§ Sample Complexity and VC Dimension :

Ch 14 M-U

Ch 20, Notes by Har-Peled.

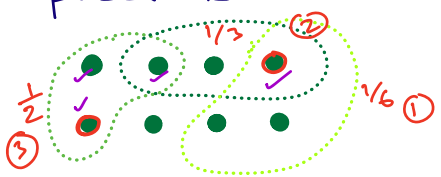
Sampling : using a small set of observations, estimate properties of an entire sample space.

sample complexity : minimum size sample to obtain the required result.

Let us consider two important problems.

① Range detection,

② Probability estimation.



- Range is just a subset of the underlying space.
- Goal is to use one set of samples to detect a set of ranges or estimate the prob. of ranges [the set of possible ranges can be really huge, even infinite].

For detection, we want the sample to intersect with each range in the set,
while for prob. estimation, we want the fraction of points in the sample that intersect with each range in the set to approximate the assoc. prob. of the range. $\frac{1}{3} : \frac{1}{2} : \frac{1}{6}$

Q. Can we obtain some sample whose size is independent of the number of ranges? and dependent on the structure of the space?

Q. Assume we are given a range $[a,b]$ with an underlying unknown prob. distr. $\mathcal{D}$ on $[0,1]$ s.t. $\mathbb{P}_{x \in \mathcal{D}}(x \in [a,b]) \geq \epsilon$.
Then how many samples do we need s.t. w.p. $\geq 1-\delta$. we have at least one sample in $[a,b]$?

Let $x_1, x_2, \ldots, x_m$ be m indep. samples in $\mathbb{R}$ from an unknown distr. $\mathcal{D}$

Given interval $[a,b]$, if $\mathbb{P}(x \in [a,b]) \geq \epsilon$, then prob that a sample of size $m = \frac{1}{\epsilon} \ln 1/\delta$ intersects $[a,b]$ is $\geq 1 - (1-\epsilon)^m \geq 1-\delta$.

Given k such intervals, union bound will show, $\mathbb{P}$[a sample of size $1/\epsilon \ln k/\delta$ intersects each of the intervals] $\geq 1 - k(1-\epsilon)^{1/\epsilon \ln k/\delta} \geq 1-\delta$.

- Say, we want select few samples from $[0,1]$ s.t. all intervals of length $1/10$ contains at least one point in the sample.

– There are infinite such sets. So union bound won't help. But <u>ten</u> equidistant points would already work.

- Indeed, we'll see for any distribution, a sample size of $\Omega(1/\epsilon \ln 1/\delta)$ intersects all intervals having prob $\geq \epsilon$, w.p. $\geq 1-\delta$.

VC dimensions & Rademacher complexity helps in evaluation of sample complexity.

§ VC Dimension: (Vapnik-Chervonenkis dimension)

Definition 14.1: *A* range space *is a pair* $(X, \mathcal{R})$ *where:*

1. *X is a (finite or infinite) set of* points*;*
2. *$\mathcal{R}$ is a family of subsets of X, called* ranges.

X is also called ground set

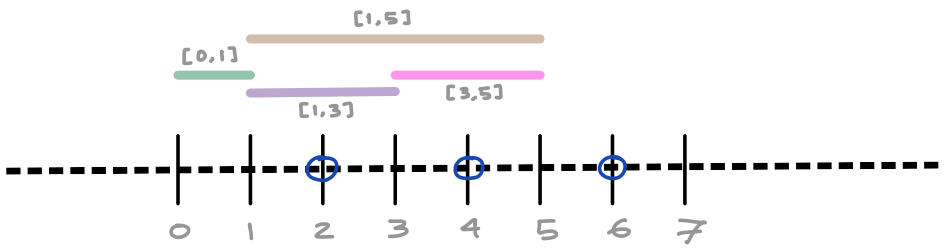
Example of range space:

$X = \mathbb{R}$, $\mathcal{R} = \{[a,b] \mid [a,b] \subseteq \mathbb{R}\}$.

set of all closed intervals

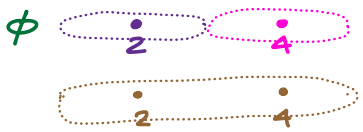
Definition 14.2: *Let $(X, \mathcal{R})$ be a range space and let $S \subseteq X$. The* projection *of $\mathcal{R}$ on S is*

$$\mathcal{R}_{\downarrow S} = \{R \cap S \mid R \in \mathcal{R}\}.$$



$S = \{2,4\}$

Any two points can be shattered.



$\mathcal{R}_{\downarrow S}$ is the set of all possible subsets of S.

$S = \{2,4,6\}$ $\mathcal{R}_{\downarrow S}$ gives seven of the eight possible subsets of S, except $\{2,6\}$.

Any 3 points can't be shattered.

– Any interval containing 2 & 6 must contain 4.

Definition 14.3: *Let $(X, \mathcal{R})$ be a range space. A set $S \subseteq X$ is* shattered *by $\mathcal{R}$ if $|\mathcal{R}_{\downarrow S}| = 2^{|S|}$. The* Vapnik–Chervonenkis (VC) dimension *of a range space $(X, \mathcal{R})$ is the maximum cardinality of a set $S \subseteq X$ that is shattered by $\mathcal{R}$. If there are arbitrarily large finite sets that are shattered by $\mathcal{R}$, then the VC dimension is infinite.*

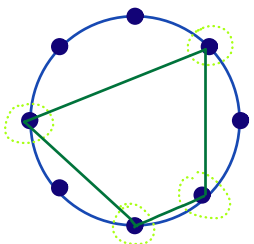
So VC Dim of above range space (with infinite points & ranges) is only 2.

Note: VC dim $(\mathcal{R}) = d$ if there is some set of cardinality d that is shattered by $\mathcal{R}$. It does not say all sets of cardinality d are shattered by $\mathcal{R}$. To show VC-dim $\leq d$, we need to show all sets of cardinality $> d$ are not shattered by $\mathcal{R}$.

- More examples:
- **Convex sets:** $X = \mathbb{R}^2$, $\mathcal{R}$ = the family of all closed convex sets on the plane.

Claim: This range space has infinite VC-dimension.

→ Need to show, for any $n \in \mathbb{N}$ there exists a set S with $|S| = n$, that can be shattered.



$S_n = \{x_1, \ldots, x_n\}$ be n points on the boundary of a circle.

Any subset $Y \subseteq S_n$, $Y \neq \phi$ defines a convex set that does not contain any points in $S_n \setminus Y$.

