Chapter 12

Combinatorics

SECTION 1 presents a simple concrete example to illustrate the use of the combinatorial method for deriving bounds on packing numbers.
SECTION 2 defines the concept of shatter dimension, for both classes of sets and binary matrices, and the VC dimension for a class of sets.
SECTION 3 derives the general combinatorial bound (the VC lemma) relating the shatter dimension to the number of distinct patterns picked out. The proof of the main theorem introduces the important technique known as downshifting.
SECTION 4 describes ways to create VC classes of sets, including methods based on o-minimality. *incomplete*
SECTION 5 establishes a connection between VC dimension and packing numbers for VC subgraph classes of functions under an L^1 metric.

Unedited fragments from here on

SECTION 6 fat shattering
SECTION 7 Mendelson & V
SECTION 8 Haussler & Long
SECTION 9 presents a refinement of the calculation from Section 5, leading to a sharper bound for packing numbers. The result is included mainly for its aesthetic appeal and for the sake of those who like to see best constants in inequalities.

1. An introductory example

Suppose $F := \{x_1, \dots, x_n\}$ is a finite subset of some $\mathcal{X}$ and $\mathcal{D}$ is some collection of subsets of $\mathcal{X}$. A set D from $\mathcal{D}$ is said to ***pick out*** the points $\{x_i : i \in \mathbb{J}\}$, where $\mathbb{J} \subseteq \{1, 2, \dots, n\}$, if

$$\begin{cases} x_i \in D & \text{for } i \in \mathbb{J} \\ x_i \in D^c & \text{for } i \in \mathbb{J}^c \end{cases}$$

The class $\mathcal{D}$ is said to ***shatter*** F if it can pick out all 2^n possible subsets of F.

The combinatorial argument often referred to as the VC method (in honor of the important contributions of Vapnik & Červonenkis) provides a bound on the number of different subsets that can be picked out from a given F in terms of the size of the largest subset F_0 of F that $\mathcal{D}$ can shatter. The combinatorial bound also leads to uniform bounds on the covering numbers of $\mathcal{D}$ as a subset of $\mathcal{L}^1(P)$, as P ranges over probability measures carried by $\mathcal{X}$. The methods can also be applied to collections of real-valued functions on $\mathcal{X}$, leading to bounds on $\mathcal{L}^\alpha(P)$ covering numbers for a wide range of α values, including $\alpha = 1$ and $\alpha = 2$. These bounds can then be fed into chaining arguments to obtain maximal inequalities for some processes indexed by collections of functions on $\mathcal{X}$.

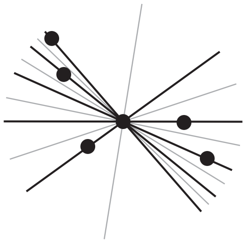
The VC method is elegant and leads to results not easily obtained in other ways. The basic calculations occupy only a few pages. Nevertheless, the ideas are subtle enough to appear beyond the comfortable reach of many would-be users. With that fact in mind, I offer a preliminary more concrete example, in the hope that the idea might seem less mysterious.

> Remark. You will notice that I make no effort to find the best upper bound. In fact, as will become clear later, any bound that grows like a polynomial in the size of the subset leads to the same sort of bound on the covering numbers.

Consider the class $\mathcal{H}_2$ 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 are there of the form $H \cap F$, with H in $\mathcal{H}_2$? Certainly there can be no more than 2^n, because F has only that many subsets.

Define $F_{\mathcal{H}_2} = \{F \cap H : H \in \mathcal{H}_2\}$, and let $\#F_{\mathcal{H}_2}$ be its cardinality. A simple argument will show that $\#F_{\mathcal{H}_2} \le 4n^2$. 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 x_0 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 runs parallel to L_0 through the point of F_0 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 $\mathcal{L}(x_0) \cup \mathcal{L}'(x_0)$ define a collection of at most $4(n-1)$ closed half-spaces, each with x_0 on its boundary. The collection $\cup_{x_0 \in F} \mathcal{H}_2(x)$ accounts for all possible nonempty subsets of F picked out by closed half-spaces. Thus there are at most



$$1 + 4n(n-1) \le 4n^2$$

subsets that can be picked out from F by $\mathcal{H}_2$. The extra 1 takes care of the empty set.

The slow increase in $\#F_{\mathcal{H}_2}$, at an $O(n^2)$ rate rather than a rapid 2^n rate, has an unexpected consequence for the packing numbers of $\mathcal{H}_2$ when equipped with an $\mathcal{L}^1(P)$ (pseudo)metric for some probability measure P on the Borel sigma-field.

> Remark. In fact, we will only need the bound when P concentrates on a finite number of points, in which case all measurability difficulties disappear.

The result is surprising because it makes no regularity assumptions about the probability measure P. The argument is due to Dudley (1978), who created the general theory for abstract empirical processes.

The $\mathcal{L}^1(P)$ distance between two (measurable) sets B and B' is defined as

$$P|B - B'| = P(B \Delta B'),$$

the probability measure of the symmetric difference. Say that the two sets are ϵ-separated if $P(B \Delta B') > \epsilon$. The packing number $\pi_1(\epsilon) := D(\epsilon, \mathcal{H}_2, \mathcal{L}^1(P))$ is defined as the largest N for which there exists a collection of N closed half-spaces, each pair ϵ-separated. We can use the polynomial bound for $\#F_{\mathcal{H}_2}$ to derive an upper bound for the packing numbers, by means of a cunningly chosen F.

Suppose there exist half-spaces $H_1, H_2, \ldots, H_N$ for which $P(H_i \Delta H_j) > \epsilon$ for $i \neq j$. The trick is to find a set F of $m = 2\log N/\epsilon$ points from which each H_i picks out a different subset. (For simplicity, I am ignoring the possibility that this m might not be an integer. A more precise calculation will be given in Section 5.) Then $\mathcal{H}_2$ will pick out at least N subsets from the m points, and thus

$$N \leq 4\left(\frac{2\log N}{\epsilon}\right)^2$$

If we bound $\log N$ by a constant multiple of $N^{1/4}$, then solve the 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.

With a little more work, the bound can even be brought to the form $C_2(\epsilon^{-1}\log(1/\epsilon))^2$, at least for ϵ bounded away from 1. At this stage there is little point in struggling to get the best bound in ϵ. The qualitative consequences of the 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.

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 $i \neq j$,

$$\begin{aligned}
&\mathbb{P}\{H_i \text{ and } H_j \text{ pick out same points from } F_0\} \\
&\quad = \mathbb{P}\{\text{no points of sample in } H_i \Delta H_j\} \\
&\quad = (1 - P(H_i \Delta H_j))^m \\
&\quad \leq (1-\epsilon)^m \\
&\quad \leq \exp(-m\epsilon).
\end{aligned}$$

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

$$\mathbb{P}\{\text{no pair } H_i, H_j \text{ pick same subset from } F_0\} \leq \binom{N}{2}\exp(-m\epsilon).$$

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

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 $\#F_{\mathcal{H}_2}$.

The class $\mathcal{H}_2$ might shatter some small F sets. For example, if F consists of 3 points, not all on the same straight line, then it can be shattered by $\mathcal{H}_2$. However no set of 9 points can be shattered, because there are $2^9 = 512$ possible subsets—the empty set included—whereas the half-spaces can pick out at most $4 \times 9^2 = 324$ subsets. More generally, $2^n > 4n^2$ for all $n \geq 9$, so that, of course, no set of more than 9 points can be shattered by $\mathcal{H}_2$.

> REMARK. You should find it is easy to improve on the 9, by arguing directly that no set of 4 points can be shattered by $\mathcal{H}_2$. Indeed, if $\mathcal{H}_2$ 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.

In summary: The size of the largest set shattered by $\mathcal{H}_2$ 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 Section 2, the class $\mathcal{H}_2$ would be said to have VC dimension equal to 3.

The argument had little to do with the choice of $\mathcal{H}_2$ as a class of half-spaces in a particular Euclidean space. It would apply to any class $\mathcal{D}$ for which $\#F_{\mathcal{D}}$ is bounded by a fixed polynomial in $\#F$. And therein lies the challenge. In general, for more complicated classes of subsets $\mathcal{D}$ of arbitrary spaces, it can be quite difficult to bound $\#F_{\mathcal{D}}$ by a polynomial in $\#F$, but it is often less difficult to prove existence of a finite VC-dim such that no set of more than VC-dim points can be shattered by $\mathcal{D}$. The miracle is that a polynomial bound then follows automatically, as will be shown in Section 2.

I need to check the assertions in the following Remark. See Dudley (1978, Section 7).

REMARK. For the set $\mathcal{H}_k$ of all closed halfspaces in $\mathbb{R}^k$, the VC method will deliver a bound

$$N \le p(n) = \sum\nolimits_{j \le k+1} \binom{n}{j}$$

for the largest number of subsets that can be picked out from a set with n points. For $k = 2$, the bound is a cubic in n, which is inferior (for large n) to the $4n^2$ obtained above. In fact, there is an even better bound,

$$N \le p(n) = 2 \sum\nolimits_{j \le k} \binom{n-1}{j}$$

which is achieved when no $k+1$ of the points lie in any hyperplane. For $k = 2$, the bound becomes $n^2 - n + 2$.

The upper bound, $p(n)$, for the number of subsets picked out from a set of n points, should not be confused with

$$\sum\nolimits_{j \le k} \binom{N}{j},$$

the largest number of regions into which $\mathbb{R}^k$ can be partitioned by N hyperplanes.

