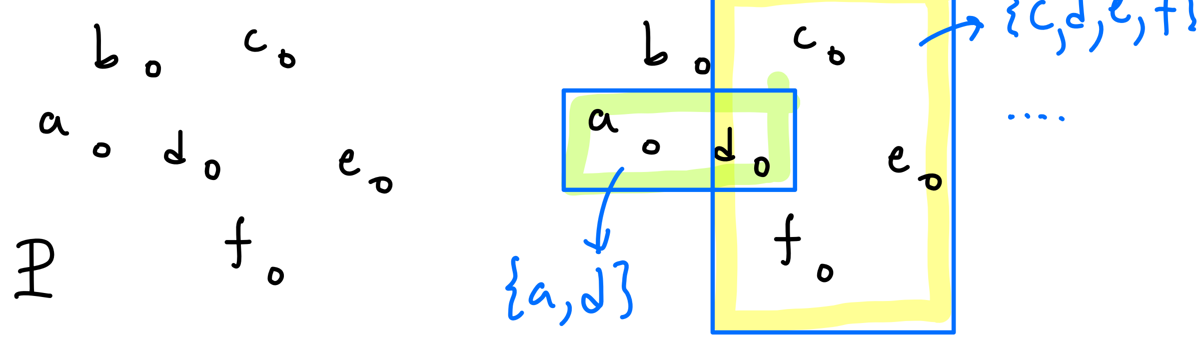
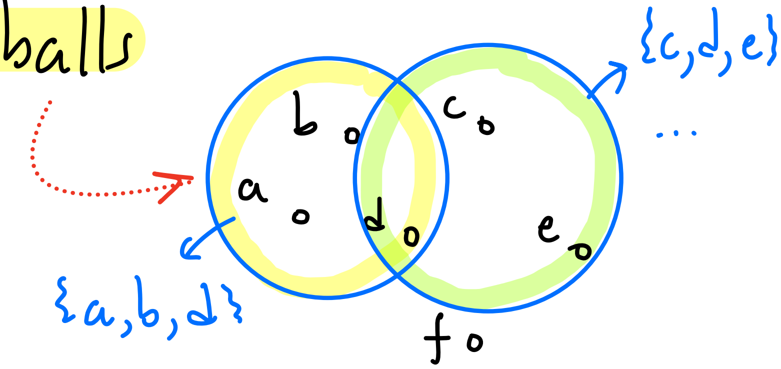
CMSC 754 - Computational Geometry
Lecture 19 - Sampling & VC-Dimension

Geometric Set Systems:

- Many problems involve sets of points that are defined by geometric objects
- Example: Given a set $P \subseteq \mathbb{R}^d$, consider all subsets of P contained in:
    - axis-aligned rectangles



    - Euclidean balls



Range Space:

Given a set P, let 2^P denote the power set of P, consisting of all subsets of P ($|2^P| = 2^{|P|}$)

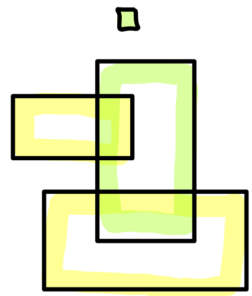
Range space is a pair $(X, \mathcal{R})$ where:

X - **domain** (a set)

$\mathcal{R}$ - **ranges** - a subset of 2^X

Eg. $X = \mathbb{R}^2$

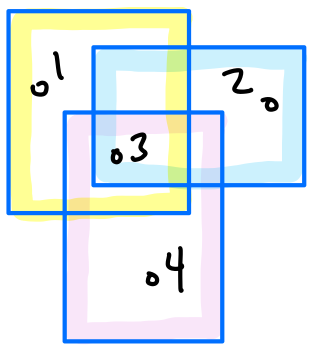
$\mathcal{R}$ = set of all axis-aligned rectangles (each is an infinite set)



Restriction: Given $P \subseteq X$, define

$$\mathcal{R}_{|P} = \{ P \cap Q \mid Q \in \mathcal{R} \}$$

the **restriction of $\mathcal{R}$ to P**



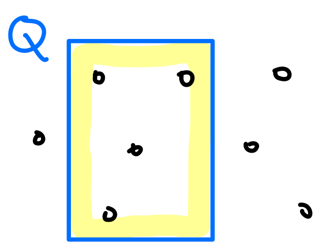
$\mathcal{R}_{|P} = \emptyset, \{1\}, \dots, \{4\}, \dots \{1,2,3,4\}\}$

But not: $\{1,4\}$ or $\{1,2,4\}$

Range space $(X, \mathcal{R})$ is **discrete** if $|X|$ finite

Given a discrete range space $(P, \mathcal{R})$ and any $Q \in \mathcal{R}$ define Q's **measure**

$$\mu(Q) = \frac{|Q \cap P|}{|P|}$$



$\mu(Q) = \frac{4}{8} = \frac{1}{2}$

Sampling: Rather than deal with entire point set (may be huge) we would like a "good" sample.

