# LP - techniques

Set - Cover : . NP- hard to approximate beyond $c \log n$.

· $O(\log n)$ - approximation algorithms.

Independent Set : NP- hard to approximate beyond $n^{1-\varepsilon}$.

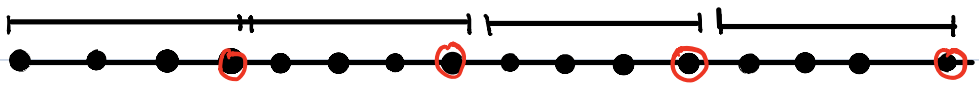
- We obtain better algorithms for geometric settings.
- Our algorithms will be via LP- rounding.
- Before we do this, we introduce a seemingly unrelated combinatorial problem, called $\varepsilon$-nets.
- These have applications in data structures, etc.

# $\varepsilon$-nets

- Let $(X, \mathcal{R})$ be a set system. $|X| = n$.
- For $\varepsilon > 0$, an $\varepsilon$-net is a set $Y \subseteq X$, such that $R \cap X \neq \emptyset \quad \forall \quad R \in \mathcal{R} : |R| > \varepsilon n$.

ie: $Y$ is a subset of points that "hits" all large subsets.

- Example: Let $X$ be a set of $n$ points on $\mathbb{R}$. Let $\mathcal{R}$ be a disjoint, contiguous sequence of $\varepsilon n$ points:



- Here $n = 16$.
- Let $\varepsilon = \frac{1}{4}$.
- Each interval drawn above is therefore "large."
- Choosing every $4^{th}$ point yields an $\varepsilon$-net for this instance.
- $|Y| = 4$.
- The example also shows that we require atleast $1/\varepsilon$ points.

- More gererally, let $X$ be a set of $n$ points on the real line, and let $R$ be a collection of intervals.

- For $\varepsilon > 0$, how can we construct an $\varepsilon$-net?

- Construct $Y$ by picking every $\lfloor 1/\varepsilon n \rfloor^{th}$ point from $X$.

- Every interval of length $\geq \varepsilon n$ will be hit by $Y$.

- We obtain an $\varepsilon$-net of size $\left\lceil \frac{1}{\varepsilon} \right\rceil$

- Note that the size of the $\varepsilon$-net is independent of $n$.

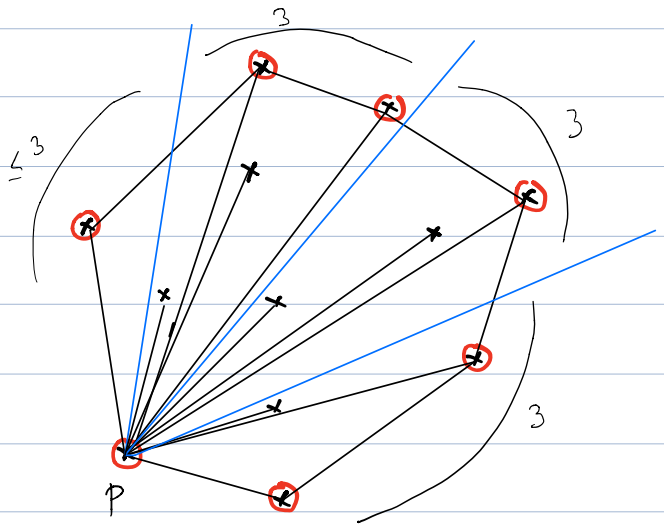
- Let us try to construct an $\varepsilon$-net for other geometric set systems:

- Let $X$ = $n$ points in $\mathbb{R}^2$

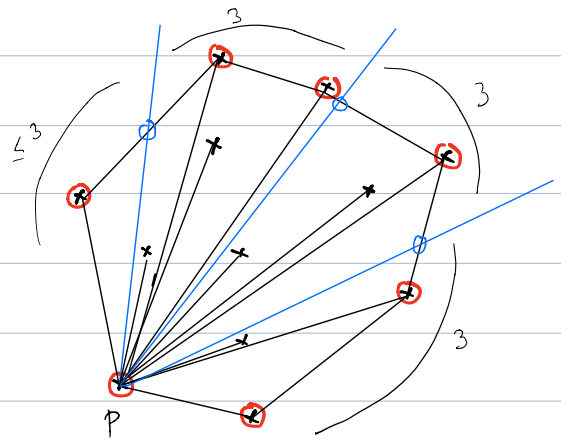
  $R$ = half-spaces in $\mathbb{R}^2$.

- The circled points are on the convex hull of $X$.
- Every halfspace containing $\geq \varepsilon n$ points must be hit.

