Printed: 22 March 2025 at 13:26

Chapter 14

VC Sets

VCsets::VCsets

Section *14.1 presents a simple concrete example to illustrate the use of the combinatorial method for deriving bounds on packing numbers.*

Section *14.2 defines the concept of a VC-class of subsets $\mathcal{D}$ of some set $\mathbb{A}$, which leads to a polynomial bound for the numbers of subsets of the form $F \cap D$ for $D \in \mathcal{D}$, uniformly over all subsets F of $\mathbb{A}$. The slightly sharper concept of shatter-dimension is obtained by restricting attention to a single F. Several Examples illustrate how shatter-dimension can be bounded and used, with control of packing numbers as the prime application.*

Section *14.3 derives the basic polynomial bound using the downshift method.*

*Section *14.4 interprets subsets of the discrete unit cube as the vertex sets of graphs. The shatter dimension gives a surprising upper bound for the number of edges of the graph. The method of proof again relies on downshifting.*

*Section *14.5 presents Haussler's refinement of the argument from Section 14.2, leading to a sharper bound for packing numbers.*

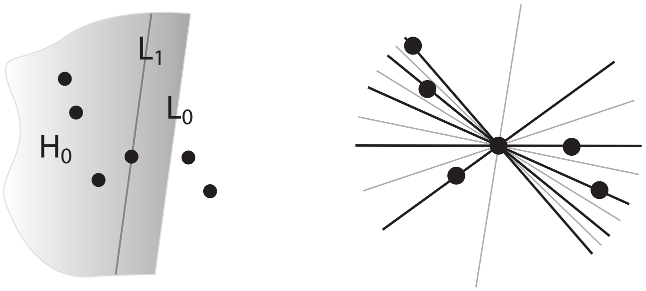
14.1 An introductory example

VCsets::S:VC.intro

For my purposes, the combinatorial argument often referred to as the VC method—in honor of the important contributions of Vapnik and Chervonenkis (1971); Vapnik and Červonenkis (1981)—is of use mainly as a step towards calculation of packing/covering numbers. The method is elegant and leads to results not easily obtained in other ways. The basic calculations occupy only a few pages. Nevertheless, the ideas are subtle enough to appear beyond the comfortable reach of many would-be users. With that fact in mind, I offer a more concrete preliminary example, in the hope that the combinatorial ideas might then seem less mysterious.

Consider the set $\mathcal{H}$ of all closed half-spaces in $\mathbb{R}^2$. Let F be a set of n points in $\mathbb{R}^2$. How many distinct subsets can $\mathcal{H}$ pick out from F? That is, how large can the set $\mathcal{H}_F := \{F \cap H : H \in \mathcal{H}\}$ be? Certainly its size (= cardinality) can be no larger than 2^n because F has only that many subsets. Even better, a simple argument shows that

$$|\mathcal{H}_F| := \text{cardinality of } \mathcal{H}_F \leq p(n) := 1 + 4n(n-1).$$



Indeed, consider a particular nonempty subset F_0 of F picked out by a particular half-space H_0. There is no loss of generality in assuming that at least one point of F_0 (call it x_0) lies on L_0, the boundary of H_0: otherwise we could replace H_0 by a smaller H_1 whose boundary, L_1, runs parallel to L_0 through the point x_0 of F_0 that is closest to L_0.

As seen from x_0, the other $n-1$ points of F all lie on a set $\mathcal{L}(x_0)$ of at most $n-1$ lines through x_0. Augment $\mathcal{L}(x_0)$ by another set $\mathcal{L}'(x_0)$ of at most $n-1$ lines through x_0, one in each angle between two lines from $\mathcal{L}(x_0)$. The lines in $\mathcal{H}(x_0) := \mathcal{L}(x_0) \cup \mathcal{L}'(x_0)$ define a collection of at most $4(n-1)$ closed half-spaces, each with x_0 on its boundary. The collection $\cup_{x\in F}\mathcal{H}(x)$ accounts for all possible nonempty subsets of F picked out by closed half-spaces. The extra 1 takes care of the empty set.

> **Remark.** Apparently (Dudley, 1978, page 921) the bound can be reduced to $n^2 - n + 2$, which is sharp in the sense that it is achieved whenever the n points are in general position, that is, no straight line runs through more than 2 of the points. Dudley also described several precursors for the key combinatorial bound (my Theorem <3>). For my purposes the sharper bound is not needed. Indeed any polynomial would suffice for the consequences described below.

The slow increase in $|\mathcal{H}_F|$, at an $O(n^2)$ rate rather than a rapid 2^n rate, has a useful consequence for the packing numbers when $\mathcal{H}$ is equipped with an $\mathcal{L}^1(P)$ metric, for some probability measure P on the plane. The $\mathcal{L}^1(P)$ distance between two Borel sets B and B' is defined as $P|B - B'| = P(B\Delta B')$, the measure of the symmetric difference. The two sets are said to be ϵ-separated (in $\mathcal{L}^1(P)$) if $P(B\Delta B') > \epsilon$. As in Chapter 10, the packing number $\text{PACK}(\epsilon, \mathcal{H}, P)$—or $\text{PACK}(\epsilon, \mathcal{H}, \mathcal{L}^1(P))$ if there is any ambiguity about the norm being used—is defined as the largest N for which there exists a collection of N closed half-spaces, each of the $\binom{N}{2}$ pairs $> \epsilon$ apart.

VCsets::poly.to.packing <1> **Example.** Here is an argument, due to Dudley (1978, Lemma 7.13), to show that the packing numbers $\text{PACK}(\epsilon, \mathcal{H}, P)$ are bounded uniformly over P by a polynomial in $1/\epsilon$, for $0 < \epsilon \leq 1$. The result is surprising because it makes no regularity assumptions about the probability measure P.

Suppose the half-spaces $H_1, H_2, \dots, H_N$ are ϵ-separated in $\mathcal{L}^1(P)$. By means of a cunningly chosen F, the polynomial bound on $|\mathcal{H}_F|$ will lead to an upper bound for N. The trick is to find a set $F_m = \{\xi_1, \dots, \xi_m\} \subset \mathbb{R}^2$, with $m = \lceil 2 \log N / \epsilon \rceil$, from which each H_α picks out a different subset. Then $\mathcal{H}$ will pick out at least N subsets from F_m, implying that

$$N \leq p(m) \leq 1 + 4\left(1 + \frac{2\log N}{\epsilon}\right)\left(\frac{2\log N}{\epsilon}\right) \leq 9\left(\frac{\log N}{\epsilon}\right)^2 .$$

Bounding the $\log N$ by a constant multiple of $N^{1/4}$ and solving the resulting inequality for N, we get an upper bound $N \leq O(1/\epsilon)^4$. With a smaller power in the bound for $\log N$ we would bring the power of $1/\epsilon$ arbitrarily close to 2.

Remark. As shown in Example <12>, the bound for the packing numbers for $\mathcal{H}$ can be sharpened to $C(\epsilon^{-1}\log(1/\epsilon))^2$, with C a universal constant, at least when $0 < \epsilon \leq 1/2$. With a lot more work, even the log term can be removed—see Section 14.5. At this stage there is little point in struggling to get the best bound in ϵ. For many applications, the qualitative consequences of a polynomial bound in $1/\epsilon$ are the same, no matter what the degree of the polynomial.

How do we find a suitable F_m? We need to place at least one point of F_m in each of the $\binom{N}{2}$ symmetric differences $H_i \Delta H_j$. It might seem we are faced with a delicate task involving consideration of all possible configurations of the symmetric differences, but here probability theory comes to the rescue.

As described in the Preface of the wonderful little book by Alon and Spencer (2000), the ***probabilistic method*** can prove existence by artificially introducing a probability measure into a problem:

> In order to prove the existence of a combinatorial structure with certain properties, we construct an appropriate probability space and show that a randomly chosen element in this space has the desired properties with positive probability. This method was initiated by Paul Erdős, who contributed so much to its development over the last fifty years, that is seems appropriate to call it "The Erdős Method."

Generate F_m as a random sample of size m from P. If $m \geq 2\log N/\epsilon$, then there is a strictly positive probability that the sample has the desired property. Indeed, for fixed α, β with $1 \leq \alpha < \beta \leq 1$,

$$\begin{aligned}
&\mathbb{P}\{H_\alpha \text{ and } H_\beta \text{ pick out same points from } F_m\} \\
&\quad = \mathbb{P}\{\text{no points of sample in } H_\alpha \Delta H_\beta\} \\
&\quad = (1 - P(H_\alpha \Delta H_\beta))^m \\
&\quad \leq (1-\epsilon)^m \leq \exp(-m\epsilon).
\end{aligned}$$

Add up $\binom{N}{2}$ such probability bounds to get a conservative estimate,

$$\mathbb{P}\{\text{some pair } H_\alpha,\ H_\beta \text{ pick same subset from } F_m\} \le \binom{N}{2}\exp(-m\epsilon).$$

□ When $m \ge 2\log N/\epsilon$ the last bound is strictly less than 1, as desired.

> **Remark.** The argument in the Example implicitly assumed that the ξ_i's sampled from P are all distinct, which need not be true if P has atoms. Thus $|F_m|$, the size of F_m, might be smaller than m. To be more precise I could have written $N \le p(|F_m|) \le p(m) \le \dots$. The added rigor is hardly worth the trouble; usually ties cause only notational difficulties. Section 14.2 will avoid the issue altogether by working with the n-tuple $(\xi_1, \dots, \xi_n)$, which might contain repeats, instead of the set $\{\xi_1, \dots, \xi_n\}$.

Notice how probability theory has been used to prove an existence result, which gives a bound for a packing number, which will be used to derive probabilistic consequences—all based ultimately on the existence of the polynomial bound for $|\mathcal{H}_F|$.