Given $S \subseteq P$ (presumably $|S| \ll |P|$) define

$$\hat{\mu}_S(Q) = \frac{|Q \cap S|}{|S|}$$

(When S is clear, we write $\hat{\mu}(Q)$)

How good is S as a sample?

Given a discrete range space (P, R) + $\varepsilon > 0$

ε-sample: $S \subseteq P$ is an ε-sample if

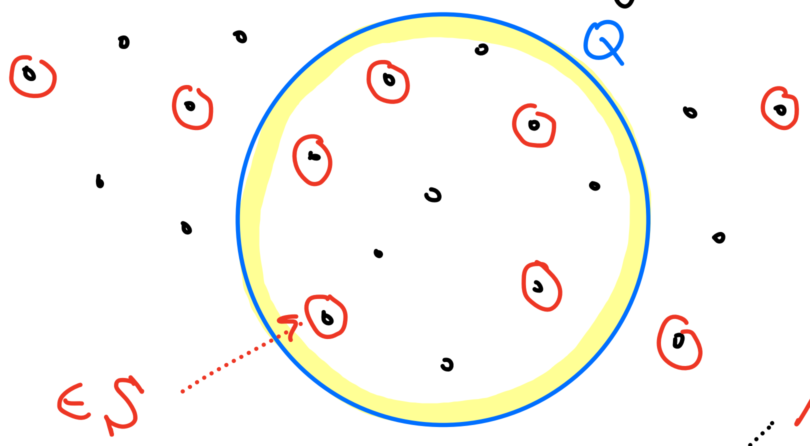
$$|\mu(Q) - \hat{\mu}(Q)| \leq \varepsilon \quad \forall Q \in R$$

ε-net: $S \subseteq P$ is an ε-net if

$$\mu(Q) \geq \varepsilon \Rightarrow S \cap Q \neq \emptyset \quad \forall Q \in R$$

Intuition:

- S is an **ε-sample** if it captures roughly the **same proportion** of elements for any range



$\mu(Q) = \frac{10}{20} = 0.5$

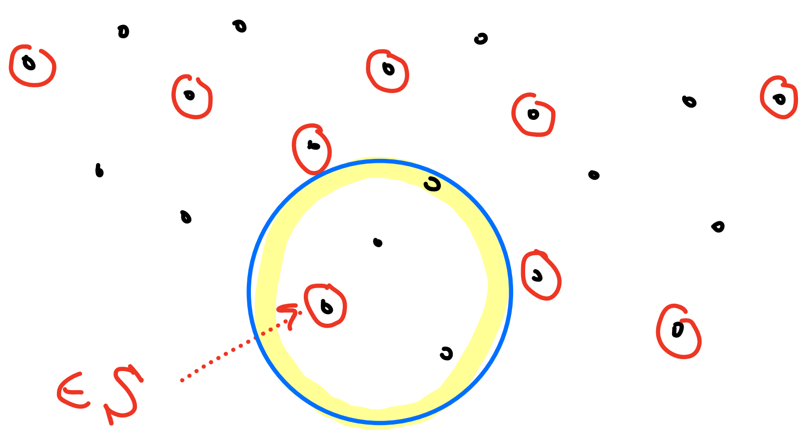
$\hat{\mu}(Q) = \frac{5}{8} = 0.625$

$|\mu(Q) - \hat{\mu}(Q)| = 0.125 = 1/8$

If this holds for all ranges in $\mathcal{R}$, S is a $1/8$-sample.

