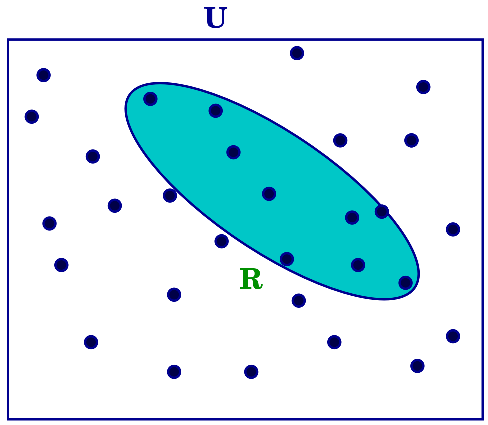
ε-Nets and VC Dimension

- **Sampling is a powerful idea applied widely in many disciplines, including CS.**
- **There are at least two important uses of sampling: estimation and detection.**
- **CNN, Nielsen, NYT etc use polling to estimate the size of a particular group in the larger population.**
- **By sampling a small segment of the population, one can predict the winner of a presidential election (with high confidence). How many prefer Bush to Gore; how many will use a new service etc.**
- **In detection, the goal is to sample so that any group with large probability measure will be caught with high confidence.**
- **Random traffic checks, for example. Frequent speeders (drinkers) are likely to get caught.**

Sampling

- A network monitoring application.
- Want to detect flows that are suspiciously big, in terms of fraction of total packets.
- Set a threshold of $\theta\%$. Any flow that accounts for more than $\theta\%$ of traffic at a router should be flagged.
- Keeping track of all flows is infeasible; millions of flows and billions of packets per second.
- By taking a number of samples that depends only on θ, we can detect offending flows with high probability.
- Track only sampled flows.

Basic Sampling Theorem

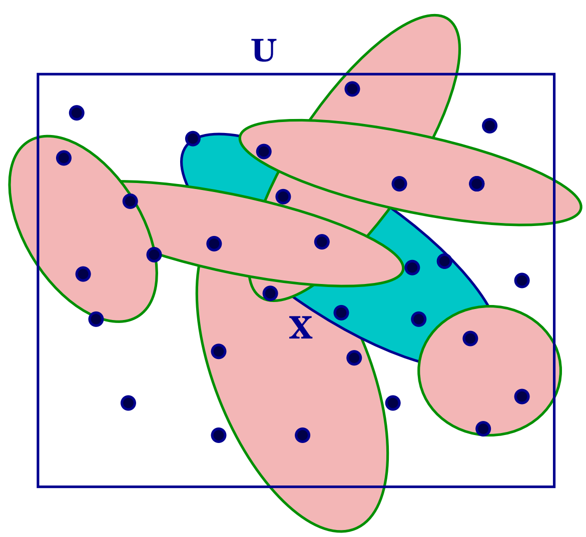


- U **is a ground set (points, events, database objects, people etc.)**
- **Let** $R \subset U$ **be a subset such that** $|R| \geq \varepsilon |U|$**, for some** $0 < \varepsilon < 1$**.**
- **Theorem: A random sample of** $\left(\frac{1}{\varepsilon} \ln \frac{1}{\delta}\right)$ **points from** U **intersects** R **with probability at least** $1 - \delta$**.**
- **Proof: A particular sample point is in** R **with prob** ε**, and not in** R **with prob.** $1 - \varepsilon$**. Prob. that none of the sampled points is in** R **is**

$$\leq \ (1-\varepsilon)^{\frac{1}{\varepsilon} \ln \frac{1}{\delta}} \leq e^{-\ln \frac{1}{\delta}} = \delta.$$