14.2 VC-dimension, shatter-dimension

VCsets::S:VCdim

The argument in the previous Section had little to do with the choice of $\mathcal{H}$ as the set of all closed half-spaces in a particular Euclidean space. It would apply to any collection $\mathcal{D}$ of (measurable) subsets of any space $\mathbb{A}$ for which the size of $\mathcal{D}_F := \{F \cap D : D \in \mathcal{D}\}$ is bounded above by a fixed polynomial in $|F|$. Unfortunately, for more complicated $\mathcal{D}$'s it can sometimes be difficult to derive such a polynomial bound directly but it is easier to prove existence of a finite k such that

$$|\mathcal{D}_F| < 2^k \text{ for every } F \subset \mathbb{A} \text{ with } |F| \le k$$

Such a $\mathcal{D}$ is usually called a ***VC-class*** of sets in honor of Vapnik & Červonenkis. This property is equivalent to the finiteness of the ***VC-dimension*** for $\mathcal{D}$, defined as follows.

VCsets::VCdef <2> **Definition.** *A collection $\mathcal{D}$ of subsets of a set $\mathbb{A}$ is said to have VC-dimension d (written* VCDIM$(\mathcal{D}) = d$*) if both the following conditions hold.*

(i) There exists at least one subset F_0 of $\mathcal{X}$ for which

$$|F_0| = d \quad \text{AND} \quad |\{F_0 \cap D : D \in \mathcal{D}\}| = 2^d.$$

(ii) $|\{F \cap D : D \in \mathcal{D}\}| < 2^{|F|}$ for every finite subset F of $\mathbb{A}$ with $|F| > d$.

> **Remark.** Some authors use the term ***shatter dimension*** instead of VC-dimension. Motivated by the extension of the concept to classes of functions, and beyond, I feel a better name would be ***surround dimension***. In fact that is the term I use in the next Chapter. I chose the name SDIM, which appears in Definition <7>, because the initial 's' could be interpreted as a reference to both 'shatter' and 'surround'.

The key fact about VC-classes is often called the VC Lemma or the Sauer/Shelah Lemma, although credit should be spread more widely. (See the Notes in Section 14.7.)

VCsets::shatter.theorem <3> **Theorem.** *If $\mathcal{D}$ is VC-class of subsets of some set $\mathbb{A}$ with* VCDIM$(\mathcal{D}) = d$ *then, for each finite subset F of $\mathbb{A}$,*

$$|\{F \cap D : D \in \mathcal{D}\}| \le \beta(n, d) := \binom{n}{0} + \binom{n}{1} + \dots + \binom{n}{d}$$

□ *where n equals $|F|$.*

This result is deduced in Section 14.3 from a slightly sharper result that allows for a more a more subtle dependence (corresponding to Definition <7>) of the bound on the set F.

Clearly $\beta(n, d)$ is a polynomial of degree d in n. More precisely, it equals 2^n if $n \le d$ and, if $n \ge d$,

$$\begin{aligned}\beta(n, d) &= \sum\nolimits_{k=0}^{d} n(n-1)\dots(n-k+1)/k! \\ &\le \sum\nolimits_{k=0}^{d} (n/d)^k d^k/k! \\ &\le (en/d)^d \qquad \text{because } (n/d)^k \le (n/d)^d \text{ when } n \ge d \ge k.\end{aligned}$$

\E@ VC.poly.bnd <4>

The upper bound can be improved slightly:

\E@ VC.poly.bnd2 <5>
$$\beta(n, d) \le \beta_1(n, d) := 1.5n^d/d! \qquad \text{if } n \ge d + 2,$$

a result apparently due to Vapnik and Červonenkis—see Dudley (2014, Proposition 4.3). As Dudley noted, the β_1 bound is “not far from optimal” because the $\binom{n}{d}$ contributes an $n^d/d!$ as the leading term in the polynomial $\beta(n, d)$. Compare with: $\beta_1(n, d) \approx 0.6d^{-1/2}(en/d)^d$ by the STIRLING approximation from Section 2.5.

As Vapnik and Chervonenkis (1971, §6) realized, the concept defined by Definition <2> needs to be sharpened when seeking necessary and sufficient conditions for uniform laws of large numbers, an improvement closely related to the symmetrization bounds from Chapter 13.

Suppose $\xi_1, \dots, \xi_n$ are independent random elements of $\mathbb{A}$ with $\xi_i \sim P_i$ for some probability measure P_i defined on a suitable sigma-field $\mathcal{A}$ on $\mathbb{A}$. Let me assume away possible measurability difficulties by supposing $\mathcal{D}$ is a countable subset of $\mathcal{A}$. Define independent $\{0, 1\}$-valued stochastic processes

$$X_i(D, \omega) := \{\xi_i(\omega) \in D\} \qquad \text{for } D \in \mathcal{D} \text{ and } i \in [[n]].$$

Let $\mathbb{Q}$ denote the uniform distribution on the discrete hypercube $\mathcal{C}_n := \{0, 1\}^n$. Then, as shown in Section 13.5, for each ORLICZ function we have

\E@ Psi.norm.xx <6>
$$\mathbb{P}\Psi\left(\sup_{D\in\mathcal{D}} |\sum\nolimits_i X_i(D, \omega) - P_i D|\right) \le \mathbb{P}^{\omega}\mathbb{Q}^{\mathfrak{s}}\Psi\left(2 \sup_{D\in\mathcal{D}} |\sum\nolimits_i \mathfrak{s}_i X_i(D, \omega)|\right).$$

The $\mathbb{Q}$ integral on the right-hand side of the inequality in <6> can also be written as $\mathbb{Q}\Psi\left(\sup_{x\in\mathcal{X}_\omega}|\langle\mathfrak{s},x\rangle|\right)$ where $\mathcal{X}_\omega$ denotes the set of all members of $\mathcal{C}_n$ of the form

$$x(D,\omega) := \Big(X_1(D,\omega),\dots,X_n(D,\omega)\Big) \qquad \text{for some } D \text{ in } \mathcal{D}.$$

Suitable bounds on the ℓ^2 packing numbers for $\mathcal{X}_\omega$ lead to an empirical process maximal inequalities via chaining arguments.

> **Remark.** As $\mathcal{X}_\omega$ is finite and $\mathcal{D}$ need not be, there will typically be many different D's that contribute the same $x(D,\omega)$.
>
> You should also take note of a subtle difference between the set $\{X_i(D,\omega) : i \in [[n]]\}$ and the vector $x(D,\omega)$ for fixed ω: even when the set $\{\xi_i(\omega) : i \in [[n]]\}$ has size $< n$, each $x(D,\omega)$ vector still has length n but with repeated coordinates.

Corresponding to Definition <2> we have an analogous property for subsets of $\mathcal{C}_n$. Instead of subsets of a finite F we now deal with subsets J of the index set $[[n]]$ and the projections they define on $\mathcal{C}_n$ by

$$x[J] := (x_i : i \in J) \in \{0,1\}^J \qquad \text{for each } x := (x_i : i \in [[n]]) \text{ in } \mathcal{C}_n.$$

> **Remark.** Here I am borrowing from the most convenient notation used by the **R** statistical language. Similarly, if M is an $n \times N$ matrix and $I \subset [[n]]$, $J \subset [[N]]$ then $M[I,J]$ denotes the $|I| \times |J|$ submatrix $(M_{i,j} : i \in I, j \in J)$ and $M[-I,J]$ denotes the $(n-|I|)\times|J|$ submatrix $(M_{i,j} : i \notin I, j \in J)$, and so on. If there is no constraint placed on the rows (or columns) just insert a "$\cdot$" symbol. For example, $M[I,\cdot\,]$ is shorthand for the $|I|\times N$ matrix $M[I,[[N]]\,]$. The downside of this notation would be the $[[[\dots]]]$ mess created if I wanted to maintain my customary convention that $(1,2,3)$ is a row vector of length 3 and $[1,2,3]$ is a column vector of length 3. Accordingly, from now on I'll happily write things like $(1,w)$ for a column vector whose first element equals 1 with the remaining coordinates coming from a vector w.

VCsets::sdim.hypercube <7> **Definition.** *A subset $\mathcal{X}$ of the discrete hypercube $\mathcal{C}_n := \{0,1\}^n$ is said to have **shatter-dimension** d (written $\textsc{sdim}(\mathcal{X}) = d$) if both the following conditions hold.*

(i) There exists at least one subset J_0 of $[[n]]$ for which

$$|J_0| = d \quad \text{AND} \quad |\{x[J_0] : x \in \mathcal{X}\}| = 2^d.$$

*That is, $\mathcal{X}$ **shatters** J_0.*

□ *(ii) $|\{x[J] : x \in \mathcal{X}\}| < 2^{|J|}$ for every subset J of $[[n]]$ with $|J| > d$.*

Remarks.

(i) To show that $\textsc{sdim}(\mathcal{X}) \le d$ we need only prove that case (ii) holds for each subset J of $[[n]]$ with $|J| = d+1$.

(i) If $\mathcal{X}$ contains at least 2 members then $1 \le \textsc{sdim}(\mathcal{X}) \le n$ with $\textsc{sdim}(\mathcal{X}) = n$ iff $\mathcal{X} = \mathcal{C}_n$. Philosphically minded readers might ponder how $\textsc{sdim}(\mathcal{X})$ should be defined if $\mathcal{X}$ is a singleton set or is empty; then they might explain how those two cases are relevant to inequality <6>.