- Start with an arbitrary point $p$ on the convex hull, and sort the points of $X$ radially with respect to $p$.




- Partition into groups containing $\lceil \varepsilon n \rceil - 1$ points in this order.
- # Groups $\leq n / (\lceil \varepsilon n \rceil - 1) \approx 1/\varepsilon$




Pick P into the $\varepsilon$-net; and restrict attention to those half-spaces not containing P. Then,

Obs 1:

- Any half space containing $\geq \varepsilon n$ points must contain points from at least 2 groups.

Obs 2:

- Any half-space containing $\geq \varepsilon n$ points, & thus points from 2 groups must contain the point intersecting the ray defining some group boundary, & the cvx hull (depicted as blue circles in the figure).

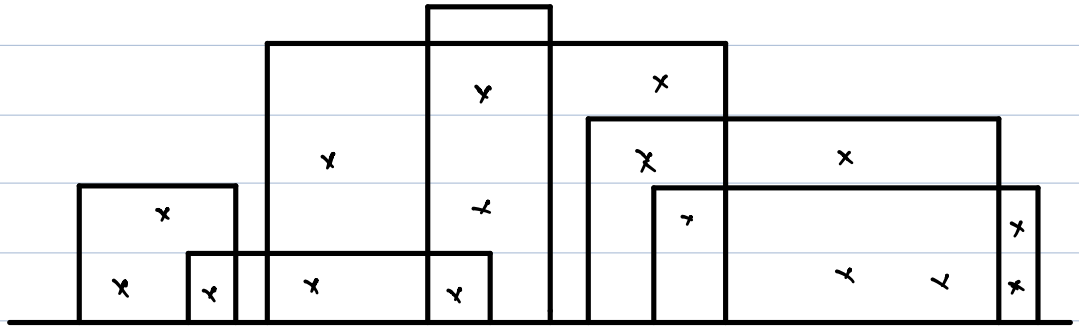
Obs 3: For each blue point, add the two vertices of the edge of the cvx hull on which it lies.

This forms an $\varepsilon$-net of size at most: $\lceil 2/\varepsilon \rceil + 1$.

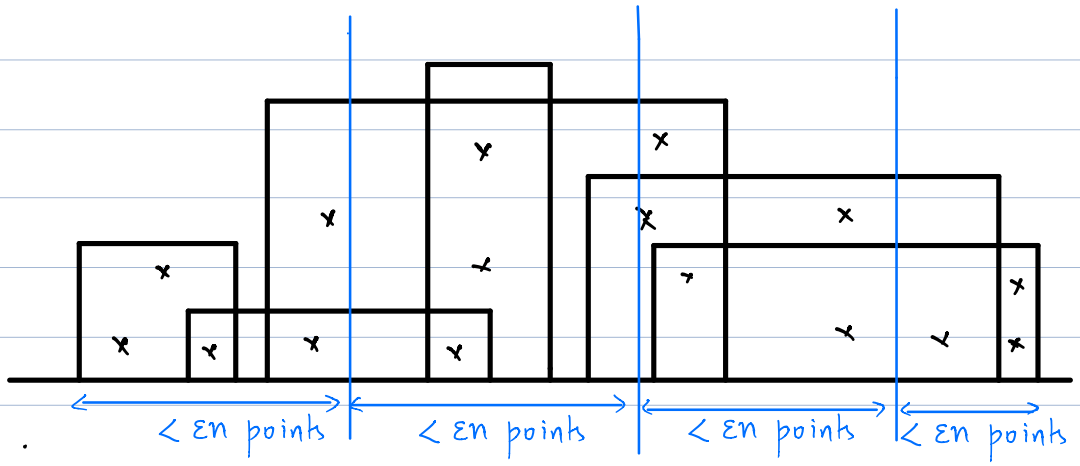
Example:

