Here the VC-dimension is infinite, since any finite convex independent set A is shattered: Each $B \subseteq A$ can be expressed as the intersection of A with a convex set, namely, $B = A \cap \operatorname{conv}(B)$.



We can now formulate the promised result about small ε-nets.

10.2.4 Theorem (Epsilon net theorem). *If X is a set with a probability measure μ, $\mathcal{F}$ is a system of μ-measurable subsets of X with $\dim(\mathcal{F}) \leq d$, $d \geq 2$, and $r \geq 2$ is a parameter, then there exists a $\frac{1}{r}$-net for $(X, \mathcal{F})$ with respect to μ of size at most $Cdr \ln r$, where C is an absolute constant.*

The proof below gives the estimate $C \leq 20$, but a more accurate calculation shows that C can be taken arbitrarily close to 1 for sufficiently large r. More precisely, for any $d \geq 2$ there exists an $r_0 > 1$ such that for all $r > r_0$, each set system of VC-dimension d admits a $\frac{1}{r}$-net of size at most $dr \ln r$. Moreover, this bound is tight in the worst case up to smaller-order terms.

For the proof (and also later on) we need a fundamental lemma bounding the number of distinct sets in a system of given VC-dimension. First we define the *shatter function* of a set system $\mathcal{F}$ by

$$\pi_{\mathcal{F}}(m) = \max_{Y \subseteq X,\ |Y|=m} \left|\mathcal{F}|_Y\right|.$$

In words, $\pi_{\mathcal{F}}(m)$ is the maximum possible number of distinct intersections of the sets of $\mathcal{F}$ with an m-point subset of X.

10.2.5 Lemma (Shatter function lemma). *For any set system $\mathcal{F}$ of VC-dimension at most d, we have $\pi_{\mathcal{F}}(m) \leq \Phi_d(m)$ for all m, where $\Phi_d(m) = \binom{m}{0} + \binom{m}{1} + \cdots + \binom{m}{d}$.*

Thus, the shatter function for any set system is either 2^m for all m (the case of infinite VC-dimension) or it is bounded by a fixed polynomial.

For d fixed and $m \to \infty$, $\Phi_d(m)$ can be simply estimated by $O(m^d)$. For more precise calculations, where we are interested in the dependence on d, we can use the estimate $\Phi_d(m) \leq \left(\frac{em}{d}\right)^d$, where e is the basis of natural logarithms. This is valid for all $m, d \geq 1$.

Proof of Lemma 10.2.5. Since VC-dimension does not increase by passing to a subsystem, it suffices to show that any set system of VC-dimension at most d on an n-point set has no more than $\Phi_d(n)$ sets. We proceed by induction on d, and for a fixed d we use induction on n.

Consider a set system $(X, \mathcal{F})$ with $|X| = n$ and $\dim(\mathcal{F}) = d$, and fix some $x \in X$. In the induction step, we would like to remove x and pass to the set system $\mathcal{F}_1 = \mathcal{F}|_{X \setminus \{x\}}$ on $n-1$ points. This $\mathcal{F}_1$ has VC-dimension at most d, and hence $|\mathcal{F}_1| \leq \Phi_d(n-1)$ by the inductive hypothesis. How many more sets can $\mathcal{F}$ have compared to $\mathcal{F}_1$? The only way that the number of sets decreases by removing x is when two sets $S, S' \in \mathcal{F}$ give rise to the same set in $\mathcal{F}_1$, which means that $S' = S \cup \{x\}$, $x \notin S$, or the other way round. This suggests that we define an auxiliary set system $\mathcal{F}_2$ consisting of all sets in $\mathcal{F}_1$ that correspond to such pairs $S, S' \in \mathcal{F}$: $\mathcal{F}_2 = \{S \in \mathcal{F}\colon x \notin S, S \cup \{x\} \in \mathcal{F}\}$.