(i) For each finite subset F of $\mathbb{A}$ and collection $\mathcal{D}$ of subsets of $\mathbb{A}$ we could define $\mathcal{X}_F$ to be the subset of $\{0,1\}^F$ consisting of all functions $\psi_D : F \to \{0,1\}$ as D ranges over $\mathcal{D}$, where $\psi_D(a) := \{a \in D\}$ for $a \in F$. Then $\textsc{vcdim}(\mathcal{D})$ could be defined as the supremum of $\textsc{sdim}(\mathcal{X}_F)$ taken over all finite subsets F of $\mathbb{A}$. Thus, to prove facts about VC-classes of sets it usually suffices to prove analogous facts about shatter dimension,

(i) Actually, Definition <7> is just a special case of Definition <2> obtained by substituting $[\![n]\!]$ for $\mathbb{A}$ then identifying $\mathcal{X}$ with the indicator functions of a collection of subsets of $[\![n]\!]$. In particular, Theorem <3> gives

\E@ sdim.bnd <8>
$$|\mathcal{X}| \le \beta(n,d) \le (en/d)^d \qquad \text{if } \textsc{sdim}(\mathcal{X}) = d.$$

The $\textsc{sdim}$ point of view will become useful in Chapter 15, when we deal with empirical processes indexed by collections of functions (or sums of more general independent stochastic processes).

Classical results about VC-dimension—as described, for example, by Dudley (2014, Chap 4) or Pollard (1984, §II.4)—usually have simpler analogs for shatter-dimension. The proofs usually require only minor modifications of the proofs for VC-classes.

VCsets::Dudley.subspace <9> **Example.** Here is result that corresponds to Dudley (1978, Theorem 7.2).

Suppose $\mathcal{X} \subset \mathcal{C}_n$ and $\mathcal{L}$ is a k-dimensional subspace of $\mathbb{R}^n$, with $k < n$. Suppose also to each $x = (x_1, \dots, x_n)$ in $\mathcal{X}$ there exists an $\ell = (\ell_1, \dots, \ell_n)$ in $\mathcal{L}$ such that $x_i = 1$ iff $\ell_i \ge 0$, for $i \in [\![n]\!]$. Then a simple piece of linear algebra shows that $\textsc{sdim}(\mathcal{X}) \le k$. Indeed, suppose J is a subset of $[\![n]\!]$ of size $k+1$. Then $\mathcal{L}_J := \{\ell[J] : \ell \in \mathcal{L}\}$ is a subspace of dimension at most k within the $(k+1)$-dimensional vector space $\mathbb{R}^J$. There must exist a nonzero vector z in $\mathbb{R}^J$ that is orthogonal to $\mathcal{L}_J$, that is, $\sum_{i \in J} z_i \ell_i = 0$ for all ℓ in $\mathcal{L}$.

Without loss of generality there is at least one i in J at which $z_i < 0$. (You might need to replace z by $-z$.) The equality

$$0 = \sum\nolimits_{i \in J} \ell_i z_i \{z_i \ge 0\} + (-\ell_i)(-z_i)\{z_i < 0\}$$

ensures there can be no ℓ in $\mathcal{L}$ for which $\ell_i \ge 0$ iff $z_i \ge 0$, for $i \in J$. From the assumed relationship between $\mathcal{X}$ and $\mathcal{L}$ it then follows that there can be no x in $\mathcal{X}$ for which $x_i = 1$ iff $z_i \ge 0$, for $i \in J$. □

The natural metric for $\mathcal{C}_n$ is the $\textsc{hamming}$ distance, defined as

\E@ Hamming.def <10>
$$\mathfrak{H}(x,y) := \sum\nolimits_{i \in [\![n]\!]} \{x_i \ne y_i\} = \sum\nolimits_i |x_i - y_i| \qquad \text{for } x, y \in \mathcal{C}_n.$$

For empirical process purposes it sometimes help to work with a weighted version of that metric, with weights coming from some (nonnegative) measure μ on $[\![n]\!]$:

\E@ weighted.Hamm <11>
$$\mathfrak{H}_\mu(x,y) := \sum\nolimits_{i \in [\![n]\!]} \{x_i \ne y_i\} \mu_i = \mu J(x,y) = \sum\nolimits_i |x_i - y_i| \mu_i$$
$$\text{where } J(x,y) := \{i \in [\![n]\!] : x_i \ne y_i\}.$$

VCsets::sdim.pack <12> **Example.** It is a straightforward exercise to extend the argument from Section 14.1 to a general subset $\mathcal{X}$ of $\mathcal{C}_n := \{0,1\}^n$ with $\textsc{sdim}(\mathcal{X}) = d$. This time I'll be slightly more careful when controlling a $\log N$ term, in order to obtain a bound on $\textsc{pack}(\epsilon, \mathcal{X}, \mathfrak{H}_\mu)$ derived by Dudley (1978, Lemma 7.13) for VC-classes of sets.

Suppose $x_1, \dots, x_N$, for some $N \ge 2$, are ϵ-separated members of $\mathcal{X}$, that is, $\mathfrak{H}_\mu(x_\alpha, x_\beta) = \mu J(x_\alpha, x_\beta) > \epsilon$ for $1 \le \alpha < \beta \le N$. Consider the projection into $\{0,1\}^m$ defined by a vector $I = (\mathcal{I}_1, \dots, \mathcal{I}_m)$ of independent observations on μ. For fixed $\alpha < \beta$,

$$\begin{aligned}\mathbb{P}\{x_\alpha[I] = x_\beta[I]\} &= \mathbb{P}\{\text{ no } \mathcal{I}_j\text{'s land in } J(x_\alpha, x_\beta)\ \}\\ &= (1 - \mu J(x_\alpha, x_\beta))^m \le \exp(-m\epsilon)\end{aligned}$$

As before, if $m := \lceil 2(\log N)/\epsilon \rceil$ then

$$\mathbb{P}\{x_\alpha[I] = x_\beta[I] \text{ for some } \alpha < \beta\ \} \le \binom{N}{2} \exp(-m\epsilon) < 1,$$

implying existence of a realization of I for which all the $x[I]$'s are distinct. It follows (via <8>) that $N^{1/d} \le em/d$.

It suffices (Why?) to bound N by a function of ϵ only for ϵ sufficiently small, say, $0 < \epsilon \le 1/2$. With that constraint we have

$$m \le \frac{\epsilon + 2\log N}{\epsilon} < \frac{3d \log N^{1/d}}{\epsilon} \qquad \text{because } \log N \ge \log 2 > 1/2,$$

so that $N^{1/d} \le em/d \le (3e/\epsilon)\log N^{1/d}$. Equivalently $G(N^{1/d}) \le w := 3e/\epsilon$ where $G(r) := e^r/r$ for $r > 0$. By an inversion inequality (Problem [3]), it then follows that $N \le c_0 w \log w$ and

$$\textsc{pack}(\epsilon, \mathcal{X}, \mathfrak{H}_\mu) \le (c_0(3e/\epsilon)\log(3e/\epsilon)) \qquad \text{for } 0 < \epsilon \le 1/2.$$

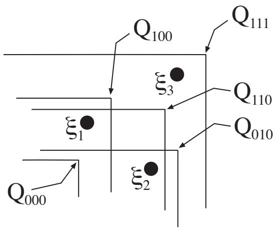
The upper bound comes tantalizingly close to $(C/\epsilon)^{\text{SDIM}(\mathcal{X})}$, which looks a lot like a packing number for a unit ball in an euclidean space of dimension $\textsc{sdim}(\mathcal{X})$. With a lot more work such a bound can be established (see Section 14.5), an aesthetically pleasing but nonessential refinement. □

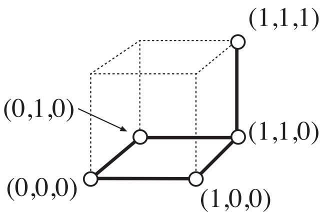
Remark. For each $p \ge 1$, note that $\left(\sum_i |x_i - y_i|^p \mu_i\right)^{1/p} = \mathfrak{H}_\mu(x,y)^{1/p}$. Thus the bound derived in the Example also leads to ℓ^p packing numbers for $\mathcal{X}$ that increase like a power of $1/\epsilon$ for smallish ϵ.

A subset $\mathcal{X}$ of $\mathcal{C}_n$ can be hard to visualize for n larger than 2 or 3. However, if $|\mathcal{X}|$ equals N then its elements can be identified with the columns of an $n \times N$ ***binary matrix*** M, a matrix with elements all belonging to $\{0,1\}$.

Remark. In the rest of the Chapter, I'll treat the general $\mathcal{X} \subset \mathcal{C}_n$ and its corresponding $n \times N$ binary matrix M as essentially the same object. (The only real distinction between the two is that M enumerates $\mathcal{X}$ as $M[\cdot,1], \dots, M[\cdot,N]$.) Thus $\textsc{sdim}(M) = \textsc{sdim}(\mathcal{X})$ and the edge set $\mathcal{E}$, a set of edges $\{x,y\}$ for $\mathcal{X}$, corresponds to a set of pairs $\{j,k\}$ with j and k distinct members of $[[N]]$.

VCsets::quadrants <13> **Example.** Let $\mathcal{Q}$ be the set of all closed quadrants with a north-east vertex in $\mathbb{R}^2$, that is, sets of the form $\{(u,v) \in \mathbb{R}^2 : u \le a,\ v \le b\}$ with $a, b \in \mathbb{R}$. You should check that $\text{VCDIM}(\mathcal{Q}) = 2$. (Hint: If a quadrant Q contains the northernmost point and the easternmost point of a finite set F, why must Q contain all of F?) The maximum number of subsets that $\mathcal{Q}$ could pick out from an F of size $|F| = 3$ is seven. (That maximum is achieved when the three points lie on straight line running from north-west to south-east.)





$$M = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

From the three points ξ_1, ξ_2, ξ_3 of $\mathbb{R}^2$ shown in the picture, $\mathcal{Q}$ picks out only five subsets. The corresponding subset $\mathcal{X}$ consists of five of the eight 'vertices' of the discrete hypercube $\mathcal{C}_3$, which are labelled in the picture to match the columns of the corresponding 3×5 binary matrix M.