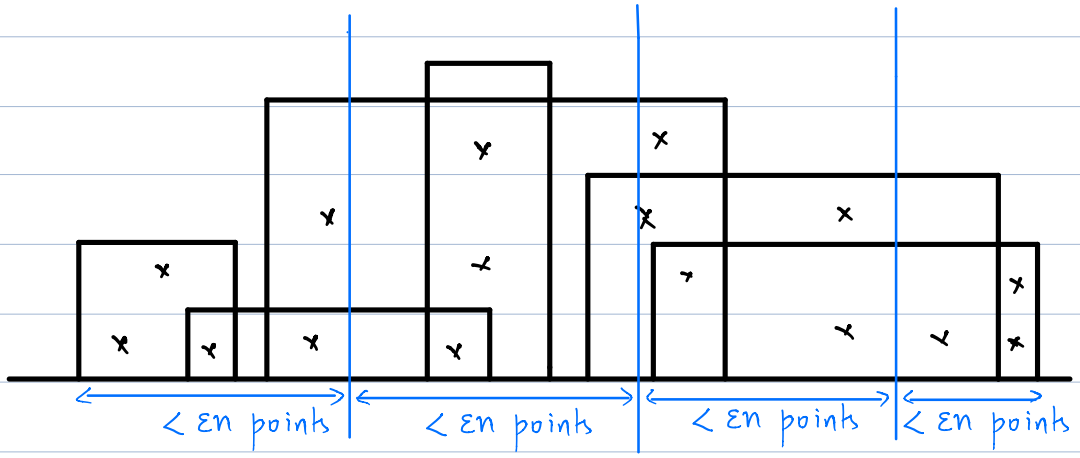
Let $X$ = points in $\mathbb{R}^2$.

$R$ = Sky-lines / rectangles on the x-axis.



- Let us use the earlier idea of splitting into slabs, each containing $< \varepsilon n$ points.

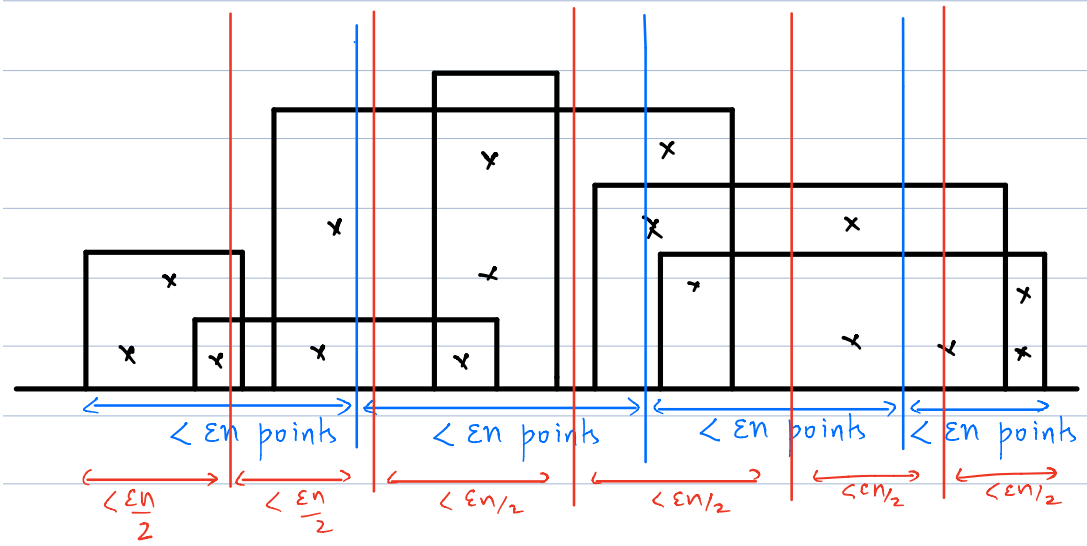




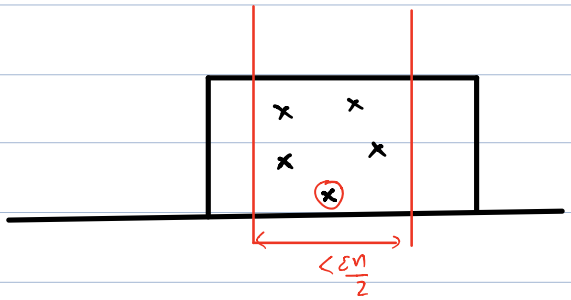

- Any slab containing $\geq \varepsilon n$ points cannot lie entirely in a single slab & must cross a slab boundary.

- Now it is not clear how to proceed.

- Suppose we make a finer partition: each slab containing $< \frac{\varepsilon n}{2}$ points:




Obs 1: Every rectangle containing $> \varepsilon n$ points must contain a point from slab it completely crosses.




Claim: Choosing the lowest point in each slab is an $\varepsilon$-net of size $\left\lceil \frac{2}{\varepsilon} \right\rceil$.

# Random Sampling

- The techniques we developed to compute $\varepsilon$-nets thus far seem very specialized to the geometric objects.

- They do not seem to generalize for objects such as disks, or rectangles, or other more complex shapes.