By the above discussion, we have $|\mathcal{F}| = |\mathcal{F}_1| + |\mathcal{F}_2|$. Crucially, we observe that $\dim(\mathcal{F}_2) \leq d-1$, since if $A \subseteq X \setminus \{x\}$ is shattered by $\mathcal{F}_2$, then $A \cup \{x\}$ is shattered by $\mathcal{F}$. Therefore, $|\mathcal{F}_2| \leq \Phi_{d-1}(n-1)$. The resulting recurrence has already been solved in the first proof of Proposition 6.1.1. □

The rest of the proof of the epsilon net theorem is a clever probabilistic argument; one might be tempted to believe that it works by some magic. First we need a technical lemma concerning the binomial distribution.

10.2.6 Lemma. *Let $X = X_1+X_2+\cdots+X_n$, where the X_i are independent random variables, X_i attaining the value 1 with probability p and the value 0 with probability $1-p$. Then* $\mathrm{Prob}\left[X \geq \frac{1}{2}np\right] \geq \frac{1}{2}$, *provided that $np \geq 8$.*

Proof. This is a routine consequence of Chernoff-type tail estimates for the binomial distribution, and in fact, considerably stronger estimates hold. The simple result we need can be quickly derived from Chebyshev's inequality for X, stating that $\mathrm{Prob}[|X-\mathbf{E}[X]| \geq t] \leq \mathrm{Var}[X]/t^2$, $t \geq 0$. Here $\mathbf{E}[X] = np$ and $\mathrm{Var}[X] = \sum_{i=1}^n \mathrm{Var}[X_i] \leq np$. So

$$\mathrm{Prob}\left[X < \tfrac{1}{2}np\right] \leq \mathrm{Prob}\left[|X-\mathbf{E}[X]| \geq \tfrac{1}{2}np\right] \leq \frac{4}{np} \leq \tfrac{1}{2}.$$

□

Proof of the epsilon net theorem. Let us put $s = Cdr\ln r$ (assuming without harm that it is an integer), and let N be a random sample picked by s independent random draws, where each element is drawn from X according to the probability distribution μ. (So the same element can be drawn several times; this does not really matter much, and this way of random sampling is chosen to make calculations simpler.) The goal is to show that N is a $\frac{1}{r}$-net with a positive probability.

To simplify formulations, let us assume that all $S \in \mathcal{F}$ satisfy $\mu(S) \geq \frac{1}{r}$; this is no loss of generality, since the smaller sets do not play any role. The probability that the random sample N misses any given set $S \in \mathcal{F}$ is at most $(1-\frac{1}{r})^s \leq e^{-s/r}$, and so if s were at least $r\ln(|\mathcal{F}|+1)$, say, the conclusion would follow immediately. But r is typically much smaller than $|\mathcal{F}|$ (it can be a constant, say), and so we need to do something more sophisticated.

Let E_0 be the event that the random sample N fails to be a $\frac{1}{r}$-net, i.e., misses some $S \in \mathcal{F}$. We bound $\mathrm{Prob}[E_0]$ from above using the following thought experiment.

By s more independent random draws we pick another random sample M.[2] We put $k = \frac{s}{2r}$, again assuming that it is an integer, and we let E_1 be the following event:

There exists an $S \in \mathcal{F}$ with $N \cap S = \emptyset$ and $|M \cap S| \geq k$.

[2] This double sampling resembles the proof of Proposition 6.5.2, and indeed these proofs have a lot in common, although they work in different settings.

Here an explanation concerning repeated elements is needed. Formally, we regard N and M as *sequences* of elements of X, with possible repetitions, so $N = (x_1, x_2, \ldots, x_s)$, $M = (y_1, y_2, \ldots, y_s)$. The notation $|M \cap S|$ then really means $|\{i \in 1, 2, \ldots, s\colon y_i \in S\}|$, and so an element repeated in M and lying in S is counted the appropriate number of times.

Clearly, $\text{Prob}[E_1] \le \text{Prob}[E_0]$, since E_1 requires E_0 plus something more. We are going to show that $\text{Prob}[E_1] \ge \frac{1}{2}\text{Prob}[E_0]$. Let us investigate the conditional probability $\text{Prob}[E_1 \,|\, N]$, that is, the probability of E_1 when N is fixed and M is random. If N is a $\frac{1}{r}$-net, then E_1 cannot occur, and $\text{Prob}[E_0 \,|\, N] = \text{Prob}[E_1 \,|\, N] = 0$.

