Section 9.1 presents a simple concrete example to illustrate the use of the combinatorial method for deriving bounds on packing numbers. Section 9.2 defines the concept of a VC-class of sets then presents the basic polynomial bound for the numbers of subsets that can be picked out by a VC-class. It also establishes Dudley’s inequality that bounds $L^1$ packing numbers for VC-classes of sets. Section 9.3 explains how the subsets picked out by a VC-class can also be thought of as a subset of the vertices of the discrete unit cube, or as the columns of a binary matrix. Section 9.4 derives a general combinatorial bound for the shatter dimension of a binary matrix, a result often called the VC Lemma or Sauer’s Lemma. The proof uses the important technique known as downshifting. *Section 9.5 interprets subsets of the discrete unit cube as 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 9.6 presents Haussler’s refinement of the calculation from Section 9.2, leading to a sharper bound for packing numbers.
9.1 An introductory example

For my purposes, the combinatorial argument often referred to as the VC method—in honor of the important contributions of Vapnik and Červonenkis (1971, 1981)—is of use mainly as a step towards calculation of covering numbers for classes of sets or functions equipped with various $L^p$ metrics. 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 seem less mysterious.

Consider the set $\mathcal{H}$ of all closed half-spaces in $\mathbb{R}^2$. Let $F = \{x_1, \ldots, x_n\}$ 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 cardinality of $\mathcal{H}_F := \{F \cap H : H \in \mathcal{H}\}$ be? Certainly it can be no larger than $2^n$ because $F$ has only that many subsets. In fact 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 $F_0$ 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 the bound can be reduced to \( n^2 - n + 2 \), which is sharp in the sense that it is achieved by some configurations of \( n \) points in \( \mathbb{R}^2 \). See the discussion by Dudley (1978, page 921) and the original sources that he cited. For my purposes the sharp bound is not needed. Indeed any polynomial would suffice for the consequences described below.

The slow increase in \( \#H_F \) at an \( O(n^2) \) rate rather than a rapid \( 2^n \) rate, has a useful consequence for the packing numbers of \( H \) when it is equipped with an \( L^1(P) \) metric. The \( L^1(P) \) distance between two Borel sets \( B \) and \( B' \) is defined as \( P|B - B'| = P(B \Delta B') \), the probability measure of the symmetric difference. The two sets are said to be \( \epsilon \)-separated (in \( L^1(P) \)) if \( P(B \Delta B') > \epsilon \). As in Chapter 4, the packing number \( \text{pack}(\epsilon, H, P) \) or \( \text{pack}(\epsilon, H, L^1(P)) \) 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 pair \( \epsilon \)-separated.

Example. Here is an argument, due to Dudley (1978, Lemma 7.13), to show that the packing numbers \( \text{pack}(\epsilon, H, P) \) 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 \( H_1, H_2, \ldots, H_N \) are \( \epsilon \)-separated in \( L^1(P) \). By means of a cunningly chosen \( F \), the polynomial bound for \( \#H_F \) will lead to an upper bound for \( N \). The trick is to find a set \( F \) of \( m = \lceil 2 \log N/\epsilon \rceil \) points from which each \( H_i \) picks out a different subset. Then \( H \) will pick out at least \( N \) subsets from the \( m \) points, 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 Lemma <9>, the bound for the packing numbers for \( H \) can be sharpened to \( C(\epsilon^{-1} \log(1/\epsilon))^2 \), with \( C \) a universal constant, at least for \( \epsilon \) bounded away from 1. With a lot more work, even the log term can be removed—see Section 9.6. At this stage there is little point in struggling to get the best bound in \( \epsilon \). 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 set $F_0 = \{x_1, \ldots, x_m\}$ of points in $\mathbb{R}^2$ from which each $H_i$ picks out a different subset? We need to place at least one point of $F_0$ 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** proves 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 it seems appropriate to call it “The Erdős Method.”

Generate $F_0$ 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 $i \neq j$,

\[
P\{H_i \text{ and } H_j \text{ pick out same points from } F_0\} = P\{\text{no points of sample in } H_i \Delta H_j\} = (1 - P(H_i \Delta H_j))^m \leq (1 - \epsilon)^m \leq \exp(-m\epsilon).
\]

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

\[
P\{\text{some pair } H_i, H_j \text{ pick same subset from } F_0\} \leq \binom{N}{2} \exp(-m\epsilon).
\]

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

\[\square\]

**Remark.** The argument in the second paragraph of the Example implicitly assumed that the $x_i$’s sampled from $P$ are all distinct, which need not be true if $P$ has atoms. To be more precise I could have written $N \leq p(|F_0|) \leq p(m) \leq \ldots$. The added rigor is hardly worth the trouble for that Example. In general, ties can cause awkward notational difficulties. Section 9.4 introduces a way to avoid such difficulties.
§9.2 VC-dimension

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 $\#H_F$.

For small $F$, the class $H$ might pick out all subsets. The established term for this property is shattering. For example, if $F$ consists of 3 points, not all on the same straight line, then it can be shattered by $H$. However no set of 8 points can be shattered, because there are $2^8 = 256$ possible subsets—the empty set included—whereas the half-spaces can pick out at most $p(8) = 225$ subsets. More generally, $2^n > p(n)$ for all $n \geq 8$, so that (of course) no set of more than 8 points can be shattered by $H$. You should find it is easy to improve on the 8, by arguing directly that no set of 4 points can be shattered by $H$. Indeed, if $H$ picks out both $F_1$ and $F_2 = F_1^c$, then the convex hulls of $F_1$ and $F_2$ must be disjoint. You have only to demonstrate that from every $F$ with at least 4 points, you can find such $F_1$ and $F_2$ whose convex hulls overlap. See Problem [1] for details.

**In summary:** The size of the largest set shattered by $H$ is 3. Note well that the assertion is not that all sets of 3 points can be shattered, but merely that there is some set of 3 points that is shattered, while no set of 4 points can be shattered. In the terminology of the next Section, the class $H$ would be said to have VC dimension equal to 3.

9.2 VC-dimension

The argument in the previous Section had little to do with the choice of $H$ as a class of half-spaces in a particular Euclidean space. It would apply to any class $D$ of sets for which $\#D_F$ (the analog of $\#H_F$) is bounded by a fixed polynomial in $\#F$. Unfortunately, for more complicated $D$'s it can sometimes be difficult to derive such a polynomial bound directly but easier to prove existence of a finite $d$ such that no set of more than $d$ points can be shattered by $D$. Such a $D$ is called a **VC-class** of sets in honor of Vapnik & Červonenkis.

**Definition.** A class $D$ of subsets of a set $X$ is said to have VC-dimension $d$ (written $\text{VCdim}(D) = d$) if both the following conditions hold.

(i) There exists at least one subset $F_0$ of $X$ for which

$$\# \{ F_0 \cap D : D \in D \} = 2^d.$$
§9.2 VC-dimension

(ii) For every finite subset $F$ of $X$ with $|F| > d$,

$$|\{F \cap D : D \in \mathcal{D}\}| < 2^{|F|}.$$ 

**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.

Miraculously, $|\mathcal{D}_F|$ is bounded above by a polynomial in $|F|$ for each VC-class $\mathcal{D}$. The key result is often called the VC Lemma or Sauer's Lemma, although credit should be spread more widely. (See the Notes in Section Notes.)