Hence, Y is included in the projection of $\mathcal{R}$ on S_n.

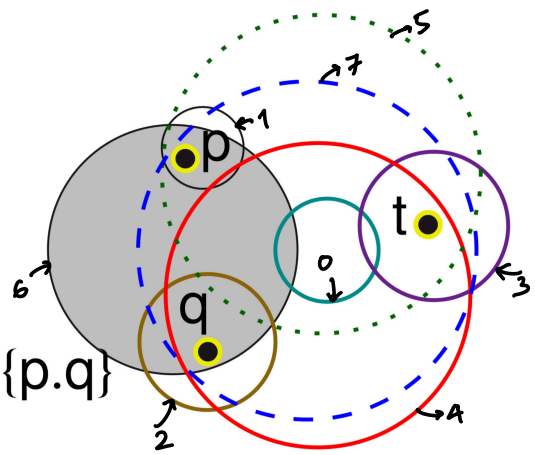
Empty set is also a projection as well.

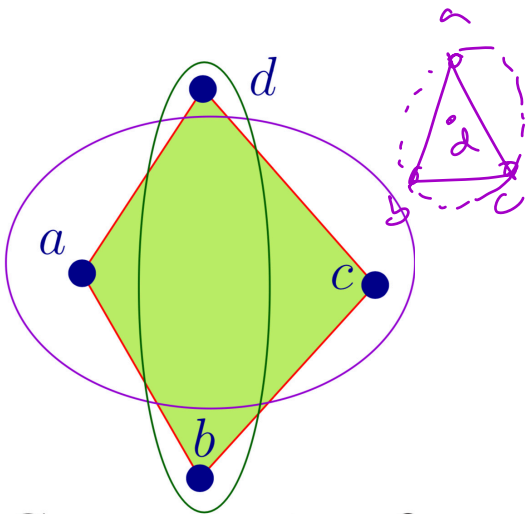
Hence, $\forall n \in \mathbb{N}$, S_n is shattered.

- **Disks:** $X = \mathbb{R}^2$, $\mathcal{R}$ = the family of all disks on the plane.

Observation: For any 3 points on the plane (in general position) one can find eight disks so that the points are shattered.

[Note: this is not true if the points are collinear. However, we just need to show one set of three points that can be shattered].





Can disks shatter a set P with four points: $\{a,b,c,d\}$?

→ Case 1: convex hull of P has only 3 points on its boundary, say a, b, c.

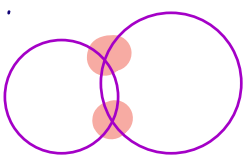
Then $X = \{a, b, c\}$ can not be obtain as a projection.

Due to convexity, any disk containing a, b, c must contain d.

→ Case 2: all 4 points are on the convex hull.

Then if we can realize $\{a,c\}$ & $\{b,d\}$ as projections, these two disks will intersect each other at 4 points — a contradiction.

Note: pseudodisks are objects that can intersect at ≤ 2 times.



§ Growth Function:

different subsets of size $\le d$ out of n elements : $g(d,n) = \sum_{i=0}^{d} \binom{n}{i}$.

For $n = d$, we have $g(d,n) = 2^d$,

for $n > d \ge 2$, $g(d,n) \le \sum_{i=0}^{d} \frac{n^i}{i!} \le n^d$.

- The number of ranges for a set of n elements grows polynomially in n (the power being dimension d), instead of exponentially.

- Observation: $g(d,n) = g(d,n-1) + g(d-1,n-1)$.
  - $g(d,n-1)$: not to include first element
  - $g(d-1,n-1)$: includes first elements

- Sauer-Shelah theorem:

Let $(X, \mathcal{R})$ be a range space with $|X| = n$, VC-dim d. Then, $|\mathcal{R}| \le g(d,n) \ (\le n^d)$.

Proof: We prove the claim by induction on d, and for each d by induction on n.

Trivially holds for $d = 0$ or $n = 0$, as then $g(d,n) = 1$.

Take $x \in X$, Define

$$\mathcal{R}_x = \{ r \setminus \{x\} \mid r \cup \{x\} \in \mathcal{R} \text{ and } r \setminus \{x\} \in \mathcal{R} \}.$$

$$\mathcal{R} \setminus x = \{ r \setminus \{x\} \mid r \in \mathcal{R} \}.$$

claim: $|\mathcal{R}| = |\mathcal{R}_x| + |\mathcal{R}\setminus x|$.

- We charge elements of $\mathcal{R}$ to their corrs. element in $\mathcal{R}\setminus x$.

The only bad case is when $\exists\, r$ s.t. both $r \cup \{x\}$ and $r \setminus \{x\}$ are in $\mathcal{R}$, as then these two distinct ranges get mapped to same range in $\mathcal{R}\setminus x$.

But such ranges contribute to exactly one element in $\mathcal{R}_x$.

We'll show VC dim of $(X\setminus\{x\}, \mathcal{R}_x) \leq d-1$ & VC-dim of $(X\setminus\{x\}, \mathcal{R}\setminus x)$ is $\leq d$.

Then we'll get:

$$\therefore |\mathcal{R}| = |\mathcal{R}_x| + |\mathcal{R}\setminus x| \leq G(d-1, n-1) + G(d, n-1)$$

(claim; induction.)

$$= G(d, n).$$

(obs.)

To conclude:

(i) VC-dim of $(X\setminus\{x\}, \mathcal{R}\setminus\{x\}) \leq d$.

→ For contradiction, assume $\mathcal{R}\setminus\{x\}$ shatters a set S of size $d+1$.

Now for every $R \in \mathcal{R}\setminus\{x\}$, there is a corrs. $R' \in \mathcal{R}$ s.t. either $R' = R$ or $R' = R \cup \{x\}$.

In either cases, $\mathcal{R}|_S$ contains $S \cap R' = S \cap R$.
Then $\mathcal{R}$ shatters S as well, contradicting VC-Dim $(X, \mathcal{R})$ to be d.

(ii) VC-dim of $(X \setminus \{x\}, \mathcal{R}_x) \leq d-1$

→ For contradiction, assume $\mathcal{R}_x$ shatters a set S of size d.

Now for every $R \in \mathcal{R} \setminus \{x\}$, both R and $R \cup \{x\}$ are in $\mathcal{R}$.

In both cases, $\mathcal{R}|_S$ contains $S \cap R' = S \cap R$.