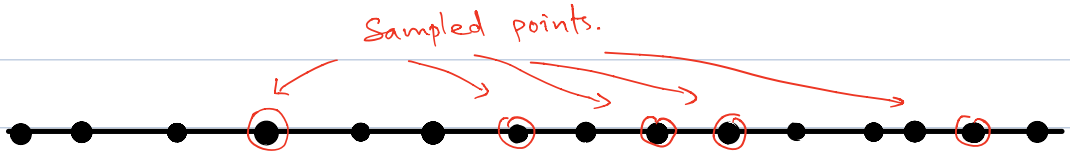
- We now revisit the $\varepsilon$-net for intervals & give a different construction of an $\varepsilon$-net that generalizes.

- Given: $X$ = Points on $\mathbb{R}$.
  $R$ = intervals of $\mathbb{R}$.

- We give a randomized construction.

- Pick each point in $X$ uniformly & independently with probability $p$ ($p$ to be determined later).

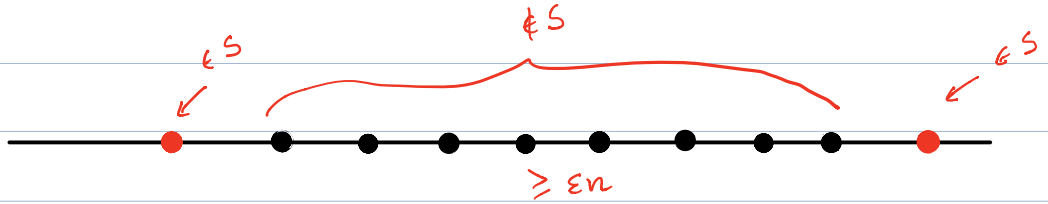
- Let $S$ be the set of sampled points.



The expected number of points sampled is:

$\mathbb{E}[|S|] = np$ ——— (1)

- Note that if we hit <u>every</u> interval of length $\varepsilon n$, we obtain an $\varepsilon$-net for $(X, \mathcal{R})$.
- The sample $S$ fails to be an $\varepsilon$-net, if there is a contiguous sequence of at least $\varepsilon n$ points between two consecutive sampled points of $S$:




- Call such an empty interval a <u>canonical interval</u>.
- The probability of obtaining such a <u>canonical interval</u> is that
  - (i) the two extreme points of the defined interval are chosen in $S$.
  - (ii) No point between them is chosen in $S$.
- The probability that we get an empty interval of length $t$ is: $p^2(1-p)^t$ → none of the points in-between are chosen in $S$.
  ↳ the two end-points are chosen in $S$

- For $t \geq \varepsilon n$, this probability is:

$$p^2(1-p)^t \leq p^2 e^{-pt} \leq p^2 e^{-p\varepsilon n} \quad \text{——— (1)}$$

- The expected number of empty intervals of length at least $\varepsilon n$ is therefore:

$$n^2 p^2 e^{-p\varepsilon n} \quad \text{——— (2)}$$

- Choose $p = \frac{c}{\varepsilon n}$.

Then, the expected number of empty intervals is at most: (plugging in the value of $p$ into (2))

$$n^2 \left(\frac{c}{\varepsilon^2 n^2}\right) \exp\left(-c\frac{\varepsilon n}{\varepsilon n}\right) = \frac{c}{\varepsilon^2} e^{-c}.$$

- what we would like is the expectation to be $< 1$.

- So, let us try $p = \frac{c}{\varepsilon n} \log \frac{1}{\varepsilon}$. Then, the expected number becomes:

$$n^2 \left(\frac{c^2}{\varepsilon^2 n^2}\right) \log^2 \frac{1}{\varepsilon} \exp\left(-\frac{c\varepsilon n}{\varepsilon n} \cdot \log\left(\frac{1}{\varepsilon}\right)\right)$$