**Theorem.** If $\mathcal{D}$ is a VC-class of subsets of $X$ with $\text{VCdim}(\mathcal{D}) \leq d$ then, for each subset $F$ of $X$, the cardinality of the set $\mathcal{D}_F := \{F \cap D : D \in \mathcal{D}\}$ is at most

$$\beta(m, d) := \binom{m}{0} + \binom{m}{1} + \cdots + \binom{m}{d} \quad \text{where } m = |F|.$$ 

This bound follows immediately from Theorem <14>, in Section downshift, which treats an analogous (more streamlined) version of the result for matrices of zeros and ones.

**Remark.** When applied to the $\mathcal{H}$ from Section 9.1, Theorem <3> gives a bound of order $O(m^3)$ rather than the $O(m^2)$ derived by direct arguments. The $\beta(m, d)$ bound is not sharp. Nevertheless it does establish a remarkable general fact: the failure to shatter for every subset $F_0$ with cardinality $d$ forces a rate of growth for $|\mathcal{D}_F|$ that increases as a polynomial in $m = |F|$, a rate much slower than the worst case $2^m$. A rate between polynomial and $2^m$, such as $2^{\sqrt{m}}$ or even $2^{m/2}$, is not possible.

Clearly $\beta(m, d)$ is a polynomial of degree $d$ in $m$. More precisely, it equals $2^m$ if $m \leq d$ and, if $m \geq d$,

$$\beta(m, d) = \sum_{k=0}^{d} \frac{m(m-1)\ldots(m-k+1)}{k!} \leq \sum_{k=0}^{d} \frac{m^k d^k}{k!} \leq (em/d)^d \quad \text{because } (m/d)^k \leq (m/d)^d \text{ when } m \geq d.$$ 

The constant $(e/d)^d$ in <4> can be improved slightly:

$$\beta(m, d) \leq 1.5m^d/d! \quad \text{if } m \geq d + 2,$$
a bound due to Vapnik and Červonenkis—see Dudley (1999, Proposition 4.15). As Dudley noted, the bound is “not far from optimal” because the \( m^d / d! \) contributes an \( m^d / d! \) as the leading term in the polynomial \( \beta(m, d) \).

**Remark.** By direct calculation, \( em/d < 2^{m/d} \) when \( m/d > x_0 \approx 3.06 \). If we were relying on \( <4> \) to detect impossibility of shattering we would overestimate the VC-dimension by a factor of about 3. For many purposes, such an overestimate would be unimportant.

It is quite easy to construct VC-classes of sets. The next Example gives a useful starting point.

**Example.** (Dudley, 1978, Theorem 7.2) Suppose \( L \) is a \( d \)-dimensional vector space of real functions on a set \( X \). Write \( D \) for the class of all subsets of \( X \) of the form \( \{ \ell \geq 0 \} \), with \( \ell \) in \( L \). Then \( D \) is a VC-class of sets, with VCdim(\( D \)) \( \leq d + 1 \).

Suppose \( F = \{ x_1, \ldots, x_m \} \subseteq X \). The set \( L_F \) of all vectors in \( \mathbb{R}^m \) of the form \( [\ell(x_1), \ldots, \ell(x_m)] \), for some \( \ell \) in \( L \), is a vector subspace of \( \mathbb{R}^m \) of dimension at most \( d \). If \( m > d \) there must exist some nonzero vector \( \alpha \) orthogonal to \( L_F \). Express the orthogonality as

\[
\sum_{i=1}^m \alpha_i \ell(x_i) = \sum_{i=1}^m (\alpha_i \ell(x_i)) \quad \text{for each } \ell \in L.
\]

Without loss of generality suppose the set \( F_0 := \{ x_i : \alpha_i < 0 \} \) is nonempty. (Replace \( \alpha \) by \( -\alpha \) if \( F_0 = \emptyset \).) No member of \( D \) can pick out the subset \( F_0^c \) from \( F \): if \( \ell(x_i) \geq 0 \) when \( \alpha_i \geq 0 \) and \( \ell(x_i) < 0 \) when \( \alpha_i < 0 \) then the right-hand side of equality \( <7> \) for that \( \ell \) would be strictly negative, while the left-hand side would be nonnegative.

Despite the simplicity of the definition of VC-dimension, it can be surprisingly tricky to check the no-shatter property directly. Instead one can rely on the polynomial bound for the number of subsets picked out and the fact that the sum or product of two polynomials is also a polynomial.

**Example.** Suppose \( D_1 \) and \( D_2 \) are VC-classes on \( X \), with VC-dimensions \( D_1 \) and \( D_2 \). Define \( D = \max(D_1, D_2) \).

Suppose \( F \) is a finite subset with \( \#F = m \). Then \( D = D_1 \cup D_2 \) has VC-dimension less than \( 5D \) because \( D \) can pick out at most \( \beta(m, D_1) + \beta(m, D_2) \) subsets from an \( F \) with \( \#F = m \) and \( 2^{1/d}(em/d) < 2^{m/d} \) when \( m \geq 5d \) because \( 2ex < 2^x \) for \( x \geq 4.67 \).

Similarly, \( \{ D_1 \cap D_2 : D_i \in D_i \} \) has VC-dimension at most \( 10D \). First note that there are at most \( (em/D)^D \) subsets picked out by \( D_1 \) and from
§9.3 Other ways to think about VC-classes

each of those sets $D_2$ picks out at most $(em/D)^D$ subsets. Then note that $(em/D)^{2D} < 2^m$ when $m \geq 10D$ because $2x < 2^{x/2}$ for $x \geq 9.33$.

A similar argument applies to other classes formed by taking the pairwise unions, or complements. The idea can be iterated to generate very fancy classes with finite VC-dimension—see Problem [2].

The connection between shatter dimension and covering numbers introduced in Section 1 extends to more general classes $D$ of measurable subsets of a space $X$ on which a probability measure $P$ is defined. This time I pay more careful attention to the inequality relating $N$ and $\log N$, for the sake of comparison with the results in Section 9.6.

Remember that the packing number $\text{pack}(\epsilon, D, P)$ is defined as the largest size of an $\epsilon$-separated (in $L^1(P)$) subset of $D$.

Lemma. (Dudley, 1978, Lemma 7.13) Let $D$ be a class of sets with VC-dimension at most $D$. Then, for each probability measure $P$,

$$\text{pack}(\epsilon, D, P) \leq \left(\frac{5e}{\epsilon} \log \frac{3e}{\epsilon}\right)^D \leq \left(\frac{15e^2}{\epsilon^2}\right)^D$$

for $0 < \epsilon \leq 1$.

Proof Suppose $D_1, \ldots, D_N$ are sets in $D$ with $P(D_i \Delta D_j) > \epsilon$ for $i \neq j$. Let $F_0 = \{x_1, \ldots, x_m\}$ be an independent sample of size $m = \lceil 2(\log N)/\epsilon \rceil$ from $P$. As in Section 9.1, for fixed $i$ and $j$,

$$P\{D_i \text{ and } D_j \text{ pick out same points from } F_0\} \leq \exp(-me).$$

The sum of $\binom{N}{2}$ such probability bounds is strictly less than 1. For some configuration $F_0$ the class $D$ picks out $N$ distinct subsets, which gives the inequality $N \leq \beta(m, D)$. Problem [4] with $\rho = 2/\epsilon$ and $\lambda \leq 3e/\epsilon$ then shows

$$N^{1/D} \leq c_0(3e/\epsilon) \log(3e/\epsilon) \quad \text{where } c_0 := (1 - e^{-1})^{-1}.$$

The stated bound merely tidies up some constants.