2. Shattered subsets, shattered columns of a matrix

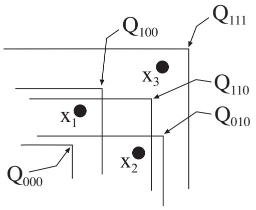
There are various ways to express the counting arguments that lead the bounds on the number of subsets picked out by a given $\mathcal{D}$ from a finite subset. Actually, when the subset is generated as a random sample from some distribution, it is better to allow for possible duplicates by counting ***patterns picked out*** from an $\mathbf{x} = (x_1, \ldots, x_n)$ in $\mathcal{X}^n$. A pattern is specified by a subset $\mathbb{J}$ of $\{1, 2, \ldots, n\}$ and a set $D_{\mathbb{J}}$ from $\mathcal{D}$ for which

<1>

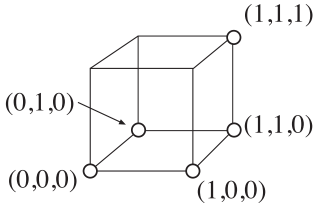
$$\begin{cases} x_i \in D_{\mathbb{J}} & \text{for } i \in \mathbb{J} \\ x_i \in D_{\mathbb{J}}^c & \text{for } i \in \mathbb{J}^c \end{cases}$$

Of course, the $D_{\mathbb{J}}$ need not be unique. In general, there will be many different $\mathcal{D}$ sets that pick out the same pattern from $\mathbf{x}$. For counting purposes, we can identify a $\mathcal{D}$ that picks out N distinct patterns from $\mathbf{x}$ with an $N \times n$ ***binary matrix*** $V = V_{\mathbf{x},\mathcal{D}}$, that is a matrix with distinct rows with $\{0, 1\}$ entries. The subset $\mathbb{J}$ from <1> would correspond to row of the matrix with a 1 in the columns picked out by $\mathbb{J}$ and 0 elsewhere. We could also identify each row of the matrix with a different vertex of $\{0, 1\}^n$.

<2> **Example.** Consider the case where $F = \{x_1, x_2, x_3\}$ is a subset of $\mathbb{R}^2$ and $\mathcal{Q}$ is the collection of all closed quadrants with a north-east vertex.



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



For the configuration shown, $F_{\mathcal{Q}}$ consists of five subsets. The labelling of quadrants that pick out these subsets corresponds to the rows of the matrix V and to a set of five vertices of the cube $\{0, 1\}^3$. Notice that the ordering of the rows and columns of the matrix is somewhat arbitrary. The subset $\{x_1, x_2\}$, which corresponds to the submatrix matrix V_0 consisting of the first two columns of V, is shattered by $\mathcal{Q}$. All four vectors $(0, 0)$, $(0, 1)$, $(1, 0)$, and $(1, 1)$ appear as rows of V_0. □

To identify various submatrices of a given $N \times n$ matrix V it is convenient to borrow notation from the S language.

(a) For positive integers p and q with $p \leq q$, write $p : q$ for the set $\{i \in \mathbb{Z} : p \leq i \leq q\}$.

(b) For subsets $\mathbb{I}$ of $1{:}N$ and $\mathbb{J}$ of $1{:}n$ define $V[\mathbb{I}, \mathbb{J}]$ as the $\#\mathbb{I} \times \#\mathbb{J}$ submatrix with rows from $\mathbb{I}$ and columns from $\mathbb{J}$.

(c) Write $V[\mathbb{I}, -\mathbb{J}]$ for the $\#\mathbb{I} \times (n - \#\mathbb{J})$ submatrix with rows from $\mathbb{I}$ and columns from $1{:}n\backslash\mathbb{J}$, and so on. Interpret a missing index set or a single dot, as in $V[\mathbb{I},]$ or $V[\mathbb{I}, \cdot]$, to mean that no constraint is placed on that coordinate.

<3> **Example.** Suppose V is the $N \times n$ matrix, for $N = 6$ and $n = 4$, with (i, j)th element $v_{i,j}$. Then

$$V[2:5, 2:4] = \begin{bmatrix} v_{2,2} & v_{2,3} & v_{2,4} \\ v_{3,2} & v_{3,3} & v_{3,4} \\ v_{4,2} & v_{4,3} & v_{4,4} \\ v_{5,2} & v_{5,3} & v_{5,4} \end{bmatrix}$$
$$= V[2:5, -1] = V[-\{1, 6\}, 2:4] = V[-\{1, 6\}, -1]$$

□

<4> **Definition.** *Let V be an $N \times n$ binary matrix.*

(i) Say that a nonempty subset $\mathbb{J}$ of $1 : n$, with $k = \#\mathbb{J}$, is shattered if each possible 2^k possible k-tuples of 0's and 1's appears at least once as a row of $V[, \mathbb{J}]$. Equivalently, there is an $\mathbb{I} \subseteq 1 : N$ with $\#\mathbb{I} = 2^k$ such that the submatrix $V[\mathbb{I}, \mathbb{J}]$ has distinct rows.

*(iii) Define the **shatter dimension** s-dim(V) of V as the largest k for which there is a shattered $\mathbb{J}$ with $\#\mathbb{J} = k$.*

(ii) If V equals $V_{\mathbf{x},\mathcal{D}}$, the matrix indicating which patterns a collection $\mathcal{D}$ of subsets picks out from $\mathbf{x} = (x_1, \ldots, x_n)$, say that $\mathcal{D}$ shatters $(x_i : i \in \mathbb{J})$ if V shatters $\mathbb{J}$. Write s-dim$(\mathbf{x}, \mathcal{D})$ for the shatter dimension of $V_{\mathbf{x},\mathcal{D}}$.

(iv) Define the VC dimension, VC-dim$(\mathcal{D})$, of $\mathcal{D}$ as the supremum of s-dim$(\mathbf{x}, \mathcal{D})$ over all $\mathbf{x} \in \mathcal{X}^n$, all $n \in \mathbb{N}$. Call $\mathcal{D}$ a VC class of sets if VC-dim$(\mathcal{D}) < \infty$.

<5> **Example.** The matrix

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

does not shatter $\{1, 2\}$ because the vector $(1, 1)$ does not appear as a row of the 6×2 submatrix $V[, \{1, 2\}]$, the first two columns of V. Each of the other five subsets of $\{1, 2, 3, 4\}$ of size two is shattered. For example, $V[\{1, 2, 4, 5\}, \{2, 3\}]$ has distinct rows. No subset of three (or four) columns is shattered. Each singleton $\{1\}$, $\{2\}$, $\{3\}$, and $\{4\}$ is shattered, because each column contains at least one 0 and one 1. (Easier: every nonempty subset of a shattered $\mathbb{J}$ is also shattered.)

The matrix V has shatter dimension s-dim(V) equal to 2. For future reference, note that the number of rows is strictly greater than $\binom{4}{0} + \binom{4}{1}$. □

<6> **Example.** For the class $\mathcal{H}_k$ of all closed half-spaces in $\mathbb{R}^k$, show that VC-dim$(\mathcal{H}_k) \le k + 1$.

It is easy to verify that $\mathcal{H}_k$ shatters the $k+1$ points consisting of the origin and the k unit vectors that make up the usual basis: consider sets of the form $\{x : \alpha \cdot x \le c\}$ for various α with components 0 or ± 1.

It remains to show that $\mathcal{H}_k$ can shatter no set $F = \{x_0, \ldots, x_{k+1}\}$ of $k+2$ points in $\mathbb{R}^k$. Linear dependence of the vectors $x_1 - x_0, \ldots, x_{k+1} - x_0$ ensures existence of coefficients α_i, not all zero, such that

$$\sum\nolimits_{i=1}^{k+1} \alpha_i (x_i - x_0) = 0.$$

Put $\alpha_0 = -\sum_{i=1}^{k+1} \alpha_i$. Then $\sum_{i=0}^{k+1} \alpha_i = 0$ and $\sum_{i=0}^{k+1} \alpha_i x_i = 0$, or

$$\sum\nolimits_{i=0}^{k+1} \alpha_i^+ x_i = \sum\nolimits_{i=0}^{k+1} \alpha_i^- x_i.$$

Divide through by the nonzero quantity $\sum_{i=0}^{k+1} \alpha_i^+ = \sum_{i=0}^{k+1} \alpha_i^-$ to recognize that we have found disjoint subsets F_0 and F_1 of F whose convex hulls overlap. There can be no closed half-space that picks out F_0 from F, for the existence of such a half-space would imply that the convex hull of F_0 is disjoint from the convex hull of $F \backslash F_0$: a contradiction. □

<7> **Example.** Suppose $\mathcal{F}$ is a k-dimensional vector space of functions on $\mathcal{X}$. Write $\mathcal{D}$ for the class of all sets of the form $\{f \ge 0\}$, with f in $\mathcal{F}$. Show that VC-dim$(\mathcal{D}) \le k$.

Consider a set of $k + 1$ points $x_0, \ldots, x_k$ in $\mathcal{X}$. The set $\mathbb{F}$ of points of the form

$$(f(x_0), \ldots, f(x_k)) \qquad \text{for } f \text{ in } \mathcal{F}$$

is a vector subspace of $\mathbb{R}^{k+1}$ of dimension at most k. There must exist some nonzero vector α orthogonal to $\mathbb{F}$. Express the orthogonality as

$$\sum_{i=0}^{k} \alpha_i^+ f(x_i) = \sum_{i=0}^{k} \alpha_i^- f(x_i) \qquad \text{for each } f \text{ in } \mathcal{F}.$$

Without loss of generality suppose $\alpha_0 > 0$. No member of $\mathcal{D}$ can pick out the subset $\{x_i : \alpha_i < 0\}$: if $f(x_i) \ge 0$ when $\alpha_i < 0$ and $f(x_i) < 0$ when $\alpha_i \ge 0$ then the left-hand side of the equality would be strictly negative, while the right-hand side would be nonnegative. The class $\mathcal{D}$ has shatter dimension at most k; it shatters no set of $k + 1$ or more points. □

3. The VC lemma for binary matrices

The key result in the area is often called the VC lemma, although credit should be spread more widely. (See the Notes in Section 11.)

<8> **Theorem.** *Let V be an $N \times n$ binary matrix. If* s-dim$(V) \le d$ *then*

$$N \le \binom{n}{0} + \binom{n}{1} + \ldots + \binom{n}{d}.$$

If $n \ge d$, the upper bound is less than $(en/d)^d$.

Proof. I will establish the contrapositive, by showing that if

$$N > \binom{n}{0} + \binom{n}{1} + \ldots + \binom{n}{d} \qquad <9>$$

then s-dim$(V) > d$.

