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
    - $P_f$ $\{a, d\}$
    - $P_g$ $\{c, d, e, f\}$
  - Euclidean balls
    - $P_1$ $\{a, b, d\}$
    - $P_2$ $\{c, d, e\}$
  - Halfplanes
    - $P_3$ $\{a\}$
    - $P_4$ $\{e, f\}$
Range Space:

Given a set \( P \), let \( 2^P \) denote the power set of \( P \), consisting of all subsets of \( P \) \((12^P = 2^{1P})\).

Range space is a pair \((X, R)\) where:
- \( X \) - domain (a set)
- \( R \) - ranges - a subset of \( 2^X \)

E.g. \( X = \{0,1,2,3,\ldots\} \)
- \( R \) = all subsets of contiguous values
  - \{3, \{0,3\}, \{1\}, \{2\}, \ldots\}
  - \{0,1,3, \{1,2\}, \{2,3\}, \ldots\}
  - \{0,1,2,3, \{1,2,3\}, \ldots\}

Restriction: Given \( P \subseteq X \), define

\[ R_{|P} = \{ P \cap Q \mid Q \in R \} \]

the restriction of \( R \) to \( P \)

E.g. \( P = \{3,4,5\} \)

\[ R_{|P} = \{ \{3\}, \{3,3\}, \{4,3\}, \{5\}, \{3,4,3\}, \{4,5\}, \{3,4,5\} \} \]
Geometric Setting:

$$(X, R): X = \mathbb{R}^2$$

$R = \text{closed axis-parallel rects}$$

Given $P = \{p_1, \ldots, p_n\} \subseteq \mathbb{R}^2$

$R_p = \{\text{all subsets of } P \text{ defined by containment in rect.}\}$

$R_{p_1} = \emptyset, \{1\}, \ldots, \{4\},\{1,2\}, \{1,3\}, \{2,3\}, \{2,4\}, \{3,4\}, \{1,2,3\}, \{1,3,4\}, \{2,3,4\}, \{1,2,3,4\}\}$

Not: $\{1,4\}$ or $\{1,2,4\}$

Range space $(X, R)$ is discrete if $|X|$ finite

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

$$\mu(Q) = \frac{|Q \cap \mathbb{R}^2|}{|\mathbb{R}^2|}$$

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

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

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

How good is $S$ as a sample?

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

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

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

$\epsilon$-net: $S \subseteq P$ is an $\epsilon$-net if

$$\mu(Q) \geq \epsilon \implies S \cap Q \neq \emptyset \quad \forall Q \in R$$
Intuition:
- $S$ is an $\varepsilon$-sample if it captures roughly the same proportion of elements for any range $Q$.

\[ \mu(Q) = \frac{10}{20} = 0.5 \]
\[ \hat{\mu}(Q) = \frac{5}{8} = 0.625 \]

\[ |\mu(Q) - \hat{\mu}(Q)| = 0.125 = \frac{1}{8} \]

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

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

An $\varepsilon$-net hits all $\varepsilon$-heavy ranges

$\varepsilon = \frac{1}{5}$

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

How to construct $\varepsilon$-nets + $\varepsilon$-samples?

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

\[
\frac{|\mathcal{P} \cap Q|}{|\mathcal{P}|} = \frac{10}{31} \approx \frac{2}{6} = \frac{|\mathcal{P} \cap Q|}{|\mathcal{Q}|} \quad \frac{|\mathcal{P} \cap Q|}{|\mathcal{P}|} = \frac{6}{31} \neq \frac{6}{6} = \frac{|\mathcal{P} \cap Q|}{|\mathcal{Q}|}
\]

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 $\mathcal{P}$ if $\mathcal{R}_{|\mathcal{P}} = 2^\mathcal{P}$ (contains all subsets of $\mathcal{P}$)

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

\[
\{1,2\}, \{2,4\}, \ldots \quad \{1,2,4\}, \{2,3,4\}, \ldots
\]
But they can't shatter everything:

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

... and they can never shatter a set of \( \geq 5 \).