Then $\mathcal{R}$ also shatters S. → A contradiction. ▨

§ VC-dimension component bounds:

Bounds VC-Dim of a complex range space as a fn of VC-dim of its simpler components.

Theorem 14.5: *Let $(X, \mathcal{R}^1), \ldots, (X, \mathcal{R}^k)$ be k range spaces each with VC dimensions at most d. Let $f : (\mathcal{R}^1, \ldots, \mathcal{R}^k) \to 2^X$ be a mapping of k-tuples $(r_1, \ldots, r_k) \in (\mathcal{R}^1, \ldots, \mathcal{R}^k)$ to subsets of X, and let*

$$\mathcal{R}^f = \{f(r_1, \ldots, r_k) \mid r_1 \in \mathcal{R}^1, \ldots, r_k \in \mathcal{R}^k\}.$$

circle squares etc.

The VC dimension of the range space $(X, \mathcal{R}^f)$ is $O(kd \ln k)$.

This yields the following corollary.

Corollary 14.6: *Let $(X, \mathcal{R}^1)$ and $(X, \mathcal{R}^2)$ be two range spaces, each with VC dimension at most d. Let*

$$\mathcal{R}^{\cup} = \{r_1 \cup r_2 \mid r_1 \in \mathcal{R}^1 \text{ and } r_2 \in \mathcal{R}^2\},$$

and

$$\mathcal{R}^{\cap} = \{r_1 \cap r_2 \mid r_1 \in \mathcal{R}^1 \text{ and } r_2 \in \mathcal{R}^2\}.$$

The VC dimensions of the range spaces $(X, \mathcal{R}^{\cup})$ and $(X, \mathcal{R}^{\cap})$ are $O(d)$.

§ Shattering dimension.

- Defn (Shatter function): Given range space $S = (X, \mathcal{R})$, its shatter function $\pi_S(m)$ is the maximum number of sets that might be created by S when restricted to subsets of size m.

$$\pi_S(m) = \max_{\substack{B \subset X \\ |B| = m}} |\mathcal{R}_{|B}|.$$

($\mathcal{R}_{|B}$: projection of $\mathcal{R}$ on B)

- Shattering dimension of S is the smallest ρ s.t. $\pi_S(m) = O(m^\rho)$, for all m.

- Note: In general $\pi_S(n) = 2^n$.

Corollary: If $S = (X, \mathcal{R})$ is a range space of VC-dim d, then $\forall$ finite $B \subseteq X$, we have

$$|\mathcal{R}_{|B}| \leq \pi_S(|B|) \leq g(d, |B|).$$

Proof: $n := |B|$, so $|\mathcal{R}_{|B}| \leq \pi_S(|B|)$ [By defn of π_S]

$\leq g(d, n)$ (By S.S. lemma)

$\leq n^d$.

So, shattering dimension of a range space is bounded by its VC-dim.

shatt. dim & VC dim are close.

Lemma (A): If $S=(X,\mathcal{R})$ is a range space with shattering dim ρ, then VCdim (S) is $O(\rho \log \rho)$.

Proof: Let $N \subseteq X$ be the largest set shattered by S and $\delta = |N|$.

We have, $2^{\delta} \overset{\text{shattering def}}{=} |\mathcal{R}_{|N}| \overset{\text{corollary}}{\leq} \pi_S(|N|) \overset{\text{shatt. dim.}}{\leq} c\delta^{\rho}$, where c is a constant.

$\Rightarrow \delta \leq \lg c + \rho \lg \delta$.

$\Rightarrow \rho \geq (\delta - \lg c)/\lg \delta$.

otherwise δ is already $O(1)$, i.e. $O(1 \log \rho)$.

Assuming, $\delta \geq \max(2, 2\lg c)$, we have

$$\frac{\delta}{2\lg\delta} \leq \rho \Rightarrow \frac{\delta}{\ln\delta} \leq \frac{2\rho}{\ln 2} \leq 6\rho$$

$\Rightarrow \delta \leq 2(6\rho)\ln(6\rho)$. [Fact: for $u \geq \sqrt{e}$, if $\frac{x}{\ln x} \leq u$ then $x \leq 2u\ln u$.] ∎

→ Advantage: shattering functions are sometimes easier to compute & gives good approximation of VC-dimension.

- Shattering dimension of disks.

Lemma: Consider range space $S = (X, \mathcal{R})$, where $X = \mathbb{R}^2$ and $\mathcal{R}$ is the set of disks. Then shattering dim of S is 3.

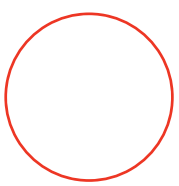
assume they are in general position.

Proof: Consider any set P of n points in the plane and set $\mathcal{F} = \mathcal{R}_{|P}$.

We claim, $|\mathcal{F}| \leq 4n^3$.

The set $\mathcal{F}$ contains only n sets with a single point in them & $\leq \binom{n}{2}$ sets with 2 points in them.

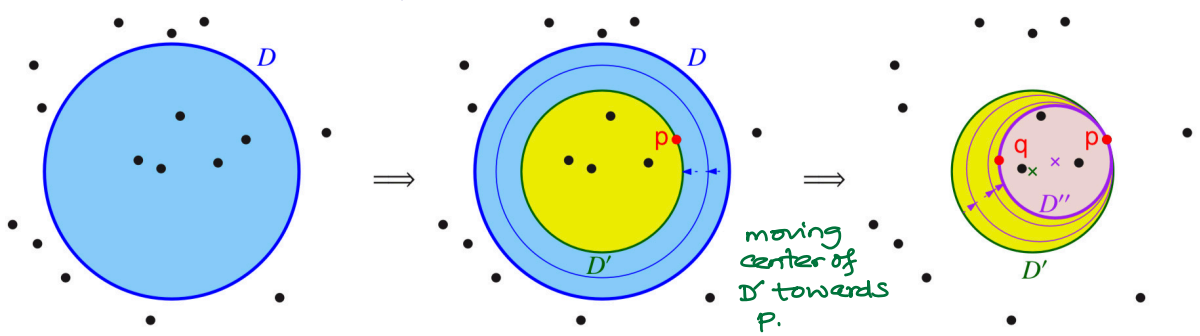
So, fix $Q \in \mathcal{F}$ s.t. $|Q| \geq 3$.