- A range Q is **ε-heavy** if $\mu(Q) \geq \varepsilon$

An **ε-net hits all ε-heavy ranges**

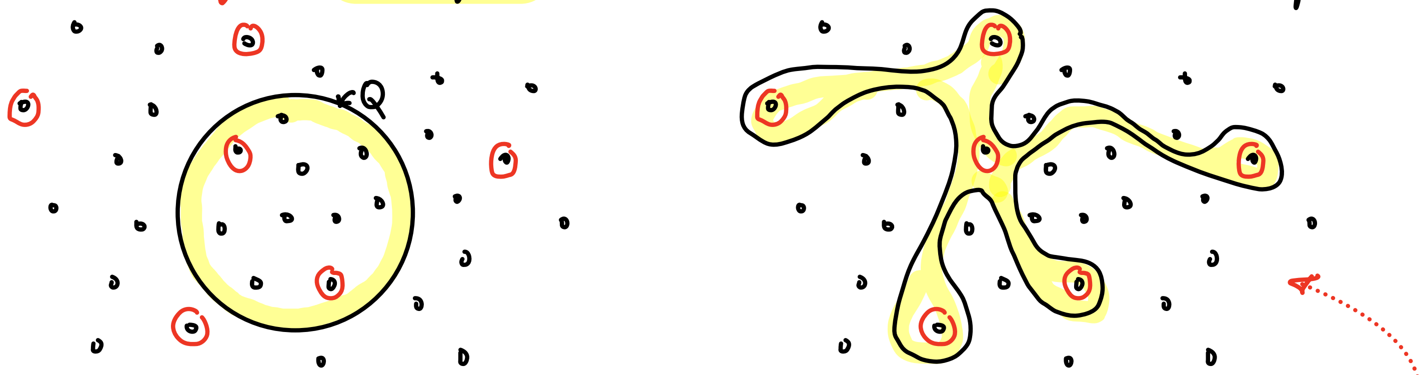


$\varepsilon = 1/5$

Any range that contains $\geq \varepsilon \cdot |P| = 4$ pts must hit a pt of S

How to construct ε-nets + ε-samples?

Intuition: Any sufficiently large random sample should work (with some prob.)



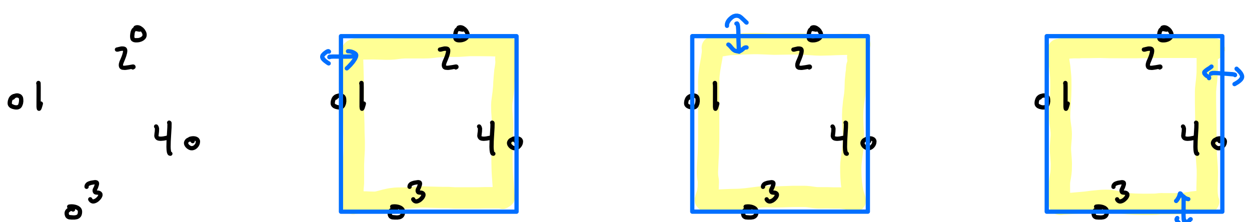
$$\frac{|P\cap Q|}{|P|} = \frac{10}{31} \approx \frac{2}{6} = \frac{|S\cap Q|}{|S|} \qquad\qquad \frac{|P\cap Q|}{|P|} = \frac{6}{31} \neq \frac{6}{6} = \frac{|S\cap Q|}{|S|}$$

But this fails if we allow very wild range shapes.
How to formally forbid such ranges?

VC-Dimension:

Shattering: A range space $(X, \mathcal{R})$ shatters a pt set P if $\mathcal{R}_{|P} = 2^P$ (contains all subsets of P)

E.g. Axis-aligned rectangles shatter the pt set below:

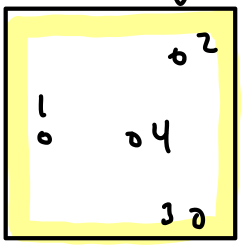


Can include or exclude 1

Can include or exclude 2

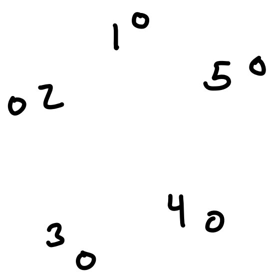
Same for 3 + 4

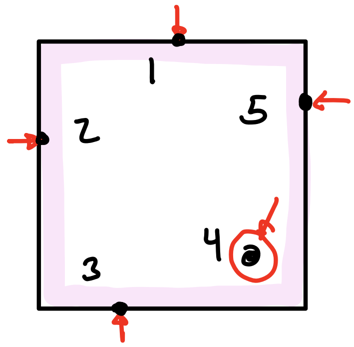
But they can't shatter everything:



Any rect. containg 1,2,3 must contain 4

... and they can never shatter a set of ≥ 5





Any rect that contains the 1,2,3,5 must contain 4

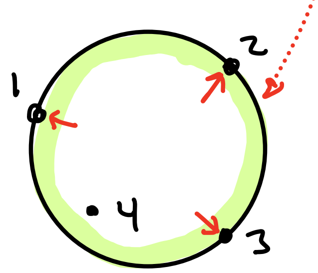
Def: The VC-dimension of a range space $(\mathcal{X}, \mathcal{R})$ is the size of the largest pt set shattered by $\mathcal{R}$.