For the moment you should ignore the 12 lines (7 dotted and 5 dark) that I added to the picture of $\mathcal{C}_3$ and its subset $\mathcal{X}$. Later, in Section 14.4, I'll reinterpret $\mathcal{X}$ as the set of vertices of a graph with 5 edges. That Section will establish a most surprising relationship between $\text{SDIM}(\mathcal{X})$ and the largest possible number of edges, which will play a key role in Section 14.5. □

For the $\mathcal{Q}$ in the previous Example there is a polynomial upper bound, of order $O(n^2)$, for the size of $|\mathcal{Q}_F|$ that hold for every F of size n. If we replace $\mathcal{Q}$ by the collection $\mathcal{K}$ of all compact convex subsets of $\mathbb{R}^2$ then the $\text{SDIM}(\cdot)$ situation changes: if F consists of n points spread around a circle then $\mathcal{K}$ picks out all 2^n subsets (consider the convex hull of F_0 for each subset F_0 of F). Thus $\text{VCDIM}(\mathcal{K}) = \infty$. However, if F consists of n points spread along a single line then $\mathcal{K}$ picks out only $\binom{n+1}{2}$ subsets. Less obviously, for observations $\xi_1, \dots, \xi_n$ from the uniform distribution on the unit square, $\text{SDIM}(\mathcal{X}_\omega)/n$ is small with high probability. See Pollard (1984, p 23).

*14.3 Proof of the shatter theorem

VCsets::S:downshift

As explained in the Remarks following Definition <7>, to prove Theorem <3> it suffices to prove an analogous result for a subset $\mathcal{X}$ of $\mathcal{C}_n := \{0,1\}^n$, namely: if $\text{SDIM}(\mathcal{X}) = d$ then

\E@ sdim.shatter <14>
$$N := \text{ the cardinality of } \mathcal{X} \ \le \beta(n,d) := \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{d}.$$

There are several ways to prove this inequality, or something essentially equivalent to it. I particularly like the ***downshift method***, which apparently is a standard combinatorial tool (Frankl, 1987). The method is easiest to understand when expressed using binary matrices, starting from the matrix M_0 constructed from the columns of $\mathcal{X}$ (as in the third little picture from Example <13>). That matrix is then transformed by a sequence of ***downshift*** operations into another binary matrix where calculations are simpler.

VCsets::down <15> **Definition.** *Let M be any $n \times N$ binary matrix with distinct columns. A downshift of the ith row is the result of applying the following operation to M.*
For $j \in [[N]]$,
if $M[i, j] = 1$ change it to a 0 unless
the resulting column would duplicate an existing column of M.
The downshift is said to succeed if it changes the row $M[i, \cdot]$ in at least one position. □

The idea is to cycle through the rows, downshifting as we go, creating new binary matrices $M_1, M_2, \dots$ until eventually we reach a matrix for which no more successful downshifts (of any row) are possible. As each successful downshift reduces the total number of 1's in the matrix it takes only finitely many steps to reach that goal.

The procedure is best understood by applying it to a simple M_0. To make it easier to keep track of which downshifts succeed and which are blocked, I'll write beside each M_i the matrix $\overline{M}_i$ consisting of the vectors from $\mathcal{C}_n$ that are missing from M_i. (I found this trick helpful when working through an example on the blackboard.)

VCsets::H2.downshift <16> **Example.** Start with an M_0 consisting of 8 of the 16 vectors in $\mathcal{C}_4$.

$$M_0 = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 \end{bmatrix} \qquad \overline{M}_0 = \begin{bmatrix} 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{bmatrix}$$

A downshift of row 1 of M_0 changes columns 4 and 6 but column 2 is blocked by column 1 and column 8 is blocked by column 7. The $\boxed{0}$'s in the matrix M_1 indicate which new columns appear and the $\boxed{1}$'s in the matrix $\overline{M}_1$ indicate the columns that disappear from M_0.

$$M_1 = \begin{bmatrix} 0 & 1 & 0 & \boxed{0} & 0 & \boxed{0} & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 \end{bmatrix} \qquad \overline{M}_1 = \begin{bmatrix} \boxed{1} & 0 & 1 & 1 & \boxed{1} & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{bmatrix}$$

Effectively, the downshifting has swapped two columns from M_0 with two columns from $\overline{M}_0$.

A downshift of the second row of M_1 again swaps two columns between M_1 and $\overline{M}_1$. This time column 1 blocks a downshift of column 4 and column 5 blocks a downshift of column 7:

$$M_2 = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & \boxed{0} & 1 & \boxed{0} \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 \end{bmatrix} \qquad \overline{M}_2 = \begin{bmatrix} 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 \\ 1 & \boxed{1} & 0 & 0 & 1 & 1 & 1 & \boxed{1} \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{bmatrix}$$

Then downshift the third row of M_2:

$$M_3 = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & \boxed{0} & \boxed{0} \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 \end{bmatrix} \qquad \overline{M}_3 = \begin{bmatrix} 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & \boxed{1} & 1 & \boxed{1} & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{bmatrix}$$

Every potential downshift of the fourth row of M_3 is blocked. No more downshifts—for any row of M_3—can succeed.

For future reference, note that no column of M_3 contains more than two 1's. □

Now back to the proof of inequality <14>. First some useful terminology. If M is an $n \times N$ binary matrix with distinct columns and $I \subset [[n]]$, say that M shatters I if $M[I, \cdot]$ contains all members of $\{0,1\}^I$ amongst its columns. Note that if $\text{SDIM}(M) = k$ then M cannot shatter any I with $|I| > k$.

In general, the downshifting operation has two important properties:

(i) No new shattered sets of rows can be created by a downshift.

Proof. For notational convenience, suppose that M^* is created from M by a downshift of the first row and that M^* shatters some I. We may assume $1 \in I$ because if $1 \notin I$ we have $M[I, \cdot] = M^*[I, \cdot]$, ensuring that M also shatters I. Again for notational convenience, we may suppose that $I = [[k]]$.

The assumption about M^* tells us that for each u in $\{0,1\}^{I \backslash \{1\}}$ there must exists vectors $v_0, v_1 \in \{0,1\}^{[[n]] \backslash I}$ such that $x_1 := (1, u, v_1)$ and $x_0 := (0, u, v_0)$ are both columns of M^*. Downshifting creates no new 1's, so x_1 must also be a column of M. And for such a column to have survived the downshift the change must have been blocked by the presence of x_0 amongst the columns of M. □

(ii) If no downshifts can succeed for a binary matrix M then it is ***hereditary***. That is, if x is a column of M and if y is obtained from x by changing some of its 1's to 0's then y is also a column of M. Consequently, if $x[i] = 1$ for each i in I, for some $I \subset [[n]]$, then M shatters I.

Proof. For simplicity suppose that $y[i] = 0$ and $x[i] = 1$ for each i in $I := [[k]]$ while $y[-I] = x[-I]$. A downshift of the first row must be blocked by the vector $x_1 = (0, x[-1])$ being a column of M. Then a downshift of the second row must be blocked by the vector $x_2 = (0, 0, x[-\{1,2\}])$ being a column of M. And so on. □

With these two facts in hand the rest of the argument is straightforward. Suppose we start with M_0 and after a R downshifts end up with an M_R (with all columns distinct) for which no more downshifts can succeed. Suppose for some j the column $\{i \in [[n]] : M_R[i,j] = 1\} = I$. By property (ii), the submatrix $M[I,\cdot]$ must contain all the vectors in $\{0,1\}^I$, that is, M shatters I. And by property (i) we have $\text{SDIM}(M_R) \le \text{SDIM}(M_0) = d$. Taken together, these facts tell us that $|\{i \in [[n]] : M_R[i,j] = 1\}| \le d$ for each J. All the colums of M_R must belong to $\{x \in \mathcal{C}_n : \sum_i x_i \le d\}$, a set of size $\beta(n,d)$. □

*14.4 Subsets of the discrete hypercube as graphs

VCsets::S:graphs

The results in this Section are included because I find them elegant and surprising, and also because they provide the main tools for the proof (in Section 14.5) of the (aesthetically pleasing but nonessential) fact that the packing numbers for a subset $\mathcal{X}$ of $\mathcal{C}_n := \{0,1\}^n$ increase like $(1/\epsilon)^{\text{SDIM}(\mathcal{X})}$.

We can also think of $\mathcal{X}$ as the set of vertices for a graph with the set of edges $\mathcal{E}$ (or $\mathcal{E}(\mathcal{X})$ if we wish to distinguish between various sets of edges) defined as all the pairs $\{x,y\}$ from $\mathcal{X}$ for which

\E@ xx.edge.def <17>
$$\mathfrak{H}(x,y) := \sum\nolimits_{i\in[[n]]} \{x_i \ne y_i\} = 1.$$

For such an edge, write $x \overset{i}{\sim} y$ if $x[i] \ne y[i]$. Define the corresponding subset of edges by $\mathcal{E}_i := \mathcal{E}_i(\mathcal{X}) := \{\{x,y\} \in \mathcal{E} : x \overset{i}{\sim} y\}$.