So suppose that there exists an $S \in \mathcal{F}$ with $N \cap S = \emptyset$. There may be many such S, but let us fix one of them and denote it by S_N. We have $\text{Prob}[E_1 \,|\, N] \ge \text{Prob}[|M \cap S_N| \ge k]$. The quantity $|M \cap S_N|$ behaves like the random variable X in Lemma 10.2.6 with $n = s$ and $p = \frac{1}{r}$, and so $\text{Prob}[|M \cap S_N| \ge k] \ge \frac{1}{2}$. Hence $\text{Prob}[E_0 \,|\, N] \le 2\,\text{Prob}[E_1 \,|\, N]$ for all N, and thus $\text{Prob}[E_0] \le 2\,\text{Prob}[E_1]$.

Next, we are going to bound $\text{Prob}[E_1]$ differently. Instead of choosing N and M at random directly as above, we first make a sequence $A = (z_1, z_2, \ldots, z_{2s})$ of $2s$ independent random draws from X. Then, in the second stage, we randomly choose s positions in A and put the elements at these positions into N, and the remaining elements into M (so there are $\binom{2s}{s}$ possibilities for A fixed). The resulting distribution of N and M is the same as above. We now prove that for every fixed A, the conditional probability $\text{Prob}[E_1 \,|\, A]$ is small. This implies that $\text{Prob}[E_1]$ is small, and therefore $\text{Prob}[E_0]$ is small as well.

So let A be fixed. First let $S \in \mathcal{F}$ be a fixed set and consider the conditional probability $P_S = \text{Prob}[N \cap S = \emptyset,\ |M \cap S| \ge k \,|\, A]$. If $|A \cap S| < k$, then $P_S = 0$. Otherwise, we bound $P_S \le \text{Prob}[N \cap S = \emptyset \,|\, A]$. The latter is the probability that a random sample of s positions out of $2s$ in A avoids the at least k positions occupied by elements of S. This is at most

$$\frac{\binom{2s-k}{s}}{\binom{2s}{s}} \le \left(1 - \frac{k}{2s}\right)^s \le e^{-(k/2s)s} = e^{-k/2} = e^{-(Cd\ln r)/4} = r^{-Cd/4}.$$

This was an estimate of P_S for a fixed $S \in \mathcal{F}$. Now, finally, we use the assumption about the VC-dimension of $\mathcal{F}$, via the shatter function lemma: The sets of $\mathcal{F}$ have at most $\Phi_d(2s)$ distinct intersections with A. Since the event "$N \cap S = \emptyset$ and $|M \cap S| \ge k$" depends only on $A \cap S$, it suffices to consider at most $\Phi_d(2s)$ distinct sets S, and so for every fixed A,

$$\text{Prob}[E_1 \,|\, A] \le \Phi_d(2s) \cdot r^{-Cd/4} \le \left(\frac{2es}{d}\right)^d r^{-Cd/4} = \left(2er\ln r \cdot r^{-C/4}\right)^d < \frac{1}{2}$$

if $d, r \ge 2$ and C is sufficiently large. So $\text{Prob}[E_0] \le 2\,\text{Prob}[E_1] < 1$, which proves Theorem 10.2.4. □

The epsilon net theorem implies that for set systems of small VC-dimension, the gap between the fractional transversal number and the transversal number cannot be too large.

10.2.7 Corollary. *Let $\mathcal{F}$ be a finite set system on a ground set X with $\dim(\mathcal{F}) \leq d$. Then we have*

$$\tau(\mathcal{F}) \leq Cd\tau^*(\mathcal{F}) \ln \tau^*(\mathcal{F}),$$

where C is as in the epsilon net theorem.