Define the ***downshift*** for the jth column of the matrix as the operation:

```
for $i = 1, \ldots, N$
        if $V[i, j] = 1$ change it to a 0 unless the resulting
        matrix $V^{(1)}$ would no longer have distinct rows
```

For example, the downshift for the 1st column of the matrix V from Example <5> generates a matrix $V^{(1)}$ with first and third rows different fro the corresponding rows of V:

$$V = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 \end{bmatrix} \qquad \text{downshifts to} \qquad V^{(1)} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 \end{bmatrix}$$

The 1 in the last row was blocked (prevented from changing to a 0) by the fifth row; if it had been changed to a 0 the row (0 0 1 1) would have appeared twice in V_1.

REMARK. The order in which we consider the 1's in column j makes no difference to the $V^{(1)}$ created from V by the downshift of column j. If it were possible to create by a downshift of a 1 in $V[i_1, j]$ a row $V^{(1)}[i_1,]$ that would block the downshift for some $V[i_2, j]$, then we would have $V[i_1,] = V[i_2, \cdot]$. We need only examine the rows of V with $V[i, j] = 0$ to determine whether a downshift is blocked.

Starting from $V^{(1)}$, select any other column for which downshifting generates a new matrix $V^{(2)}$. And so on. It is possible that the downshift of a particular 1 in column j that is initially blocked by some row might succeed at a later stage, because the blocking row might itself be changed by some downshift carried out between two downshift operations on column j. Stop when no more changes can be made by downshifting, leaving a binary matrix $V^{(m)}$.

For example, a downshift on the 2nd column of the $V^{(1)}$ shown above generates

$$V^{(2)} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 \end{bmatrix},$$

the fourth row being blocked from changing by the third row. Then, just for a change in the routine, downshift on the 4th column then the 3rd column:

$$V^{(3)} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix}, \quad V^{(4)} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}.$$

No more changes can be generated by downshifts; every possible change of a 1 to a 0 is blocked by some row already present. Notice that columns $\{3, 4\}$ are shattered by $V^{(4)}$.

The downshifting operation has two important properties:

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

(ii) When no more downshifting is posssible, there is a very simple way to choose a set of $d + 1$ columns that are shattered by the final matrix, and hence must have been shattered by the original V.

Let me establish (i) and then (ii).

For simplicity of notation, consider a downshift for the 1st column of the matrix V. Suppose $V^{(1)}$ shatters $\mathbb{J}$. We need to show that V also shatters $\mathbb{J}$. If $1 \notin \mathbb{J}$ the submatrix $V^{(1)}[, \mathbb{J}]$ is the same as $V[, \mathbb{J}]$, which makes the assertion about V trivially true. So suppose, for notational simplicity, that $\mathbb{J} = 1{:}k$.

For every $(k-1)$-tuple x of 0's and 1's, the row $(1, x)$ appears somewhere in $V^{(1)}[\cdot, \mathbb{J}]$; that is, for some $(n-k)$-tuple w, the row $v := (1, x, w)$ appears in $V^{(1)}$. As the downshift creates no new 1's, the row v must already have been present in V. Moreover, the row $(0, x, w)$ must also have appeared in V to block the downshift of the leading 1 in v. Thus both $(0, x)$ and $(1, x)$ are rows of $V[\cdot, \mathbb{J}]$.

For (ii), note that the final matrix, $V^{(m)}$, has more rows than the $\binom{n}{0} + \binom{n}{1} + \ldots + \binom{n}{d}$ that could be created by allowing d or fewer 1's per row. There must be a row v_0 of $V^{(m)}$ with 1's in columns $\mathbb{J}$ for a $\mathbb{J}$ with $\#\mathbb{J} > d$. Each possible downshift of a 1 in v_0 must be blocked some other row v_1 of $V^{(m)}$. And each possible downshift of a 1 in v_1 must be blocked by some row v_2 of $V^{(m)}$. And so on. In fact, every row vector obtainable from v_0 by changing some subset of its 1's to 0's must appear amongst the rows of $V^{(m)}$. The matrix $V^{(m)}$ must shatter $\mathbb{J}$.

The weaker upper bound for N when $n \geq d$ will follow from a simple calculation with a random variable Z distributed Bin(n, $1/2$). The sum of binomial coefficients equals

$$2^n \mathbb{P}\{Z \leq d\} \leq 2^n \mathbb{P}(d/n)^{Z-d} = (d/n)^{-d}\left(1 + \frac{d}{n}\right)^n \leq (n/d)^d e^d,$$

□ the asserted upper bound.

4. How to generate VC classes of set

Incomplete

<10> **Theorem.** *If a class $\mathcal{D}$ has finite shatter dimension $\mathbb{D}$, then, for each finite F,*

$$\#F_{\mathcal{D}} \leq \binom{n}{0} + \binom{n}{1} + \ldots + \binom{n}{\mathbb{D}}, \qquad \text{where } n = \#F.$$

If $n \geq \mathbb{D}$, the sum of binomial coefficients is bounded above by $(en/\mathbb{D})^{\mathbb{D}}$.

<11> **Example.** If $\mathcal{D}$ is a class of sets with shatter dimension at most $\mathbb{D}$ then the class $\mathcal{U} = \{D_1 \cup D_2 : D_i \in \mathcal{D}\}$ has shatter dimension at most $10\mathbb{D}$. (The bound is crude, but adequate for our purposes.) From a set F of $n = 10\mathbb{D}$ points, there are at most $(10\mathbb{D}e/\mathbb{D})^{\mathbb{D}}$ distinct sets FD, with $D \in \mathcal{D}$. The trace class $F_{\mathcal{U}}$ consists of at most $(10e)^{n/5}$ subsets. The class $\mathcal{U}$ does not shatter F, because $(10e)^{1/5} < 2$.

A similar argument applies to other classes formed from $\mathcal{D}$, such as the pairwise intersections, or complements. The idea can be iterated to generate □ very fancy classes with finite shatter dimension (Problem [2]).

Think of all regions of $\mathbb{R}^k$ that can be represented as unions of at most ten million sets of the form $\{f \geq 0\}$, with f a polynomial of degree less than a million, if you want to get some idea of how complicated a class with finite shatter dimension can be.

Describe cross sections from o-minimal structure.
Stengle & Yukich (1989)
van den Dries (1998, Chapter 5)

5. Packing numbers for function classes

The connection between shatter dimension and covering numbers introduced in Section 1 extends to more general classes $\mathcal{D}$ of measurable subsets of a space $\mathcal{X}$ on which a probability measure P is defined. I will derive an upper bound slightly more precise than before, for the sake of comparison with the results that will be derived in Section 9. Remember that the packing number $D(\epsilon, \mathcal{D}, \mathcal{L}^1(P))$ is defined as the largest number of sets in $\mathcal{D}$ separated by at least ϵ in $\mathcal{L}^1(P)$ distance: $P(D_i \Delta D_j) > \epsilon$ for $i \neq j$.

<12> **Lemma.** *Let $\mathcal{D}$ be a class of sets with* VC-dim$(\mathcal{D}) \leq d$. *Then for each probability measure P,*

$$D(\epsilon, \mathcal{D}, \mathcal{L}^1(P)) \leq \left(\frac{5}{\epsilon}\log\frac{3e}{\epsilon}\right)^d \qquad \text{for } 0 < \epsilon \leq 1.$$

REMARK. If P is concentrated on a finite number of points we need not worry about measurability.

Proof. Suppose $D_1, \ldots, D_N$ are sets in $\mathcal{D}$ with $P(D_i \Delta D_j) > \epsilon$ for $i \neq j$. The asserted bound holds trivially when $N \leq (5\log(3e))^d$. It is more than enough to treat the case where $\log N > d$.

Let $\mathbf{x} = (x_1, \ldots, x_m)$ be an independent sample of size $m = \lceil 2(\log N)/\epsilon \rceil$ from P. Notice that $3(\log N)/\epsilon > m \geq d$. As in Section 1, for fixed i and j,

$$\mathbb{P}\{D_i \text{ and } D_j \text{ pick out same pattern from } \mathbf{x}\} \leq \exp(-m\epsilon).$$

The sum of $\binom{N}{2}$ such probability bounds is strictly less than 1. For some realization $\mathbf{x}$, the class $\mathcal{D}$ picks out N distinct patterns from $\mathbf{x}$, which gives the inequality

$$N \leq \binom{m}{0} + \binom{m}{1} + \ldots + \binom{m}{d} \leq \left(\frac{em}{d}\right)^d \qquad \text{because } m \geq d.$$

Tus

$$N^{1/d} \leq \frac{3e\log N}{d\epsilon}$$

From Problem [1], the function $g(x) = x/\log x$ is increasing on the range $e \le x < \infty$, and if $y \ge g(x)$ then $x \le (1-e^{-1})^{-1} y \log y$. Put $x = N^{1/d}$ and $y = 3e/\epsilon$ to deduce that

$$N^{1/d} \le (1-e^{-1})^{-1}(3e/\epsilon)\log(3e/\epsilon).$$

□ The stated bound merely tidies up some constants.

The Lemma has a most useful analog for classes of functions equipped with various $\mathcal{L}^1(\mu)$ pseudometrics. Let $\mathcal{F}$ be a class of real-valued functions on $\mathcal{X}$. Let F be a ***measurable envelope*** for $\mathcal{F}$, that is, a measurable function for which $\sup_{f\in\mathcal{F}} |f(x)| \le F(x)$ for every x. Suppose μ is a measure on $\mathcal{X}$ for which $\mu F < \infty$. The packing number $D(\epsilon, \mathcal{F}, \mathcal{L}^1(\mu))$ is defined as the largest N for which there exist functions $f_1, \dots, f_N$ in $\mathcal{F}$ with

$$\mu|f_i - f_j| > \epsilon \qquad \text{for } i \ne j.$$

If we replace ϵ by $\epsilon\mu F$, the definition becomes invariant to rescalings of μ and the functions in $\mathcal{F}$. More importantly, in one very common case there will exist a bound on the corresponding packing numbers that does not depend on μ, a property of great significance in empirical process theory.