If we are using the $n \times N$ binary matrix M notation to represent $\mathcal{X}$ we should identify vertices with members of $[[N]]$ and the set of edges $\mathcal{E}(M)$ should consist of pairs $\{j,k\}$ from $[[N]]$ for which

\E@ M.edge.def <18>
$$\mathfrak{H}(j,k) := \sum\nolimits_{i\in[[n]]} \{M[i,j] \ne M[i,k]\} = 1.$$

In what follows I'll treat $\mathcal{X}$ and M as the same object, using whichever name seems most convenient to decribe any particular operation.

There is a most surprising general relationship between the shatter dimension of $\mathcal{X}$ and the number of edges in the corresponding graph.

VCsets::edges <19> **Theorem.** *(Haussler, 1995, Lemma 2) If* $\text{SDIM}(M) = d$ *then* $|\mathcal{E}| \le Nd$ *edges.*

The proof will again make use of downshifting using the following pleasant fact about the effect of a single downshift operation.

VCsets::edges.downshift <20> **Lemma.** *If a single downshift transforms* $(M, \mathcal{E})$ *into* $(M^*, \mathcal{E}^*)$ *then* $|\mathcal{E}| \le |\mathcal{E}^*|$.

Proof. For simplicity, consider just the function σ defined by a downshift of the first row: $\sigma M[\cdot,j] := M^*[\cdot,j]$. The proof works by constructing a one-to-one map (also called an injection) $S : \mathcal{E}\backslash\mathcal{E}^* \to \mathcal{E}^*\backslash\mathcal{E}$ such that $S : \mathcal{E}_i\backslash\mathcal{E}^* \to \mathcal{E}_i^*\backslash\mathcal{E}$ for each i. That property implies

$$|\mathcal{E}\backslash\mathcal{E}^*| = \sum\nolimits_{i\in[[n]]} |\mathcal{E}_i\backslash\mathcal{E}^*| \le \sum\nolimits_{i\in[[n]]} |\mathcal{E}_i^*\backslash\mathcal{E}| = |\mathcal{E}^*\backslash\mathcal{E}|.$$

The argument will involve careful bookkeeping while defining S on each $\mathcal{E}_i\backslash\mathcal{E}^*$.

For $i = 1$ everything is easy, because $\mathcal{E}_1\backslash\mathcal{E}^* = \emptyset$: if $x = (1, w) \in \mathcal{X}$ and $y = (0, w) \in \mathcal{X}$ then y blocks x from changing, so that $\{x, y\} \in \mathcal{E}^*$.

The argument for $i = 2$ is typical of the remaining cases. Suppose $\{x, y\} \in \mathcal{E}_2$ with $x = (a, 1, v)$ and $y = (a, 0, v)$ for some $a \in \{0, 1\}$ and some $v \in \{0, 1\}^{[\![n]\!]\backslash\{1,2\}}$. If $a = 0$ then neither x nor y is affected by σ, implying $\{x, y\} \in \mathcal{E}^*$. Thus we need only consider the case where $x = (1, 1, v) \in \mathcal{X}$ and $y = (1, 0, v) \in \mathcal{X}$. For such x and y,

(i) If $x \neq \sigma x$ then $\sigma x = w := (0, 1, v)$ cannot belong to $\mathcal{X}$, for otherwise w would have blocked the downshift.

(ii) if $x = \sigma x$ then the downshift must have been blocked by w:=$(0, 1, v)$, which must belong to $\mathcal{X}$. Also $w = \sigma w \in \mathcal{X}^*$ because $w[1] = 0$.

(iii) If $y \neq \sigma y$ then $\sigma y = z := (0, 0, v)$ cannot belong to $\mathcal{X}$, for otherwise z would have blocked the downshift.

(iv) if $y = \sigma y$ then the downshift must have been blocked by z:=$(0, 0, v)$, which must belong to $\mathcal{X}$. Also $\sigma z = z$ because $z[1] = 0$.

There are now four possibilities to consider.

case 1. If $x = \sigma x$ and $y = \sigma y$ then $\{x, y\} \in \mathcal{E}^*$, so $\{x, y\} \notin \mathcal{E}\backslash\mathcal{E}^*$

case 2. If $x = \sigma x$ and $y \neq \sigma y = (0, 0, v) \notin \mathcal{X}$ define $S\{x, y\} = \{(0, 1, v), \sigma y\}$.

case 3. If $x \neq \sigma x = (0, 1, v) \notin \mathcal{X}$ and $y = \sigma y$ define $S\{x, y\} = \{\sigma x, (0, 0, v)\}$.

case 4. If $x \neq \sigma x = (0, 1, v)$ and $y \neq \sigma y = (0, 0, v)$ define $S\{x, y\} = \{\sigma x, \sigma y\}$.

By properties (i) through (iv), in cases 2,3,4 we have $S\{x, y\} \in \mathcal{E}_2^*\backslash\mathcal{E}$, as desired. Moreover, in those cases $\{x, y\}$ can be uniquely recovered from $S\{x, y\}$ by changing initial 0's in the vectors back to 1's, which ensures that S is one-to-one. □

Proof (of Theorem <19>). Let M and $\mathcal{E}$ be the binary matrix and the edge set corresponding to $\mathcal{X}$. Apply a sequence of downshifts until arriving at a matrix $M^\dagger$, with column set $\mathcal{X}^\dagger$ and edge set $\mathcal{E}^\dagger$, for which no more downshifts can succeed. By repeated appeals to Lemma <20> we know that $|\mathcal{E}| \leq |\mathcal{E}^\dagger|$ and, from Section 14.3, that $\textsc{sdim}(M^\dagger) \leq \textsc{sdim}(M) = d$. As $M^\dagger$ is hereditary, it follows that the number of 1's in each column of $M^\dagger$ must be $\leq d$.

Consider an edge $\{x, y\}$ from $\mathcal{E}^\dagger$. There must be exactly one α for which $x[\alpha] \neq y[\alpha]$. Without loss of generality suppose $x[\alpha] = 1$ and $y[\alpha] = 0$, which means that y is derived from x by changing one of its 1's to a 0. That can happen in at most d different ways. Put another way, every edge in $\mathcal{E}^\dagger$ appears on the list of pairs obtained by working through the set of N members of $\mathcal{X}^\dagger$ and pairing each with one of d other vertices, which gives $|\mathcal{E}^\dagger| \leq Nd$. □

The final paragraph of the previous Proof effectively bounded the size of $\mathcal{E}^\dagger$ by assigning an orientation to each edge: if $\mathfrak{e} = \{x, y\}$ and $\sum_i x[i] = 1+\sum_i y[i]$ we could define $\text{HEAD}(\mathfrak{e}) := x$ and $\text{TAIL}(\mathfrak{e}) = y$, giving the edge an orientation from tail to the head. The bound $\sum_i x_i \le \text{SDIM}(\mathcal{X}^\dagger)$ then gives

$$\text{indegree}(x) := |\{\mathfrak{e} \in \mathbb{E}^\dagger : \text{HEAD}(\mathfrak{e}) = x\}| \le \text{SDIM}(\mathcal{X}).$$

The proof in the next Section requires a similar ordering for the original graph $(\mathcal{X}, \mathcal{E})$. The existence of such an ordering can be deduced from Theorem <20> via the classical "marlem" (=marriage lemma: UGMTP Problem 10.5). Recall the statement of that result:

> If for each s in a finite set S we are given a nonempty subset K_s of some finite set $\mathcal{K}$ and for each subset A of S we have $|\cup_{s\in A} K_s| \ge |A|$ then there exists a one-to-one function ψ mapping S into $\mathcal{K}$ with $\psi(s) \in K_s$ for each $s \in S$.

VCsets::directedVC *<21>* **Corollary.** *(Based on Haussler (1995, Lemma 3), who applied a result from Alon and Tarsi (1992, Lemma 3.1) to orient edges.) Suppose $(\mathcal{X}, \mathcal{E})$ is a graph with $\text{SDIM}(\mathcal{X}) \le d$, as in Theorem <19>. Then there exists an orientation of the edges for which $|\sum_{\mathfrak{e}\in\mathcal{E}}\{\text{HEAD}(\mathfrak{e}) = x\}| \le d$ for every x in $\mathcal{X}$.*

Proof. Apply marlem with $\mathcal{E}$ playing the role of the set S. We want a map HEAD from $\mathcal{E}$ into the set $\mathcal{X}$ that allows as many as d edges to map to the same vertex. As marlem produces a one-to-one map we need $\mathcal{K}$ to contain d copies of each x in $\mathcal{X}$. That can be arranged by defining

$$\mathcal{K} := \mathcal{X} \times [d] := \{(x, j) : x \in \mathcal{X}, j \in [d]\}.$$

Near the end of the argument we can discard the j's to turn a one-to-one map $\psi : \mathcal{E} \to \mathcal{K}$ into many-to-one map HEAD for which $\text{HEAD}(\mathfrak{e}) = x$ iff $\psi(\mathfrak{e}) = (x, j)$ for some j in $[d]$.

Define

$$K(\mathfrak{e}) := \{(x_0, j) : j \in [[d]\} \cup \{(y_0, j) : j \in [[d]\} \qquad \text{if } \mathfrak{e}_0 = \{x_0, y_0\} \in \mathcal{E}.$$

If $A \subset \mathcal{E}$ let $\mathcal{X}_0 := \cup A$, the set of all members of $\mathcal{X}$ that appear as a vertex of some edge in A, and $\mathcal{E}_0 := \{\{x, y\} : x, y \in \mathcal{X}_0 \text{ and } \mathfrak{H}(x, y) = 1\}$. Then we have $A \subseteq \mathcal{E}_0$ and $\text{SDIM}(\mathcal{X}_0) \le \text{SDIM}(\mathcal{X}) \le d$, so that $|\mathcal{E}_0| \le d|\mathcal{X}_0|$ by Theorem <19>, and $\cup_{\mathfrak{e}\in A} K(\mathfrak{e}) = \mathcal{X}_0 \times [[d]]$. It follows that

$$|A| \le |\mathcal{E}_0| \le d|\mathcal{X}_0| = |\cup_{\mathfrak{e}\in A} K(\mathfrak{e})|,$$

as required by marlem to provide the one-to-one map $\psi : \mathcal{E} \to \mathcal{K}$. If $\mathfrak{e}_0 = \{x_0, y_0\} \in \mathcal{E}$ then $\psi(\mathfrak{e})$ equals either (x_0, j) or (y_0, j) for some j in $[[d]]$. Strip off the j to define $\text{HEAD}(\mathfrak{e})$ then take $\text{TAIL}(\mathfrak{e})$ as the other vertex of $\mathfrak{e}$.