("VC" – Vapnik-Chervonenkis - 1971)

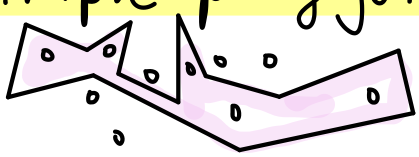
Examples:

→ VC-dim of axis-aligned rects in $\mathbb{R}^2 = 4$

→ VC-dim of Euclidean disks in $\mathbb{R}^2 = 3$



→ VC-dim of simple polygons in $\mathbb{R}^2 = \infty$



Intuitively: Range spaces of constant VC-dim have a constant num. of degrees of freedom

Sauer's Lemma: If $(X, \mathcal{R})$ is a range space of VC-dim d in $|X| = n$, then

$$|\mathcal{R}| = \mathcal{O}(n^d)$$

More precisely:

$$|\mathcal{R}| \leq \Phi_d(n)$$

where:

$$\Phi_d(n) = \binom{n}{0} + \binom{n}{1} + \dots + \binom{n}{d}$$

Observe: Φ satisfies the recurrence:

$$\Phi_d(n) = \Phi_d(n-1) + \Phi_{d-1}(n-1)$$

(Exercise)

Proof: (of Sauer's Lemma) Induction on $d + n$.

Basis: $n = 0$ or $d = 0$ – trivial $\mathcal{R} = \{\emptyset\}$

Step: Fix any $x \in X$

Consider two new range spaces over $X \setminus \{x\}$

$\mathcal{R}_x = \{ Q \setminus \{x\} : Q \cup \{x\} \in \mathcal{R} \text{ \& } Q \setminus \{x\} \in \mathcal{R} \}$

$\hookrightarrow$ Pairs that differ only on x

$\mathcal{R} \setminus \{x\} = \{ Q \setminus \{x\} : Q \in \mathcal{R} \}$

$\hookrightarrow$ Just remove x

Example: $X = \{1, 2, 3, 4\}$ let $x = 4$

Suppose $\mathcal{R}$ has: $\qquad\qquad$ $\mathcal{R}_x$ has:

$\{2,3\}$ & $\{2,3,4\}$ $\longrightarrow$ $\{2,3\}$

$\{1\}$ & $\{1,4\}$ $\longrightarrow$ $\{1\}$

$\{\}$ & $\{4\}$ $\longrightarrow$ $\{\}$

and $\mathcal{R}$ has: $\{1,3\}$ but not $\{1,3,4\}$

$\{2,4\}$ but not $\{2\}$

Then: $\mathcal{R}_x = \{ \{\}, \{1\}, \{2,3\} \}$

$\mathcal{R} \setminus \{x\} = \{ \{\}, \{1\}, \{2,3\}, \{1,3\}, \{2\} \}$

Observe:

- $|\mathcal{R}| = |\mathcal{R}_x| + |\mathcal{R} \setminus \{x\}|$
- $\mathcal{R}_x$ has VC-dim $d-1$
- Both over domain of size $n-1$

$\Rightarrow |\mathcal{R}| \leq \Phi_{d-1}(n-1) + \Phi_d(n-1) = \Phi_d(n)$ $\square$

Recall:

Given a discrete range space $(P, \mathcal{R})$ + $\varepsilon > 0$

ε-sample: $S \subseteq P$ is an ε-sample if

$$|\mu(Q) - \hat{\mu}(Q)| \leq \varepsilon \quad \forall Q \in \mathcal{R}$$

ε-net: $S \subseteq P$ is an ε-net if

$$\mu(Q) \geq \varepsilon \Rightarrow S \cap Q \neq \phi \quad \forall Q \in \mathcal{R}$$

Range spaces of low VC-dimension have ε-samples + ε-nets of small size:

ε-Sample Theorem: Given range space $(X, \mathcal{R})$ of VC-dim d, let P be finite subset of X. There exists constant c s.t. with probability $\geq 1-\varphi$, a random sample of P of size $\geq$

$$\frac{c}{\varepsilon^2}\left(d \cdot \log \frac{d}{\varepsilon} + \log \frac{1}{\varphi}\right)$$

is an ε-sample for $(P, \mathcal{R})$.

ε-Net Theorem: Given range space (X, R) of VC-dim d, let P be finite subset of X. There exists constant c s.t. with probability $\geq 1-\varphi$, a random sample of P of size $\geq$

$$\frac{c}{\varepsilon}\left(d \log \frac{1}{\varepsilon} + \log \frac{1}{\varphi}\right)$$

is an ε-net for (P, R).