<13> **Definition.** *Say that $\mathcal{F}$ is a VC-subgraph class if the collection $\mathcal{S}(\mathcal{F})$ of **subgraphs** $S(f) := \{(x,r) \in \mathcal{X} \otimes \mathbb{R} : f(x) \ge r\}$ is a VC class.*

<14> **Lemma.** *Let $\mathcal{F}$ be a VC-subgraph class with envelope F in $\mathcal{L}^1(\mu)$. Then*

$$D(2\epsilon\mu F, \mathcal{F}, \mathcal{L}^1(\mu)) \le \left(\frac{5}{\epsilon}\log\frac{3e}{\epsilon}\right)^d \qquad \text{for } 0 < \epsilon \le 1,$$

if VC-dim $\big(\mathcal{S}(\mathcal{F})\big) \le d$.

Proof. Suppose $\{f_1, \dots, f_N\} \subseteq \mathcal{F}$ with $\mu|f_i - f_j| > 2\epsilon\mu F$ for $i \ne j$. Notice that each $S(f_i)\Delta S(f_j)$ is a subset of

$$\mathbb{B} := \{(x,t) \in \mathcal{X} \times \mathbb{R} : |t| \le F(x)\}$$

Let $\mathfrak{m}$ denote Lebesgue measure on $\mathcal{B}(\mathcal{R})$. Write P for the probability measure defined by $PA = \mu \otimes \mathfrak{m}(A\mathbb{B})/\mu \otimes \mathfrak{m}(\mathbb{B})$. By Fubini, $\mu \otimes \mathfrak{m}\mathbb{B} = 2\mu F$. For $i \ne j$ it follows that

$$P\big(S(f_i)\Delta S(f_j)\big) = \frac{\mu \otimes \mathfrak{m}|\{f_i(x) \ge t\} - \{f_j(x) \ge t\}|}{2\mu F} > \epsilon.$$

□ Conclude that $N \le D(\epsilon, \mathcal{S}(\mathcal{F}), \mathcal{L}^1(P))$, which gives the asserted upper bound.

For many applications it is enough that the covering numbers are uniformly of order $O(\epsilon^{-W})$ for some W.

<15> **Definition.** *A collection of (measurable) functions is said to be **Euclidean** for an envelope F if there exist constants A and W for which*

$$D(\epsilon\mu F, \mathcal{F}, \mathcal{L}^1(\mu)) \le A(1/\epsilon)^W \qquad \text{for } 0 < \epsilon \le 1$$

for every measure μ with $F \in \mathcal{L}^1(\mu)$.

In particular, every VC subgraph class is Euclidean, with constants A and W that depend only on the VC dimension of the set of subgraphs.

For many purposes, an empirical process indexed by a Euclidean class of functions behaves like a process smoothly indexed by a compact subset of some (finite dimensional) Euclidean space.

The existence of packing bounds that work for many different measures μ allows us to calculate bounds for $\mathcal{L}^\alpha(\mu)$ packing numbers, for values of α greater than 1, if $F \in \mathcal{L}^\alpha(\mu)$. For suppose $\{f_1, \dots, f_N\} \subseteq \mathcal{F}$ with

$$\left(\mu|f_i - f_j|^\alpha\right)^{1/\alpha} > \epsilon \left(\mu(F^\alpha)\right)^{1/\alpha} \qquad \text{for } i \neq j.$$

Define a new measure λ by $d\lambda/d\mu = F^{\alpha-1}$. Then, for $i \neq j$,

$$\begin{aligned} 2^{\alpha-1}\lambda|f_i - f_j| &= \mu(2F)^{\alpha-1}|f_i - f_j| \\ &\geq \mu(|f_i| + |f_j|)^{\alpha-1}|f_i - f_j| \\ &\geq \mu|f_i - f_j|^\alpha \\ &> \epsilon^\alpha \lambda F \end{aligned}$$

Thus

$$\begin{aligned} D(\epsilon \left(\mu(F^\alpha)\right)^{1/\alpha}, \mathcal{F}, \mathcal{L}^\alpha(\mu)) &\leq D(\epsilon^\alpha 2^{1-\alpha}\lambda(F), \mathcal{F}, \mathcal{L}^1(\lambda)) \qquad \text{if } F \in \mathcal{L}^\alpha(\mu) \\ &\leq A_\alpha (1/\epsilon)^{W_\alpha} \qquad \text{if } \mathcal{F} \text{ is Euclidean,} \end{aligned}$$

where A_α and W_α are functions of α and of the Euclidean constants, A and W. In particular, if $F \in \mathcal{L}^2(\mu)$ then

$$D(\epsilon \left(\mu(F^2)\right)^{1/2}, \mathcal{F}, \mathcal{L}^2(\mu)) \leq D\left(\tfrac{1}{2}\epsilon^2\lambda(F), \mathcal{F}, \mathcal{L}^1(\lambda)\right) \leq 4A(1/\epsilon)^{2W}.$$

When specialized to the measure that puts mass $1/n$ at each of $x_1, \dots, x_n$, the last inequality gives bounds for covering numbers under ordinary Euclidean distance.

6. Fat shattering

The VC-subgraph property of Definition <13> imposes a subtle micro-constraint of the functions in a class $\mathcal{F}$. The property can be destroyed by arbitrarily small perturbations in the values of the functions, even by perturbations that have a negligible effect on the packing numbers. The concept of ***fat shattering*** makes the shattering property of the functions more robust by requiring a small margin of error for the property that $\mathcal{F}$ can shatter a finite set of points.

<16> **Definition.** *Say that a class $\mathcal{F}$ of functions ϵ-surrounds a point $x = (x_1, \dots, x_n)$ at levels $(\xi_1, \dots, \xi_n)$ if for each subset $\mathbb{K}$ of $\mathbb{J} := \{1, \dots, n\}$ there exists a function $f_\mathbb{K} \in \mathcal{F}$ for which*

$$f_\mathbb{K}(x_j) \begin{cases} \geq \xi_j + \epsilon/2 & \text{if } j \in \mathbb{K} \\ \leq \xi_j - \epsilon/2 & \text{if } j \in \mathbb{J}\backslash\mathbb{K} \end{cases}$$

Define ϵ-shattering dimension $\mathbb{D}(\mathcal{F}, \epsilon)$ of $\mathcal{F}$ as the largest n for which there exists an x that is ϵ-surrounded (at some level) by $\mathcal{F}$.

If $\mathcal{F}$ consists of indicator functions of sets, the ϵ-surround property is the same for all ϵ in $(0, 1]$ if $\xi_j \equiv 1/2$.

An approximation argument will let us study the consequences of fat shattering as a combinatorial problem. Suppose $0 < \epsilon \leq 1$. Define $p := \lfloor 1/\epsilon \rfloor$, the largest positive integer for which $\epsilon \leq 1/p$. Suppose also that the functions in $\mathcal{F}$ take values in the interval $[0, 1]$. For each function f in $\mathcal{F}$ define $v(\epsilon, x, f)$ to be the n-tuple of integers with jth element

$$v(\epsilon, x, f)_j := \lfloor f(x_j)/\epsilon \rfloor \qquad \text{for } j = 1, \dots, n.$$

That is, $v(\epsilon, x, f)_j$ is the unique integer v_j in the set $\mathbb{S}_p := \{0, 1, \dots, p\}$ for which

$$f(x_j) = \epsilon v_j + \epsilon_j \qquad \text{with } 0 \leq \epsilon_j < \epsilon$$

Distinct functions f and g might correspond to the same n-tuple. We can identify the set of all distinct $v(\epsilon, x, f)$ with the rows of an $N \times n$ matrix $V_{\epsilon,x,\mathcal{F}}$ with elements from $\mathbb{S}_p$. Necessarily, $1 \le N \le (p+1)^n$.

If the point x is 2ϵ-surrounded by $\mathcal{F}$ at levels ξ then the integers $k_j := \lfloor \xi_j/\epsilon \rfloor$ have the property: for each subset $\mathbb{K}$ of $\mathbb{J} := \{1, \dots, n\}$ there exists a function $f_{\mathbb{K}} \in \mathcal{F}$ for which

<17> $$v(\epsilon, x, f_{\mathbb{K}})_j \begin{cases} \ge k_j + 1 & \text{if } j \in \mathbb{K} \\ \le k_j - 1 & \text{if } j \in \mathbb{J}\backslash\mathbb{K} \end{cases}.$$

Conversely, if <17> holds then

$$f_{\mathbb{K}}(x_j) = \epsilon v(\epsilon, x, f_{\mathbb{K}})_j + \epsilon_j \begin{cases} \ge \epsilon k_j + \epsilon & \text{if } j \in \mathbb{K} \\ \le \epsilon k_j - (\epsilon - \epsilon_j) < \epsilon k_j & \text{if } j \in \mathbb{J}\backslash\mathbb{K} \end{cases},$$

which implies that x is $\epsilon/2$-surrounded by $\mathcal{F}$ at levels $\epsilon(k_j + \frac{1}{2})$.

In reducing the possibly infinite set of functions $\mathcal{F}$ to the finite matrix $V_{\epsilon,x,\mathcal{F}}$ we sacrifice only a factor of 2 is our study of fat shattering.

<18> **Definition.** *Let $\mathbb{L}_n := \cup_{\mathbb{J}} \mathbb{Z}^{\mathbb{J}}$, the union running over all nonempty subsets of $\{1, \dots, n\}$, and say that a lattice point $\zeta \in \mathbb{Z}^{\mathbb{J}}$ has degree $\#\mathbb{J}$.*

*Write $\mathbb{M}(n, p)$ for the set of all matrices with n columns and distinct rows with elements from $\mathbb{S}_p := 0{:}p$. Let $\mathbb{J}$ be a nonempty subset of $\{1, \dots, n\}$. Say that a lattice point $\zeta \in \mathbb{Z}^{\mathbb{J}}$ is **2-surrounded** by a V in $\mathbb{M}(n, p)$ if for each of the $2^{\#\mathbb{J}}$ subsets $\mathbb{K}$ of $\mathbb{J}$ there is a row $i_{\mathbb{K}}$ of V for which*

$$V[i_{\mathbb{K}}, j] \begin{cases} \ge \zeta_j + 1 & \text{if } j \in \mathbb{K} \\ \le \zeta_j - 1 & \text{if } j \in \mathbb{J}\backslash\mathbb{K} \end{cases}.$$

*Define the **2-shatter dimension** $\mathbb{D}_2(V)$ to be the largest degree of a lattice point 2-surrounded by V. Define the **2-surround number** $\mathbb{S}_2(V)$ as the number of distinct lattice points from $\mathbb{L}_n$ that are 2-surrounded by V.*

FALSE: The set $\mathbb{M}_1$ is the set of binary matrices. The definitions for $p = 1$ are just slight reformulation of properties of binary matrices.