Remark. The Lemma also gives bounds for $L^p$ packing numbers, via

$$\text{pack}(\epsilon, D, L^p(P)) = \text{pack}(\epsilon^p, D, L^1(P)),$$

because $\|D - D'\|_p = \|D - D'\|_1^{1/p}$ for each $p \geq 1.$

Draft: 26July13 ©David Pollard
For Definition <2> a VC-class is a set $\mathcal{D}$ of subsets of some set $\mathcal{X}$. The defining property involves subsets $D \cap F$, with $D \in \mathcal{D}$, picked out from a finite subset $F$ of $\mathcal{X}$. In classical empirical process applications, $F$ is often generated as a sample from some probability distribution $P$ on $\mathcal{X}$.

If $P$ has atoms then a random sample from $P$ can involve ties, points of $\mathcal{X}$ that are selected more than once in the sample. As you saw in Example <1>, ties cause awkward (but minor) notational difficulties. Those difficulties can be avoided by thinking of $x_1, \ldots, x_n$ as the coordinates of a single vector $x$ in $\mathcal{X}^n$, rather than as members of a subset $F$ of $\mathcal{X}$. Instead of subsets picked out from $F$ by $D$ we have a set of “patterns”

$$D_x := \{D_x : D \in \mathcal{D}\} \subseteq \{0, 1\}^n \subset \mathbb{R}^n$$

where $D_x := \{\{x_1 \in D\}, \ldots, \{x_n \in D\}\}$. That is, for each $D$ in $\mathcal{D}$ the indicators $\{x_i \in D\}$ define the coordinates of an $n$-vector $D_x$ of 0’s and 1’s.

**Remark.** If $\#D_x = N$ then $\mathcal{D}$ picks out exactly $N$ distinct patterns from $x$. If the vector $x$ contains repeated coordinates then $N$ cannot equal $2^n$. For example, if $x_1 = x_2$ then, for each $D$, the first two coordinates of $D_x$ are the same: $x_1 \in D$ if and only if $x_2 \in D$.

The sets $D_x$ for various $x$ capture all the useful properties of VC-classes. Subpatterns are given by projections: If $I$ is a subset of $\{1, \ldots, n\}$, the pattern picked out by $D$ from the subsequence $(x_i : i \in I)$ is equal to the projection $\pi_I D_x$ of $D_x$ onto the coordinate subspace $\mathbb{R}^I$.

**Definition.** Say that a subset $\mathcal{V}$ of $\{0, 1\}^n$ shatters a subset $I$ of $\{1, \ldots, n\}$ if the coordinate projection $\pi_I \mathcal{V}$ onto $\mathbb{R}^I$ equals $\{0, 1\}^I$. Define the **shatter dimension** $\text{sdim}(\mathcal{V})$ of $\mathcal{V}$ to be the largest $\#I$ for which $\mathcal{V}$ shatters $I$.

As a small exercise in notation, you should convince yourself that

$$\text{VCdim}(\mathcal{D}) = \max\{\text{sdim}(D_x) : x \in \mathcal{X}^n, n \in \mathbb{N}\}.$$
is a random variable. As Vapnik and Červonenkis (1971, Theorem 4) showed, the uniform (over \( D \)) law of large numbers holds for that \( P \) if and only if \( n^{-1} \log \text{sdim}(D_x) \) converges in \( P \) probability to zero.

The \( \text{sdim} \) approach also fits well with the stochastic process setting where \( \{X_t : t \in T\} \) is an \( \mathbb{R}^n \)-valued stochastic process. As you saw in Chapter 8, the random set

\[
X_\omega := \{X(\omega, t) : t \in T\} \subset \mathbb{R}^n \quad \text{for each fixed } \omega
\]

indexes an important stochastic process, \( \{s \cdot z : z \in X_\omega\} \), with \( s \) a vector of independent Rademacher variables. The process has subgaussian increments controlled by Euclidean (\( \ell^2 \)) distance. If each coordinate \( X_i(\omega, t) \) takes values in \( \{0, 1\} \) then \( \text{sdim}(X_\omega) \) can be used to bound the \( \ell^2 \) packing numbers for \( X_\omega \). See Chapter 10 for the more interesting case where the range is not restricted to \( \{0, 1\} \).

You might be wondering why I even bothered discussing VC-classes of sets when the most important properties involve the shatter dimension of subsets of \( \{0, 1\}^n \). In fact, the shatter dimension is actually a special case of VC-dimension. In a formal mathematical sense, if \( \mathcal{V} \subseteq \{0, 1\}^n \) then each \( v \) in \( \mathcal{V} \) is a function from the set \( \mathcal{X}^* = \{1, \ldots, n\} \) to \( \{0, 1\} \); and each such function can be identified with a subset of \( \mathcal{X}^* \). That is, we could think of \( \mathcal{V} \) as a set of subsets of \( \mathcal{X}^* \). As another small exercise, you should now convince yourself that \( \text{VCdim}(\mathcal{V}) \), with \( \mathcal{V} \) regarded as set of subsets of \( \{1, \ldots, n\} \), is the same as \( \text{sdim}(\mathcal{V}) \), with \( \mathcal{V} \) regarded as a subset of \( \{0, 1\}^n \).

For some arguments I find it is easier to visualize what is going on when dealing with finite subsets \( \mathcal{V} \) (of size \( N \)) of \( \{0, 1\}^n \) if I write each vector in \( \mathcal{V} \) as a column of an \( n \times N \) binary matrix \( V \), a matrix with distinct columns whose elements all belong to the “binary” set \( \{0, 1\} \).

For dealing with submatrices there is a convenient notation that should be familiar to anyone used to working with programming languages that treat vectors and matrices as basic objects.

(i) For each pair of integers \( p, q \) with \( p \leq q \) write \([p : q]\) for the set of integers \( \{i \in \mathbb{Z} : p \leq i \leq q\} \). For example, \([1 : n]\) = \( \{1, \ldots, n\} \).

(ii) If \( V \) is an \( n \times N \) matrix, and if \( I \subseteq [1 : n] \) with \( \#I = r \) and \( J \subseteq [1 : N] \) with \( \#J = c \), define \( V[I, J] \) as the \( r \times c \) matrix obtained by discarding from \( V \) all rows except those in \( I \) and all columns except those in \( J \). More precisely, if \( I \) is enumerated as \((i_1, \ldots, i_r)\) and \( J \) as \((j_1, \ldots, j_c)\) then the \((\alpha, \beta)\)th element of \( V[I, J] \) is \( V[i_\alpha, j_\beta] \), for \( 1 \leq \alpha \leq r \) and \( 1 \leq \beta \leq c \).
(iii) Abbreviate \( V[I, [1 : N]] \) to \( V[I, \cdot] \), the submatrix with only the rows in \([1 : n]\backslash I\) removed. Interpret \( V[\cdot, J] \) similarly.

(iv) For each \( I \subset [1 : n] \) write \( V[\cdot, I] \) for \( V[[1 : n] \backslash I, \cdot] \). In particular write \( V[\cdot, i] \) for \( V[[1 : n] \backslash \{i\}, \cdot] \). And so on.