Finally, we can argue for any given x in $\mathcal{X}$ that an edge $\mathfrak{e}$ has $\text{HEAD}(\mathfrak{e}) = x$ iff $\mathfrak{e} = \{x, y\}$ for some $y \in \mathcal{X}$ with $\mathfrak{H}(x, y) = 1$ and $\psi(\mathfrak{e}) = (x, j)$ for some j in $[[d]]$. For each such j there can be at most one edge $\mathfrak{e}$ for which $\psi(\mathfrak{e}) = (x, j)$. Thus $\sum_{\mathfrak{e}\in\mathcal{E}}\{\text{HEAD}(\mathfrak{e}) = x\} \le d$, as asserted. □

For future reference, the following table summarizing the correspondences between the two representations.

$\mathcal{X} \subset \{0,1\}^n$ with $\lvert\mathcal{X}\rvert = N$	$n \times N$ binary matrix M, distinct columns
vertex $x \in \mathcal{X}$	vertex $j \in [[N]]$ with $M[\cdot, j] = x$
edge $e \in \mathcal{E}$	edge $\mathfrak{e} = \{j_1, j_2\}$ with $\mathfrak{H}(M[\cdot, j_1], [\cdot, j_2]) = 1$
edge $\mathfrak{e} \in \mathcal{E}_i$	$\mathfrak{e} = \{j_1, j_2\}$ and $\lvert M[i, j_1] - M[i, j_2] \rvert = 1$

We should also reinterpret HEAD and TAIL from Corollary <21> as functions from $\mathcal{E}$ into $[[N]]$ for which

$$|\sum\nolimits_{\mathfrak{e}\in\mathcal{E}} \{\textsc{head}(\mathfrak{e}) = j\}| \le \textsc{sdim}(M) \qquad \text{for every } j \text{ in } [[N]].$$

*14.5 Haussler's improvement of the packing bound

VCsets::S:HausslerVC

For the case of a weighted HAMMING distance with uniform weights Haussler (1995) improved the bound on packing numbers from Example <12> by using a more subtle randomization argument to remove the $\log(3e/\epsilon)$ factor. The rearrangement of Haussler's argument that are presented in this Section resulted from discussions I had in 1994 with Aad van der Vaart and Jon Wellner. For their version of the rearrangement see van der Vaart and Wellner (1996, §2.6).

VCsets::Haussler.pack <22> **Theorem.** *Let $\mathcal{X}$ be a subset of $\mathcal{C}_n := \{0,1\}^n$ with $\textsc{sdim}(\mathcal{X}) \le d$ and let μ be the uniform distribution on $[[n]]$. Then*

$$\textsc{pack}(\epsilon, \mathcal{X}, \mathfrak{H}_\mu) \le \mathcal{H}_d(\epsilon) := Cd(2e/\epsilon)^d \qquad \text{for } 0 < \epsilon < 1.$$

□ *for a universal constant C.*

> **Remark.** The assumption about μ can be removed by means of a simple approximation trick. See Problem [4].

The proof of the Theorem is easier to understand when written using binary matrix representations. The following Lemmas covers all the tricky bits. They both involve an integer m, which will be chosen during the Proof of the Theorem to optimize a bound.

VCsets::Haussler.lower <23> **Lemma.** *Let M be an $n \times N$ binary matrix with distinct columns for which $\textsc{sdim}(M) \le d$ and let $\mathfrak{J}$ be a random variable that is uniformly distributed over $[[N]]$. Suppose, for some positive ϵ,*

\E@ M.sep <24>
$$\sum\nolimits_{i\in[[n]]} \{M[i, j_1] \ne M[i, j_2]\} > n\epsilon \qquad \text{for all distinct } j_1, j_2 \text{ in } [[N]].$$

Then for each subset K of $[[n]]$ of size m we have

$$\sum\nolimits_{i\in[[n]]\backslash K} \mathbb{P}\text{var}\left(M[i, \mathfrak{J}] \mid M[K, \mathfrak{J}]\right) \ge \tfrac{1}{2} n\epsilon \left(1 - \beta(m, d)/N\right),$$

□ *with $\beta(\cdot, n)$ as in Theorem <3>.*

VCsets::Haussler.upper <25> **Lemma.** *Let B be an $(m+1)\times L$ binary matrix with distinct columns for which* $\text{SDIM}(B)\le d$. *If $\mathcal{L}$ is a random variable taking values in $[[L]]$ then we have*

$$\square \qquad \sum\nolimits_{i\in[[m+1]]} \mathbb{P}\mathrm{var}\left(B[i,\mathcal{L}] \mid B[-i,\mathcal{L}]\right) \le d.$$

Prove the Theorem first, assuming the results from the Lemmas, then prove those Lemmas.

Proof (of Theorem <22>). For a given ϵ in $(0,1)$ and $N = \text{PACK}(\epsilon,\mathcal{X},\mathfrak{H}_\mu)$ let M be a binary matrix of dimension $n\times N$ whose columns form an ϵ-packing set for $\mathcal{X}$ and let $R := M[\cdot,\mathcal{J}]$ for a random variable $\mathcal{J}$ distributed uniformly on $[[N]]$.

From Lemma <23> we get

\E@ Ki.lower <26>
$$\sum\nolimits_{|K|=m}\sum\nolimits_{i\in[[n]]\backslash K} \mathbb{P}\,\mathrm{var}(R[i]\mid R[K]) \ge \binom{n}{m}\tfrac{1}{2}n\epsilon\left(1-\beta(m,d)/N\right).$$

Now consider a fixed subset I of $[[n]]$ with size $m+1$. The columns of the submatrix $M[I,\cdot]$ need not be unique. Let B be the $(m+1)\times L$ binary matrix constructed from the set of vectors from $\{0,1\}^I$ that appear at least once amongst the columns of $M[I,]$. We must have $\text{SDIM}(B)\le d$ because B is a submatrix of M.

For each ℓ in $[[L]]$ define $A_\ell := \{j\in[[N]] : M[I,j] = B[\cdot,\ell]$ and then define a probability measure P on $[[L]]$ by $P\{\ell\} := \mathbb{P}\{\mathcal{J}\in A_\ell\}$. Then let $\mathcal{L}$ be a random variable with distribution P. These choices ensure that the random vector $B[\cdot,\mathcal{L}]$ has the same distribution as $R[I]$. Consequently, for each i in I we have

$$\mathbb{P}\mathrm{var}(B[i,\mathcal{L}] \mid B[I\backslash\{i\},\mathcal{L}] = \mathbb{P}\mathrm{var}(R[i]\mid R[I\backslash\{i\}])$$

Sum over i in I then over all I with $|I| = m+1$, then invoke Lemma <25> to deduce that

\E@ sum.Ii <27>
$$\sum\nolimits_{|I|=m+1}\sum\nolimits_{i\in I} \mathbb{P}\,\mathrm{var}(R[i]\mid R[K]) \le \binom{n}{m+1}d.$$

Remark. This inequality assumes that $m+1 > d$. Soon we will be choosing $m = \lceil 2(d+1)/\epsilon\rceil$.

We need can transform <27> into a form comparable to <26> by noting that each pair (i,K) with $|K| = m$ and $i\in[[n]]\backslash K$ corresponds to exactly one pair (i,I) with $|I| = m+1$ and $i\in I$ if we define $I = K\cup\{i\}$. Thus

\E@ Ki.upper <28>
$$\sum\nolimits_{|K|=m}\sum\nolimits_{i\in[[n]]\backslash K} \mathbb{P}\,\mathrm{var}(R[i]\mid RK) \le \binom{n}{m+1}d.$$

Remark. As a check, note that the double sums in <27> and <28> both involve $(m+1)\binom{n}{m+1} = (n-m)\binom{n}{m}$ terms.

Together <26> and <28> imply

$$\binom{n}{m}\tfrac{1}{2}n\epsilon\left(1-\beta(m,d)/N\right),\le\binom{n}{m+1}d$$

which rearranges to

$$N\le\beta(m,d)\left(1-\frac{2(n-m)d}{n\epsilon(m+1)}\right)^{-1}.$$

We could try to optimize over m immediately, as Haussler (1995, page 225) did, to bound N by a function of ϵ, d, and n, then take a supremum over n. However, it is simpler to note that all the conditions of the Theorem apply if n is replaced by qn, for some positive integer q, and each x in $\mathcal{X}$ is replaced by the concatenation of q copies of x. (The new vectors are the columns of the $(qn)\times N$ binary matrix obtained by stacking q copies of M.) Letting q tend to infinity (with m fixed) eliminates n from the bound, leaving

$$N\le\beta(m,d)\left(1-\frac{2d}{\epsilon(m+1)}\right)^{-1}\le(em/d)^d\left(1-\frac{2d}{\epsilon(m+1)}\right)^{-1}.$$

As a function of m, the first factor is increasing and the second is decreasing. If we treated m as a continuous variable and ignored the difference between m and $m+1$ the minimizing value would be $2(d+1)/\epsilon$. That suggests we choose $m=\lceil 2(d+1)/\epsilon\rceil$, which gives the bound

$$N\le\left(\frac{2e(d+1)}{\epsilon d}\right)^d\frac{2(d+1)}{2(d+1)-2d}\le(2e/\epsilon)^d(1+d^{-1})^{d+1}.$$

Remark. For comparison's sake, note that Haussler got the bound

$$e(d+1)\left(2e(n+1)/(n\epsilon+2d+2)\right)^d.$$

Proof (of Lemma <23>). As the set K is fixed throughout the argument I'll omit it from the notation, even though everything—such as the random vector $R:=M[\cdot,\mathcal{J}]$—depends on the choice of K.