For a given $\mathbb{J}$ with $\#\mathbb{J} = k$ and $\zeta \in \mathbb{Z}^{\mathbb{J}}$, there are at most $p - 1$ choices for each ζ_j if the lattice point is to be surrounded. There are at most $\binom{n}{k}(p-1)^k$ lattice points of degree k that could possibly be 2-surrounded by a matrix $V \in \mathbb{M}(n, p)$. If

$$\mathbb{S}_2(V) > \sum\nolimits_{k=1}^{d} \binom{n}{k} (p-1)^k,$$

the pigeon-hole principle shows that there must be at least one lattice point of degree at least $(d+1)$ that is 2-surrounded by V; and if $\mathbb{D}_2(V) \le d$ then

$$\mathbb{S}_2(V) \le \sum\nolimits_{k=1}^{d} \binom{n}{k} (p-1)^k.$$

In particular, if x is 2ϵ-surrounded by $\mathcal{F}$ then $\mathbb{D}_2\left(V_{\epsilon,x,\mathcal{F}}\right) \ge d$

has 2-shatter dimension d For a matrix

????

7. Mendelson and Vershynin

Based on Mendelson & Vershynin (2003).

<19> **Theorem.** *Suppose V is an $N \times n$ matrix in $\mathbb{M}_p$. For $\alpha \in [1, \infty)$ define $K_\alpha := \left(\sum_{k\ge2} k^\alpha 2^{-k}\right)^{1/\alpha}$ and $C_\alpha := 2^{(1+\alpha)/\alpha} 3K_\alpha$. Suppose the rows of V are C_α-separated, in the sense that*

$$\left(n^{-1}\sum\nolimits_{j\le n} |V[i_1, j] - V[i_2, j]|^\alpha\right)^{1/\alpha} \ge C_\alpha \qquad \text{for all } 1 \le i_1 < i_2 \le n.$$

Then

(i) $\mathbb{S}_2(V) \ge \sqrt{N} - 1$.

(ii) if $\mathbb{D}_2(V) \le d$ *then* $\sqrt{N} \le \sum_{k=0}^{d} \binom{n}{k} p^k \le (epn/d)^d$.

<20> **Lemma.** *Let X be a random variable with zero median for which $\infty > \mathbb{P}|X|^\alpha \ge \tau^\alpha$. Then there exists a $\beta \in (0, 1/2]$ and an interval $[a, b]$ of length at least τ/K_α such that either*

$$\mathbb{P}\{X \le a\} \ge \beta/2 \qquad \text{and} \qquad \mathbb{P}\{X \ge b\} \ge 1 - \beta$$

$$\text{or}$$

$$\mathbb{P}\{X \le a\} \ge 1 - \beta \qquad \text{and} \qquad \mathbb{P}\{X \ge b\} \ge \beta/2$$

Proof. Represent X as $q(U)$ where q is an increasing function with $q(1/2) = 0$ and U has a Uniform(0, 1) distribution. With no loss of generality, suppose $\sigma^\alpha := 2\mathbb{P}|q(U)|^\alpha\{U > 1/2\} \ge \tau^\alpha$. Suppose a constant c has the property that

<21> $$q(1 - \beta) + c\sigma > q(1 - \beta/2) \qquad \text{for } 0 < \beta \le 1/2.$$

Repeated appeals to this inequality for $\beta = 2^{-k}$ for $k = 1, 2, \ldots$ followed by a telescoping summation give $kc\sigma > q\left(1 - 2^{-k-1}\right)$, and hence

$$\sigma^\alpha = 2\sum\nolimits_{k=1}^{\infty} \int_{1-2^{-k}}^{1-2^{-k-1}} q(u)^\alpha du \le \sum\nolimits_{k=1}^{\infty} 2^{-k} q\left(1 - 2^{-k-1}\right)^\alpha < \left(c\sigma K_\alpha\right)^\alpha.$$

Inequality <21> must therefore fail if we choose $c = 1/K_\alpha$: there must exist some β in (0, 1/2] for which $q(1 - \beta) + \sigma/K_\alpha \le q(1 - \beta/2)$. For that β we have

$$\mathbb{P}\{X \ge q(1 - \beta/2)\} \ge \mathbb{P}\{U \ge 1 - \beta/2\} = \beta/2$$
$$\mathbb{P}\{X \le q(1 - \beta)\} \ge \mathbb{P}\{U \le 1 - \beta\} = 1 - \beta.$$

□ Analogous inequalities hold if we shrink the interval to have length τ/K_α.

<22> **Corollary.** *For the matrix V from Theorem <19> there exists a constant $\beta \in (0, 1/2]$, a column j_0, and an integer η such that both subsets*

<23> $$\mathbb{I}_1 := \{i : V[i, j_0] \ge \eta + 1\} \qquad \text{and} \qquad \mathbb{I}_2 := \{i : V[i, j_0] \le \eta - 1\}$$

are nonempty, with

$$\max\left(\#\mathbb{I}_1, \#\mathbb{I}_2\right) \ge (1 - \beta)N \qquad \text{and} \qquad \min\left(\#\mathbb{I}_1, \#\mathbb{I}_2\right) \ge \beta N/2.$$

Proof. Independently select I_1 and I_2 from the uniform distribution on $\{1, 2, \ldots, N\}$. We have $\mathbb{P}\{I_1 = I_2\} = 1/N$. When $I_1 \ne I_2$ the corresponding rows of V are C_α separated. Thus

$$\mathbb{P}\sum\nolimits_{j\le n} |V[I_1, j] - V[I_2, j]|^\alpha \ge nC_\alpha^\alpha\left(1 - N^{-1}\right) \ge \tfrac{1}{2}nC_\alpha^\alpha,$$

which implies existence of at least one j_0 for which

$$\mathbb{P}|V[I_1, j_0] - V[I_2, j_0]|^\alpha \ge \tfrac{1}{2}C_\alpha^\alpha.$$

Let m_0 be a median for the distribution of the random variable $V[I_1, j_0]$. Via the inequality $|x + y|^\alpha \le 2^{\alpha-1}\left(|x|^\alpha + |y|^\alpha\right)$ for real numbers x and y deduce that

$$2^\alpha\mathbb{P}|V[I_1, j_0] - m_0|^\alpha \ge \tfrac{1}{2}C_\alpha^\alpha$$

The random variable $X := V[I_1, j_0] - m_0$ has $\mathbb{P}|X|^\alpha \geq \frac{1}{2}(C_\alpha/2)^\alpha = (3K_\alpha)^\alpha$. A gap of length $3K_\alpha/K_\alpha$ must contain at least one interval $(\eta-1, \eta+1)$ with η an integer.
□

Proof of Theorem <19>. Assertion (ii) follows from assertion (i), as explained in the previous Section.

Assertion (i) is trivial for $N = 1$. For the purposes of an inductive proof, suppose that $N \geq 2$ and that the assertion is true for matrices with smaller numbers of rows.

For simplicity of notation, suppose the j_0 from Corollary <22> equals 1. Write N_r for $\#\mathbb{I}_r$ and V_r for $V[\mathbb{I}_r,]$, for $r = 1, 2$.

Consider a $\mathbb{J}$ with $1 \in \mathbb{J}$. If V_1 2-surrounds a lattice point $\zeta \in \mathbb{Z}^{\mathbb{J}}$ we must have $\zeta_1 \geq \eta + 1$; and if V_2 2-surrounds ζ we must have $\zeta_1 \leq \eta - 1$. It is therefore impossible for both V_1 and V_2 to 2-surround the same lattice point in this $\mathbb{Z}^{\mathbb{J}}$. In particular, neither V_r can 2-surround η in its role as a lattice point from $\mathbb{Z}^{\{1\}}$.

Define

$$\begin{aligned}
\mathbb{L}^{(1)} &:= \{\zeta \in \mathbb{L} : \text{ only } V_1 \text{ 2-surrounds } \zeta\} \\
\mathbb{L}^{(2)} &:= \{\zeta \in \mathbb{L} : \text{ only } V_2 \text{ 2-surrounds } \zeta\} \\
\mathbb{L}^{(1,2)} &:= \{\zeta \in \mathbb{L} : \text{ both } V_1 \text{ and } V_2 \text{ 2-surround } \zeta\}
\end{aligned}$$

Clearly V surrounds every point in $\mathbb{L}^{(1)} \cup \mathbb{L}^{(2)} \cup \mathbb{L}^{(1,2)}$. If $\zeta \in \mathbb{L}_{\mathbb{J}} \cap \mathbb{L}^{(1,2)}$ we must have $1 \notin \mathbb{J}$. Neither V_r can surround the lattice point $(\eta, \zeta) \in \mathbb{L}_{\{1\}\cup\mathbb{J}}$ but V does: the submatrix V_1 provides all those $i_{\mathbb{K}}$ for which $V[i_{\mathbb{K}}, 1] > \eta$ and the submatrix V_2 provides all those $i_{\mathbb{K}}$ for which $V[i_{\mathbb{K}}, 1] < \eta$. In short, for each ζ 2-surrounded by both V_1 and V_2 there are two centers, $\zeta \in \mathbb{L}^{(1,2)}$ and $(\eta, \zeta) \notin \mathcal{S}(V_1) \cup \mathcal{S}(V_2)$, 2-surrounded by V. It follows that

<24> $$\mathcal{S}_2(V) \geq 1 + \#\mathbb{L}^{(1)} + \#\mathbb{L}^{(2)} + 2 \times \#\mathbb{L}^{(1,2)} = 1 + \mathcal{S}_2(V_1) + \mathcal{S}_2(V_2).$$

Invoke the inductive hypothesis for both submatrices V_1 and V_2 to conclude that

$$\mathcal{S}_2(V) \geq 1 + (\sqrt{N}_1 - 1) + (\sqrt{N}_2 - 1) = \sqrt{N}\left(\sqrt{\beta/2} + \sqrt{1-\beta}\right) - 1 \geq \sqrt{N} - 1.$$

□ Assertion (i) follows by induction on N.