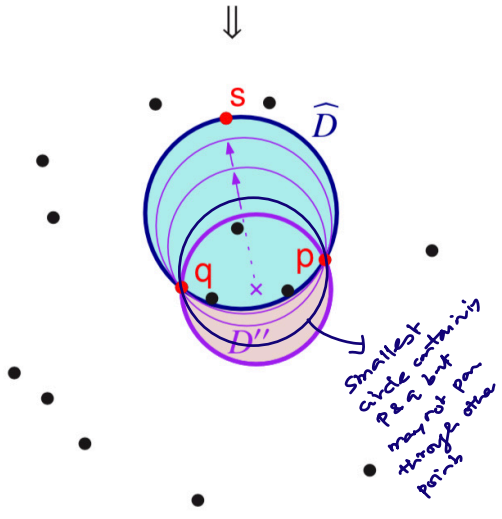
(1) Let disk D realizes Q, i.e. $P \cap D = Q$.

(2) Shrink D till its boundary passes through a point p.

(3) Continue shrinking until the boundary passes through two points p & $q \in Q$.



④ We continuously deform D'' s.t. it has both p & q on its boundary.



This can be done by moving center of D'' along the bisector line between p & q.

We continue till we hit a third point $s \in P$.

$\hat{D}$ is a unique circle passing through p, q, s.

Also, $D \cap (P \setminus \{s\}) = \hat{D} \cap (P \setminus \{s\})$.

Thus we can specify the point set $P \cap D$ by specifying (p, q, s, x_p, x_q, x_s); (p, q, s) are points defining D and $x_* \in \{0,1\}$ states whether point $*$ is in Q or not. In the above case it is $(1,1,0)$.

There are $\leq 8\binom{n}{3}$ different subsets in $\mathcal{F}$ containing more than 3 points, as each such subset maps to a 'canonical' disk, there are $\binom{n}{3}$ such disks & each disk defines at most 8 different subsets.

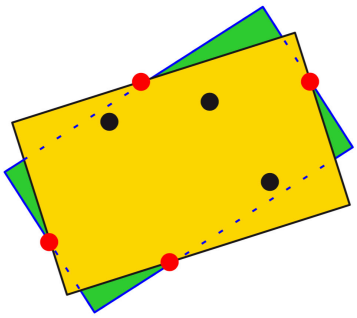
Similar argumentation implies that there are at most $4\binom{n}{2}$ subsets that are defined by a pair of points that realizes the diameter of the resulting disk.

$\therefore |\mathcal{R}| = 1 + n + 4\binom{n}{2} + 8\binom{n}{3} \leq 4n^3.$ ∎

Insight: Above argumentation gives a powerful tool → shattering dim of a range space defined by a family of shapes is always bounded by the number of points that determine a family.

This sometimes makes it more convenient to work with shattering dim. instead of VC dimension.

Example: shattering dimension of arbitrarily oriented rectangles is bounded by 5.



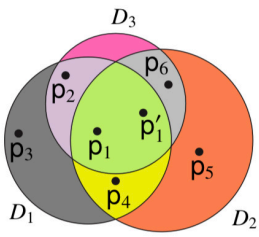
Dual shattering dimension.

Ranges ⟷ Points

Definition 20.2.8. The ***dual range space*** to a range space $S = (X, \mathcal{R})$ is the space $S^\star = (\mathcal{R}, X^\star)$, where $X^\star = \{\mathcal{R}_p \mid p \in X\}$.

$S^{**} = S$

(A)

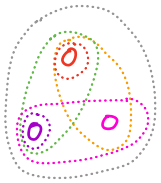


(B)

	p_1	p_1'	p_2	p_3	p_4	p_5	p_6
D_1	1	1	1	1	1	0	0
D_2	1	1	0	0	1	1	1
D_3	1	1	1	0	0	0	1

(C)

	D_1	D_2	D_3
p_1	1	1	1
p_1'	1	1	1
p_2	1	0	1
p_3	1	0	0
p_4	1	1	0
p_5	0	1	0
p_6	0	1	1



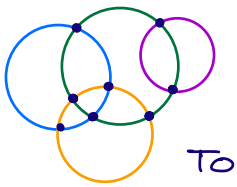
Let the ***dual shatter function*** of the range space S be $\pi_S^\star(m) = \pi_{S^\star}(m)$, where $S^\star$ is the dual range space to S.

Definition 20.2.9. The ***dual shattering dimension*** of S is the shattering dimension of the dual range space $S^\star$. $\pi_{S^\star}(m)$

Alternately, dual shattering fn for S is the maximum number of points that are created when restricted to m sets F. Dual shattering dim of S is the smallest ρ s.t. $\pi_{S^\star}(m) = O(m^\rho)$ $\forall$ m.

Claim : Dual shattering dim of disks is 2.

→ Disks intersect each other at most 2 times.



The complexity of arrangement of n disks is $O(2 \cdot \binom{n}{2})$, i.e., $O(n^2)$.

To maximize X^*, we need at least one point in every intersection combination of ranges in $\mathcal{R}$.

Hence, number of ranges in $X^* \leq$ the complexity of arrangement of ranges in $\mathcal{R} = O(n^2)$. ■

Lemma: Consider a range space $S = (X, \mathcal{R})$ with VC-dim d. Then the dual range space $S^* = (\mathcal{R}, X^*)$ has VC-dim $\leq 2^{d+1}$.

Lemma: If a range space $S = (X, \mathcal{R})$ has dual shattering dimension δ, then its VC-dim is $\leq \delta^{O(\delta)}$.

Proof: The shattering dim of dual range space S^* is $\leq \delta$.

By lemma Ⓐ, VC-dim of S^* δ' is $O(\delta \log \delta)$.

Since the dual range space to S^* is S, we have from the previous lemma :

$$\text{VC-dim}(S) \leq 2^{\delta'+1} = \delta^{O(\delta)}.$$ ■

→ This is very useful when shapes in $\mathcal{R}$ are simple. If we show the dual shattering dim of S is $O(1)$, we also obtain VC-dim is $O(1)$.

§ ε-nets and ε-samples.

- ε-nets are combinatorial object that catches or intersects with every range of sufficient size.

