CHAPTER 4

Clustering – Definitions and Basic Algorithms

In this chapter, we will initiate our discussion of ***clustering***. Clustering is one of the most fundamental computational tasks but, frustratingly, one of the fuzziest. It can be stated informally as: "Given data, find an interesting structure in the data. Go!"

The fuzziness arises naturally from the requirement that the clustering should be "interesting", which is not a well-defined concept and depends on human perception and hence is impossible to quantify clearly. Similarly, the meaning of "structure" is also open to debate. Nevertheless, clustering is inherent to many computational tasks like learning, searching, and data-mining.

Empirical study of clustering concentrates on trying various measures for the clustering and trying out various algorithms and heuristics to compute these clusterings. See the bibliographical notes in this chapter for some relevant references.

Here we will concentrate on some well-defined clustering tasks, including k-center clustering, k-median clustering, and k-means clustering, and some basic algorithms for these problems.

4.1. Preliminaries

A clustering problem is usually defined by a set of items, and a distance function between the items in this set. While these items might be points in $\mathbb{R}^{\mathsf{d}}$ and the distance function just the regular Euclidean distance, it is sometime beneficial to consider the more abstract setting of a general metric space.

4.1.1. Metric spaces.

Definition 4.1. A ***metric space*** is a pair $(\mathcal{X}, \mathbf{d})$ where $\mathcal{X}$ is a set and $\mathbf{d} : \mathcal{X} \times \mathcal{X} \to [0, \infty)$ is a ***metric*** satisfying the following axioms: (i) $\mathbf{d}_{\mathcal{M}}(x, y) = 0$ if and only if $x = y$, (ii) $\mathbf{d}_{\mathcal{M}}(x, y) = \mathbf{d}_{\mathcal{M}}(y, x)$, and (iii) $\mathbf{d}_{\mathcal{M}}(x, y) + \mathbf{d}_{\mathcal{M}}(y, z) \geq \mathbf{d}_{\mathcal{M}}(x, z)$ (triangle inequality).

For example, $\mathbb{R}^2$ with the regular Euclidean distance is a metric space. In the following, we assume that we are given ***black-box access*** to $\mathbf{d}_{\mathcal{M}}$. Namely, given two points $\mathsf{p}, \mathsf{q} \in \mathcal{X}$, we assume that $\mathbf{d}_{\mathcal{M}}(\mathsf{p}, \mathsf{q})$ can be computed in constant time.

Another standard example for a finite metric space is a graph G with non-negative weights $\omega(\cdot)$ defined on its edges. Let $\mathbf{d}_{\mathsf{G}}(x, y)$ denote the shortest path (under the given weights) between any $x, y \in V(\mathsf{G})$. It is easy to verify that $\mathbf{d}_{\mathsf{G}}(\cdot, \cdot)$ is a metric. In fact, any ***finite metric*** (i.e., a metric defined over a finite set) can be represented by such a weighted graph.

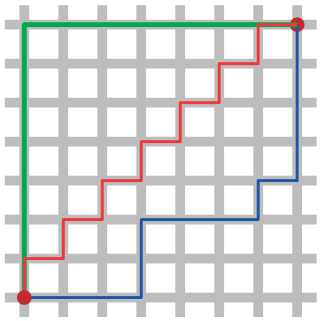
The L_p***-norm*** defines the distance between two points $\mathsf{p}, \mathsf{q} \in \mathbb{R}^{\mathsf{d}}$ as

$$\left\| \mathsf{p} - \mathsf{q} \right\|_p = \left(\sum_{i=1}^{\mathsf{d}} |\mathsf{p}_i - \mathsf{q}_i|^p \right)^{1/p},$$

for $p \geq 1$. The L_2***-norm*** is the regular Euclidean distance.

The L_1-*norm*, also known as the ***Manhattan distance*** or ***taxicab distance***, is

$$\left\|\mathsf{p}-\mathsf{q}\right\|_1 = \sum_{i=1}^{\mathsf{d}} |\mathsf{p}_i - \mathsf{q}_i| .$$

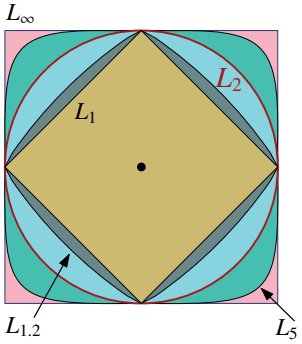


The L_1-norm distance between two points is the minimum path length that is axis parallel and connects the two points. For a uniform grid, it is the minimum number of grid edges (i.e., blocks in Manhattan) one has to travel between two grid points. In particular, the shortest path between two points is no longer unique; see the picture on the right. Of course, in the L_2-norm the shortest path between two points is the segment connecting the two points.

