In the next few lectures we will be looking at techniques of approximation, namely, random sampling, \( \epsilon \)-approximations, \( \epsilon \)-nets, and discrepancy. Random sampling methods are useful in obtaining fast algorithms which approximate the desired solution rather than exactly compute it. The notions of \( \epsilon \)-approximations and \( \epsilon \)-nets, and discrepancy are used to bound the error in approximation.

## 9.1 Random Sampling

The need for sampling arises in many scenarios such as performing computations on large data sets, where a fast approximate solution is preferred over a slow exact one. For example, suppose we want to find the median of a set \( S = \{s_1, \ldots, s_n\} \) of \( n \) numbers. Algorithm 1 finds an element whose rank is close to \( \lfloor n/2 \rfloor \).

**Input:** \( S = \{s_1, \ldots, s_n\} \)

**Output:** \( s_i \in S \) such that \( \text{rank}(s_i) \approx \text{rank} \lfloor n/2 \rfloor \), where \( \text{rank}(\alpha) \) denotes the rank of \( \alpha \) as an element of \( S \)

Choose random subset \( A \subseteq S \) of size \( O \left( \frac{1}{\epsilon} \log n \right) \);

**return** the median of \( A \)

**Algorithm 1:** Finding median of \( n \) numbers using random sampling

In general, one chooses a random subset of suitable size (which depends upon the desired accuracy) and uses computations on the subset as an approximation to the solution to the original problem. Here is another example - computation of a centerpoint of a point set, a generalization of median to higher dimensions. Let \( P \subseteq \mathbb{R}^d \) be a set of \( n \) points and let \( \Gamma \) be the set of halfspaces in \( \mathbb{R}^d \). The depth of a point \( x \in \mathbb{R}^d \) is defined as:

\[
\text{Depth}(x, P) = \min_{\gamma \in \Gamma} |x \cap P|.
\]

The depth of the point set \( P \) is:

\[
\text{Depth}(P) = \max_{x \in \mathbb{R}^d} \text{Depth}(x, P).
\]

A centerpoint of \( P \) is any point \( x \in \mathbb{R}^d \) such that, \( \text{Depth}(x, p) \geq \left\lfloor \frac{n}{2 + \epsilon} \right\rfloor \). It is known that a center point always exists; see e.g [2].

We can adapt the paradigm of random sampling to compute the centerpoint of a point set in \( \mathbb{R}^d \) as follows:

The general result on random sampling discussed in the next section will imply that the above algorithm returns a center point with probability at least \( 1/2 \).

**Claim 1** \( \text{Depth}(X, S) \geq (1 - \epsilon) \frac{n}{2 + \epsilon} \) with probability at least \( 1/2 \).
1: Choose a random subset \( A \subseteq P \), with \( |A| = O(\frac{1}{\epsilon} \log \frac{1}{\epsilon}) \)
2: Compute the centerpoint \( X \) of \( A \)
3: Return \( X \)

**Algorithm 2:** Return a center point of \( A \)

An application of the notion of centerpoint is in the proof of the Lipton-Tarjan separator theorem. Let \( G = (V, E) \) be a planar graph with \( n \) vertices. \( V \) can be partitioned into \( A, B, C \) such that \( |A|, |B| \leq \frac{3}{4}n \) and \( |C| < 2\sqrt{n} \) and there is no edge between a vertex of \( A \) and a vertex of \( B \). This result has been the basis of many divide-and-conquer algorithms for planar graphs. Miller and Thurston [3] gave proof of the above theorem using the notions of centerpoint and stereographic projection combined with the following theorem by Andreev (1936).

**Theorem 1** For a planar graph \( G \), there exist a set \( D = D_1, \ldots, D_n \) of \( n \)-disks with pairwise disjoint interiors in \( \mathbb{R}^2 \) such that \( D_i \) and \( D_j \) touch each other iff the corresponding vertices \( V_i, V_j \) have an edge between them.