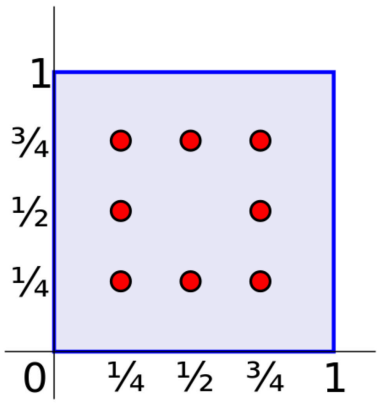
Definition 14.4 [combinatorial definition]: *Let $(X, \mathcal{R})$ be a range space, and let $A \subseteq X$ be a finite subset of X. A set $N \subseteq A$ is a* combinatorial ϵ-net *for A if N has a nonempty intersection with every set $R \in \mathcal{R}$ such that $|R \cap A| \geq \epsilon |A|$.*

Definition 14.5: *Let $(X, \mathcal{R})$ be a range space, and let $\mathcal{D}$ be a probability distribution on X. A set $N \subseteq X$ is an ϵ-net for X with respect to $\mathcal{D}$ if for any set $R \in \mathcal{R}$ such that $\Pr_{\mathcal{D}}(R) \geq \epsilon$, the set R contains at least one point from N, i.e.,*

$$\forall R \in \mathcal{R}, \ \Pr_{\mathcal{D}}(R) \geq \epsilon \Rightarrow R \cap N \neq \emptyset.$$

Here, $\Pr_{\mathcal{D}}(R)$ is the prob. that a point chosen according to $\mathcal{D}$ is in R.

Note, combinatorial defn. corrs. to the setting when $\mathcal{D}$ is uniform over A.



An ε-net with ε = 1/4 of the unit square in the range space where the ranges are closed filled rectangles.

- The minimum sample size that contains an ε-net (or ε-sample) can be bounded in terms of the VC-dimension of the range space.

§ The ε-net theorem:

Naive approach:

Let $(X, \mathcal{R})$ be a range space with VC-dim $d \geq 2$, let $A \subseteq X$ with $|A| = n$.
then there exists a combinatorial ε-net N for A of size at most $\lceil d \ln n / \varepsilon \rceil$.

Proof: Let $\mathcal{R}'$ be the projection of $\mathcal{R}$ on A.
By Sauer-Shelah theorem, $|\mathcal{R}'| \leq n^d$. ⊛

We take a sample of $\kappa = \lceil d \ln n / \varepsilon \rceil$ points of A, independently & uniformly at random.

For each set $S \in \mathcal{R}$ with $|S \cap A| \geq \varepsilon |A|$, there is a corresponding set $S' \in \mathcal{R}'$.

So a point in our sample is in S' with prob $\geq \varepsilon$.

Then $\mathbb{P}$[our sample misses a given set S']
$\leq (1-\varepsilon)^\kappa$.

There are n^d such sets to consider. [From ⊛]

Applying union bound, the prob that the sample misses at least one such S' is

$$\leq n^d (1-\varepsilon)^\kappa < n^d e^{-\varepsilon \kappa} < n^d e^{-d \ln n} = 1.$$

Thus by probabilistic method, there is a set N of size κ that misses no set $S' \in \mathcal{R}'$.

Hence, N is an ε-net for A. ∎

Now we prove the main theorem that shows existence of ε-net of size $O(\frac{d}{\varepsilon} \ln \frac{d}{\varepsilon})$. (independent of n.)

• **Theorem:** Let $(X, \mathcal{R})$ be a range space with VC-dim d and let $\mathcal{D}$ be a prob. distribution on X. For any $0 < \delta, \varepsilon \leq 1/2$, there is an $m = O(\frac{d}{\varepsilon} \ln \frac{d}{\varepsilon} + \frac{1}{\varepsilon} \ln \frac{1}{\delta})$ such that a random sample from $\mathcal{D}$ of size m is an ε-net for X with probability at least $1-\delta$.

Proof:

Let M be a set of m independent samples from X acc. to $\mathcal{D}$.

Let $E_1 := \{\exists S \in \mathcal{R} \mid Pr_{\mathcal{D}}(S) \geq \varepsilon \text{ and } |S \cap M| = 0\}$, i.e. E_1 is the event that M is not an ε-net for X w.r.t. $\mathcal{D}$.

We want to show, $\mathbb{P}(E_1) \leq \delta$.

For this, we go beyond the union bound approach.

Choose a second set T of m indep. samples from X acc. $\mathcal{D}$.

Define, $E_2 := \{\exists S \in \mathcal{R} \mid Pr_{\mathcal{D}}(S) \geq \varepsilon, |S \cap M| = 0, |S \cap T| \geq \varepsilon m/2\}$.

Following lemma shows E_1 & E_2 have similar probabilities.

Lemma 1: For $m \geq 8/\varepsilon$, $\mathbb{P}(E_2) \leq \mathbb{P}(E_1) \leq 2\,\mathbb{P}(E_2)$. (trivial as $E_2 \subseteq E_1$.)

Proof: If event E_1 holds, there is some S' s.t. $|S' \cap M| = 0$ and $Pr_{\mathcal{D}}(S') \geq \varepsilon$.

Hence, $\frac{\mathbb{P}(E_2)}{\mathbb{P}(E_1)} = \frac{\mathbb{P}(E_1 \cap E_2)}{\mathbb{P}(E_1)} = \mathbb{P}(E_2 | E_1) \geq \mathbb{P}(|T \cap S'| \geq \varepsilon m/2)$. ...(1)

Now for a fixed range S' and a random sample T, the random variable $|T \cap S'|$ has a binomial distr. $B(m, Pr_D(S'))$.

As $Pr_D(S') \geq \varepsilon$, using Chernoff bounds:

$$\mathbb{P}(|T \cap S'| < (1-\delta)\,\mathbb{E}[|T \cap S'|]) \leq e^{-\frac{\delta^2}{2} \cdot \mathbb{E}[|T \cap S'|]} \quad [\because m \geq 8/\varepsilon]$$

$$\Rightarrow \mathbb{P}(|T \cap S'| < (1-\tfrac{1}{2})m\varepsilon) \leq e^{-\frac{1}{4} \cdot \frac{1}{2} \cdot m\varepsilon} = e^{-\varepsilon m/8} \leq e^{-1} < \frac{1}{2}$$ ($\delta = 1/2$) ...(2)

Hence, from (1) & (2), $\mathbb{P}(E_1) \leq 2\,\mathbb{P}(E_2)$. ●

Now we bound $\mathbb{P}[E_2]$ by prob. of a larger event E_2'.

$E_2' := \{\exists S \in \mathcal{R} \mid |S \cap M| = 0 \text{ and } |S \cap T| \geq \varepsilon m/2\}$.

- Lemma 2. $\mathbb{P}(E_1) \leq 2\,\mathbb{P}(E_2) \leq 2\,\mathbb{P}(E_2') \leq 2\,(2m)^d\, 2^{-\varepsilon m/2}$.