**Remark.** The submatrix notation also makes sense if \( I = (i_1, \ldots, i_r) \) and \( J = (j_1, \ldots, j_c) \) are sequences with possible repeats, or if the elements are not enumerated in increasing order. The matrix \( V[I, J] \) will then have some rows or columns repeated and rows or columns in orders different from \( V[I, J] \). This extension will prove convenient when \( I \) is chosen as a random sample with repeats from \([1 : n]\). For example:

\[
V = \begin{bmatrix}
1 & 2 & 3 & 4 & 5 & 6 \\
7 & 8 & 9 & 10 & 11 & 12 \\
13 & 14 & 15 & 16 & 17 & 18 \\
19 & 20 & 21 & 22 & 23 & 24
\end{bmatrix}
\]

if

\[
V[(1, 1, 3), (2, 1, 4)] = \begin{bmatrix} 2 & 1 & 4 \\ 2 & 1 & 4 \\ 14 & 13 & 16 \end{bmatrix}
\]

and

\[
V[-2, -(1, 3, 5)] = \begin{bmatrix} 2 & 4 & 6 \\ 14 & 16 & 18 \\ 20 & 22 & 24 \end{bmatrix}
\]

Notice the subtle difference between the \(3 \times 3\) matrix \(V[(1, 1, 3), (2, 1, 4)]\) and—if we follow the usual mathematical convention that \([1, 1, 3]\) is a set with only two members—the \(2 \times 3\) matrix \(V[[1, 1, 3], [2, 1, 4]]\). To avoid such mathematical quicksand, and to be explicit about the row or column order, I typically use \((\ldots)\) when extracting submatrices, even for cases where \((\ldots)\) would have the same effect.

**Definition.** If \( V \) is an \( n \times N \) binary matrix and \( I \subset [1 : n] \), say that \( V \) shatters the set of rows \( I \) if the submatrix \( V[I, \cdot] \) contains all \(2^{|I}|\) elements of \(\{0, 1\}^{|I|}\) amongst its columns. Define the **shatter dimension** of \( V \), denoted by \(\text{sdim}(V)\), as the largest \(|I|\) for which \( V \) shatters \( I \).

**Remark.** For a binary matrix \( V \) with each column corresponding to a unique member of a set \( \mathcal{V} \subset \{0, 1\}^n \), the shatter dimension is also equal to the largest \(|I|\) for which the \(\pi_I V = \{0, 1\}^I\). Thus \(\text{sdim}(V) = \text{sdim}(\mathcal{V})\). The columns of \( V \) are unique but the columns of the submatrix \(V[I, \cdot]\) need not be unique; some vectors in \( V \) might project onto the same vector in \(\{0, 1\}^I\).
Example. The matrix

\[ V = \begin{bmatrix}
1 & 0 & 1 & 0 & 0 & 1 \\
0 & 1 & 0 & 1 & 0 & 0 \\
0 & 1 & 0 & 0 & 1 & 1 \\
0 & 0 & 1 & 1 & 1 & 1
\end{bmatrix} \]

does not shatter rows \{1, 2\} because the column vector \([1, 1]\) does not appear as a column of the \(6 \times 2\) submatrix \(V[(1, 2),]\), the first two rows of \(V\). Each of the other five subsets of two rows is shattered. For example,

\[ V[(2, 3), (1, 2, 4, 5)] = \begin{bmatrix}
0 & 1 & 1 & 0 \\
0 & 1 & 0 & 1
\end{bmatrix} \]

has distinct columns, the four elements of \(\{0, 1\}^2\). No subset of three (or four) rows can be shattered because \(V\) has fewer than \(2^3\) columns. Each singleton \(\{1\}, \{2\}, \{3\},\) and \(\{4\}\) is shattered, because each row contains at least one 0 and one 1. (Easier: every nonempty subset of a shattered \(I\) is also shattered.)

The matrix \(V\) has shatter dimension 2. Note that the number of columns in \(V\) is strictly larger than \(\binom{4}{0} + \binom{4}{1}\), which Theorem <14> shows is enough to force \(\text{sdim}(V) > 1\).

Example. Let \(Q\) be the set of all closed quadrants with a north-east vertex in \(\mathbb{R}^2\), that is, sets of the form \(\{(x, y) \in \mathbb{R}^2 : x \leq a, y \leq b\}\) with \(a, b \in \mathbb{R}\). You should check that \(\text{VCdim}(Q) = 2\). (Hint: If a quadrant \(Q\) contains the northern-most point and the eastern-most point of a finite set \(F\), why must \(Q\) contain all of \(F\)?)

For the three points \(x_1, x_2, x_3\) of \(\mathbb{R}^2\) shown in the picture, the set \(Q_x\) consists of five vertices of the discrete cube \(\{0, 1\}^3\) and \(\text{sdim}(V_x) = 2\). The columns of \(V[(1, 2), \cdot]\) include all four elements of \(\{0, 1\}^2\), with one duplicate because \((1, 1, 1)\) and \((1, 1, 0)\) both project to \((1, 1)\).
If point $x_3$ were removed, leaving $x = (x_1, x_2)$, then

$$V_x = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 \end{bmatrix} \quad \text{with } \text{sdim}(V_x) = 2.$$  

If instead of the three $x_i$'s in the picture the point $x_2$ were moved to coincide with the $x_3$ then

$$V_x = \begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{with } \text{sdim}(V_x) = 1.$$  

Note that the second and third rows of this $V_x$ are the same because $x_2 = x_3$.

\[\square\]

### 9.4 A combinatorial inequality

The following Lemma—sometimes called the VC Lemma and sometimes called Sauer’s Lemma—plays a key role in the VC theory. See the Notes for a brief account of its complicated history.

**Theorem.** Let $V$ be an $m \times N$ binary matrix with all columns distinct. If

$$N > \beta(m,d) := \binom{m}{0} + \binom{m}{1} + \cdots + \binom{m}{d}$$

then $\text{sdim}(V) > d$.

**Corollary.** If $\text{sdim}(V) \leq d$ then $N \leq \beta(m,d)$.

**Proof (of the Theorem).** Define the **downshift** for the $i$th row of the matrix as the operation:

- for $j \in [1 : N]$,
  - if $V[i,j] = 1$ change it to a 0
  - unless the resulting column already appears in $V$.

**Remark.** The binary matrix $V$ defines a subset $\mathcal{V}$ of the discrete hypercube $\{0,1\}^n$. The downshifting of $V$ corresponds to a movement of the points of $\mathcal{V}$ along the edges of the hypercube.
For example, the downshift for the 1st row of the matrix
\[
V = \begin{bmatrix}
1 & 0 & 1 & 0 & 0 & 1 \\
0 & 1 & 0 & 1 & 0 & 0 \\
0 & 1 & 0 & 0 & 1 & 1 \\
0 & 0 & 1 & 1 & 1 & 1 \\
\end{bmatrix}
\]
generates
\[
V_1 = \begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 1 \\
0 & 1 & 0 & 1 & 0 & 0 \\
0 & 1 & 0 & 0 & 1 & 1 \\
0 & 0 & 1 & 1 & 1 & 1 \\
\end{bmatrix}
\].
The 1 in the last column was blocked (prevented from changing to a 0) by the fifth column of \(V\); if it had been changed to a 0 the column \([0,0,1,1]\) would have appeared twice in \(V_1\).

Continuing from \(V_1\), downshift on any other row to generate a new matrix \(V_2\). And so on. The order in which the rows are selected is unimportant. Stop when no more changes can be made by downshifting.

**Remark.** It is plausible that the downshift of a particular 1 in row \(i\) that is initially blocked by some column might succeed at a later stage, because the blocking column might itself be changed between the first and second downshift operations on row \(i\). In fact, as Problem [5] shows, unblocking cannot occur. The whole process actually requires only a single downshift attempt for each row of the matrix. (I learned that fact from Dana Yang and Sören Künzel.)

Such possibilities have no effect on the argument that follows. It does no harm to imagine that the process requires some arbitrary, but finite, number of downshift steps.

For example, a downshift on the 2nd row of
\[
V_1 = \begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 1 \\
0 & 1 & 0 & 1 & 0 & 0 \\
0 & 1 & 0 & 0 & 1 & 1 \\
0 & 0 & 1 & 1 & 1 & 1 \\
\end{bmatrix}
\]
generates
\[
V_2 = \begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 1 & 0 & 0 \\
0 & 1 & 0 & 0 & 1 & 1 \\
0 & 0 & 1 & 1 & 1 & 1 \\
\end{bmatrix}
\],
the fourth column being blocked from changing by the third column of \(V_1\).