Too many parameters! :(

tl;dr : - Constant VC-dim
- Constant prob. of success

Size of ε-sample is $O\left(\frac{1}{\varepsilon^2} \log \frac{1}{\varepsilon}\right)$

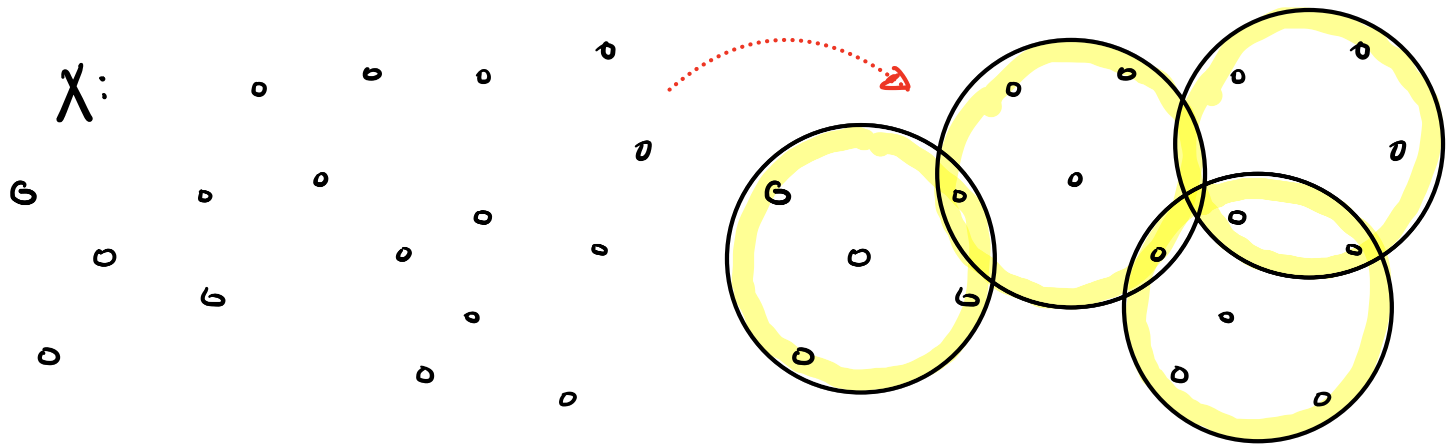
ε-net is $O\left(\frac{1}{\varepsilon} \log \frac{1}{\varepsilon}\right)$

Proofs? See Har-Peled's book

Application: Geometric Set Cover

Given a pt set X + a collection of sets $\mathcal{R}$ over X, a cover is a collection of sets from $\mathcal{R}$ that contain every pt of X

E.g. X is a set of n pts in $\mathbb{R}^d$
$\mathcal{R}$ = set of all unit Euclidean balls in $\mathbb{R}^d$



Set cover Problem: Given X and $\mathcal{R}$, find the smallest cover of X

- Set cover is NP-hard ☹
- No known constant factor approximation ☹
- Simple greedy algorithm computes a cover of size