Proof: As M & T are random samples, we assume to choose $2m$ random samples and partition randomly into two equal sized sets M & T.

For a fixed $S \in \mathcal{R}$, $K = \varepsilon m/2$, let

$E_S := \{|S \cap M| = 0,\ |S \cap T| \geq K\}$.

This event means $(M \cup T) \cap S \geq K$, but all these elements were placed in T and not in M.

So, out of $\binom{2m}{m}$ possible partitions of $M \cup T$, we choose one of $\binom{2m-K}{m}$ partitions where no element of S is in M.

Hence, $\mathbb{P}(E_S) \le \mathbb{P}(|M \cap S| = 0 \mid |S \cap (M \cup T)| \ge k)$

$$= \binom{2m-k}{m} / \binom{2m}{m}$$

$$= (2m-k)!\, m! / (2m)!\, (m-k)!$$

$$= \frac{m(m-1)\cdots(m-k+1)}{(2m)(2m-1)\cdots(2m-k+1)} \le 2^{-k} \le 2^{-\varepsilon m/2}.$$

By Sauer-Shelah theorem, the projection of $\mathcal{R}$ on $M \cup T$ has $\le (2m)^d$ ranges.

Hence, using union bounds:

$$\mathbb{P}(E_2') \le (2m)^d 2^{-\varepsilon m/2}.$$

■

To complete the proof of ε-net theorem, we need to show $\mathbb{P}(E_1) \overset{\text{Lem 1}}{\le} 2(2m)^d 2^{-\varepsilon m/2} \le \delta$, for $m \ge \frac{8d}{\varepsilon} \ln \frac{16d}{\varepsilon} + \frac{4}{\varepsilon} \ln \frac{2}{\delta}$.

eqv., we need $\varepsilon m/2 \ge \ln(2/\delta) + d \ln(2m)$.

As $m > \frac{4}{\varepsilon} \ln \frac{2}{\delta}$, $\varepsilon m/4 \ge \ln(2/\delta)$.

To finish we show $\varepsilon m/4 \ge d \ln(2m)$.

[Fact: if $y \ge x \ln x \ge e$, then $2y/\ln y \ge x$]

Use $y = 2m \ge \frac{16d}{\varepsilon} \ln \frac{16d}{\varepsilon}$, $x = \frac{16d}{\varepsilon}$, we have

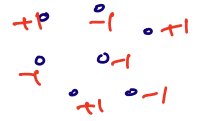
$$\frac{4m}{\ln(2m)} \ge \frac{16d}{\varepsilon} \Rightarrow \frac{\varepsilon m}{4} \ge d \ln(2m).$$

■

§ Application:

Probably Approximately Correct (PAC) Learning.

We are given a set of items X and a prob distr. $\mathcal{D}$ is defined on X.



A binary classification is a subset $C \subseteq X$ s.t. all items in C are labeled 1 and in $X \setminus C$ are labeled -1.

The concept class $\mathcal{C}$ is the set of all possible classifications defined by the problem.

- Learning algo has access to ORACLE$(C, \mathcal{D})$ that produces a pair $(x, c(x))$, where $x \sim \mathcal{D}$ and $c(x) = 1$ if $x \in C$ and -1 otherwise.
- We also assume the classification problem is realizable, i.e. $\exists h \in \mathcal{C}$, $Pr_{\mathcal{D}}(h(x) \neq c(x)) = 0$.

Definition 14.9 [PAC Learning]: *A concept class C over input set X is* PAC learnable[1] *if there is an algorithm L, with access to a function* ORACLE$(C, \mathcal{D})$*, that satisfies the following properties: for every correct concept $C \in \mathcal{C}$, every distribution $\mathcal{D}$ on X, and every $0 < \epsilon, \delta \leq 1/2$, the number of calls that the algorithm L makes to the function* ORACLE$(C, \mathcal{D})$ *is polynomial in ϵ^{-1} and δ^{-1}, and with probability at least $1 - \delta$ the algorithm L outputs a hypothesis h such that* $\Pr_{\mathcal{D}}(h(x) \neq c(x)) \leq \epsilon$.

probably

approximately

Theorem: Any finite concept class $\mathcal{C}$ can be PAC-learned with $m = \frac{1}{\epsilon}(\ln|\mathcal{C}| + \ln\frac{1}{\delta})$ samples.

Proof: Let $c^* \in \mathcal{C}$ be the correct classification.

A hypothesis h is "bad" if $\Pr_{\mathcal{D}}[h(x) \neq c^*(x)] \geq \epsilon$.

$\mathbb{P}[$ a bad h is consistent with m random samples$]$
$\leq (1-\epsilon)^m$.

Using union bound,

$\mathbb{P}[\exists$ bad h is consistent with m random samples$]$
$\leq |\mathcal{C}|(1-\epsilon)^m$.
$\leq \delta \quad [\text{as } m = \frac{1}{\epsilon}(\ln|\mathcal{C}| + \ln\frac{1}{\delta})]$.

So, we return whatever h is consistent with all m random sample. — all these are "non-bad."

By assumption, as correct classification $c^* \in \mathcal{C}$, at least one such h exists. ■

→ Note that X can be infinite.
So, it is interesting that we can PAC-learn $\mathcal{C}$, with sample complexity independent of n.

Note: A concept class is efficiently PAC learnable if the algorithm runs in time polynomial in the size of problem, $1/\epsilon$, $1/\delta$.
Here, we are only interested in sample complexity, computational complexity may not necessarily be polynomial in sample size.

- Can we make sample complexity indep of $|\mathcal{C}|$?
- Can we extend this to infinite concept classes?

Say, we are learning interval $[a,b] \in \mathbb{R}$.
Concept class is collection of all closed intervals in $\mathbb{R}$: $\mathcal{C} = \{[x,y] \mid x \le y\} \cup \phi$.

Let $c^* \in \mathcal{C}$ be the concept to be learned.
h be the hypothesis returned by our algo.

Training set T is collection of n points drawn from D.

Let $x \in T$, if $x \in [a,b]$ it is a +ve example,
else a -ve example.