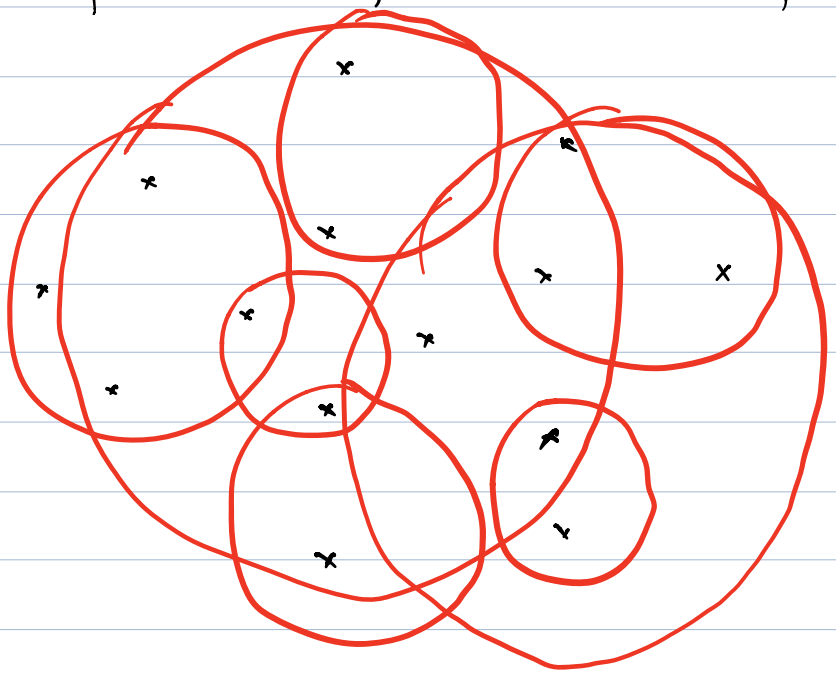
$$= \frac{c^2}{\varepsilon^2} \log^2 \frac{1}{\varepsilon} \exp\left(-c \log \frac{1}{\varepsilon}\right)$$

$$= \frac{c^2}{\varepsilon^2} \log^2 \frac{1}{\varepsilon} \, \varepsilon^c < 1 \quad \text{for } c \text{ large enough.}$$

- The expected size of the sample we picked is obtained by plugging the value of $p$ into (1).
- This is: $np = O\left(\frac{1}{\varepsilon} \log \frac{1}{\varepsilon}\right)$.

□

Let us try to use this technique to obtain $\varepsilon$-nets for the set system $X \subseteq \mathbb{R}^2$, $R$ = disks in the plane.

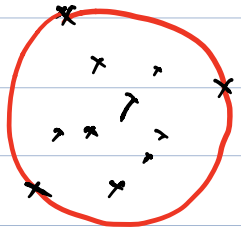


- We proceed as before: Let $S \subseteq X$ be a set of points chosen uniformly & independently with probability $p$ (to be determined later).

- If our set $S$ hits <u>every</u> disk containing $\geq \varepsilon n$ points, then we obtain an $\varepsilon$-net.

- Just as we defined a canonical interval, as one starting at a sampled point, we can define a canonical disk.

- A canonical disk can be defined by 3 input points.



<u>Claim</u>: Any input disk is contained in the union of 2 canonical disks.

<u>Obs</u>:

- If we construct an $\varepsilon/2$-net hitting the canonical disks, we obtain an $\varepsilon$-net for the original set of disks.

- Let us use $\varepsilon$ as before.

- The probability of obtaining an empty canonical disk with $t$ points inside is:

$$p^3(1-p)^t$$

→ No point in the interior is chosen.

↳ The three points defining the canonical disk are chosen.

- Hence, the expected number of empty canonical disks is at most:

$$n^3 p^3 e^{-p\varepsilon n}.$$

- Put $p = \frac{c}{\varepsilon n}\log\left(\frac{1}{\varepsilon}\right)$, & we obtain that the expected number of empty disks is:

$$n^3\left(\frac{c^3}{\varepsilon^3 n^3}\right)\log^3\left(\frac{1}{\varepsilon}\right)\exp\left(-c\frac{\varepsilon n}{\varepsilon n}\log\frac{1}{\varepsilon}\right)$$

$$\leq \frac{c^3}{\varepsilon^3}\log^3\frac{1}{\varepsilon}\,\varepsilon^c < 1 \quad \text{for } c \text{ large enough.}$$

- Therefore, there is a sample $S$ of size: $np = O\left(\frac{1}{\varepsilon}\log\frac{1}{\varepsilon}\right)$ that is an $\varepsilon$-net.

- The key reason why the above argument worked for disks is that we could define $n^3$ canonical disks of large size.

- Suppose we have a set system, where we could likewise define $n^d$ canonical sets, for some constant $d$.

- Taking $p = \frac{cd}{\varepsilon n}\log\left(\frac{d}{\varepsilon}\right)$, we get that the expected number of empty sets of size $\geq \varepsilon n$ is: $p^d n^d \exp(-\varepsilon n p)$

$$= O\left(\frac{d}{\varepsilon}\log\frac{d}{\varepsilon}\right)$$

- Thus, what we require for the argument to work is that a canonical object can be determined by a constant # points.
- In fact, we will only require that for any large subset $S$ of points, only an $O(|S|^d)$ subsets can be obtained by intersecting with $R$.
- Shatter dimension: Let $(X, R)$ be a set system, $Y \subseteq X$.
Let $R|_Y = \{Y \cap R : R \in R\}$. ie: the induced sets in $Y$.