$$\leq (\ln |X|) \cdot \text{opt}$$

Select set that covers the most uncovered pts

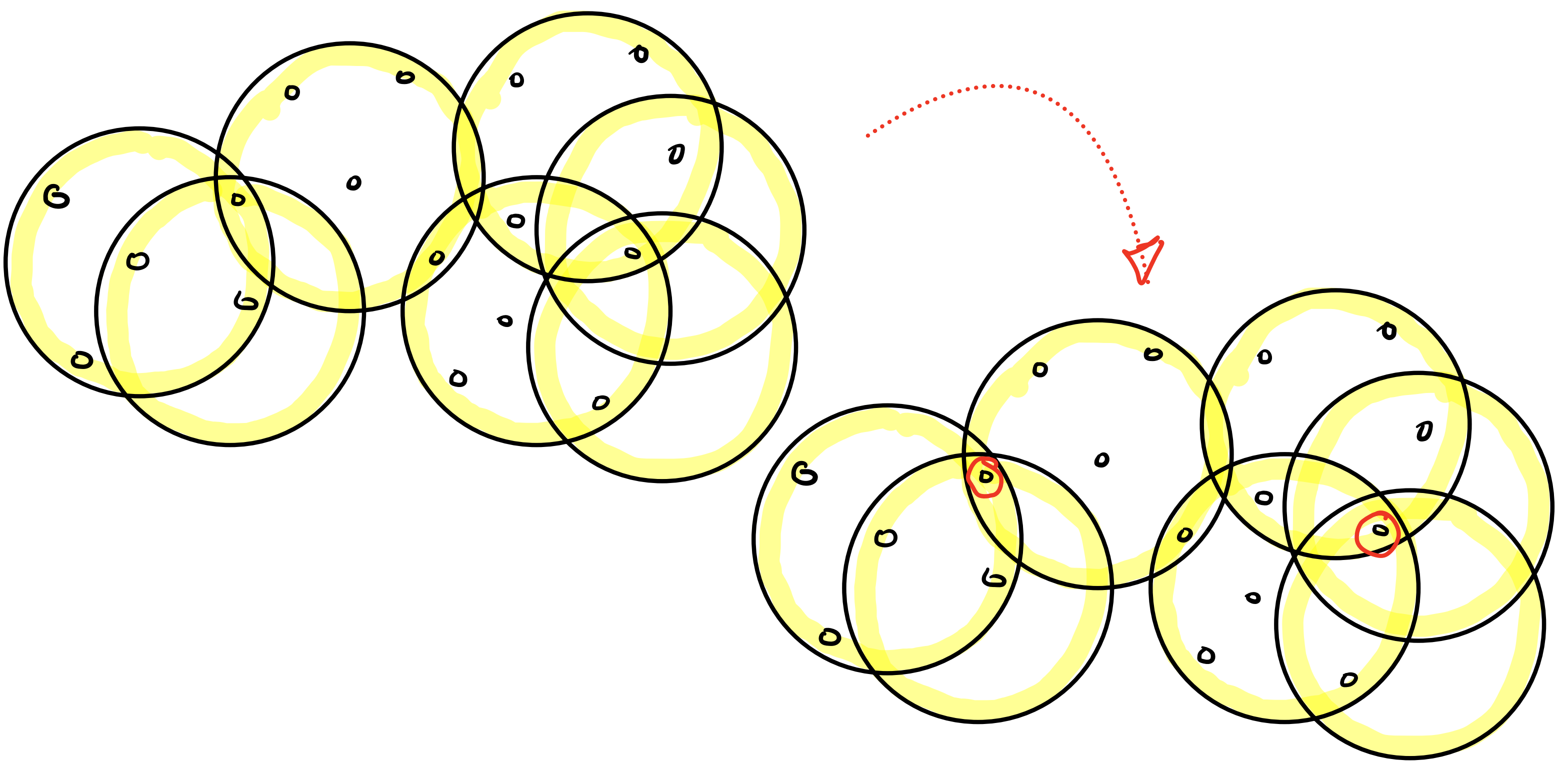
We'll show that if $(X, \mathcal{R})$ is a set system of constant VC-dimension, it is possible to compute an approx. solution of size

$\leq (\log k) \cdot opt$

where k is number of sets in opt. cover (Note $k < |X|$, so this is always better)

Set cover $\longleftrightarrow$ Hitting Set Duality

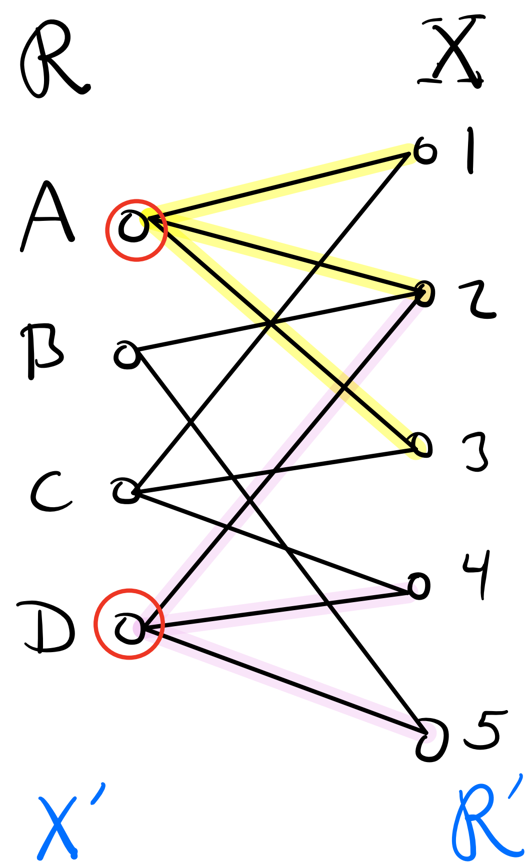
Hitting Set: Given a collection of sets $\mathcal{R}$ over some domain X, a hitting set is a subset of X such that every set of $\mathcal{R}$ contains at least one of them.