Proof. Let $r = \tau^*(\mathcal{F})$. Since $\mathcal{F}$ is finite, we may assume that an optimal fractional transversal $\varphi: X \to [0,1]$ is concentrated on a finite set Y. This φ, after rescaling, defines a probability measure μ on X, by letting $\mu(\{y\}) = \frac{1}{r}\varphi(y)$, $y \in Y$. Each $S \in \mathcal{F}$ has $\mu(S) \geq \frac{1}{r}$ by the definition of fractional transversal, and so a $\frac{1}{r}$-net for $(X, \mathcal{F})$ with respect to μ is a transversal. By the epsilon net theorem, there exists a transversal of size at most $Cdr \ln r$. □

We mention a concrete application of the corollary in the next section, where we collect examples of set systems of bounded VC-dimension.

Bibliography and remarks. The notion of VC-dimension originated in statistics. It was introduced by Vapnik and Chervonenkis [VC71]. Under different names, it has also appeared in other papers (Sauer [Sau72] and Shelah [She72]), but the work [VC71] was probably the most influential for subsequent developments. The name VC-dimension and some other, by now more or less standard, terminology were introduced by Haussler and Welzl [HW87]. VC-dimension and the related theory play an important role in several mathematical fields, such as statistics (the theory of empirical processes), computational learning theory, computational geometry, discrete geometry, combinatorics of hypergraphs, and discrepancy theory.

The shatter function lemma was independently discovered in the three already mentioned papers [VC71], [Sau72], [She72].

The shatter function, together with the dual shatter function (defined as the shatter function of the dual set system) was introduced and applied by Welzl [Wel88]. Implicitly, these notions were used much earlier, and they appear in the literature under various names, such as *growth functions.*

The notion of ε-net and the epsilon net theorem (with X finite and μ uniform) are due to Haussler and Welzl [HW87]. Their proof is essentially the one shown in the text, and it closely follows an earlier proof by Vapnik and Chervonenkis [VC71] concerning the related notion of ε-*approximations.* In the same setting as in the definition of

ε-nets, a set $A \subseteq X$ is an ε-approximation for $(X, \mathcal{F})$ with respect to μ if for all $S \in \mathcal{F}$,

$$\left|\mu(S) - \frac{|A \cap S|}{|A|}\right| < \varepsilon.$$

So while an ε-net intersects each large set at least once, an ε-approximation provides a "proportional representation" up to the error of ε. Vapnik and Chervonenkis [VC71] proved the existence of $\frac{1}{r}$-approximations of size $O(dr^2 \log r)$ for all set system of VC-dimension d.

Komlós, Pach, and Wöginger [KPW92] improved the dependence on d in the Haussler–Welzl bound on the size of ε-nets. The improvement is achieved by choosing the second sample M of size t somewhat larger than s and doing the calculations more carefully. They also proved an almost matching lower bound using suitable random set systems. The proofs can be found in [PA95] as well.

The proof in the Vapnik–Chervonenkis style, while short and clever, does not seem to convey very well the reasons for the existence of small ε-nets. Somewhat longer but more intuitive proofs have been found in the investigation of deterministic algorithms for constructing ε-approximations and ε-nets; one such proof is given in [Mat99a], for instance.

Exercises

1. Show that for any integer d there exists a convex set C in the plane such that the family of all isometric copies of C has VC-dimension at least d. $\boxed{4}$
2. Show that the shatter function lemma is tight. That is, for all d and n construct a system of VC-dimension d on n points with $\Phi_d(n)$ sets. $\boxed{2}$

10.3 Bounding the VC-Dimension and Applications

The VC-dimension can be determined without great difficulty in several simple cases, such as for half-spaces or balls in $\mathbf{R}^d$, but for only slightly more complicated families its computation becomes challenging. On the other hand, a few simple steps explained below show that the VC-dimension is bounded for any family whose sets can be defined by a formula consisting of polynomial equations and inequalities combined by Boolean connectives (conjunctions, disjunctions, etc.) and involving a bounded number of real parameters. This includes families like all ellipsoids in $\mathbf{R}^d$, all boxes in $\mathbf{R}^d$, arbitrary intersections of pairs of circular disks in the plane, and so on. On the other hand, arbitrary convex polygons are not covered (since a general convex polygon cannot be described by a bounded number of real parameters) and indeed, this family has infinite VC-dimension.