- Suppose there exists $d \in \mathbb{N}$ such that $|R|_Y| = O(|Y|^d)$, $\forall Y \subseteq X$ then we say that the shatter dimension of $(X, R) = d$.
- We only require that large sets are not shattered:

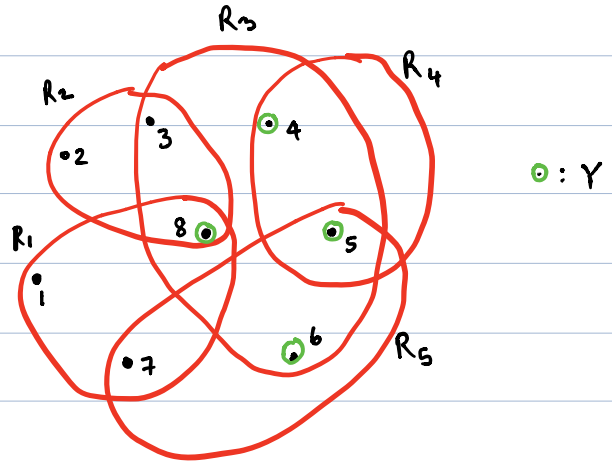
We next introduce a notion that will bound the shatter dimension:

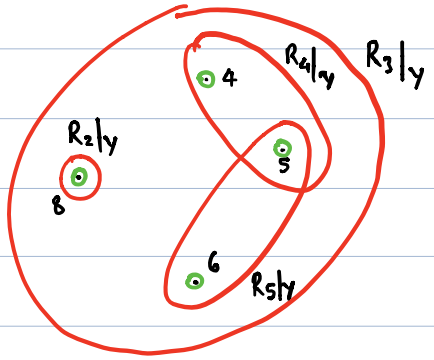
# VC-dimension

- VC-dimension, or Vapnik-Chervonenkis dimension is a measure of the complexity of a set system
- The measure was introduced by Vapnik & Chervonenkis in the context of learning theory in '71.
- This notion has since found wide applicability in several areas, most notably geometry.

- Let $(X, \mathcal{R})$ be a set-system.
- Here we allow both $X$ & $\mathcal{R}$ to be infinite.
- Set-systems are also called range-spaces, & $\mathcal{R}$, the set of ranges.
- Let $Y \subseteq X$ be a subset of size $n$.
- $\mathcal{R}|_Y = \{ R \cap Y : R \in \mathcal{R} \}$. ie: $\mathcal{R}|_Y$ is the induced set-system on $Y$.

- For example:




Observe that $R_1$ & $R_2$ map to the same set in $R|_Y$.

- Given a set $Y \subseteq X$, $|Y| = n$, the total number of potential subsets is $2^{|Y|}$.

- A set $Y \subseteq X$ is said to be <u>shattered</u> if $R_{|Y} = 2^Y$.
  ie: the sets $R$ induce all possible subsets of $Y$.

- In the example above, if we had set $Y = \{5, 8\}$, then
  $R_{|Y} = \{\emptyset, \{8\}, \{5\}, \{8,5\}\}$
  ↳ $R_4$ ↳ $R_2, R_1$ ↳ $R_5$ ↳ $R_3$

- For a set system $(X, \mathcal{R})$, the largest set $Y \subseteq X$ that can be shattered is the VC-dimension of the set system.

Example: $X = \mathbb{R}$, $\mathcal{R} = \{[a, \infty), a \in \mathbb{R}\}$.

Ie: the ground set is the real line, & the sets are all semi-infinite lines to the right.

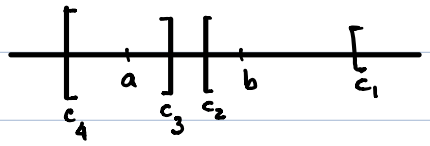
Let $Y = \{a\}$: This set can easily be shattered.

Claim: No set of size 2 can be shattered.

Thus VC-dimension of this set system = 1.

Example: $X = \mathbb{R}$, $\mathcal{R} = \{[a, \infty), (-\infty, b]\}$.

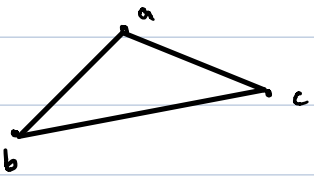
- A set of size 2 can be shattered:




Claim: No set of size 3 can be shattered.

Example: $X = \mathbb{R}^2$, $R$ = all halfspaces.