Define $\mathcal{Z}$ to be the set of all vectors in the discrete hypercube $\{0,1\}^K$ that appear at least once as a column of $M[K,\cdot]$. From the fact that $\text{SDIM}(M[K,\cdot])\le\text{SDIM}(M)\le d$ we get $|\mathcal{Z}|\le\beta(m,d)$.

For each z in $\mathcal{Z}$ define $J_z:=\{j\in[[N]]:M[K,j]=z\}$, so that we have $\{R[K]=z\}=\{\mathcal{J}\in J_z\}$. The distribution, P_z, of $\mathcal{J}$ conditional on $\mathcal{J}\in J_z$ is just the uniform distribution on J_z. A very well known symmetrization argument (cf. the variance of a difference of two independent random variables) will give a simple expression for $\text{var}(R[i]\mid R[K]=z)$. Let $\widetilde{P}_z$ be a copy of P_z. For each fixed i in $[[n]]\backslash K$,

$$\begin{aligned}\text{var}(R[i]\mid R[K]=z)&=P_zM[i,j]^2-P_zM[i,j]\,\widetilde{P}_zM[i,\widetilde{j}]\\&=\tfrac{1}{2}\left(P_zM[i,j]^2-2P_zM[i,j]\,\widetilde{P}_zM[i,\widetilde{j}]+\widetilde{P}_zM[i,\widetilde{j}]^2\right)\\&=\tfrac{1}{2}P_z\otimes\widetilde{P}_z\left(M[i,j]-M[i,\widetilde{j}]\right)^2=\tfrac{1}{2}P_z\otimes\widetilde{P}_z\{M[i,j]\ne M[i,\widetilde{j}]\}\end{aligned}$$

because $M[i,j]-M[i,\widetilde{j}]$ takes only values $0,\pm1$.

Sum over i in $[[n]]\backslash K$, remembering property <24> and the fact that that $M[K,j]=M[K,\widetilde{j}]=z$ when $j,\widetilde{j}\in J_z$, to deduce that

$$\begin{aligned}\sum\nolimits_{i\notin K}\text{var}(R[i]\mid R[K]=z)&=\sum\nolimits_{i\in[[n]]}\tfrac12 P_z\otimes\widetilde{P}_z\{j\ne\widetilde{j}\}\{M[i,j]\ne M[i,\widetilde{j}]\}\\&\ge\tfrac12 P_z\otimes\widetilde{P}_z\{j\ne\widetilde{j}\}n\epsilon=\tfrac12 n\epsilon(1-1/|J_z|).\end{aligned}$$

Finally, average out over $\mathcal{Z}$, remembering that $\mathbb{P}\{\mathcal{J}\in J_z\}=|J_z|/N$ and $\sum_z|J_z|=N$ and $|\mathcal{Z}|\le\beta(m,d)$ to conclude that

$$\begin{aligned}&\sum\nolimits_{i\in[[n]]\backslash K}\mathbb{P}\text{var}\left(R[i]\mid R[K]\right)\\&\quad=\sum\nolimits_{z\in\mathcal{Z}}\mathbb{P}\{\mathcal{J}\in J_z\}\sum\nolimits_i\text{var}\left(R[i]\mid\mathcal{J}\in J_z\right)\\&\quad\ge\tfrac12 n\epsilon\sum\nolimits_{z\in\mathcal{Z}}\left(|J_z|/N-1/N\right)=\tfrac12 n\epsilon\left(1-\beta(m,d)/N\right).\end{aligned}$$

□

Proof (of Lemma <25>). To simplify notation define $I:=[[m+1]]$ and $R:=B[\cdot,\mathcal{L}]$. As in Section 14.4, let $\mathcal{E}$ denote the set of edges of B, the set of pairs $\{\ell,\ell'\}$ from $[[L]]$ for which $\sum_i\{B[i,\ell]\ne B[i,\ell']\}=1$. And write P for the distribution of $\mathcal{L}$.

For each i in I define

$$\begin{aligned}\mathcal{Z}_i&:=\{B[-i,\ell]:\ell\in[[L]]\}\subset\{0,1\}^{I\backslash\{i\}},\\\mathcal{Z}_i^{(1)}&:=\{z\in\mathcal{Z}_i:z=B[-i,\lambda(z)]\text{ for a unique }\lambda(z)\text{ in }[[L]]\ \},\\\mathcal{Z}_i^{(2)}&:=\{z\in\mathcal{Z}_i:\exists\lambda_0(z),\lambda_1(z)\in[[L]]\text{ for which}\\&\qquad\qquad B[-i,\lambda_\alpha(z)]=z,\quad B[i,\lambda_\alpha(z)]=\alpha\text{ for }\alpha=0,1\ \}.\end{aligned}$$

Each z in $\mathcal{Z}_i^{(2)}$ defines an edge $\mathfrak{e}_z:=\{\lambda_0(z),\lambda_1(z)\}$ in $\mathcal{E}_i$ and

$$\{R[-i]=z\}=\begin{cases}\{\mathcal{L}=\lambda(z)\}&\text{if }z\in\mathcal{Z}_i^{(1)}\\\{\mathcal{L}\in\mathfrak{e}_z\}&\text{if }z\in\mathcal{Z}_i^{(2)}\end{cases}.$$

If $z\in\mathcal{Z}_i^{(1)}$ the conditional distribution of $R[i]$ given $R[-i]=z$ is degenerate at a single value, which implies $\text{var}(R[i]\mid R[-i]=z)=0$.

If $z\in\mathcal{Z}_i^{(2)}$ the conditional distribution of $R[i]$ given $R[-i]=z$ is $\textsc{Ber}(p_z)$, where $p_z:=P\{\lambda_1(z)\}/P(\mathfrak{e}_z)$, so that

$$\text{var}(R[i]\mid R[-i]=z)=p_z(1-p_z)\le\min(p_z,1-p_z).$$

By Lemma <21>, the edges in $\mathcal{E}$ can be oriented by a pair of maps $\textsc{head},\textsc{tail}$ from $\mathcal{E}$ into $[[L]]$, for which

\E@ B.indegree <29>
$$\sum\nolimits_{\mathfrak{e}\in\mathcal{E}}\{\textsc{head}(\mathfrak{e})=\ell\}\le d\qquad\text{for each }\ell\text{ in }[[L]].$$

For the edge $\mathfrak{e}_z$ we have $\min(p_z,1-p_z)\le P\{\textsc{head}(\mathfrak{e}_z)\}/P(\mathfrak{e}_z)$ and

$$\mathbb{P}\{R[-i]=z\}=\mathbb{P}\{\mathcal{L}\in\mathfrak{e}_z\}=P(\mathfrak{e}_z).$$

Thus, for each i in I,

$$\begin{aligned}\mathbb{P}\text{var}\,(R[i] \mid R[-i]) &= \sum\nolimits_{z\in\mathcal{Z}_i} \mathbb{P}\{R[-i]=z\} \times \text{var}(R[i] \mid R[-i]=z)\\ &\le \sum\nolimits_{z\in\mathcal{Z}_i^{(2)}} P\{\text{HEAD}(\mathfrak{e}_z)\} = \sum\nolimits_{\mathfrak{e}\in\mathcal{E}_i} \mathbb{P}\{\mathcal{L} = \text{HEAD}(\mathfrak{e})\}.\end{aligned}$$

Sum out over i to deduce that

$$\sum\nolimits_{i\in I} \mathbb{P}\text{var}\,(R[i] \mid R[-i]) \le \mathbb{P}\sum\nolimits_{\mathfrak{e}\in\mathcal{E}}\{\mathcal{L} = \text{HEAD}(\mathfrak{e})\}.$$

Inequality <29> ensures that the expression on the right-hand side is $\le d$, the desired inequality. □

14.6 Problems

VCsets::S:Problems

VCsets::P:convex.hulls

[1] With points $x_1, x_2, \dots, x_{n+2}$ in $\mathbb{R}^n$ define $y_i := x_i - x_{n+2}$ for $1 \le i \le n+1$.

(i) Use a linear dependence argument to show that there exist constants $\{\alpha_i\}$, not all zero, for which $\sum_{i=1}^{n+1} \alpha_i y_i = 0$.

(ii) Define $\beta_i := \alpha_i$ for $1 \le i \le n+1$ and $\beta_{n+2} := -\sum_{i\in[\![n+1]\!]} \alpha_i$. Show that $\sum_{i\in[\![n+2]\!]} \beta_i = 0$ and $\sum_{i\in[\![n+2]\!]} \beta_i x_i = 0$.

(iii) Define $A := \{i \in [\![n+2]\!] : \beta_i > 0\}$ and $B := \{i \in [\![n+2]\!] : \beta_i < 0\}$. Show that $c := \sum_{i\in A} \beta_i = -\sum_{i\in B} \beta_i > 0$ and $w := \sum_{i\in A}(\beta_i/c)x_i = \sum_{i\in B}(-\beta_i/c)x_i$. Deduce that $w \in \text{co}\{x_i : i \in A\} \cap \text{co}\{x_i : i \in B\} \ne \emptyset$.