Universal Samples

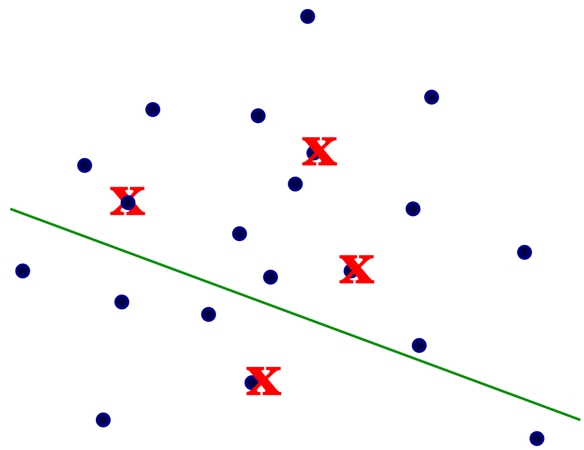
- **Sample size is independent of $|U|$.**
- **Basic sampling theorem guarantees that for a given set R, a random sample set works.**
- **If we want to hit each of the sets R_1, R_2, $\ldots$, R_m, then this idea is too limiting. It requires a separate sample for each R_i.**
- **Can we get a single universal sample set, which hit all the R_i's?**



- **ε-Nets and VC dimension characterize when this is possible.**

ε-Nets

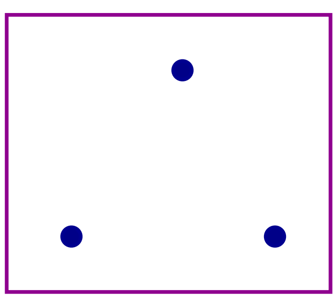
- Let $(\mathcal{U}, \mathcal{R})$ be a finite set system, and let $\varepsilon \in [0, 1]$ be a real number.
- A set $N \subseteq \mathcal{U}$ is called an ε-net for $(\mathcal{U}, \mathcal{R})$ if $N \cap R \neq \emptyset$ for all $R \in \mathcal{R}$ whenever $|R| \geq \varepsilon|\mathcal{U}|$.

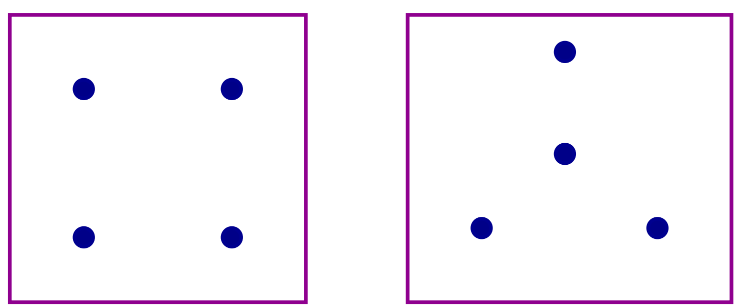


- A more general form of ε-net can be defined using probability measures. Think of this as endowing points of $\mathcal{U}$ with weights.

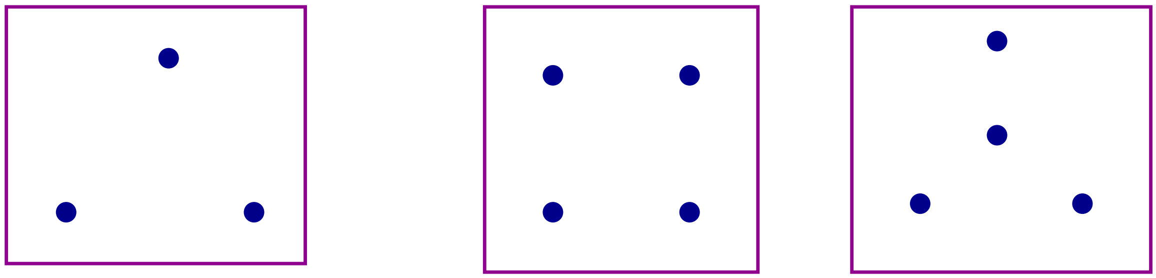
Shatter Function

- A set system $(\mathcal{U}, \mathcal{R})$, where $\mathcal{U}$ is the ground set and $\mathcal{R}$ is a family of subsets.
- $\mathcal{R} = \{R_1, \ldots, R_m\}$, with $R_i \subset \mathcal{U}$, are ranges that we want to hit.
- A subset $X \subset \mathcal{U}$ is shattered by $\mathcal{R}$ if all subsets of X can be obtaind by intersecting X with members of $\mathcal{R}$.
- That is, for any $Y \subseteq X$, there is some $A \in \mathcal{R}$ such that $Y = X \cap A$.
- Examples: $\mathcal{U} =$ points in the plane. $\mathcal{R} =$ half-spaces.