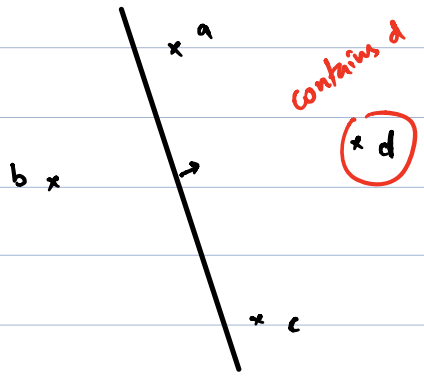
VC-dim $\geq 3$: Given 3 non-collinear points $a, b, c$.



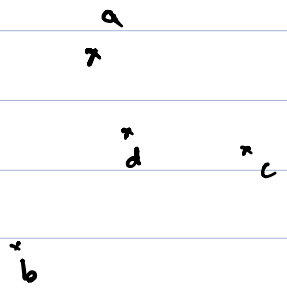

it is easy to find half-spaces forming all possible subsets.

Claim: VC-dim $\leq 3$.

- Let $\{a, b, c, d\}$ be 4 points.
- Either the points are in cvx. position, or one point, say d is contained in the cvx. hull of $\{a, b, c\}$.
- If the four points are in cvx. position:




Any halfspace containing the diagonally opposite points contains one of the other points.

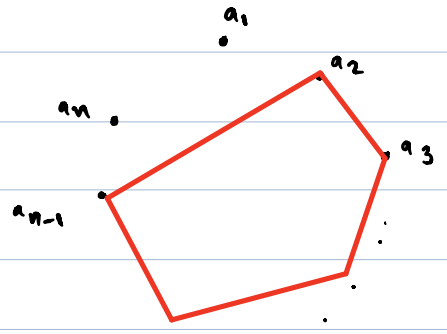



No halfspace contains only $\{d\}$.

Example: $X = \mathbb{R}^2$, $\mathcal{R}$ = all convex polygons.

Claim: VC-dim. of this set system is $\infty$.

Consider $n$ points in cvx. position.




To form a set $\{a_{i_1}, a_{i_2}, \ldots, a_{i_k}\}$ construct a convex polygon whose vertices are $\{a_{i_1}, a_{i_2}, \ldots, a_{i_k}\}$.

- Sets of any size $n$ can be shattered.
- Hence VC-dimension $= \infty$.

- In many geometric situations we will see, VC-dimension will be bounded.

# Shatter function & Shatter dimension.

Consider the following recurrence:

$$\Phi_d(n) = \begin{cases} 1, & n > 0,\ d = 0 \\ 1, & n = 0,\ d \geq 0 \\ \Phi_{d-1}(n-1) + \Phi_d(n-1), & n > 0,\ d > 0 \end{cases}$$

**Lemma:** $\Phi_d(n) = \sum_{i=0}^{d} \binom{n}{i} = O(n^d)$, $d < n$ & $\Phi_d(n) = 2^n$ otherwise.

VC-dim

**Lemma** (Sauer-Shelah) Let $\mathcal{H} = (X, R)$ be a set-system, $\dim(\mathcal{H}) = d$, $|X| = n$. Then, $|R| = O(n^d)$.

**Proof:**

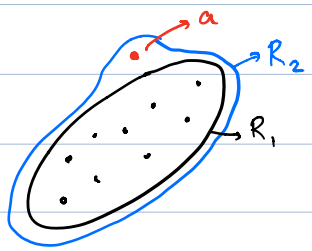
- By induction on $d$, $n$. (We only show the inductive step)
- Suppose the statement is true for all $n$ & $\dim < d$, and for all $n' < n$ & all dim. $d$.
- Let $a \in X$ be an arbitrary element.
- We can define two set systems:

$\mathcal{H}_1 = (X - a, R_{-a})$, where $X - a = X \setminus \{a\}$.

$R_{-a} = \{R \setminus \{a\} : R \in R\}$.

- Note that two sets in $\mathcal{R}$ can be mapped to the same set in $\mathcal{H}$.

Eg:




ie: $R_2 \setminus R_1 = \{a\}$.

Let $\mathcal{R}' = \{ R \in \mathcal{R} : a \in R, \text{ and } R \setminus \{a\} \in \mathcal{R} \}$, ie: all sets of type $R_2$ in the example above.

Thus: $|\mathcal{R}| = |\mathcal{R}_{-a}| + |\mathcal{R}'|$

Define $\mathcal{H}_2 = (X - a, \mathcal{R}')$.