A downshift on the 4th row followed by a downshift on the 3rd row then complete the process:
\[
V_3 = \begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 1 & 1 \\
0 & 0 & 1 & 0 & 1 & 0 \\
\end{bmatrix}
\]
and
\[
V_4 = \begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 1 & 0 & 0 \\
0 & 1 & 0 & 0 & 1 & 0 \\
0 & 0 & 1 & 0 & 1 & 0 \\
\end{bmatrix}
\].
No more changes can be generated by downshifts; every possible change of a 1 to a 0 in \(V_4\) is blocked by some column already present. Notice that rows \(\{3,4\}\) are shattered by \(V_4\).

In general, the downshifting operation has two important properties:
(i) No new shattered sets of rows can be created by a downshift.

(ii) When no more downshifting is possible, the final matrix, $V^*$, is hereditary. That is, if $v_1$ is a column of $V^*$ and if $v_2$ is obtained from $v_1$ by changing some of its 1’s to 0’s then $v_2$ is also a column of $V^*$.

With these two properties in hand the rest of the proof then becomes a simple pigeonhole argument. Let me first establish (i) and then (ii), and then deal with the pigeons.

For simplicity of notation, consider a downshift for the 1st row of a matrix $V$, which generates a matrix $V_1$. Suppose $V_1$ shatters a set of rows $I$. If $1 \not\in I$ the submatrix $V_1[I, \cdot]$ is the same as $V[I, \cdot]$, so $V$ also shatters $I$. For the other case, suppose (again for notational simplicity) that $1 \in I = [1 : k]$. For each $x$ in $\{0, 1\}^{k-1}$ the vector $[1, x]$ appears as column of $V_1[I, \cdot]$ so that, for some $w \in \{0, 1\}^{m-k}$, the vector $v = [1, x, w]$ is a column of $V_1$. As the downshift creates no new 1’s, the vector $v$ must also have been a column of $V$ and have been blocked from changing by a column $[0, x, w]$ of $V$. Thus both $[0, x]$ and $[1, x]$ are columns of $V[I, \cdot]$. The matrix $V$ must also have shattered $I$.

For (ii), suppose $v_1$ is a column of $V^*$. If $v_1[i] = 1$ the vector $v_2$ obtained by changing that $i$th element to a 0 must also be a column of $V^*$, for otherwise a downshift on row $i$ would change the $v_1$ column to $v_2$. And so on.

The rest of the proof for the Theorem is easy. Remember that $V^*$ must have all $N$ of its columns distinct. At most $\binom{m}{k}$ of those columns can contain exactly $k$ ones and $m - k$ zeros. Consequently, at most

$$\binom{m}{0} + \binom{m}{1} + \cdots + \binom{m}{d} = \beta(m, d)$$

of the columns can contain $d$ or fewer ones. As $N > \beta(m, d)$, there must be some column $v_0$ of $V^*$ containing at least $d + 1$ ones, in rows $I_0$ for some $I_0 \subseteq [1 : m]$ with $\#I_0 \geq d + 1$. By the hereditary property, the matrix $V^*$ must shatter $I_0$. By (i), the original matrix $V$ must also shatter $I_0$, so that $\text{sdim}(V) \geq \#I_0 > d$. \[\square\]

### 9.5 Patterns as subgraphs of the discrete cube

The results from this Section are needed in Section 9.6 to establish an elegant refinement of Lemma <9>. I feel it is helpful to separate out the
downshifting part of the argument because it reveals yet another way of thinking about VC-classes.

Every nonempty subset $V$ of the discrete hypercube $\{0, 1\}^n$ has a natural interpretation as the vertex set of a graph with edge set $E$ defined by pairs at Hamming distance equal to one: the pair of vertices $\{v_1, v_2\}$ form an (undirected) edge if and only if

\[
\text{hamm}(v_1, v_2) := \sum_{i \leq n} \{v_1[i] \neq v_2[i]\} = \sum_{i \leq n} |v_1[i] - v_2[i]| = 1.
\]

That is, endpoints of an edge must differ in exactly one coordinate position.

**Theorem.** For a graph with vertex set $V \subseteq \{0, 1\}^n$ and edge set $E$ defined by $\langle 16 \rangle$, $\#E/\#V \leq \text{sdim}(V)$.

**Proof.** Write $V$ for the $n \times N$ binary matrix that corresponds to the set $V$. Identify the $N$ vertices with the column numbers: vertex 1 is $V[\cdot, 1]$, and so on. Also decompose the set of edges $E = E(V)$ into disjoint subsets according to the single coordinate at which the endpoints disagree: that is, $\{j_1, j_2\} \in E_i(V)$ if and only if

\[
\text{hamm}(V[\cdot, j_1], V[\cdot, j_2]) = 1 = |V[i, j_1] - V[i, j_2]|
\]

Repeat the downshifting procedure used for proving Theorem $\langle 14 \rangle$, leading to an hereditary binary matrix $V^*$ when no more downshifting changes are possible. The important new idea is that a downshift cannot decrease the number of edges. To understand why, consider a downshift of row 1 (for example) of $V$ that generates a new binary matrix $V_1$. To prove that $\#E(V) \leq \#E(V_1)$ it suffices to show that $\#E_i(V) \leq \#E_i(V_1)$ for every $i$.

For $i = 1$ the inequality is actually an equality because $E_1(V) = E_1(V_1)$. Indeed, if $\{j_1, j_2\} \in E_1(V)$ then, without loss of generality,

\[
V[\cdot, (j_1, j_2)] = B := \begin{bmatrix} 1 & 0 \\ w & w \end{bmatrix}
\]

for some $w \in \{0, 1\}^{n-1}$.

Column $V[\cdot, j_1]$ cannot change because it is blocked by column $V[\cdot, j_2]$. Conversely, if $V_1[\cdot, (j_1, j_2)] = B$ then $[1, w]$ must have been a column of $V$ blocked by $[0, w]$.

The argument gets more exciting for $i \neq 1$. It suffices to construct a one-to-one map $\psi_i$ from $\text{OLD}_i := E_i(V) \setminus E_i(V_1)$ into $\text{NEW}_i := E_i(V_1) \setminus E_i(V)$ for each $i \geq 2$. Consider an edge $e = \{j_1, j_2\}$ in $\text{OLD}_i$. That is, $v_1 = V[\cdot, j_1]$ and $v_2 = V[\cdot, j_2]$ are the endpoints of $e$ in $\text{OLD}_i$.\[Draft: 26july13 ©David Pollard]
Write \( K \) for \([1 : n]\) \( \setminus \{1, i\} \). I assert that the downshift must change exactly one of \( v_1 \) and \( v_2 \), for otherwise \( e \) would also be an edge in \( \mathcal{E}_i(V_1) \). For concreteness, suppose \( v_1 \) is changed. We must have \( V[1, j_1] = V[1, j_2] = 1 = V_1[1, j_1] \) and \( V_1[1, j_2] = 0 \). The possible change to \( v_2 \) must have been blocked by some column \( V_{·, j} \) that agrees with \( V_{·, j_2} \) except for a zero in the first position. The relevant part of \( V \) must look like

\[
V[(1, i, K), (j_1, j_2, j)] = \begin{bmatrix}
1 & 1 & 0 \\
a & b & b \\
w & w & w
\end{bmatrix}
\text{ with } a \neq b \text{ and } w \in \{0, 1\}^K,
\]

which is downshifted to

\[
V_1[(1, i, K), (j_1, j_2, j)] = \begin{bmatrix}
0 & 1 & 0 \\
a & b & b \\
w & w & w
\end{bmatrix}.
\]

The edge \( e = \{j_1, j_2\} \) in \texttt{OLD}_i has been destroyed but has been replaced by an edge \( \{j_1, j\} \) in \texttt{NEW}_i. The new edge is uniquely determined by \( v_1 \) and \( v_2 \). Define \( \psi(\{j_1, j_2\}) = \{j_1, j\} \).