VC Dimension



(i) (ii) (iii)

Shattered by R **Not Shattered by R**

- **The shatter function measures the complexity of the set system.**
- **If instead of half-spaces, we used ellipses, then (ii) and (iii) can be shattered as well.**
- **So, the set system of ellipses has higher complexity than half-spaces.**

VC Dimension: **The VC dimension of a set system $(\mathcal{U}, \mathcal{R})$ is the maximum size of any set $X \subset \mathcal{U}$ shattered by $\mathcal{R}$.**

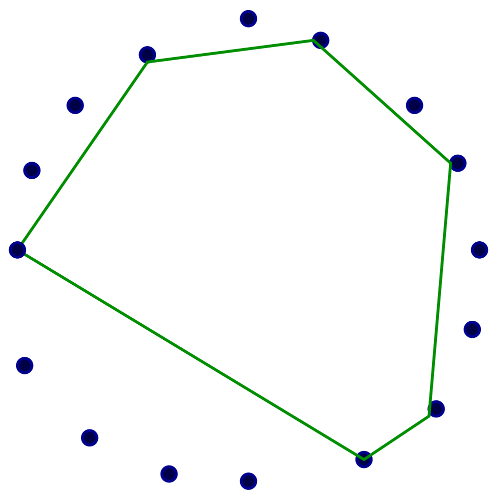
- **Thus, the half-spaces system has VC dimension 3.**

Other Examples

- Set system where $\mathcal{U}$ = points in d-space, and $\mathcal{R}$ = half-spaces, has VC-dimension $d + 1$.
- A simplex is shattered, but no $(d + 2)$-point set is shattered (by Radon's Lemma).
- Set system where $\mathcal{U}$ = points in the plane, and $\mathcal{R}$ = circles, has VC-dimesion 4.

Convex Set System

- **Consider $(\mathcal{U}, \mathcal{R})$, where $\mathcal{U}$ is set of points in the plane, and $\mathcal{R}$ is family of convex sets.**
- **Members of $\mathcal{R}$ are subsets that can be obtained by intersecting $\mathcal{U}$ with a convex polygon.**



Set system of convex polygons

- **Any subset $X \subseteq \mathcal{U}$ can be obtained by intersecting $\mathcal{U}$ with an appropriate convex polygon.**
- **Thus, entire set $\mathcal{U}$ is shattered.**
- **VC dimension of this set system is ∞.**

ε-Net Theorem

- **Suppose $(\mathcal{U}, \mathcal{R})$ is a set system of VC dimension d, and let ε, δ be real numbers, where $\varepsilon \in [0, 1]$ and $\delta > 0$.**

- **If we draw**

$$O\left(\frac{d}{\varepsilon}\log\frac{d}{\varepsilon} + \frac{1}{\varepsilon}\log\frac{1}{\delta}\right)$$

  points at random from $\mathcal{U}$, then the resulting set N is an ε-net with probability $\geq \delta$.