8. Haussler and Long

Based on Haussler & Long (1995)
Define

$$\psi(i, j) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i < j \\ \star & \text{if } i > j \end{cases}$$

For $u, v \in \mathbb{Z}^k$ define

$$\psi(u, v) = \big(\psi(u_1, v_1), \ldots, \psi(u_k, v_k)\big) \in \{0, 1, \star\}^k$$

Let V be an $N \times n$ matrix in $\mathbb{M}(n, p)$. For a nonempty subset $\mathbb{J}$ of $\{1, \ldots, n\}$ say that V ψ-shatters $\mathbb{J}$ if there exists a $\zeta \in \mathbb{Z}^{\mathbb{J}}$ such that

$$\{\psi\left(V[i, \mathbb{J}], \zeta\right) : 1 \leq i \leq N\} \supseteq \{0, 1\}^{\mathbb{J}}.$$

That is, for each subset $\mathbb{K}$ of $\mathbb{J}$ there exists an $i_{\mathbb{K}}$ such that

<25> $$\psi\left(V[i_{\mathbb{K}}, j], \zeta_j\right) = \begin{cases} 1 & \text{if } j \in \mathbb{K} \\ 0 & \text{if } j \in \mathbb{J}\backslash\mathbb{K} \end{cases}$$

Write $\mathbb{D}_\psi(V)$ for the largest d for which some set $\mathbb{J}$ of d columns of V is ψ-shattered.

For $0 \le d \le n$ and $n \ge 1$ define $\Gamma(n, d, p)$ as the smallest integer for which: if V is an $N \times n$ matrix in $\mathbb{M}(n, p)$ for which $N > \Gamma(n, d, p)$ then there is some subset $\mathbb{J}$ of at least $d+1$ columns that is ψ-shattered by V. Show

<26> $$\Gamma(n, d, p) = \begin{cases} 1 & \text{for } d = 0 \\ (1+p)^n & \text{for } d = n \end{cases} \qquad \text{for } n = 1, 2, \dots$$

The second equality corresponds to the fact that no matrix in $\mathbb{M}(n, p)$ has more than $(1+p)^n$ rows.

<27> **Theorem.** *For* $0 \le d \le n$,

$$\Gamma(n, d, p) = \sum_{k=0}^{d} \binom{n}{k} p^k$$

Proof. Notice that the asserted value for $\Gamma(n, d, p)$ is exactly equal to

$$G(n, d, p) := \#\{x \in \mathbb{S}_p^n : x \text{ has at most } d \text{ nonzero entries}\}$$

By counting separately for those x that end with a zero and those that end with one of the p integers $1, \dots, p$, we get a recursive expression for G,

<28> $$G(n+1, d, p) = G(n, d, p) + pG(n, d-1, p),$$

which together with the equality

$$G(1, d, p) = \begin{cases} 1 & \text{for } d = 0 \\ 1 + p & \text{for } d = 1 \end{cases}$$

uniquely determines G.

Argue by induction on n. True for $n = 1$. Suppose true for values up to n, for some $n \ge 1$. Prove it for $n+1$.

Suppose $V \in \mathbb{M}(n+1, p)$. For each i, call $V[i, 1{:}n]$ the prefix of the row and $V[i, n+1]$ the suffix. Write $\mathcal{W}$ for set of all distinct prefixes in V. For each $w \in \mathcal{W}$, write V_w for the submatrix consisting of all rows with prefix w. Write k_w for the smallest suffix amongst the rows in V_w. All other rows in V_w have suffix strictly greater than k_w.

Suppose $\#\mathcal{W} = N_0$. Let $\mathbb{I}_0$ denote the set of N_0 rows of the form (w, k_w) for $w \in \mathcal{W}$. The suffix for every remaining row must be at least 1. Write $\mathbb{I}_s$ for the set of rows in $(1{:}N)\backslash\mathbb{I}_0$ with suffix s, for $s = 1, 2, \dots, p$. Note that $N = \sum_{s=0}^{p} N_s$. Now suppose that

$$N > \Gamma(n, d, p) + p\Gamma(n, d-1, p) \qquad \text{for some } d \text{ with } 1 \le d \le n.$$

Then either

(i) $N_0 > \Gamma(n, d, p)$

or

(ii) $N_s > \Gamma(n, d-1, p)$ for some $s \ge 1$

In the first case, the $N_0 \times (n+1)$ matrix $V[\mathbb{I}_0,]$, which has distinct rows, must shatter some set of $d+1$ columns, by the inductive hypothesis. In the second case, the $N_s \times n$ matrix $V[\mathbb{I}_s, 1{:}n]$, which also has distinct rows (otherwise two rows of V with suffix s would have the same prefix), must shatter some set $\mathbb{J}$ of d columns from $1{:}n$. That is, there exists some $\zeta \in \mathbb{Z}^{\mathbb{J}}$ such that for each subset $\mathbb{K}$ of $\mathbb{J}$ there exists an $i_{\mathbb{K}} \in \mathbb{I}_s$ with

$$\psi\left(V[i_{\mathbb{K}}, j], \zeta_j\right) = \begin{cases} 1 & \text{if } j \in \mathbb{K} \\ 0 & \text{if } j \in \mathbb{J}\backslash\mathbb{K} \end{cases} .$$

Let $i'_{\mathbb{K}} \in \mathbb{I}_0$ be such that $V[i_{\mathbb{K}},]$ and $V[i'_{\mathbb{K}},]$ have the same prefix. Define $\xi := (\zeta, s)$. Then V shatters $\overline{\mathbb{J}} := \mathbb{J} \cup \{n+1\}$:

$$\psi\left(V[i_{\mathbb{K}}, j], \xi_j\right) = \begin{cases} 1 & \text{if } j \in \overline{\mathbb{K}} \\ 0 & \text{if } j \in \overline{\mathbb{J}} \backslash \overline{\mathbb{K}} \end{cases} \qquad \text{where } \overline{\mathbb{K}} := \mathbb{K} \cup \{n+1\}$$

and

$$\psi\left(V[i'_{\mathbb{K}}, j], \xi_j\right) = \begin{cases} 1 & \text{if } j \in \mathbb{K} \\ 0 & \text{if } j \in \overline{\mathbb{J}} \backslash \mathbb{K} \end{cases} .$$

In either case, we have a set of $d+1$ columns shattered by V.

It follows that

<29> $$\Gamma(n+1, d, p) \le \Gamma(n, d, p) + p\Gamma(n, d-1, p).$$

Together with <26>, this inequality determines an upper bound for $\Gamma(n, d, p)$. Indeed, if we define $\Delta(n, d, p) = G(n, d, p) - \Gamma(n, d, p)$ then <30> and <28> give the recursive inequality

<30> $$\Delta(n+1, d, p) \ge \Delta(n, d, p) + p\Delta(n, d-1, p)$$

with the initial condition $\Delta(1, d, p) = 0$ for $d = 0, 1$. It follows that $\Gamma(n, d, p) \le G(n, d, p)$.

In fact, the last inequality must be an equality, because the $G(n, d, p) \times n$ matrix V consisting of all rows from $\mathbb{S}_p^n$ with at most d nonzero elements cannot shatter any set $\mathbb{J}$ of more than d columns: to get <25> with $\mathbb{K} = \emptyset$ we would need $\zeta_j > 0$ for all j in $\mathbb{J}$; but then at most d elements of $\psi\left(V[i, \mathbb{J}], \zeta\right)$ could equal 1. □

9. An improvement of the packing bound

By means of a more subtle randomization argument, it is possible to eliminate the $\log(6e/\epsilon)$ factor from the bound in Theorem <14>, with a change in the constant. The improvement is due to Haussler (1995).

<31> **Theorem.** *Let V be an $N \times n$ binary matrix for which*

$$\sum\nolimits_{j \le n} \{V[i_1, j] \ne V[i_2, j]\} \ge n\epsilon \qquad \text{for all } 1 \le i_1 < i_2 \le n.$$

for some $0 < \epsilon \le 1$. Then

Not what Haussler got.

$$N \le e(1 + \text{VC-dim})\,(2e/\epsilon)^{\mathbb{D}}$$

where $\mathbb{D} = \mathbb{D}(V)$, the shatter dimension of V.

For Haussler's method, we generate random variables $X_j := V[I, j]$ and random vectors $X_{\mathbb{K}} := V[I, \mathbb{K}]$ by means of an I that is uniformly distributed on $1{:}N$. The separation assumption of the Theorem provides (via Lemma <32>) a lower bound for

$$\sum\nolimits_{\#\mathbb{K}=m} \sum\nolimits_{j \notin \mathbb{K}} \mathbb{P}\text{var}(X_j \mid X_{\mathbb{K}}),$$

where the first sum runs over all subsets $\mathbb{K}$ of size $m-1$ from $1:n$, for a strategically chosen value of m. An elegant extension of Theorem <8> provides (via Lemma <35>) an upper bound for the same quantity. The pair of bounds gives an inequality involving N, n, ϵ and $\mathbb{D}$, which leads to the inequality asserted by the Theorem.

REMARK. My m corresponds to $m-1$ in Haussler's paper.

To keep the notation simple, I will prove unconditional forms of the two Lemmas, then deduce the conditional forms by applying the Lemmas to submatrices of V.

<32> **Lemma.** *Let I be uniformly distributed on $1{:}N$,*

$$\sum\nolimits_{j\le n} \operatorname{var}(X_j) \ge \tfrac{1}{2}n\epsilon\left(1-N^{-1}\right).$$

Proof. Let I' be chosen independently of I, with the same uniform distribution. Write X'_j for $V[I',j]$. Note that the difference $X_j - X'_j$ takes the values ± 1 when $V[I,j] \ne V[I',j]$ and is otherwise zero. Then

$$\begin{aligned}\sum\nolimits_j \operatorname{var}(X_j) &= \sum\nolimits_j \tfrac{1}{2}\mathbb{P}|X_j - X'_j|^2 \qquad \text{by independence}\\ &= \tfrac{1}{2}\mathbb{P}\sum\nolimits_j\{V[I,j]\ne V[I',j]\}.\end{aligned}$$

Whenever $I \ne I'$ which happens with probability $1-N^{-1}$, the last sum is
□ greater than $n\epsilon$.