After enough downshifts we reach the conclusion that \( \#\mathcal{E}(V) \leq \#\mathcal{E}(V^*) \).

By definition of edges, for each \( e = \{j_1, j_2\} \in \mathcal{E}(V^*) \) we must have

\[
\sum_{i \leq n} (V^*[i, j_1] - V^*[i, j_2]) = \pm 1.
\]

Orient \( e \) from the vertex with fewer 1’s to the vertex with more 1’s. That is, \( e \) becomes the directed edge \( (j_1, j_2) \) if the last sum equals \(-1\) and \((j_2, j_1)\) if it equals \(+1\).

As you saw in the proof of Theorem \textless 14\textgreater, for every \( k \) at most \( \mathbb{D} = \text{sdim}(V) \) of the coordinates of \( v = V^*[·, k] \) can equal 1. The vertices \( w = V^*[·, k'] \) for which \((w, v)\) is a directed edge of the \( V^* \) graph are all obtained by changing exactly one of those 1’s to a 0. The indegree of \( v \) must be \( \leq \mathbb{D} \).

The number of edges in the (directed) graph is obtained by summing the indegrees for all vertices, which gives the bound \( N\mathbb{D} \geq \#\mathcal{E}(V^*) \geq \#\mathcal{E}(V) \). \qed

In the previous proof, the graph with vertex set determined by the columns of the hereditary \( V^* \) was oriented to produce a directed graph with the indegree of every vertex bounded by \( \text{sdim}(V) \). The proof in the next Section requires a similar ordering for the graph on the original vertex set, which can be achieved by an appeal to the classical Marriage Lemma (UGMTP Problem 10.5). Recall the statement of that Lemma. For each \( \sigma \)
§9.6 Haussler’s improvement of the packing bound

In a finite set $S$ we are given a nonempty subset $K(σ)$ of some finite set $K$. We seek a function $ψ : S → K$ for which $ψ(σ) ∈ K(σ)$ for every $σ$ in $S$. The Lemma asserts that such a function exists if and only if

$$\#(∪_{σ ∈ A} K(σ)) ≥ \#A$$

for every nonempty subset $A$ of $S$.

**Corollary.** (Haussler, 1995 following Alon and Tarsi, 1992) For $V$ and $E$ as in Theorem §9.5, there exists an orientation of the edges for which the indegree of every vertex is at most $sdim(V)$.

**Proof** Once again identify the members of $V$ with the columns of an $n × N$ binary matrix $V$ with distinct columns and label the vertices in $V$ with the corresponding column numbers of $V$. For each nonempty subset $J$ of $[1 : N]$ the subset $V_J$ of $\{0, 1\}^n$ corresponding to the columns of $V[·, J]$ has $sdim(V_J) ≤ D : = sdim(V)$. By the Theorem, the corresponding edge set $E_J : = E(V_J)$ satisfies $\#E_J/\#V_J ≤ D$.

Define $S = E(V)$ and $X = [1 : D] × V$. (That is, $X$ consists of $D$ “copies” of each vertex in $V$.) For each $e = \{j_1, j_2\}$ in $S$ define

$$K(e) : = [1 : D] × \{j_1, j_2\}.$$ 

For a nonempty subset $A$ of $S$ define $J$ to be the sets of vertices that appear as an endpoint of at least one edge from $A$. Note that $A ⊆ E_J$. Then

$$∪_{e ∈ A} K(e) = [1 : D] × J,$$

a set with cardinality $D × \#J = D × \#V_J ≥ \#E_J ≥ \#A$.

The Marriage Lemma gives a map $ψ : E(V) → X$ for which $ψ(e) = (i, j)$ with $1 ≤ i ≤ D$ and $j$ one of the two vertices that define edge $e$. If $e = \{j', j\}$ then orient it as $(j', j)$. For each vertex $j$ there are at most $D$ edges $e$ for which $ψ(e) = (i, j)$ with $1 ≤ i ≤ D$. Vertex $j$ has indegree at most $D$.

□

*9.6 Haussler’s improvement of the packing bound

By means of a more subtle randomization argument, it is possible to eliminate the $\log(3e/ε)$ factor from the bound in Theorem §9.5. The improvement is due to Haussler (1995).
§9.6  Haussler’s Improvement of the Packing Bound

Throughout the Section, \( \| \cdot \|_1 \) denotes the \( \ell^1 \) norm on \( \mathbb{R}^n \), that is, \( \| v \|_1 = \sum_{i \leq n} |v[i]| \). For vectors \( v_1 \) and \( v_2 \) in \( \{0, 1\}^n \) the \( \ell^1 \) distance \( \| v_1 - v_2 \|_1 \) coincides with Hamming distance.

**Theorem.** Let \( \mathcal{V} = \{v_1, \ldots, v_N\} \) be a subset of \( \{0, 1\}^n \) for which

\[ \| v_j - v_{j'} \|_1 \geq n\epsilon \quad \text{for all } 1 \leq j < j' \leq N, \text{ for some } 0 < \epsilon \leq 1. \]

If \( \text{sdim}(\mathcal{V}) \leq D \) then \( N \leq e(1 + D)(2e/\epsilon)^D \).

**Proof** Let \( X \) be a random vector distributed uniformly on \( \mathcal{V} \). As usual, write \( X_1, \ldots, X_n \) for the coordinates, each of which takes values in \( \{0, 1\} \).

The proof depends on two inequalities involving conditional variances, for nonempty, proper subsets \( I \) and \( K \) of \([1 : n] \):

\[ \sum_{i \in I} \mathbb{P} \text{var}(X_i \mid X_{I \setminus \{i\}}) \leq D \]

and

\[ \sum_{i \notin K} \mathbb{P} \text{var}(X_i \mid X_K) \geq \frac{1}{2} n\epsilon \left( 1 - \frac{\text{\#}_K}{\text{\#}_I} \right) \]

with \( \text{\#}_K := \pi_K \mathcal{V} \)

\[ \geq \frac{1}{2} n\epsilon \left( 1 - \beta(\text{\#}_K, D)/N \right), \]

the final inequality by Corollary <15>. Inequality <21> depends on Corollary <18>; inequality <20> depends on the assumed separation of the vectors in \( \mathcal{V} \). The proofs of both are given as Corollaries to Lemma <24> at the end of this Section.

Let \( M \) equal \( \lceil 2(D + 1)/\epsilon \rceil - 1 \). (The reason for this choice becomes evident at the end of the proof.) Without loss of generality assume that \( M < n \). (Soon \( n \) is replaced by \( \ell n \) for an arbitrarily large integer \( \ell \).) Sum inequality <20> over all possible subsets \( I \) of \([1 : n] \) with \( \#I = M + 1 \) to get

\[ \sum_{\#I = M + 1} \sum_{i \in I} \mathbb{P} \text{var}(X_i \mid X_{I \setminus \{i\}}) \leq \left( \frac{n}{M + 1} \right)^D. \]

Equivalently,

\[ \sum_{\#K = M} \sum_{i \notin K} \mathbb{P} \text{var}(X_i \mid X_K) \leq \left( \frac{n}{M + 1} \right)^D. \]

(To see the equivalence, write \( K \) for \( I \setminus \{i\} \). Every \((i, K)\) pair with \( i \notin K \) and \( \#K = M \) appears exactly once in the first double sum. As a check,
§9.6 Haussler’s improvement of the packing bound

note that both double sums involve \((M + 1)\binom{n}{M + 1} = (n - M)\binom{n}{M}\) terms.