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

**Def:** The VC-dimension of a range space \((X, R)\) is the size of the largest point set shattered by \(R\).

("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 \((X, R)\) is a range space of VC-dim \(d\) in \(|X| = n\), then
\[|R| = O(n^d)\]
More precisely:
\[|R| \leq \Phi_d(n)\]
where:
\[\Phi_d(n) = \binom{n}{0} + \binom{n}{1} + \ldots + \binom{n}{d}\]

Observe: \(\Phi\) 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: \(n = 0 \lor d = 0\) - trivial \(R = \{\emptyset\}\)

Step: Fix any \(x \in X\)
Consider two new range spaces:
over \(X \setminus \{x\}\)
\[ R_x = \{ \{ x \} : \emptyset \cup \{ x \} \in R \cup \{ x \} \in R \} \]
\[ \Rightarrow \text{Pairs that differ only on } x \]

\[ R \setminus \{ x \} = \{ \{ x \} : \emptyset \in R \} \]
\[ \Rightarrow \text{Just remove } x \]

**Example:** \( X = \{1, 2, 3, 4\} \) let \( x = 4 \)

Suppose \( R \) has:

\[
\begin{align*}
\{2, 3\} & \mapsto \{2, 3\} \\
\{1\} & \mapsto \{1, 4\} \\
\emptyset & \mapsto \{4\}
\end{align*}
\]

and \( R \) has:

\[
\begin{align*}
\{1, 3\} & \text{ but not } \{1, 3, 4\} \\
\{2, 4\} & \text{ but not } \{2\}
\end{align*}
\]

Then:

\[
R_x = \{\emptyset, \{1\}, \{2, 3\}\}
\]

\[
R \setminus \{ x \} = \{\emptyset, \{1\}, \{2, 3\}, \{1, 3\}, \{2\}\}
\]

**Observe:**

- \(|R| = |R_x| + |R \setminus \{ x \}|\)
- \(R_x \) has VC-dim d-1
- Both over domain of size n-1

\[ |R| \leq \Phi_{d-1}(n-1) + \Phi_d(n-1) = \Phi_d(n) \]
Recall:
Given a discrete range space \((P, R)\) with \(\epsilon > 0\)

**\(\epsilon\)-sample:** \(S \subseteq P\) is an \(\epsilon\)-sample if \[|\mu(Q) - \hat{\mu}(Q)| \leq \epsilon \quad \forall Q \in R\]

**\(\epsilon\)-net:** \(S \subseteq P\) is an \(\epsilon\)-net if \[
\mu(Q) \geq \epsilon \implies S \cap Q \neq \emptyset \quad \forall Q \in R
\]

Range spaces of low VC-dimension have \(\epsilon\)-samples and \(\epsilon\)-nets of small size:

**\(\epsilon\)-Sample Theorem:** Given range space \((X, R)\) of VC-dim \(d\), let \(P\) be a finite subset of \(X\). There exists a constant \(c\) s.t. with probability \(\geq 1 - \varphi\), a random sample of \(P\) of size \(\geq \frac{\epsilon^2}{\epsilon^2} (d \cdot \log \frac{d}{\epsilon} + \log \frac{1}{\varphi})\) is an \(\epsilon\)-sample for \((P, R)\).
**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 $X$ and a collection of sets $R$ over $X$, a cover is a collection of sets from $R$ that contain every pt of $X$.

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

Set cover Problem: Given $X$ and $R$, find the smallest cover of $X$.

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

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

\[ \leq (\log k) \cdot \text{opt} \]

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

**Set cover \leftrightarrow** Hitting Set Duality

**Hitting Set:** Given a collection of sets \(R\) over some domain \(X\), a hitting set is a subset of \(X\) such that every set of \(R\) contains at least one of them.
Set cover and hitting set are the same problem in disguise.

