# Sauer-Shelah Lemma and its Application to Codes

Stefano Della Fiore

Università degli studi di Brescia

November 26, 2020

# Overview

# Learning Theory - VC Dimension

## Definition

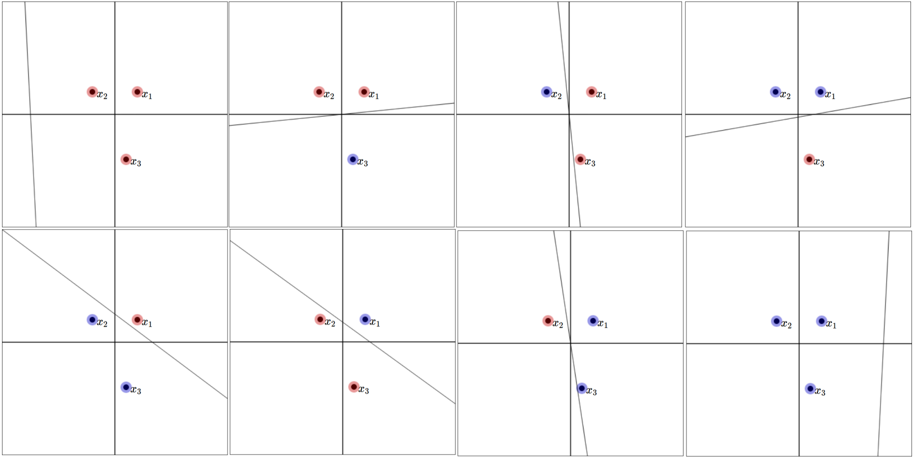
In a binary classification setting with labels $\mathcal{Y} = \{-1, +1\}$, a set of $n$ points $S = \{x_1, \dots, x_n\}$ is said to be **shattered** by a function class $\mathcal{F}$ if

$$\forall y \in \mathcal{Y}^n, \exists f \in \mathcal{F} \text{ such that } f(x_i) = y_i \text{ for } i = 1, \dots, n.$$

The **VC-dimension** of a function class $\mathcal{F}$ is the size of the largest set of points that can be shattered by $\mathcal{F}$.

# Learning Theory - VC Dimension - 2

Here, we illustrate how the class of linear classifiers shatters a set of 3 points in $\mathbb{R}^2$. No set of 4 points is shattered by a linear classifier in $\mathbb{R}^2$ then the VC-Dimension of these classifiers is equal to 3.



# Density of a Family

## Definition

The **density** of a family $\mathcal{F}$ of subsets of a set $S$ is the largest number $d$ such that there exists a set $A$ with $|A| = d$ and $|\mathcal{F} \cap A| = |\{F \cap A : F \in \mathcal{F}\}| = 2^d$.

$$\mathcal{F} = \boxed{\begin{array}{c} \{1\} \\ \{3\} \\ \{2,3\} \\ \{1,2,3\} \end{array}}$$

$$A = \{1,2\}$$
$$A \cap \{3\} = \emptyset$$
$$A \cap \{1\} = \{1\}$$
$$A \cap \{2,3\} = \{2\}$$
$$A \cap \{1,2,3\} = \{1,2\}$$

Figure: Example of a family with density equal to 2

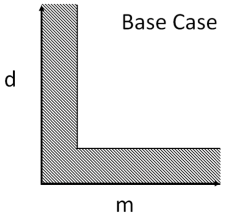
# Formulation

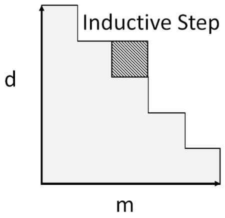
**Lemma (Sauer-Shelah)**

*If the density of the family $\mathcal{F}$ of subsets of a set $S$ with $|S| = m$ is equal to $d$ then*

$$|\mathcal{F}| \leq \sum_{i=0}^{d} \binom{m}{i}.$$

The proof is done by induction on $m + d$. In the inductive step we show the lemma holds for any m, d with $m + d = k$ for some constant k assuming that it holds for all $m, d$ with $m + d < k$



# Proof by induction on m+d

Let $\Phi_d(m) := \sum_{i=0}^{d} \binom{m}{i}$. Note that $\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}$.
If the density of the family $\mathcal{F}$ of subsets of a set $S$ with $|S| = m$ is $d$ then $|\mathcal{F}| \leq \Phi_d(m)$.

## Proof.

The proof is done by induction on $m + d$, the $m = 0$ and $d = 0$ cases are trivial. Consider $m, d > 0$ and fix an arbitrary element $p \in S$. Define

$$\mathcal{F}_p = \{F \in \mathcal{F} : p \notin F, \{p\} \cup F \in \mathcal{F}\}$$

Then,

$$|\mathcal{F}| = |\mathcal{F} \cap \{S - p\}| + |\mathcal{F}_p|.$$

Since the density of $\mathcal{F}_p$ is at most $d - 1$ we have by induction

$$|\mathcal{F}| \leq \Phi_d(m-1) + \Phi_{d-1}(m-1) = \Phi_d(m).$$

$\square$

# Asymptotically is better

If $\mathcal{F}$ is a family of subsets of a set $|S| = n$ with $n$ large enough has density $d_n \leq n/2$ then

$$|\mathcal{F}| \leq \sum_{i=0}^{d_n} \binom{n}{i} \leq 2^{n \cdot h(d_n/n) + o(n)}$$

rewritten in term of the "rate" of $\mathcal{F}$ we have

$$1/n \log |\mathcal{F}| \leq h(d_n/n) + o(1)$$

where $h(\alpha) = -\alpha \log \alpha - (1-\alpha) \log(1-\alpha)$ is the binary entropy function.