- **Size of ε-Net is independent of the size of $\mathcal{U}$.**

- **Example: Consider set system of points in the plane with half-space ranges. It has VC-dim = 3. Assuming ε, δ constant, we have an ε-net of $O(1)$ size.**

Consequences

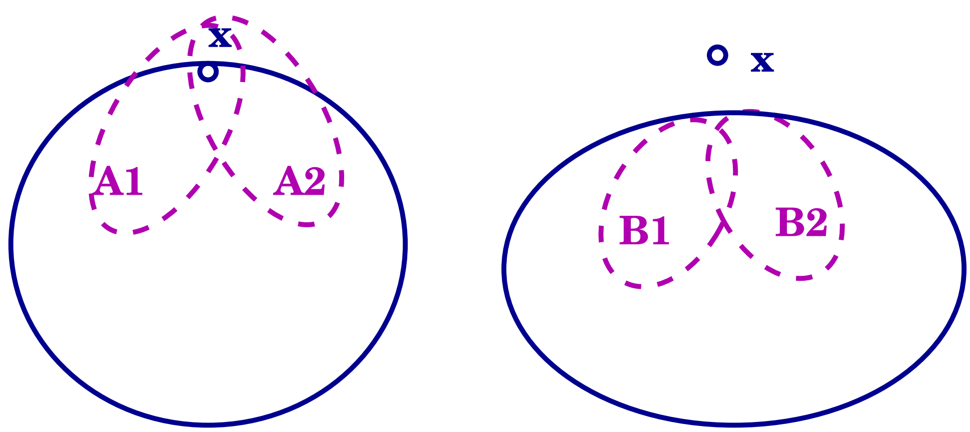
- **We will not prove the ε-net theorem, but look at some applications, and prove a related result, bounding the size of the set system.**
- **Suppose the set system $(\mathcal{U}, \mathcal{R})$, where $|\mathcal{U}| = n$, has VC dimension d. How many sets can be in the family $\mathcal{R}$?**
- **Naively, the best one can say is that $|\mathcal{R}| \leq 2^n$.**
- **We will show that**

$$|\mathcal{R}| \leq \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{d} \leq n^d$$

- **This is the best bound one can prove in general, but it's not necessarily the best for individual set systems.**
- **E.g., for points and half-spaces in the plane, this theorem gives n^3, while we can see that the real bound is n^2.**

Proof

- **Define** $g(d,n) = \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{d}$**.**
- **Proof by induction. Base case trivial:** $n = d = 0$ **and** $\mathcal{U} = \mathcal{R} = \emptyset$**.**
- **Choose an arbitrary point** $x \in \mathcal{U}$**, and consider** $\mathcal{U}' = \mathcal{U} - \{x\}$**.**
- **Let** $\mathcal{R}'$ **be the projection of** $\mathcal{R}$ **onto** $\mathcal{U}'$**. That is.** $\mathcal{R}' = \{A \cap \mathcal{U}' | A \in \mathcal{R}\}$**.**
- **VC-dim of** $(\mathcal{U}', \mathcal{R}')$ **is at most** d**—if** $\mathcal{R}'$ **shatters a** $(d+1)$**-size set, so does** $\mathcal{R}$**.**
- **By induction,** $|\mathcal{R}'| \leq g(d, n-1)$**.**



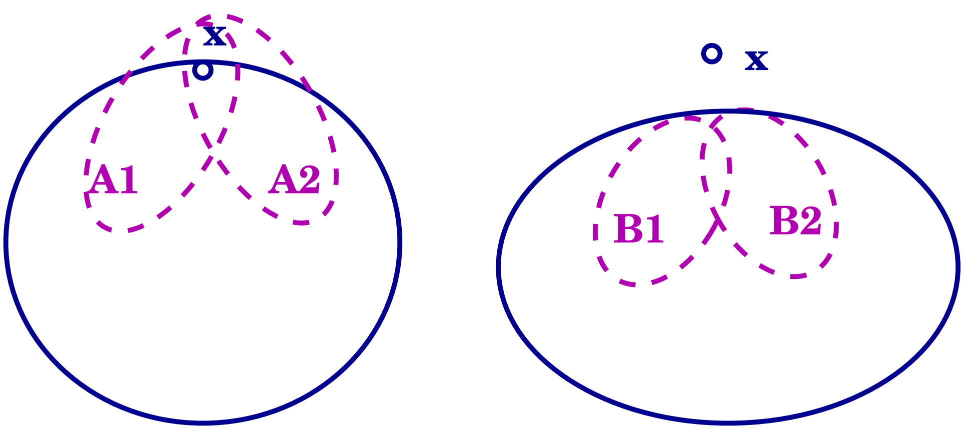
System (U, R) **System (U', R')**

Proof

- **What's the difference between $\mathcal{R}$ and $\mathcal{R}'$?**
- **Two sets $A, A' \in \mathcal{R}$ map to same set in $\mathcal{R}'$ only if $A = A' \cup \{x\}$ and $x \notin A'$.**
- **Define a new set system $(\mathcal{U}, \mathcal{R}'')$ where**