- $\dim(\mathcal{H}_1) \leq d$.
- We show that $\dim(\mathcal{H}_2) \leq d-1$.

  Suppose not: There is a set $A \subseteq X - a$, $|A| = d$, that is shattered by $\mathcal{R}'$.
- This implies: $A \cup \{a\}$ is shattered by $\mathcal{R}$.

- But this implies there is a set $A \cup \{a\}$, $|A \cup \{a\}| = d+1$ that is shattered by $\mathcal{R}$, contradicting the assumption that $\dim(\mathcal{H}) = d$.
- Thus, the number of sets: $\overline{\Phi}_d(n) = \overline{\Phi}_d(n-1) + \overline{\Phi}_{d-1}(n-1)$
  $= O(n^d)$ ∎

# Set Cover for Set Systems with bounded VC-dimension.

- Let $(X, \mathcal{R})$ be a set system with VC-dimension $d$.
- Recall that for the Set Cover problem, we could obtain an $O(\log n)$-approximation.
- Here, we will obtain a better approximation factor for sets of bounded VC-dimension:
- For the sake of uniformity with the earlier presentation, we will consider instead the hitting set problem.
- For a set system $(X, \mathcal{R})$, the hitting set problem is the problem of picking $Y \subseteq X$ of smallest cardinality such that $\forall\, R \in \mathcal{R} \quad R \cap Y \neq \emptyset$.
- We define a <u>dual</u> set system as $(\mathcal{R}, \mathcal{R}_X)$, where $\mathcal{R}$ is the ground-set, and $\mathcal{R}_x = \{R \in \mathcal{R} : x \in R\}$, $x \in X$ & $\mathcal{R}_X = \bigcup_{x \in X} \mathcal{R}_x$.

- The hitting set problem is the set cover problem for the dual set system.

- Further, if the VC-dimension of $(X, \mathcal{S})$ is $d$, then the VC-dimension of $(R, R_X)$ is $O(2^d)$.

- Consider the LP-relaxation for the hitting set problem:

$$\min \sum_{v \in X} x_v$$

subject to,

$$\sum_{v \in S} x \geq 1 \quad \forall S \in \mathcal{S}$$

$$x_v \geq 0. \quad \forall v \in X.$$

- Let $x$ be an optimal solution to the LP, and
- Let $W = OPT_{LP}$ be the optimal solution of this linear program.

- Assign a weight of $\frac{x_v}{W}$ to each $v \in X$
- Since $x$ is a feasible solution to the LP, $\sum x_v \geq 1$.

- Thus, $\sum \frac{x_v}{W} \geq \frac{1}{W} \quad \forall R \in \mathcal{R}$.

ie: Each set has weight $\geq \frac{1}{W}$.

- Since the VC-dimension is $d$, $\exists$ an $\varepsilon$-net of size

$$O\left(\frac{d}{\varepsilon} \log \frac{d}{\varepsilon}\right).$$

- Put $\varepsilon = \frac{1}{W}$, & we obtain a hitting set of size

$$O(dW \log dW) = O(d\,OPT \log OPT)$$

ie: an $O(\log OPT)$-approximation.

Remarks: Note that the algorithm also works in the weighted setting -

- Here, each set $R \in \mathcal{R}$ has a weight $w_R > 0$, and we want to minimize the total weight of the sets chosen that is still a cover of the points.

- This is because each set in the sample is chosen with uniform probability. [Show this]

- The constructions of $\varepsilon$-nets we saw earlier did not pick each set with uniform probability.

- Therefore, the rounding algorithm presented does not yield an $O(1/2)$-approximation even when we have an $\varepsilon$-net of size $O(1/\varepsilon)$.

- Quasi-Uniform Sampling yields an $O(1/\varepsilon)$-approx. in some settings.

# Packing problems:

- Given a set of disks $\mathcal{D}$ in the plane, disk $D \in \mathcal{D}$ has weight $w_D > 0$.

- We want to compute a maximum weight independent set.

  ie: A set $S \subseteq \mathcal{D}$ of disks that are pairwise disjoint, and: $\sum_{D \in S} w(D)$ is as large as possible.

- Let us write an LP-relaxation for the problem:

$$\max \sum_{D \in \mathcal{D}} w_D x_D$$

subject to:

$$\sum_{D : p \in D} w_D \leq 1 \quad \forall p \in P$$

$$w_D \in [0, 1].$$

- Let $x = (x_{D_1}, \ldots, x_{D_n})$ be an optimal LP solution.

- $OPT_{LP} = \sum_{D \in \mathcal{D}} w_D x_D$.