Set cover + hitting set are the same problem in disguise

E.g. $A = \{1,2,3\}$ $B = \{2,5\}$
$C = \{1,3,4\}$ $D = \{2,4,5\}$

Cover: $\{A,D\}$



Let's reinterpret: sets $\longrightarrow X'$; pts $\longrightarrow R'$

1: $\{A,C\}$ 2: $\{A,B,D\}$ 3: $\{A,C\}$ 4: $\{C,D\}$ 5: $\{B,D\}$

Hitting set: $\{A,D\}$

$\boxed{\text{Obs: } (X,R) \text{ has set cover of size } k \text{ iff } (X',R') \text{ has hitting set of size } k}$

Theorem: Given a set system (X, R) of constant VC-dimension, in polynomial time it is possible to compute a hitting set of size $O(k^* \log k^*)$ where k^* = size of optimal hitting set.

Note: A set has constant VC-dim iff its dual has constant VC-dim.

Iterative Reweighting:

Weighted ε-Nets: Given a set system (X, R) where each $x \in X$ has a positive weight $w(x)$. Let $w(X)$ be total weight:

$$w(X) = \sum_{x \in X} w(x)$$

A set $S \subseteq X$ is an ε-net if

$$\forall Q \subseteq R \text{ if } \frac{w(Q \cap P)}{w(P)} \geq \varepsilon \text{ then } Q \cap S \neq \emptyset$$

Standard ε-net ≡ all pts have $w(x) = 1$

Weighted sampling:

ε-Net Theorem still holds, but rather than random sample of size $O(\frac{1}{\varepsilon}\log\frac{1}{\varepsilon})$ sample each point with probability proportionate to its weight to get a set of this size.

Iterative Reweighting:

- Guess the size k of opt hitting set (binary search to get best k)
- Set all weights to 1
- Repeat:
  - $S \leftarrow$ weighted ε-net of X
  - Is this a hitting set? yes → success
  - No? Find any set $Q \subseteq R$ not hit + double weights of all $x \in Q$
- Too many iterations? Fail → try larger k

Intuition: If we fail to hit we double weights of unhit object — more likely to hit next time.

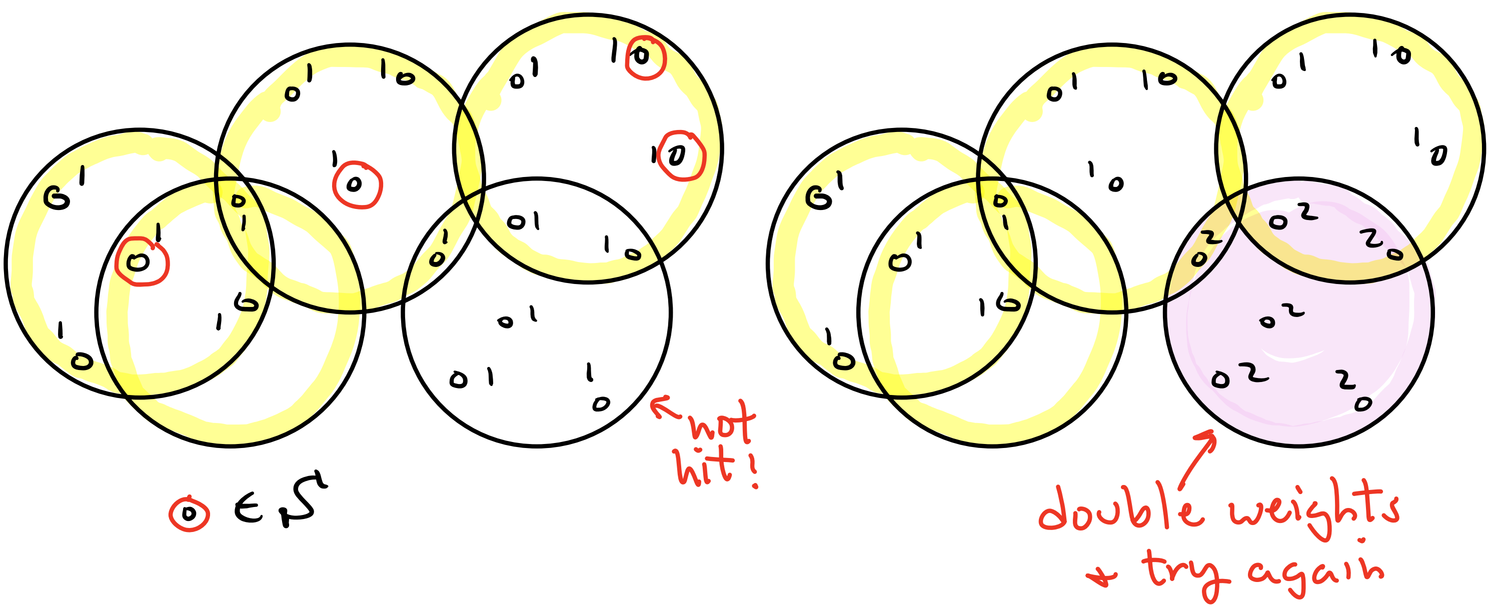
Why it works: Critical items (in opt. solution) increase in weight rapidly - eventually they are all selected.