For the second Lemma we need another inequality involving the shatter dimension. Write $\mathcal{E}_j$ for the set of all pairs (i_1,i_2) for which the two rows $V[i_1,\,]$ and $V[i_2,\,]$ differ only in the jth position. Define $\mathcal{E} = \cup_{j\le n}\mathcal{E}_j$. Call the pairs in $\mathcal{E}$ ***edges***: If we were to identify each row of V with a vertex of the hypercube $\{0,1\}^n$, then $\mathcal{E}$ would correspond to a set of edges between pairs of occupied vertices a distance 1 apart.

By means of a slight modification of the downshifting argument from Theorem <8>, Problem [5] shows that

<33> $$\#\mathcal{E} \le N\mathbb{D}.$$

Using the analogous property for subsets of $\mathcal{E}$, Problem [6], then shows that it is possible to provide an orientation for each edge, so that edge $\mathfrak{e}$ points from vertex $i_1^{\mathfrak{e}}$ to vertex $i_2^{\mathfrak{e}}$, in such a way that no vertex has in-degree greater than $\mathbb{D}$:

<34> $$\#\{\mathfrak{e} : i_2^{\mathfrak{e}} = i\} \le \mathbb{D} \qquad \text{for } 1 \le i \le N.$$

When we apply the Lemma to submatrices of V the distribution of I will not be uniform. We will need the Lemma for more general distributions.

<35> **Lemma.** *For $X_j := V[I,j]$ and $X_{-j} := V[I,-j]$,*

$$\sum\nolimits_{j=1}^{n} \mathbb{P}\operatorname{var}\left(X_j \mid X_{-j}\right) \le \mathbb{D}.$$

under every distribution P for I.

Proof. The random vector X_{-j} takes values in $\mathcal{V}_{-j}$, the set of distinct rows of $V[\,,-j]$. For each $v \in \mathcal{V}_{-j}$, the set $\{i : V[i,-j] = v\}$ contains either one or two values: either v uniquely determines the i for which $V[i,-j]=v$, or there is an edge $\mathfrak{e} = (i_1^{\mathfrak{e}}, i_2^{\mathfrak{e}})$ in $\mathcal{E}_j$ for which $V[i_1^{\mathfrak{e}},-j] = V[i_2^{\mathfrak{e}},-j] = v$. There is a partition of $\mathcal{V}_{-j}$ into two subsets, $\mathcal{V}_{-j}^1$ and $\mathcal{V}_{-j}^2$, corresponding to these two possibilities. There is a one-to-one correspondence between $\mathcal{E}_j$ and pairs of rows from $\mathcal{V}_{-j}^2$; and if $v \in \mathcal{V}_j^2$ corresponds to the edge $\mathfrak{e}$ then $\mathbb{P}\{X_{-j} = v\} = P\mathfrak{e}$.

Once we know $X_{-j} = v$, the value of I is either uniquely determined (if $v \in \mathcal{V}_{-j}^1$) or is one of the two vertices of an edge $\mathfrak{e}$ in $\mathcal{E}_j$ with conditional probabilities $P\{i_1^{\mathfrak{e}}\}/P\mathfrak{e}$ and $P\{i_2^{\mathfrak{e}}\}/P\mathfrak{e}$. Thus

$$\operatorname{var}(X_j \mid X_{-j} = v) = \begin{cases} 0 & \text{if } v \in \mathcal{V}_j^1\\ P\{i_1^{\mathfrak{e}}\}P\{i_2^{\mathfrak{e}}\}/(P\mathfrak{e})^2 & \text{if } v \in \mathcal{V}_j^2.\end{cases}$$

Averaging over the choice of edge corresponding to v we get

$$\sum_{j=1}^{n}\mathbb{P}\operatorname{var}(X_j \mid X_{-j}) = \sum_{j=1}^{n}\sum_{\mathfrak{e}\in\mathcal{E}_j} P\mathfrak{e}\left(P\{i_1^{\mathfrak{e}}\}P\{i_2^{\mathfrak{e}}\}/(P\mathfrak{e})^2\right) \le \sum_{e\in\mathcal{E}} P\{i_2^{\mathfrak{e}}\}.$$

As $\mathfrak{e}$ ranges over all edges, $i_2^{\mathfrak{e}}$ visits each vertex at most $\mathbb{D}$ times. The last sum
□ is at most $\mathbb{D}\sum_{i\le N} P\{i\} = \mathbb{D}$.

Proof of Theorem <31>. Let $\mathbb{J}$ be a subset of $1{:}n$. The uniform distribution on $1{:}N$ induces a distribution P on $\mathcal{V}_{\mathbb{J}}$, with $Pv = N_v/N$ if v appears N_v times as a row of $V[,\mathbb{J}]$. A submatrix of $V[,\mathbb{J}]$ with only one copy of each v from $\mathcal{V}_{\mathbb{J}}$ cannot have a shatter dimension larger than $\mathbb{D}(V)$. Applying Lemma <35> to that matrix we get

$$\sum\nolimits_{j\in\mathbb{J}} \mathbb{P}\text{var}\left(X_j \mid X_{\mathbb{J}\backslash\{j\}}\right) \le \mathbb{D}.$$

> Remark. You should not be worried by the fact that the rows of the new matrix might not be ϵ-separated. In fact, the proof of the Lemma made no use of the separation assumption. It would have been more precise to state the Lemma as an assertion about binary matrices with a given shatter dimension.

Sum the last inequality over all possible subsets $\mathbb{J}$ of $1{:}n$ with $\#\mathbb{J} = m+1$, for a value of m that will be specified soon.

$$\sum_{\#\mathbb{J}=m+1}\sum_{j\in\mathbb{J}} \mathbb{P}\text{var}(X_j \mid X_{\mathbb{J}\backslash\{j\}}) \le \binom{n}{m+1}\mathbb{D}.$$

Write $\mathbb{K}$ for $\mathbb{J}\backslash\{j\}$. Every $(j,\mathbb{K})$ pair with $j \notin \mathbb{K}$ and $\#\mathbb{K} = m$ appears exactly once in the double sum. The inequality is equivalent to

<36> $$\sum_{\#\mathbb{K}=m}\sum_{j\notin\mathbb{K}} \mathbb{P}\text{var}(X_j \mid X_{\mathbb{K}}) \le \binom{n}{m+1}\mathbb{D}.$$

As a check, note that the double sums involve $(m+1)\binom{n}{m+1} = (n-m)\binom{n}{m}$ terms.

For a fixed $\mathbb{K}$ and a fixed v in $\mathcal{V}_{\mathbb{K}}$, the distribution of I conditional on $X_K = v$ is uniform over the rows of the $N(v,\mathbb{K}) \times (n-m)$ submatrix $V_{v,\mathbb{K}}$ of $V[,-\mathbb{K}]$ consisting of those rows for which $V[i,\mathbb{K}] = v$. Moreover,

$$\sum\nolimits_{j\notin\mathbb{K}} |V[i_1,j] - V[i_2,j]| = \sum\nolimits_{j\le n} |V[i_1,j] - V[i_2,j]| \ge n\epsilon$$

for rows $i_1 < i_2$ of $V_{v,\mathbb{K}}$. Apply Lemma <32> to that submatrix.

$$\sum\nolimits_{j\notin\mathbb{K}} \text{var}(X_j|X_{\mathbb{K}} = v) \ge \tfrac{1}{2}n\epsilon\left(1 - N(v,\mathbb{K})^{-1}\right).$$

> Remark. The proof of Lemma <32> is valid when $N(v,\mathbb{K}) = 1$, even though the ϵ-separation property is void for a matrix with only one row. In any case, the lower bound becomes zero when $N(v,\mathbb{K}) = 1$.

Average over the possible values for $X_{\mathbb{K}}$, using the fact that $\mathbb{P}\{X_{\mathbb{K}} = v\} = N(v,\mathbb{K})/N$ for each v in $\mathcal{V}_{\mathbb{K}}$.

<37> $$\sum_{j\notin\mathbb{K}} \mathbb{P}\text{var}(X_j|X_{\mathbb{K}}) \ge \tfrac{1}{2}n\epsilon \sum_{v\in\mathcal{V}_{\mathbb{K}}} \left(N(v,\mathbb{K})/N - N^{-1}\right) = \tfrac{1}{2}n\epsilon\left(1 - \#\mathcal{V}_{\mathbb{K}}/N\right)$$

From Theorem <8>,

$$\begin{aligned}\#\mathcal{V}_{\mathbb{K}} \le p(m,\mathbb{D}) &:= \binom{m}{0} + \ldots + \binom{m}{\mathbb{D}} \\ &\le (em/\mathbb{D})^{\mathbb{D}} \qquad \text{if } m \ge \mathbb{D}.\end{aligned}$$

Sum over all $\binom{n}{m}$ subsets $\mathbb{K}$ of size m to get the companion lower bound to <36>.

$$\binom{n}{m}\tfrac{1}{2}n\epsilon\left(1 - p(m,\mathbb{D})/N\right) \le \sum_{\#\mathbb{K}=m}\sum_{j\notin\mathbb{K}} \mathbb{P}\text{var}(X_j \mid X_{\mathbb{K}})$$

Together the two bounds imply

$$\binom{n}{m}\tfrac{1}{2}n\epsilon\left(1 - \frac{p(m,\mathbb{D})}{N}\right) \le \binom{n}{m+1}\mathbb{D},$$

which rearranges to

$$N \le p(m,\mathbb{D}) \Big/ \left(1 - \frac{2(n-m)}{n\epsilon(m+1)}\mathbb{D}\right).$$

We could try to optimize over m immediately, as Haussler did, to bound N by a function of ϵ, $\mathbb{D}$, and n, then take a supremum over n. However, it is simpler to note that all the conditions of the Theorem apply to the $N \times \ell n$ matrix obtained by binding together ℓ copies of V. Thus the last inequality also holds if we replace n by ℓn, for an arbitrarily large ℓ. Letting ℓ tend to infinity with m fixed, we then eliminate n from the bound.

$$N \le p(m,\mathbb{D}) \Big/ \left(1 - \frac{2\mathbb{D}}{\epsilon(m+1)}\right)$$

If we choose $m = \lfloor 2(\mathbb{D}+1)/\epsilon \rfloor$ then $m \ge \mathbb{D}$ and the upper bound for N is smaller than

$$\left(\frac{em}{\mathbb{D}}\right)^{\mathbb{D}} \frac{\epsilon(m+1)}{\epsilon(m+1) - 2\mathbb{D}} \le (2e/\epsilon)^{\mathbb{D}} \left(1 + \mathbb{D}^{-1}\right)^{\mathbb{D}} \frac{2\mathbb{D}+2}{2\mathbb{D}+2-2\mathbb{D}},$$

which is smaller than the bound stated in the Theorem.