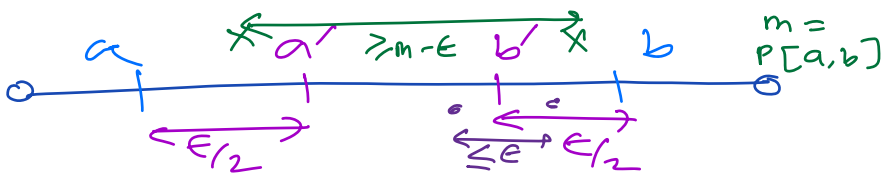
Algo:
If no sample is positive
return trivial hypothesis.
else return $[c,d]$ where c,d are smallest and largest +ve examples among the samples.

Q. What is the prob ALGO makes an error?

ALGO can only make an error on an input x if $x \in [a,b]$. for $x \notin [a,b]$, it always returns -1.

Case 1. $Pr_D(x \in [a,b]) \le \epsilon$.
From the above fact, prob of error $\le \epsilon$

Case 2. $Pr_D(x \in [a,b]) > \epsilon$.



Say, $a' \geq a$ be smallest val s.t $Pr_D([a,a']) \geq \epsilon/2$.
and $b' \leq b$ be largest val s.t. $Pr_D([b',b]) \geq \epsilon/2$.

So, $a' \leq b'$.
For simplicity assume $a' < b'$.

If ALGO returns a bad hypothesis then error $\geq \epsilon$,
Now if sample points fell in $[a,a']$ and $[b,b']$ then we would have returned $[c,d] \supseteq [a',b']$. → a good hypothesis

So, $\mathbb{P}[\text{bad hypothesis}] \leq \mathbb{P}\left(\begin{array}{c}\text{sample points didn't}\\ \text{fall, in either } [a,a'] \text{ or } [b,b']\end{array}\right)$.

The prob. that a training set of n points does not contain any examples from either $[a,a']$ or $[b,b']$ is

$$\leq 2\left(1-\frac{\epsilon}{2}\right)^n \leq 2e^{-\epsilon n/2} \leq \delta.$$

by choosing $n \geq 2\ln(2/\delta)/\epsilon$.

- ϵ-net & VC-dim generalizes this idea.

Theorem 14.13: *Let $\mathcal{C}$ be a concept class that defines a range space with VC dimension d. For any $0 < \delta, \epsilon \leq 1/2$, there is an*

over a set of points X

$$m = O\left(\frac{d}{\epsilon}\ln\frac{d}{\epsilon} + \frac{1}{\epsilon}\ln\frac{1}{\delta}\right)$$

such that $\mathcal{C}$ is PAC learnable with m samples.

Proof: Let $c \in \mathcal{C}$ be the correct classification.

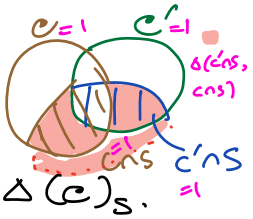
For $c' \in \mathcal{C}$, $\Delta(c', c) = \{x \mid c(x) \neq c'(x)\}$.

$\Delta(\mathcal{C}) := \{\Delta(c', c) \mid c' \in \mathcal{C}\}$ ← collection of all possible sets of points of disagreement with the correct classification.

Surprising lemma! motivate more!

Lemma: $\text{VC-Dim}(X, \Delta(\mathcal{C})) = \text{VC-Dim}(X, \mathcal{C})$. (for any $c \in \mathcal{C}$!)

→ For any $S \subseteq X$, let $\mathcal{C}_S$, $\Delta(\mathcal{C})_S$ be projection of $(X, \mathcal{C})$ and $(X, \Delta(\mathcal{C}))$ on S.



We define bijection $b: \mathcal{C}_S \to \Delta(\mathcal{C})_S$ by mapping $c' \cap S \in \mathcal{C}_S$ to $\Delta(c' \cap S, c \cap S) \in \Delta(\mathcal{C})_S$.

Then for $S \subseteq X$, $|\mathcal{C}_S| = |\Delta(\mathcal{C})_S|$, and S is shattered by $\mathcal{C}$ iff it is shattered by $\Delta(\mathcal{C})$.

Hence, the range spaces have same VC-dim.

HW: Show for $S \subseteq X$, S is shattered by $\mathcal{C}$ iff S is shattered by $\Delta(\mathcal{C})$.

To complete the proof of lemma, we need to show b is bijection.

Take $c', c'' \in \mathcal{C}$ with $c' \cap S \neq c'' \cap S$.
Then $\exists y \in S$ s.t. $c'(y) \neq c''(y)$.
W.l.o.g. assume $c'(y) \neq c(y)$ but $c''(y) = c(y)$.
Then, $y \notin \Delta(c' \cap S, c) \cap S$
but $y \in \Delta(c'' \cap S, c \cap S)$.
Hence, $\Delta(c' \cap S, c \cap S) \neq \Delta(c'' \cap S, c \cap S)$.

For the other direction, if for $c, c' \in \mathcal{C}$,
s.t. $\Delta(c' \cap S, c \cap S) \neq \Delta(c'' \cap S, c \cap S)$,
then $\exists y \in S$ s.t. $c'(y) \neq c''(y)$, so $c' \cap S \neq c'' \cap S$. ∎

Thus as VC-Dim$[(X, \Delta(\mathcal{C}))] = d$, there exists
$m = O(d/\epsilon \ln d/\epsilon + 1/\epsilon \ln 1/\delta)$ s.t. sample of size $\geq m$
is ϵ-net for this range-space w.p. $\geq 1-\delta$.
Hence, w.p. $\geq 1-\delta$ it has nonempty intersection
with every set $\Delta(c', c)$ that has prob $\geq \epsilon$,
i.e. ALG0 can exclude any hyp. w. error prob $\geq \epsilon$. ∎

So any c' far from correct c is anyway bit. & thus can be excluded

As we only select hypotheses that are consistent w. c in all the samples.

- Σ-sample provides stronger guarantees.
- it maintains relative probability weight all sets $R \in \mathcal{R}$ within error of ϵ, and needs just additional $O(1/\epsilon)$ factor in sample size..

Definition 14.6: *Let $(X, \mathcal{R})$ be a range space, and let $\mathcal{D}$ be a probability distribution on X. A set $S \subseteq X$ is an* ϵ-sample *for X with respect to $\mathcal{D}$ if for all sets $R \in \mathcal{R}$,*

$$\left| \Pr_{\mathcal{D}}(R) - \frac{|S \cap R|}{|S|} \right| \leq \epsilon.$$

Relative freq. of a range.

Again, by fixing the distribution $\mathcal{D}$ to be uniform over a finite set $A \subseteq X$, we obtain the combinatorial version of this concept.