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

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

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

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

Obs: $(X, R)$ has set cover of size $k$ iff $(X', R')$ has hitting set of size $k$
Theorem: Given a set system \((X, \mathcal{R})\) 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 \(\varepsilon\)-Nets: Given a set system \((X, \mathcal{R})\) 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 \(\varepsilon\)-net if

\[
\forall Q \in \mathcal{R} \text{ if } \frac{w(Q \cup P)}{w(P)} \geq \varepsilon \text{ then } Q \cap S \neq \emptyset
\]

Standard \(\varepsilon\)-net = all pts have \(w(x) = 1\)
Weighted sampling:
\[\varepsilon\text{-Net Theorem still holds, but rather than random sample of size } O(\frac{1}{\varepsilon} \log \frac{1}{\varepsilon})\]
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 \(\varepsilon\text{-net of } X\)
  - Is this a hitting set? yes \(\rightarrow\) success
  - No? Find any set \(Q \subseteq R\) not hit + double weights of all \(x \in Q\)
- Too many iterations? Fail \(\rightarrow\) try larger \(k\)

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

Huh? This can’t work! Just chasing our tail.
Algorithm: Given \((X, R)\)

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 \(\varepsilon \leftarrow \frac{1}{4k}\) (for suitable constant \(c\))

- Repeat until success or \(2k \cdot \log \frac{n}{k}\) iterations

- \(S \leftarrow \text{wgt } \varepsilon\text{-net of size } c \cdot k \cdot \log k\)

- Are all sets of \(R\) hit by \(S\)?
  - yes \(\rightarrow\) return with success!
  - no \(\rightarrow\) find any set \(Q \in R\) not hit

\(\forall x \in Q, w(x) \leftarrow 2 \cdot w(x)\)

\(\circ \in S\)

not hit!

\(\text{double weights + try again}\)
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 sets 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 $2k \cdot \lg^{1/k}$ iterations. ($\lg \equiv \log_2$)

Proof: Let $n = |X|$ $m = |R|$
- Let $H$ be hitting set of size $k$
  - $W_i(X) =$ total weight after $i$th iteration
  - $W_i(H) =$ weight of $H$
- Note: $W_0(X) = |X| = n$
- Since $S$ is an $\epsilon$-net, if we fail to hit a set $Q$, then $w_i(Q) < \epsilon W_i(X)$
\[ \Rightarrow \ W_i(X) = W_{i-1}(X) + \omega_{i-1}(Q) \]
\[ \leq W_{i-1}(X) + \varepsilon \cdot W_{i-1}(X) \]
\[ = (1 + \varepsilon) W_{i-1}(X) \]

\[ \Rightarrow \ W_i(X) \leq (1 + \varepsilon)^2 W_{i-2}(X) \]
\[ \leq (1 + \varepsilon)^3 W_{i-3}(X) \]
\[ \vdots \]
\[ \leq (1 + \varepsilon)^i W_0(X) = (1 + \varepsilon)^i n \]

Fact: \[ 1 + x \leq e^x \]

\[ \Rightarrow \ W_i(X) \leq n \cdot e^{i \cdot \varepsilon} \]

Since \( H \) hits all sets, it hits \( Q \)

\[ \Rightarrow \text{in each (unsuccessful) iteration, at least one element of } H \text{ doubles} \]

\[ \Rightarrow \text{growth rate of } W_i(H) \text{ is slowest if all its members double at same rate (Jensen's Ineq.)} \]

\[ \Rightarrow \text{After } i^{\text{th}} \text{ iteration, each of the } k \text{ elements of } H \text{ doubled } \frac{i}{k} \text{ times} \]

\[ \Rightarrow \ W_i(H) \geq k \cdot 2^{i/k} \]
Since $H \subseteq X$, we know $W_i(H) \leq W_i(X)$

\[ k \cdot 2^{i/k} \leq n \cdot e^{i\varepsilon} \]

Recall, we set $\varepsilon \leftarrow \frac{1}{4k}$

\[ k \cdot 2^{i/k} \leq n \cdot e^{i/4k} \]

\[ \Rightarrow \quad \lg k + \frac{i}{k} \leq \lg n + \frac{i}{4k} \leq \lg n + \frac{i}{2k} \]

\[ \Rightarrow \quad \frac{i}{k} - \frac{i}{2k} = \frac{i}{2k} \leq \lg n - \lg k = \lg \frac{n}{k} \]

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

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

Total time:

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

\[ = O(n^2 \cdot m \cdot \log n) \quad \text{since } k \leq n \]