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 \((X, R)\) where:

- \(X\) - domain (a set)
- \(R\) - ranges - a subset of \(2^X\)

Eg. \(X = \mathbb{R}^2\)

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

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

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

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

\[ R_{|P} = \emptyset, \{1\}, \{1,2\}, \{1,2,3\} \]

But 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 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 $\varepsilon$-sample if it captures roughly the same proportion of elements for any range

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

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

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

If this holds for all ranges in $\mathcal{R}$, $S$ 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 $\geq \varepsilon \cdot |\mathcal{P}| = 4$ 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.)

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

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

**VC-Dimension:**

**Shattering:** A range space \((X, R)\) shatters a pt set \(P\) if \(R(P) = 2^P\) (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. containing 1, 2, 3 must contain 4

... and they can never shatter a set of 3, 5

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

("VC" - Vapnik-Chervonenkis - 1971)

Examples:

\[
\rightarrow \text{VC-dim of axis-aligned rects in } \mathbb{R}^2 = 4
\]

\[
\rightarrow \text{VC-dim of Euclidean disks in } \mathbb{R}^2 = 3
\]

\[
\rightarrow \text{VC-dim of simple polygons in } \mathbb{R}^2 = \infty
\]
Intuitively: Range spaces of constant VC-dim have a constant num. of degrees of freedom

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

More precisely:
\[|\mathcal{R}| \leq \Phi_d(n)\]

where:
\[\Phi_d(n) = \binom{n}{0} + \binom{n}{1} + \ldots + \binom{n}{d}\]

Observe: \(\Phi\) satisfies the recurrence:
\[\Phi_d(n) = \Phi_d(n-1) + \Phi_{d-1}(n-1)\]

\(\Leftrightarrow\text{(Exercise)}\)

\textbf{Proof:} (of Sauer's Lemma) Induction on \(d + n\).

\textbf{Basis:} \(n=0 \text{ or } d=0\) - trivial \(\mathcal{R} = \{\emptyset\}\)

\textbf{Step:} Fix any \(x \in X\)

Consider two new range spaces: over \(X \setminus \{x\}\)
\[ R_x = \{ Q \setminus \{x\} : Q U \{x\} \in R + Q \setminus \{x\} \in R \} \]

→ Pairs that differ only on \( x \)

\[ R \setminus \{x\} = \{ Q \setminus \{x\} : Q \in R \} \]

→ Just remove \( x \)

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

Suppose \( R \) has:

\[
\begin{align*}
\{2,3\} &+ \{2,3,4\} \quad \rightarrow \quad \{2,3\} \\
\{1\} &+ \{1,4\} \quad \rightarrow \quad \{1\} \\
\{3\} &+ \{4\} \quad \rightarrow \quad \{3\}
\end{align*}
\]

and \( R \) has:

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

Then:

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

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

**Observe**:

\[
\begin{align*}
|R| &= |R_x| + |R \setminus \{x\}| \\
R_x &\text{ has VC-dim } d-1 \\
\text{Both over domain of size } n-1
\end{align*}
\]

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

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

\[\epsilon\text{-net: } S \subseteq P \text{ is an } \epsilon\text{-net if } \mu(Q) \geq \varepsilon \Rightarrow S \cap Q \neq \emptyset \quad \forall Q \in R\]

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

\[\epsilon\text{-Sample Theorem: } \text{Given range space } (X, R) \text{ of VC-dim } d, \text{ let } P \text{ be finite subset of } X. \text{ There exists constant } c \text{ s.t. with probability } \geq 1 - \varphi, \text{ a random sample of } P \text{ of size } \geq \frac{c}{\varepsilon^2} \left( d \cdot \log \frac{d}{\varepsilon} + \log \frac{1}{\varphi} \right) \text{ is an } \epsilon\text{-sample for } (P, R).\]
\( \varepsilon\)-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 \(\varepsilon\)-net for \((P, R)\).

Too many parameters!  

\text{tl; dr :} \quad - \text{Constant VC-dim}  
\quad - \text{Constant prob. of success}  

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

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

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 $X$.

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

**Set cover ↔ 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.  
\[ A = \{1, 2, 3\} \quad B = \{2, 5\} \]
\[ C = \{1, 3, 4\} \quad D = \{2, 4, 5\} \]

Cover: \{A, D\}

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\}

Hitting 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 \((X, R)\) 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 \((X, R)\) 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

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

Standard \(\varepsilon\)-net = all pts have \(w(x) = 1\)
Weighted sampling:
- \(\varepsilon\)-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 \(\varepsilon\)-net of \(X\)
  - Is this a hitting set? \(\text{yes} \rightarrow \text{success}\)
  - No? Find any set \(Q \subseteq R\) not hit + double weights of all \(x \in Q\)
  - Too many iterations? \(\text{Fail} \rightarrow \text{try larger } k\)

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 \((X, R)\)

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 const. \(c\))

- 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

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

\(\circ \in S'\)

\(\text{not hit!}\)

\(\text{double weights} + \text{try again}\)
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 $2k \cdot 1g^{\frac{1}{k}}$ iterations. ($1g = \log_2$)

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 $\varepsilon$-net, if we fail to hit a set $Q$, then $w_i(Q) < \varepsilon W_i(X)$
\[ W_i(X) = W_{i-1}(X) + \omega_{i-1}(Q) \leq W_{i-1}(X) + \varepsilon \cdot W_{i-1}(X) = (1 + \varepsilon) W_{i-1}(X) \]

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

**Fact:** \( 1 + x \leq e^x \)

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

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

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

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

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

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

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

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

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

\[ \varepsilon = \frac{i}{4k} \]

\[ \Rightarrow \quad 1 + \frac{i}{k} \leq \log n + \frac{i}{4k} \]

\[ \leq \log n + \frac{i}{2k} \]

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

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

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

Total time:

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

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