We present a very high level view of this proof below. A planar graph \( G \), corresponds to a system of disks \( D \) in \( \mathbb{R}^2 \), as implied by the theorem above. Let \( P = \{P_1, \ldots, P_n\} \) be the centers of the disks as points of \( \mathbb{R}^3 \). The plane in which the disks lie is chosen as the \( xy \)-plane. A stereographic projection is a function that maps a point \( q \) on the \( xy \)-plane to a point \( q' \) on the unit sphere \( S \) centered at \((0, 0, 1)\) such that \( q' \) is the point of intersection with of the segment joining \( q \) and the north pole with the sphere \( S \). Let \( \sigma \) be the center point of \( P \). Miller and Thurston used a transformation which was the composition of an inverse stereographic projection, a dilation(scaling) of the \( xy \)-plane and a stereographic projection and such that the center-point of \( P \) becomes the center of sphere \( S \). Under these conditions they proved that any random plane separates the system of disks. Using this key fact, they then proved the Lipton-Tarjan theorem. A more detailed proof can be found in [4].

### 9.2 Approximations

In the computation of centerpoint, the set of halfspaces \( \Gamma \) was used to “generate” subsets of \( P \). Such a set is usually referred to as “ranges”. In this section, we formalize these notions and study range spaces, \( \epsilon \)-approximations, and \( \epsilon \)-nets.

#### 9.2.1 Range spaces and VC-dimension

Let \( X \) be a finite set of objects and \( R \subseteq 2^X \) be a subset of the power set of \( X \) referred to as the family of ranges. The pair \( \Sigma = (X, R) \) is referred to as a range space. For a subset \( A \subseteq X \), the subspace of \( \Sigma \) induced by \( A \) is \( \Sigma_A = (A, R_A) \), where \( R_A = \{r \cap A : r \in R\} \). Furthermore, we say \( A \) is shattered by \( \Sigma \) if \( R_A = 2^A \).

The VC-dimension of \( \Sigma \), denoted by \( \text{VC-dim}(\Sigma) \), is an integer \( d \) such that, there is a shattered set of size \( d \) but not of size \( d + 1 \). We illustrate these notions with a few examples below. In the examples, we assume \( S \) to be a finite set of points (in general they can be objects) in \( \mathbb{R}^d \), we denote by \( R_i \) the family of ranges and by \( \Sigma_i \) the corresponding range space.
Example: \( R_1 = \{ S \cap \gamma : \gamma \text{ is a half space} \} \). In this example \( |R_1| = O(n^d) \), and the range space is \( \Sigma_1 = (S, \mathbb{R}^d) \). The VC-dim(\( \Sigma_1 \)) = 3, as illustrated in Figure 9.1.

![Figure 9.1: The points with filled interiors cannot be separated from those with hollow interiors by halfspaces.](image1)

Example: \( R_2 = \{ S \cap B : B \text{ is a ball} \} \) and \( \Sigma_2 = (S, R_2) \) and \( |R_2| = O(n^{d+1}) \).

Example: \( S' = \{ \text{Set of } n \text{ lines in } \mathbb{R}^2 \} \) and \( R_3 = \{ \{ l \in S' : l \cap r \neq \emptyset \} : r \text{ is a segment} \} \), \( |R_3| = O(n^4) \), \( \Sigma_3 = (S', R_3) \). This is shown in Figure 9.2.

![Figure 9.2: The thick segment indicates a range, while the dashed lines are the input objects.](image2)

All the range spaces presented above, have finite VC-dimension. The following example shows a geometric range space whose VC-dimension is infinite.

Example: \( R_4 = \{ S \cap \rho : \rho \text{ is a convex polygon} \} \), \( |R_4| \leq 2^n \), \( \Sigma = (R_4, \rho) \). The VC-dim(\( \Sigma \)) = \( \infty \) as shown in Figure 9.3.

Radon’s theorem, stated below, gives an upper bound on the size of the family of ranges.