# Applied to codes is even better

A family $\mathcal{F}$ of subsets of $S = \{1, \dots, n\}$ can be seen as a code $C_n$.

ex. $F = \{2, 5, n-1\} \in \mathcal{F} \longleftrightarrow (0,1,0,0,1,0,\dots,0,1,0) \in C_n$.

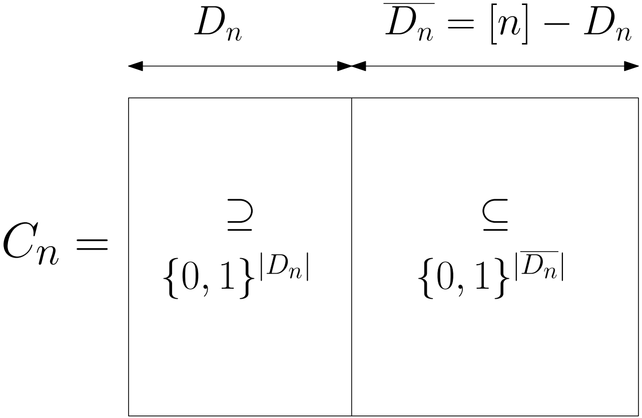
By Sauer-Shelah Lemma there is a set of coordinate $D_n$ satisfying

$$\lim_{n\to\infty} |D_n|/n \geq h^{-1}(R),$$

where $R = \limsup_{n\to\infty} 1/n \log |C_n|$.

# Applied to codes is even better - 2

The set of coordinates $D_n$ has the property that




i.e., the union of all the projections in $D_n$ are $2^{|D_n|}$.

# Near sunflowers and String quartets

## Definition

We say that the binary code $C_n$ with codewords of length $n$ is **r-near-sunflower free** if for all $r$ distinct codewords of $C_n$ there exists a coordinate in which the numbers of 1's is between 2 and $r-2$.

$$
\begin{array}{c}
1 \quad \cdots \quad n \\
\longleftrightarrow
\end{array}
$$

$$
C_n = \begin{array}{c|ccccc|}
 & \vdots & \vdots & \cdots & \vdots & \vdots \\
x_1 & 0 & 1 & \cdots & 0 & 0 \\
x_2 & 0 & 1 & \cdots & 1 & 1 \\
x_3 & 0 & 0 & \cdots & 1 & 0 \\
x_4 & 1 & 1 & \cdots & 0 & 0 \\
 & \vdots & \vdots & \cdots & \vdots & \vdots
\end{array}
$$

Example
of a
4-near sunflower free

# An upper bound on the rate of 4-near-sunflower-free codes

Let $C_n$ be a 4-near-sunflower-free code with maximum cardinality. Sauer's lemma gives us a set of coordinates $D_n$ with $\lim_{n\to\infty} |D_n|/n \geq h^{-1}(R)$ with all the $2^{|D_n|}$ projections.

Suppose that $|C_n| = 2^{nR} > 2^{n(1-h^{-1}(R))}$, by the pigeonhole principle and Sauer's Lemma we have

$$\begin{array}{cc} \overleftrightarrow{\quad D_n \quad} & \overleftrightarrow{\quad \overline{D_n} = [n] - D_n \quad} \end{array}$$

$$C_n = \begin{array}{|l|c|}
\hline
x \quad 0 \quad 0 \; \cdots \; 0 & a \\
y \quad 1 \quad 0 \; \cdots \; 0 & a \\
z \quad 1 \quad 1 \; \cdots \; 0 & a \\
t \quad 1 \quad 0 \; \cdots \; 1 & b \\
\qquad \vdots & \vdots \\
\hline
\end{array}$$

# An upper bound on the rate of 4-near-sunflower free codes

Let $C_n$ be a 4-near-sunflower free with maximum cardinality then its cardinality must satisfy the following inequality

$$|C_n| \leq 2^{n(1-h^{-1}(R))}$$

that in terms of rates is

$$R \leq 1 - h^{-1}(R)$$

which restated is

$$R \leq h(1-R).$$

The largest value of $R$ for which the previous inequality holds is $\approx 0.773$.

# Focal families and further improvements

## Theorem (Alon et al. 2020)

*Let $R$ be the rate of the largest 4-near-sunflower-free code. Then*

$$R \leq 2/3 = 0.\overline{6}$$

## Problem

*Try to mix the two ideas. Sauer's Lemma + Focal families.*

# References


- N. Sauer, "On the density of families of sets", JCT Ser. A, 13 pp. 145-147, *1972*.
- S. Shelah, "A combinatorial problem: Stability and order for models and theories in infinitary languages", Pacific J. Math. 41, pp. 247-261, *1972*.
- Vapnik, V. N and Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities, Theory of Probability and its Applications 16", pp. 264-280, *1971*.
- N. Alon, J. Korner, A. Monti, "String quartets in binary", Combinatorics Probability and Computing, *2002*.
- N. Alon, R. Holzman, "Near-sunflowers and focal families", ArXiv, https://arxiv.org/abs/2010.05992, *2020*.