10. Problems

[1] Define $g(x) = x/(\log x)$ for $x > 1$.

(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 $y \ge g(x)$ for some $x \ge e$. Show that $\log y \ge (1 - e^{-1}) \log x$. Deduce that $x \le (1 - e^{-1})^{-1} y \log y$.

[2] Let $\mathcal{D}_1, \dots, \mathcal{D}_k$ be classes of sets each with VC dimension at most VC-dim. 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 $\beta(k)$VC-dim.

[3] (generalized marriage lemma) Let S and T be finite sets and R be a nonempty subset of $S \times T$. Let μ be a finite, nonnegative measure on S and ν be a finite, nonnegative measure on T. Say that a nonnegative measure λ on $S \times T$ is a solution to the (μ, ν, R, S, T) problem if

(a) $\lambda R^c = 0$

(b) $\lambda\left(\{i\} \times T\right) \le \mu\{i\}$ for all $i \in S$

(c) $\lambda\left(S \times \{j\}\right) \le \nu\{j\}$ for all $j \in T$.

Write R_j for $\{i \in T : (i, j) \in R\}$ and $R_{\mathbb{J}}$ for $\cup_{j \in \mathbb{J}} R_j$ for subsets $\mathbb{J}$ of S. By the following steps, show that there exists a solution for which all the inequalities in (c) are actually equalities if and only if

$$(*) \qquad \nu(\mathbb{J}) \le \mu\left(R_{\mathbb{J}}\right) \qquad \text{for all } \mathbb{J} \subseteq T.$$

(i) If λ is a solution with equalities in (c) for every j, then

$$\nu(\mathbb{J}) = \lambda(S \times \mathbb{J}) = \lambda\left(R_{\mathbb{J}} \times \mathbb{J}\right) \le \mu(R_{\mathbb{J}}) \qquad \text{for each } \mathbb{J} \subseteq T.$$

(ii) Now suppose $(*)$ holds. Let λ be a maximal solution to the (μ, ν, R, S, T) problem, that is, a solution for which λR is as large as possible. Show that there cannot exist an (i, j) in R for which $\lambda\left(\{i\} \times T\right) < \mu\{i\}$ and $\lambda\left(S \times \{j\}\right) < \nu\{j\}$, for otherwise λR could be increased by adding some more mass at (i, j).

(iii) Deduce that there must exist at least one j for which equality holds in (c), for otherwise $\mu R_T = \sum_i \lambda\left(\{i\} \times T\right) = \lambda R < \nu T$, contradicting $(*)$.

(iv) Without loss of generality, suppose $\lambda\left(S \times \{1\}\right) = \nu\{1\}$. Define $\overline{R} := R \backslash \left(T \times \{1\}\right)$. Let $\overline{\lambda}$ be the restriction of λ to $\overline{R}$ and $\overline{\nu}$ be the restriction of ν to $S\backslash\{1\}$. Define λ_1 to be the measure on S for which $\lambda_1\{i\} = \lambda\{(i,1)\}$. Define $\overline{\mu} = \mu - \lambda_1$. Show that $\overline{\lambda}$ is a maximal solution to the $(\overline{\mu}, \overline{\nu}, \overline{R}, S\backslash\{1\}, T)$ problem. Hint: If there were another solution γ with $\gamma(\overline{R}) > \overline{\lambda}(\overline{R})$, then the measure obtained by pasting together γ and λ_1 would give a solution to the original problem with total mass strictly greater than λR.

(v) Repeat the argument from (iv), but starting from $\overline{\lambda}$ as a maximal solution to the $(\overline{\mu}, \overline{\nu}, \overline{R}, S\backslash\{1\}, T)$ problem, to deduce equality in (c) for another column. And so on.

[4] Show that assertion of Problem [3] is still valid if the measures λ, μ, and ν are retricted to take nonnegative integer values.

[5] Suppose an $N \times n$ binary matrix V has shatter dimension VC-dim. Let $\mathcal{E}$ be the corresponding edge set, as defined in Section 9. Show that $\#\mathcal{E} \le N$VC-dim by following these steps.

First show that the downshift operation used for the proof of Theorem <8> cannot decrease the number of edges. Suppose V is transformed to V^* by a downshift of the first column. Suppose (i_1, i_2) is an edge of V but not of V^*.

(i) Suppose $V[i_1,]$ and $V[i_2,]$ differ only in the jth position. Show that $j > 1$, for otherwise $V^*[i_1,]$ and $V^*[i_2,]$ would differ only in the first position.

(ii) Show that $V[i_1, 1] = V[i_2, 1] = 1$, for otherwise the downshift could not change either row.

(iii) Suppose $V[i_1,] = (1, w)$ and $V^*[i_1,] = (0, w)$ and $V[i_2,] = V^*[i_2,] = (1, y)$, where w and y differ only in the jth position. Show that $(0, y) = V[i_0,] = V^*[i_0,]$ for some i_0.

(iv) Deduce that (i_0, i_1) is an edge of V^* but not an edge of V.

(v) Explain why every downshift that destroys an edge creates a new one.

Now suppose that V^* is the result of not just one downshift, but that it is the matrix that remains when no more downshifting is possible. Let $\mathcal{E}^*$ be its set of edges.

(vi) Explain why $\#\mathcal{E}^* \ge \#\mathcal{E}$.

(vii) Argue as in the proof of Theorem <8> to show that no row of V^* can contain more than $\mathbb{D}$ ones.

(viii) Define $\psi : \mathcal{E}^* \to 1\!:\!N$ by taking $\psi(\mathfrak{e})$ as the row corresponding to the vertex of $\mathfrak{e}$ with the larger number of ones. Show that $\psi^{-1}(i)$ contains at most $\mathbb{D}$ edges, for every i. Hint: How many different edges can be created by discarding a 1 from $V^*[i,]$?

(ix) Deduce that $\#\mathcal{E}^* \le N\mathbb{D}$.

[6] Suppose V be an $N \times n$ binary matrix with shatter dimension $\mathbb{D}$ and edge set $\mathcal{E}$. Show that there exists a map $\psi : \mathcal{E} \to 1\!:\!N$ such that

(a) $\psi(\mathfrak{e})$ is one of the two vertices on the edge $\mathfrak{e}$,

(b) $\#\psi^{-1}(i) \le$ VC-dim for every i,

by following these steps.

(i) Let $\mathcal{E}_0$ be a subset of $\mathcal{E}$, with vertices all contained in $\mathbb{I} \subseteq 1{:}N$. Apply the result from Problem [5] to show that

$$\#\mathcal{E}_0 \leq \text{ number of edges of } V[\mathbb{I},] \leq \mathbb{D}\#\mathbb{I}.$$

(ii) Invoke the result from Problem [4] with

$$S = 1{:}N \qquad T = \mathcal{E} \qquad R_{\mathfrak{e}} = \text{the pair of vertices of } \mathfrak{e}$$

and $\nu\{\mathfrak{e}\} = 1$ for each $\mathfrak{e}$ and $\mu\{i\} = \mathbb{D}$ for each i. Show that the measure λ puts a single atom of mass 1 in each $R_{\mathfrak{e}}$.

(iii) Let $\psi(\mathfrak{e})$ be the vertex in $R_{\mathfrak{e}}$ where λ puts its mass. Show that

$$\#\psi^{-1}(i) = \lambda\left(\{i\} \times \mathcal{E}\right) \leq \mu\{i\} \leq \mathbb{D}.$$

11. Notes

Get the history on VC Lemma straight. VC? Sauer (1972) ? Frankl?

Section 3: Dudley for sets; Pollard (1982) via Le Cam for functions.

Section 4: Haussler. Explain why result is interesting.

Downshift technique: compare with original VC argument. Talagrand? Haussler, and refs. Compare with Ledoux & Talagrand (1991, p. 420) and different explanation in Talagrand (1987). Check the 1987 survey article of Frankl, cited by Haussler.

Cover (1965) for exact bound for half-spaces. More comments on suboptimality of VC bound? What does the cubic vs quadratic say about attempts to squeeze the best results from the VC bound?

Vapnik & Červonenkis (1971) Cite other VC paper too.

Steele (1975) Cite Steele paper, and Sauer, and Frankl.

Talagrand (2003)

References

Cover, T. M. (1965), 'Geometric and statistical properties of systems of linear inequalities with applications to pattern recognition', *IEEE Transactions on Elec. Comp.*

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

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.

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

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

Mendelson, S. & Vershynin, R. (2003), 'Entropy and the combinatorial dimension', *Inventiones mathematicae* **152**, 37–55.

Pollard, D. (1982), 'A central limit theorem for k-means clustering', *Annals of Probability* **10**, 919–926.

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

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

Stengle, G. & Yukich, J. (1989), 'Some new Vapnik-Chervonenkis classes', *Annals of Statistics* **17**, 1441–1446.

Talagrand, M. (1987), ‘Donsker classes and random geometry’, *Annals of Probability* **15**, 1327–1338.

Talagrand, M. (2003), ‘Vapnik-Chervonenkis type conditions and uniform Donsker classes of functions’, *Annals of Probability* **31**, 1565–1582.

van den Dries, L. (1998), *Tame Topology and O-minimal Structures*, Cambridge University Press.

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