Similarly, a summation over all subsets \(K\) of \([1 : n]\) of size \(M\) in \(<21>\) gives

\[
\sum_{\#K = M} \sum_{i \notin K} \mathbb{P} \text{var}(X_i \mid X_K) \geq \binom{n}{M} \frac{1}{2n} \epsilon (1 - \beta(M, \mathbb{D})/N).
\]

Together \(<22>\) and \(<23>\) imply

\[
\frac{\beta(M, \mathbb{D})}{N} \geq 1 - \frac{2(n - M)}{n \epsilon (M + 1)} \mathbb{D}
\]

or

\[
N \leq \beta(M, \mathbb{D}) \left( 1 - \frac{2(n - M)}{n \epsilon (M + 1)} \mathbb{D} \right)^{-1}.
\]

I could try to optimize over \(M\) immediately, as Haussler (1995, page 225) did, to bound \(N\) by a function of \(\epsilon, \mathbb{D},\) 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 \(\ell n\), for some positive integer \(\ell\), and each \(v_j\) in \(V\) is replaced by the concatenation of \(\ell\) copies of \(v_j\). (That is, if \(V\) is represented by the \(n \times N\) binary matrix \(V\), the new vectors are the columns of the \((\ell n) \times N\) binary matrix obtained by stacking \(\ell\) copies of \(V\).) Letting \(\ell\) tend to infinity (with \(M\) fixed) eliminates \(n\) from the bound, leaving

\[
N \leq \beta(M, \mathbb{D}) \left( 1 - \frac{2 \mathbb{D}}{\epsilon (M + 1)} \right)^{-1}
\]

The first factor is an increasing function of \(M\), the second a decreasing function. If \(M\) is chosen so that \(M + 1 \geq 2(\mathbb{D} + 1)/\epsilon \geq M\) then \(M \geq \mathbb{D}\) and the upper bound for \(N\) is less than

\[
\left( \frac{eM}{\mathbb{D}} \right)^\mathbb{D} \frac{\epsilon (M + 1)}{\epsilon (M + 1) - 2 \mathbb{D}} \leq \left( \frac{2e(\mathbb{D} + 1)}{\epsilon \mathbb{D}} \right)^\mathbb{D} \frac{2(\mathbb{D} + 1)}{2(\mathbb{D} + 1) - 2 \mathbb{D}} \leq (2e/\epsilon)^\mathbb{D} (1 + \mathbb{D}^{-1})^\mathbb{D} (\mathbb{D} + 1),
\]

which is smaller than the bound stated in the Theorem.

\[\square\]

**Remark.** My \(M\) corresponds to \(m - 1\) in Haussler’s paper. To see the reason for the choice of \(M\), note that the function \(x^\mathbb{D} (1 - (2\mathbb{D})/(\epsilon x))^{-1}\) is minimimized over \(x\) in \(\mathbb{R}^+\) by \(x = 2(\mathbb{D} + 1)/\epsilon\).
It remains only to justify inequalities \(<20\>) and \(<21\>). To keep the notation simple, I first prove a general result then deduce the inequalities as Corollaries.

**Lemma.** Suppose \(W\) is a subset of \(\{0,1\}^R\) of size \(L\). Suppose also that \(Y\) is a random \(R\)-vector distributed according to some distribution \(P\) on \(W\). Then

(i) \(\sum_{i \leq R} \mathbb{E} \text{Var}(Y_i | Y_{-i}) \leq \text{sdim}(W)\).

(ii) If \(P\) is the uniform distribution on \(W\) and the elements are \(\delta\)-separated,

\[ \|w_j - w_{j'}\|_1 \geq \delta \quad \text{for } 1 \leq j < j' \leq L, \]

for some \(\delta > 0\) then \(\sum_{i \leq R} \text{Var}(Y_i) \geq \frac{1}{2} \delta (1 - L^{-1})\).

**Proof** For (i) let \(E\) denote the edge set, where \(e = \{w, w'\}\) is an edge if and only if \(\|w - w'\|_1 = 1\). Write \(D\) for \(\text{sdim}(W)\). By Lemma \(<18\>\), the edges can be directed in such a way that every vertex in \(W\) has indegree at most \(D\). Put another way,

\[ \sum_{e \in E} \{\text{head}(e) = w\} \leq D \quad \text{for each } w \in W. \]

Consider the case \(i = 1\). The set \(W_{-1} := \pi_{-1} W\) can be partitioned into two subsets: the set \(W^1_{-1}\) of those \(x\)'s for which there is a unique \(w \in W\) that projects onto \(x\), and the set \(W^2_{-1}\) of those \(x\)'s for which there is an edge \(e = \{[0,x],[1,x]\}\) in \(E_1(W)\). If \(x \in W^1_{-1}\) then the distribution of \(Y\) given \(Y_{-1} = x\) is degenerate at a single point and \(\text{var}(Y_1 | Y_{-1} = x) = 0\). If \(x \in W^2_{-1}\), the conditional distribution of \(Y_1\) given \(Y_{-1} = x\) is \(\text{Ber}(p_1(x))\) where \(p_1(x) := P\{[1,x]\}/(P e)\) and

\[ \text{var}(Y_1 | Y_{-1} = x) = p_1(x)[1 - p_1(x)] = P\{\text{tail}(e)\}P\{\text{head}(e)\}/(Pe)^2. \]

The analysis for other values of \(i\) in \([1: R]\) is similar. Thus

\[ \sum_{i = 1}^R \mathbb{E} \text{Var}(Y_i | Y_{-i}) = \sum_{i = 1}^R \sum_{e \in E_i(W)} Pe \frac{P\{\text{tail}(e)\}P\{\text{head}(e)\}}{(Pe)^2} \leq \sum_{e \in E} P\{\text{head}(e)\} = \sum_{w \in W} \sum_{e \in E} \{\text{head}(e) = w\}P\{w\}. \]

As \(e\) ranges over all edges, \(\text{head}(e)\) visits each vertex at most \(D\) times. The last sum is at most \(D \sum_{w \in W} P\{w\} = D\).
§9.7 Problems

Let (ii) let \( Y' = (Y'_i : 1 \leq i \leq R) \) be chosen independently of \( Y \), with the same uniform distribution. Note that the difference \( Y_i - Y'_i \) takes the values \( \pm 1 \) when \( Y_i \neq Y'_i \) and is otherwise zero. Thus \( (Y_i - Y'_i)^2 = |Y_i - Y'_i| \) and

\[
\sum_{i \leq R} \text{var}(Y_i) = \sum_{i} \frac{1}{2} \mathbb{P} |Y_i - Y'_i|^2 \quad \text{by independence}
\]

\[
= \frac{1}{2} \mathbb{P} \|Y - Y'\|_1
\]

\[
\geq \frac{1}{2} \delta \mathbb{P} \{ Y \neq Y' \} = \frac{1}{2} \delta (1 - L^{-1})
\]

\[
\square
\]

Corollary. (Inequality <20>) \( \sum_{i \in I} \mathbb{P} \text{var}(X_i \mid X_{I \setminus \{i\}}) \leq D \).