VCsets::P:halfspaces

[2] Let $\mathcal{H}_d$ denote the set of all closed half-spaces in $\mathbb{R}^d$. Show that $\text{VCDIM}(\mathcal{H}_d) = d+1$ by the following arguments.

(i) Show that $\mathcal{H}_d$ shatters the set $F = \{0, e_1, \dots, e_d\}$, where e_i is the unit vector with $+1$ in the ith position. Hint: Consider closed halfspaces of the form $\{x : \theta\cdot x \ge r\}$ with $\theta \in \{-1,+1\}^d$.

(ii) Suppose $F = \{x_i : i = 0, 1, \dots, d+1\}$ is a set of $d+2$ distinct points in $\mathbb{R}^d$. If H picks out a subset $F_K := \{x_i : i \in K\}$ from F, show that that the convex hull of F_K is a subset of H, which is disjoint from the convex hull of F_{K^c}.

(iii) Show that $\sum_{1\le i\le d+1} \alpha_i(x_i - x_0) = 0$ for some constants α_i that are not all zero. Define $\alpha_0 = -\sum_{1\le i\le d+1}\alpha_i$. Show that

$$\sum\nolimits_{0\le i\le d+1} \alpha_i x_i = 0 \qquad \text{AND} \qquad \sum\nolimits_{0\le i\le d+1} \alpha_i = 0.$$

Define $J = \{i \in [0 : d+1] : \alpha_i > 0\}$. Show that the convex hulls of the subsets $F_J = \{x_i : i \in J\}$ and F_J^c have a nonempty intersection. Thus $\mathcal{H}_d$ cannot pick out the subset F_J from F.

VCsets::P:G.ineq

[3] Example <12> described one way to bound packing numbers using shatter dimension. The bound relied on an inversion inequality for the function $G(r) = e^r/r$ for $r > 0$: if $w \ge e$ and $G(r) \le w$ then $e^r \le c_0 w \log w$ for

a universl constant c_0. Follow these steps to etablish that inequality with constant $c_0 := (1-e^{-1})^{-1} \approx 1.58$.

(i) Show that G is convex, with $\inf_{r>0} G(r) = G(1) = 1/e$. For a fixed $w \geq e$, let R be the unique value for which $R \geq 1$ and $G(R) = w$. Deduce that $\{r > 0 : G(r) \leq w\} \subset (0, R]$.

(ii) Show that

$$\log w = R - \log(R) \geq R\left(1 - \sup_{r\geq 1}(\log r)/r\right) = (1-e^{-1})R.$$

(iii) Deduce that $e^r \leq e^R = RG(R) \leq c_0 w \log w$ for $0 < r \leq R$.

VCsets::P:Haussler

[4] Suppose M is an $n \times N$ binary matrix with distinct columns and $\textsc{sdim}(M) \leq d$. For some probability measure θ on $[[n]]$, suppose $\mathcal{X}$ is ϵ-separated under the weighted HAMMING metric $\mathfrak{H}_\theta$, for some ϵ in $(0,1)$. That is,

$$\sum\nolimits_{i\in[[n]]} \theta_i |M[i,j] - M[i,j']| > \epsilon \qquad \text{for each pair } \{j, j'\} \text{ with } 1 \leq j < j' \leq N.$$

Let c be any constant that is > 1. Show that $N \leq \mathcal{H}_d(\epsilon/c)$, where $\mathcal{H}_d$ is the function from Theorem <22>. Argue as follows.

(i) For a positive integer L, which can depend on M and θ, define $L_i := \lfloor L\theta_i \rfloor$ and $m := \sum_{i\in[[n]]} L_i$. Show that L can be chosen large enough that $\theta_i/c \leq L_i/m \leq \theta_i$ for each i.

(ii) Let B be the $m \times N$ binary matrix obtained by stacking together L_i copies on $M[i,\cdot]$ for each i. Show that $\textsc{sdim}(B) \leq d$ and, for $1 \leq j < j' \leq N$,

$$m^{-1} \sum_{i\in[[m]]} |B[i,j] - B[i,j']| = \sum_{i\in[[n]]} (L_i/m)|M[i,j] - M[i,j']| > \epsilon/c.$$

Deduce that $N \leq \mathcal{H}_d(\epsilon/c)$.

14.7 Notes

VCsets::S:Notes

The bound stated in Theorem <3> appeared in the paper by Sauer (1972). In an online blog (http://leon.bottou.org/news/vapnik-chervonenkis_sauer), Bottou made a case that the credit should go to Vapnik and Červonenkis, who first published a short summary (Vapnik and Červonenkis, 1968) of their result and then a longer version (Vapnik and Chervonenkis, 1971). Dudley (2014, page 211) (see also Dudley 1978, Section 7) noted that the 1971 bound was a little weaker than the Sauer result.

In his online blog (https://gilkalai.wordpress.com/) for 28 September 2008, Kalai presented another version of the VC Lemma then commented:

> It was mentioned (with an algebraic proof by Frankl and Pach) in Gowers' blog and also, in another context, in Kowalski's blog. Sauer proved it in response to a problem of Erdos. Shelah (with Perles) proved it as a useful lemma for Shelah's theory of stable

> models. (At some later time, Benjy Weiss asked Perles about such a result in the context of ergodic theory and Perles who forgot that he proved it once proved it again.) Vapnik and Chervonenkis proved it in the context of statistical learning theory.

Steele (1975, 1978) generalized the result described in Theorem <3> to matrices whose entries come from a finite alphabet. Haussler and Long (1995) generalized further.

I learned about the downshift method for proving Theorem <3> from Michel Talagrand. Ledoux and Talagrand (1991, pp. 411-412) used the set-theoretic version of the downshift method, attributing it (page 420) to Frankl (1983). Independently, both Alon (1983) and Frankl (1983) used the shifting method to prove a result that implies Theorem <3>. Frankl (1995, page 1298) noted that the shifting technique was introduced by Erdős, Ko, and Rado (1961).

The results in Sections 14.4 and 14.5 are based on Haussler (1995), who acknowledged (page 220) Linial for the downshifting method of proof for Theorem <19>, with the comment that the result had already been proved by Haussler, Littlestone, and Warmuth (1994). For an illuminating Bayesian interpretation of the probability method used to prove Theorem <22> see Haussler (1995, Section 3).

References

Alon1983DiscMath Alon, N. (1983). On the density of sets of vectors. *Discrete Mathematics 46*, 199–202.

AlonSpencer2000 Alon, N. and J. H. Spencer (2000). *The Probabilistic Method* (second ed.). Wiley.

AlonTarsi1992Combinatorica Alon, N. and M. Tarsi (1992). Colorings and orientations of graphs. *Combinatorica 12(2)*, 125–134.

Dudley78clt Dudley, R. M. (1978). Central limit theorems for empirical measures. *Annals of Probability 6*, 899–929.

Dudley2014UCLT Dudley, R. M. (2014). *Uniform Central Limit Theorems* (2nd ed.), Volume 142 of *Cambridge studies in advanced mathematics*. Cambridge University Press. (First edition, 1999).

ErdosKoRado1995QJM Erdős, P., C. Ko, and R. Rado (1961). Intersection theorems for systems of finite sets. *Quart. J. Math. Oxford 12 (2)*, 313–320.

Frankl1983JCT Frankl, P. (1983). On the trace of finite sets. *Journal of Combinatorial Theory, Series A 34*, 41–45.

Frankl1987surveys Frankl, P. (1987). *Surveys in combinatorics]*, Volume 123 of *London Math. Soc. Lecture Note Series*, pp. 81–110. Cambridge Univ. Press.

Frankl1995handbook

Frankl, P. (1995). Extremal set systems. In R. L. Graham, M. Grötschel, and L. Lovász (Eds.), *Handbook of Combinatorics*, Volume 2, Chapter 24, pp. 1293–1329. Elsevier.

Haussler:95jct

Haussler, D. (1995). Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik-Chervonenkis dimension. *Journal of Combinatorial Theory 69*, 217–232.

erLittlestoneWarmuth1994IC

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

HausslerLong:95jct

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

LedouxTalagrand91book

Ledoux, M. and M. Talagrand (1991). *Probability in Banach Spaces: Isoperimetry and Processes*. New York: Springer.

Pollard84book

Pollard, D. (1984). *Convergence of Stochastic Processes*. New York: Springer.

Sauer72jct

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

Steele75PhD

Steele, J. M. (1975). *Combinatorial Entropy and Uniform Limit Laws*. Ph. D. thesis, Stanford University.

Steele1978JCT

Steele, J. M. (1978). Existence of submatrices with all possible columns. *Journal of Combinatorial Theory, Series A 24*, 84–88.

vaartwellner96book

van der Vaart, A. W. and J. A. Wellner (1996). *Weak Convergence and Empirical Process: With Applications to Statistics*. Springer-Verlag.

VapnikCervonenkis1968SMD

Vapnik, V. N. and A. J. Červonenkis (1968). The uniform convergence of frequencies of the appearance of events to their probabilities. *Dokl. Akad. Nauk SSSR 181*, 781–783. In Russian. English translation: Soviet Math. Dokl. 9 (1968), 915–918. See Mathematical Review MR0231431 by R. M. Dudley.

VapnikCervonenkis71events

Vapnik, V. N. and A. Ya. Chervonenkis (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications 16*, 264–280. Russian original received by the editors on May 7, 1969.

VapnikCervonenkis81fns

Vapnik, V. N. and A. Ya. Červonenkis (1981). Necessary and sufficient conditions for the uniform convergence of means to their expectations. *Theory of Probability and Its Applications 26*, 532–553. Russian original received by the editors on July 28, 1978.