Algorithm: Given $(X, \mathcal{R})$

for $k = 1, 2, 4, \ldots, 2^i, \ldots$ until success

// Guess that $\exists$ hitting set of size k

- $\forall x \in X$ set $w(x) \leftarrow 1$
- Set $\varepsilon \leftarrow 1/4k$
- Repeat until success or $2k \cdot \lg n/k$ iterations
    - $S \leftarrow$ wgt ε-net of size $c \cdot k \cdot \log k$ (for suitable const. c)
    - are all sets of $\mathcal{R}$ hit by S?
        - yes $\rightarrow$ return with success!
        - no $\rightarrow$ find any set $Q \in \mathcal{R}$ not hit
          $\forall x \in Q$, $w(x) \leftarrow 2 \cdot w(x)$



Why this works? Assume k is correct

- Since opt hitting set hits all sets, at least one point of opt doubles in weight
- Weight of opt hitting set grows exponentially fast
- Total weight of pt set grows much more slowly
- Soon, opt hitting set's weight is so high we must sample it.

Lemma: If (X, R) has hitting set of size k, then the repeat-loop has success within $2k \cdot \lg n/k$ iterations. ($\lg \equiv \log_2$)

Proof: Let $n = |X|$ $m = |R|$

- Let H be hitting set of size k

  $\varpi_i(X)$ = total weight after i^{th} iteration

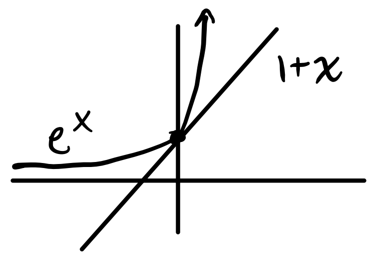
  $\varpi_i(H)$ = weight of H " " "

- Note: $\varpi_0(X) = |X| = n$

- Since S is an ε-net, if we fail to hit a set Q, then $w_i(Q) < \varepsilon \varpi_i(X)$

$$\Rightarrow \quad w_i(X) = w_{i-1}(X) + w_{i-1}(Q)$$
$$\leq w_{i-1}(X) + \varepsilon \cdot w_{i-1}(X)$$
$$= (1+\varepsilon) w_{i-1}(X)$$

$$\Rightarrow \quad w_i(X) \leq (1+\varepsilon)^2 w_{i-2}(X)$$
$$\leq (1+\varepsilon)^3 w_{i-3}(X)$$
$$\vdots$$
$$\leq (1+\varepsilon)^i w_0(X) = (1+\varepsilon)^i \cdot n$$



Fact: $1+x \leq e^x$

$$\Rightarrow \quad w_i(X) \leq n \cdot e^{i \cdot \varepsilon}$$

Since H hits all sets, it hits Q

$\Rightarrow$ in each (unsucessful) iteration, at least one element of H doubles

$\Rightarrow$ growth rate of $w_i(H)$ is slowest if all its members double at same rate (Jensen's Ineq.)

$\Rightarrow$ After i^{th} iteration, each of the k elements of H doubled i/k times

$$\Rightarrow \quad w_i(H) \geq k \cdot 2^{i/k}$$

Since $H \subseteq X$, we know $W_i(H) \leq W_i(X)$

$$\Rightarrow \quad k \cdot 2^{i/k} \leq n \cdot e^{i \cdot \varepsilon}$$

Recall, we set $\varepsilon \leftarrow 1/4k$

$$\Rightarrow \quad k \cdot 2^{i/k} \leq n \cdot e^{i/4k}$$

$\lg e < 2$

$$\Rightarrow \quad \lg k + \frac{i}{k} \leq \lg n + \frac{i}{4k} \lg e$$

$$\leq \lg n + \frac{i}{2k}$$

$$\Rightarrow \frac{i}{k} - \frac{i}{2k} = \frac{i}{2k} \leq \lg n - \lg k = \lg \frac{n}{k}$$

$$\Rightarrow \quad \text{No. of iterations } i \leq 2k \cdot \lg \frac{n}{k}$$

(If we exceed this number, we know $|H| > k$, and we fail) $\square$

Total time:

$$\left(2k \cdot \log \frac{n}{k}\right) \cdot \left[(k \cdot \log k) + m \cdot k\right]$$

since $k \leq n$

$$= O(n^2 \cdot m \cdot \log n)$$