**Theorem 2** Let \( \Sigma = (X, R) \) be a finite range space such that VC-dim(\( \Sigma \)) = \( d \) then \( |R| = O(|X|^d) \)

We now introduce the notion of \( \epsilon \)-approximations and \( \epsilon \)-nets.
Figure 9.3: A family of convex polygons as ranges has infinite VC-dimension as it can separate any subset of points of arbitrary size lying on a circle.

### 9.2.2 \( \epsilon \)-nets and \( \epsilon \)-approximations

**Definition** A set \( N \subseteq X \) is an \( \epsilon \)-approximation of \( \Sigma \) if for each \( r \in R \),

\[
\left| \frac{|r|}{|X|} - \frac{|r \cap N|}{|N|} \right| \leq \epsilon
\]

**Definition** A set \( N \subseteq X \) is an \( \epsilon \)-net of \( \Sigma \) if for each \( r \in R \), with \(|r| \geq \epsilon |X|\) the intersection \( r \cap N \) is non-empty.

Using Chernoff’s bound one can prove that the size of an \( \epsilon \)-approximation is bounded by \( O\left(\frac{1}{\epsilon^2} \log |R|\right) \), and the size of an \( \epsilon \)-net has an upper bound \( O\left(\frac{d}{\epsilon} \log |R|\right) \). For range spaces of bounded VC-dimension, however the bounds can be improved:

**Theorem 3** Let \( \Sigma = (X, R) \) be a range space, let \( VC\text{-}dim(\Sigma) = d \), and let \( 0 < \epsilon, \delta < 1 \) be parameters. A random subset of size \( \frac{8d}{\epsilon^2} \left( \ln\left(\frac{1}{\epsilon \delta}\right) \right) \) is an \( \epsilon \)-approximation of \( \Sigma \) with probability \( \geq 1 - \delta \).

**Theorem 4** For a range space \( \Sigma = (X, R) \) of bounded VC-dimension \( d \), a random subset of size \( \frac{2d}{\epsilon^2} \left( \ln\left(\frac{1}{\epsilon \delta}\right) \right) \) is an \( \epsilon \)-net of \( \Sigma \) with probability \( 1 - \delta \), where \( \delta, \epsilon \in (0, 1) \).

The upper bound on the size of \( \epsilon \)-approximations is not tight, the tight bound being \( O\left(\frac{2^d}{\epsilon^3} \log \frac{1}{\epsilon \delta}\right) \), whereas the upper bound on the size of \( \epsilon \)-nets is tight. The following theorems give bounds on the size of a random subset if it to be an \( \epsilon \)-approximation or an \( \epsilon \)-net.

The notion of \( \epsilon \)-approximations and \( \epsilon \)-nets have a wide range of applications such as machine learning, computation of center point etc.
9.2.3 Discrepancy

In this section we study discrepancy, a tool which is used to obtain good $\epsilon$-approximations to a range space. Let $\Sigma = (X, R)$ be a range space. A coloring $\chi$ of the set $X$ is a map $\chi : X \to \{-1, +1\}$. The discrepancy (denoted $\text{disc}$) of $\Sigma$ is defined below:

$$\text{disc}(r, \chi) = \sum_{a \in r} \chi(a)$$

$$\text{disc}(\Sigma, \chi) = \max_{r \in R} \text{disc}(r, \chi)$$

$$\text{disc}(\Sigma) = \min_{\chi} \text{disc}(\Sigma, \chi)$$

Let $X$ be the set of $n$ points in $\mathbb{R}^2$ and $R = \{ x \cap r : r \text{ is an orthogonal rectangle } \}$, then $\text{disc}((X, R)) = O(\log n)$. For the range space $\Sigma^2_1$ described above, $\text{disc}(\Sigma^2_1) = \Theta(n^{3/4})$.

Intuitively, discrepancy measures imbalance in a set with respect to a set of ranges. A balanced set will have low discrepancy. Range spaces with low discrepancy have good $\epsilon$-approximations. Relationship between $\epsilon$-approximations and discrepancy can be found in [1].

References