$$\mathcal{R}'' = \{A' | A' \in \mathcal{R}, \quad x \notin A', \quad A' \cup \{x\} \in \mathcal{R}\}$$

- **$|\mathcal{R}| = |\mathcal{R}'| + |\mathcal{R}''|$—sets in $\mathcal{R}''$ are exactly those that are counted only once in $\mathcal{R}'$.**
- **Claim: VC-dim of $\mathcal{R}''$ is $\leq \; d - 1$.**
- **We show that whenever $\mathcal{R}''$ shatters Y, $\mathcal{R}$ shatters $Y \cup \{x\}$.**



System (U, R) **System (U', R')**

Proof

- **Two cases: Consider $A \subseteq Y \cup \{x\}$.**

  1. **If $A \subseteq Y$, then since Y is shattered, $\exists$ $S \in \mathcal{R}''$ so that $S \cap Y = A$.**
  2. **Since $x \notin S$, but $S \in \mathcal{R}$, it follows that $S \cap (Y \cup \{x\}) = A$.**
  3. **If $x \in A$, then $\exists$ $S \in \mathcal{R}''$ so that $S \cap Y = A - \{x\}$.**
  4. **By definition of $\mathcal{R}''$, $S \cup \{x\} \in \mathcal{R}$, and so $(S \cup \{x\}) \cap (Y \cup \{x\}) = A \cup \{x\} = A$.**

- **Thus, $Y \cup \{s\}$ is shattered.**

- **Thus, VC-dim of $\mathcal{R}''$ is at most $d - 1$, and by induction, $|\mathcal{R}''| \leq g(d-1,\ n-1)$.**

Proof

- **Since** $|\mathcal{R}| \; = \; |\mathcal{R}'| + |\mathcal{R}''|$**, we have**

$$
\begin{aligned}
|\mathcal{R}| \;\; &\leq \;\; g(d,\; n-1) \; + \; g(d-1,\; n-1) \\
&= \;\; \sum_{i=0}^{d} \binom{n-1}{i} \; + \; \sum_{i=0}^{d-1} \binom{n-1}{i} \\
&= \;\; \binom{n-1}{0} + \sum_{i=1}^{d} \left(\binom{n-1}{i} \; + \; \binom{n-1}{i-1} \right) \\
&= \;\; \binom{n}{0} + \sum_{i=1}^{d} \binom{n}{i} \\
&= \;\; g(d,\; n)
\end{aligned}
$$

ε-Approximation

- **Suppose $(\mathcal{U}, \mathcal{R})$ is a set system of VC dimension d, and let ε, δ be real numbers, where $\varepsilon \in [0, 1]$ and $\delta > 0$.**

- **A set $N \subseteq \mathcal{U}$ is called an ε-approximation for $(\mathcal{U}, \mathcal{R})$ if for any $A \in \mathcal{R}$,**
$$\left| \frac{|N \cap A|}{|N|} - \frac{|A|}{|\mathcal{U}|} \right| \leq \varepsilon$$

- **If we draw**
$$O\left(\frac{d}{\varepsilon^2} \log \frac{d}{\varepsilon} + \frac{1}{\varepsilon^2} \log \frac{1}{\delta} \right)$$
points at random from $\mathcal{U}$, then the resulting set N is an ε-approximation with probability $\geq \delta$.

- **An ε-approximation is also an ε-net, but not vice versa.**