Proof Define \( W = \pi_I \mathcal{V} \) and \( Y = X_I \), whose distribution need not be uniform on \( W \).

\[
\square
\]

Corollary. (Inequality <21>)

\[
\sum_{i \notin K} \mathbb{P} \text{var}(X_i \mid X_K) \geq \frac{1}{2} n \epsilon \left( 1 - \frac{N(K, z)}{N(K, z)} \right)
\]

where \( N(K, z) = \# W_z \). Also \( \mathbb{P} \{ X_K = z \} = N(K, z)/N \). Thus

\[
\sum_{i \notin K} \mathbb{P} \text{var}(X_i \mid X_K) \geq \frac{1}{2} n \epsilon \sum_{z \in \mathcal{V}_K} \frac{N(K, z)}{N} \left( 1 - \frac{1}{N(K, z)} \right),
\]

which reduces to the stated expression.

\[
\square
\]

9.7 Problems

[1] 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 \( F = \{0, e_1, \ldots, e_d\} \), where \( e_i \) is the unit vector with +1 in the \( i \)th position. Hint: Consider closed halfspaces of the form \( \{ x : \theta \cdot x \geq r \} \) with \( \theta \in \{-1, +1\}^d \).
§9.7 Problems

(ii) Suppose $F = \{x_i : i = 0, 1, \ldots, d + 1\}$ 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 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 $\sum_{1 \leq i \leq d + 1} \alpha_i (x_i - x_0) = 0$ for some constants $\alpha_i$ that are not all zero. Define $\alpha_0 = -\sum_{1 \leq i \leq d + 1} \alpha_i$. Show that

$$\sum_{0 \leq i \leq d + 1} \alpha_i x_i = 0 \quad \text{and} \quad \sum_{0 \leq i \leq d + 1} \alpha_i = 0.$$  

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 $F$.

[2] Suppose $\mathcal{D}, \mathcal{D}_1, \ldots, \mathcal{D}_k$ are VC-classes of subsets of $\mathcal{X}$.

(i) (Dudley, 1978, Proposition 7.12) Write $A(B_1, \ldots, B_k)$ for the algebra of subsets of $\mathcal{X}$ generated by subsets $B_1, \ldots, B_k$ of $\mathcal{X}$. For each finite $k$, show that $\cup \{A(D_1, \ldots, D_k) = D_1, \ldots, D_k \in \mathcal{D}\}$ is also a VC-class.

(ii) Suppose $\text{VCdim}(\mathcal{D}_i) \leq \mathcal{D}$ for each $i$. Let $\mathcal{B}_k$ denote the class of all sets expressible by means of at most $k$ union, intersection, or complement symbols. Find an increasing, integer-valued function $\beta(k)$ such that the VC-dimension of $\mathcal{B}_k$ is at most $\beta(k)$.

[3] Define $G(x) = x/(\log x)$ for $x \geq 1$, with $G(1) = +\infty$.

(i) Show that $G$ achieves its minimum value of $e$ at $x = e$ and that $G$ is an increasing function on $[e, \infty)$.

(ii) Suppose $y \geq G(x)$ for some $x \geq e$. Show that

$$\log y \geq (1 - 1/G(\log x)) \log x \geq (1 - e^{-1}) \log x.$$  

Deduce that $x = G(x) \log x \leq c_0 y \log y$, where $c_0 := (1 - e^{-1})^{-1} \approx 1.58 < 2$.

[4] Suppose $N$ and $d$ are positive integers for which

$$N \leq \beta(m, d) := \sum_{k=0}^{d} \binom{m}{k} \quad \text{with} \quad m = \lceil \rho \log N \rceil \text{ for some } \rho \geq 1.$$  

Show that, for the constant $c_0$ defined in Problem [3],

$$N^{1/d} \leq c_0 \lambda \log \lambda \leq c_0 \lambda^2 \quad \text{where} \quad \lambda := (1 + \rho) e$$  

by the following steps.
(i) First check that the inequality is trivially true if $N \leq e^d$.
(ii) For $N > e^d$ check that $m \geq d$, which implies $\beta(m, d) \leq (em/d)^d$. Deduce that
\[
N^{1/d} \leq \frac{em}{d} \leq \frac{e(1 + \rho \log N)}{d} \leq e(1 + \rho \log N^{1/d}),
\]
so that $G(N^{1/d}) \leq (1 + \rho)e = \lambda$.

For the downshift argument described in Section 9.4, show that every element that is blocked from changing to 0 during at least one downshift must be blocked for every downshift; it must stay equal to 1 in the final $V^*$. Generalize the following argument.

Consider the form of $V$ after the first attempted downshift on row 1 at step $m$. If $k$ of the 1’s were blocked, the new $V_m$ must contain an $n \times (2k)$ submatrix like
\[
W_m = \begin{bmatrix}
1 & 0 & 1 & 0 & \ldots & 1 & 0 \\
1 & 1 & a_2 & a_2 & \ldots & a_k & a_k \\
w_1 & w_1 & w_2 & w_2 & \ldots & w_k & w_k
\end{bmatrix}
\]
for distinct column vectors $[a_i, w_i] \in \{0, 1\}^{n-1}$. Suppose row 2 is the target for the next downshift. Consider the effect on the first two columns of $W_m$.

If $a_1 = 0$ then neither column can change; the 1 will still be blocked on any subsequent attempt to downshift row 1. A similar conclusion holds if $a_1 = 1$ and both downshifts of the $a_1$ are blocked. Only if $a_1 = 1$ and exactly one of the downshifts succeeds will the future of the 1 be in any doubt. In that case the first two columns of $W_m$ will be changed to either
\[
\begin{bmatrix}
1 & 0 \\
1 & 0 \\
w_1 & w_1
\end{bmatrix}
\quad \text{or} \quad
\begin{bmatrix}
1 & 0 \\
0 & 1 \\
w_1 & w_1
\end{bmatrix}
\]
In the first case the matrix $W_m$ must already have had a column $[1, 0, w_1]$, which must have been blocked by a column $[0, 0, w_1]$ when $V_m$ was created. But that same $[0, 0, w_1]$ should have blocked the change in the second column of $W_m$. Thus the first case cannot occur.

In the second case the matrix $W_m$ must already have had a column $[0, 0, w_1]$, which then will block future attempts to change 1.

In short, the $V_{m+1}$ will again contain a submatrix like $W_m$. And so on.
9.8 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 Červonenkis, 1971). Dudley (1999, page 167) (see also Dudley 1978, Section 7) noted that the 1971 bound was a little weaker than the Sauer result. Inequality <5> is proved on page 154 of the Wapnik and Tscherwonenkis (1979) book, a translation into German of a 1974 Russian edition, which I have not seen.


I learned about the downshift method for proving Theorem <14> from Michel Talagrand. Ledoux and Talagrand (1991, pp. 411-412) used the set-theoretic version of the technique, attributing it (page 420) to Frankl (1983). (See also Talagrand 1988, Proposition 8 for another interesting application of downshifting.) Independently, both Alon (1983) and Frankl (1983) used the shifting method to prove a result that implies Theorem <14>. Frankl (1995, page 1298) noted that the shifting technique was introduced by Erdős et al. (1961).

The results in Section 9.5 come from Haussler (1995), who acknowledged (page 220) Linial for the downshifting method of proof for Theorem <17>, 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 <19> see Haussler (1995, Section 3).

Stengle and Yukich (1989) showed how to obtain VC-classes of subsets of a Euclidean space by taking cross-sections of semi-algebraic subsets (defined by finitely many inequalities for polynomials) of higher dimensional Euclidean spaces. See van den Dries (1998, Chapter 5) for more general results about cross-sections of definable sets in o-minimal structures.

References

§9.8 Notes