Definition 14.7 [combinatorial definition]: *Let $(X, \mathcal{R})$ be a range space, and let $A \subseteq X$ be a finite subset of X. A set $N \subseteq A$ is a* combinatorial ϵ-sample *for A if for all sets $R \in \mathcal{R}$,*

$$\left| \frac{|A \cap R|}{|A|} - \frac{|N \cap R|}{|N|} \right| \leq \epsilon.$$

Definition 14.8: *A range space $(X, \mathcal{R})$ has the* uniform convergence property *if for every $\epsilon, \delta > 0$ there is a sample size $m = m(\epsilon, \delta)$ such that for every distribution $\mathcal{D}$ over X, if S is a random sample from $\mathcal{D}$ of size m then, with probability at least $1 - \delta$, S is an ϵ-sample for X with respect to $\mathcal{D}$.*

ϵ-sample theorem :

Theorem 14.15: *Let $(X, \mathcal{R})$ be a range space with VC dimension d and let $\mathcal{D}$ be a probability distribution on X. For any $0 < \epsilon, \delta < 1/2$, there is an*

$$m = O\left(\frac{d}{\epsilon^2} \ln \frac{d}{\epsilon} + \frac{1}{\epsilon^2} \ln \frac{1}{\delta} \right)$$

such that a random sample from $\mathcal{D}$ of size greater than or equal to m is an ϵ-sample for X with probability at least $1 - \delta$.

§ Application: Agnostic Learning.

In PAC learning, we assumed there is a $c^* \in \mathcal{C}$ that is correct on all items in X and so conforms with all examples in training set.

→ But training set can have error & there may not be any correct classification in $\mathcal{C}$. In agnostic learning, the goal is find a nearly best classification c' s.t.

$$Pr_{\mathcal{D}}(c'(x) \neq c(x)) \leq \inf_{h \in \mathcal{C}} Pr_{\mathcal{D}}(h(x) \neq c(x)) + \epsilon.$$

($c'(x) \neq c(x)$: correct classif, may not be in $\mathcal{C}$)

If the training set define an $\epsilon/2$-sample for $(X, \Delta(c))$ then algo has sufficiently many examples to estimate the error prob of each $c' \in \mathcal{C}$ to within an additive error $\epsilon/2$.

Using ϵ-sample theorem, agnostic learning of a concept class with VC-dim d requires $O\left(\frac{d}{\epsilon^2} \ln \frac{d}{\epsilon^2} + \frac{1}{\epsilon^2} \ln \frac{1}{\delta}\right)$ samples.

Theorem 14.19: *The following three conditions are equivalent:*

1. *A concept class C over a domain X is agnostic PAC learnable.*
2. *The range space (X, C) has the uniform convergence property.*
3. *The range space (X, C) has a finite VC dimension.*

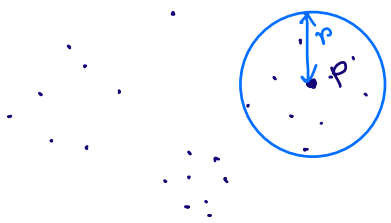
§ Applications : Data mining.

① Estimating dense neighborhoods.

Given : n (very large) points in $\mathbb{R}^2$, $p = (x, y)$, $r \in \mathbb{R}$.

Goal : What fraction of points are distance $\leq r$ from (x, y).

[Application : opening new facility / business]



We define range space $(\mathbb{R}^2, \mathcal{R})$ where $\mathcal{R}$ includes $\forall (x, y) \in \mathbb{R}^2$ and $r \in \mathbb{R}^+$, set of all points inside the disk of radius r centered at (x, y).

VC-Dim of the set of all disks = 3. ← constant

We can sample a random set of $O\left(\frac{1}{\varepsilon^2} \ln \frac{1}{\varepsilon\delta}\right)$ points and give fast approximate answers to all the queries by scanning only the sample.

ϵ-sample theorem guarantees w.p. $\geq 1 - \delta$, we can answer all queries within ϵ of the correct value.

We can also it for other purposes, such as (approx) identifying k densest disks.

- Similar example : Range searching.
(How many points are included in a query rectangle?)

② Mining frequent itemsets.

→ Given: A set of items $\mathcal{I}$, a collection of transactions $\mathcal{T}$, where each transaction $t \in \mathcal{T}$ is $\subseteq \mathcal{I}$. — Both $|\mathcal{I}|, |\mathcal{T}|$ are large.

Goal: Find set of items that appear in $\geq \theta$ fraction of transactions.

Say we want to characterize freq $\geq \theta$ to be frequent items, freq $\leq \theta - \epsilon$ to be infrequent. Freq. $[\theta - \epsilon, \theta]$ can be ambiguous.

The number of possible of transactions = subsets of $\mathcal{I}$ is huge!

Even for transactions of size $\leq \ell$, there can be $O(|\mathcal{I}|^\ell)$ of them which could be frequent.

(HW: Using Chernoff + union bound would give $\Omega\left(\frac{\theta}{\epsilon^2}\left(\ell \ln|\mathcal{I}| + \ln\frac{1}{\delta}\right)\right)$ samples are needed.)

ε-sample does better!

For each $s \subseteq \mathcal{I}$, $T(s) := \{t \in \mathcal{T}, s \subseteq t\}$ (transactions that include s.)

let $\mathcal{R} = \{T(s) \mid s \subseteq \mathcal{I}\}$.

Claim: VC-Dim $[(\mathcal{T}, \mathcal{R})] \leq \ell$. ($\ell$: the max size of any transaction in the data set)

- A transaction of size ℓ has 2^ℓ subsets, is therefore included in $\leq 2^\ell$ ranges.

Now, consider $S \subseteq \mathcal{T}$, $\mathcal{R}_S = \{R \cap S \mid R \in \mathcal{R}\}$.

Then, $|\mathcal{R}_S| \leq 2^\ell$, as s can belong to atmost 2^ℓ ranges.

Thus no more than l transactions can be shattered. $\Rightarrow$ VC-dim $= l$.

Hence, by ε-sample theorem, w.p. $\geq 1-\delta$, a sample of size $O(l/\varepsilon^2 \ln l/\varepsilon + 1/\varepsilon^2 \ln 1/\delta)$ guarantee all itemsets are accurately determined to within $\varepsilon/2$ of their true proportion

- This is enough to identify frequent itemsets.

§ Rademacher complexity.

- Bounds can depend on the training set distribution
- Generalizes to nonbinary functions.