The L_∞-*norm* is

$$\left\|\mathsf{p}-\mathsf{q}\right\|_\infty = \lim_{p\to\infty}\left\|\mathsf{p}-\mathsf{q}\right\|_p = \max_{i=1}^{\mathsf{d}} |\mathsf{p}_i - \mathsf{q}_i| .$$

The triangle inequality holds for the L_p-norm, for any $p \geq 1$ (it is called the ***Minkowski inequality*** in this case). In particular, L_p is a metric for $p \geq 1$. Specifically, $\mathbb{R}^{\mathsf{d}}$ with any L_p-norm (i.e., $p \geq 1$) is another example of a metric space.



It is useful to consider the different unit balls of L_p for different value of p; see the figure on the right. The figure implies (and one can prove it formally) that for any point $\mathsf{p} \in \mathbb{R}^{\mathsf{d}}$, we have that $\|\mathsf{p}\|_p \leq \|\mathsf{p}\|_q$ if $p > q$.

Lemma 4.2. *For any* $\mathsf{p} \in \mathbb{R}^{\mathsf{d}}$, *we have that* $\|\mathsf{p}\|_1 / \sqrt{\mathsf{d}} \leq \|\mathsf{p}\|_2 \leq \|\mathsf{p}\|_1$.

Proof. Indeed, let $\mathsf{p} = (\mathsf{p}_1, \ldots, \mathsf{p}_{\mathsf{d}})$, and assume that $\mathsf{p}_i \geq 0$, for all i. It is easy to verify that for a constant α, the function $f(x) = x^2 + (\alpha - x)^2$ is minimized when $x = \alpha/2$. As such, setting $\alpha = \|\mathsf{p}\|_1 = \sum_{i=1}^{\mathsf{d}} |\mathsf{p}_i|$, we have, by symmetry and by the above observation on $f(x)$, that $\sum_{i=1}^{\mathsf{d}} \mathsf{p}_i^2$ is minimized under the condition $\|\mathsf{p}\|_1 = \alpha$, when all the coordinates of p are equal. As such, we have that $\|\mathsf{p}\|_2 \geq \sqrt{\mathsf{d}(\alpha/\mathsf{d})^2} = \|\mathsf{p}\|_1 / \sqrt{\mathsf{d}}$, implying the claim. ■

4.1.2. The clustering problem. There is a metric space $(\mathcal{X}, \mathbf{d})$ and the input is a set of n points $\mathsf{P} \subseteq \mathcal{X}$. Given a set of centers $\mathbf{C}$, every point of P is assigned to its nearest neighbor in $\mathbf{C}$. All the points of P that are assigned to a center $\overline{\mathsf{c}}$ form the ***cluster*** of $\overline{\mathsf{c}}$, denoted by

$$\text{cluster}(\mathbf{C}, \overline{\mathsf{c}}) = \left\{ \mathsf{p} \in \mathsf{P} \;\middle|\; \mathbf{d}_{\mathcal{M}}(\mathsf{p}, \overline{\mathsf{c}}) = \mathbf{d}(\mathsf{p}, \mathbf{C}) \right\}, \tag{4.1}$$

where

$$\mathbf{d}(\mathsf{p}, \mathbf{C}) = \min_{\overline{\mathsf{c}} \in \mathbf{C}} \mathbf{d}_{\mathcal{M}}(\mathsf{p}, \overline{\mathsf{c}})$$

denotes the ***distance*** of p to the set $\mathbf{C}$. Namely, the center set $\mathbf{C}$ partition P into clusters. This specific scheme of partitioning points by assigning them to their closest center (in a given set of centers) is known as a ***Voronoi partition***.

In particular, let $\mathsf{P} = \{\mathsf{p}_1, \ldots, \mathsf{p}_n\}$, and consider the n-dimensional point

$$\mathsf{P}_{\mathbf{C}} = \Big(\mathbf{d}(\mathsf{p}_1, \mathbf{C}), \mathbf{d}(\mathsf{p}_2, \mathbf{C}), \ldots, \mathbf{d}(\mathsf{p}_n, \mathbf{C}) \Big).$$

The ith coordinate of the point $\mathsf{P}_{\mathbf{C}}$ is the distance (i.e., cost of assigning) of p_i to its closest center in $\mathbf{C}$.

4.2. On k-center clustering

In the ***k-center clustering problem***, a set $\mathsf{P} \subseteq \mathcal{X}$ of n points is provided together with a parameter k. We would like to find a set of k points, $\mathbf{C} \subseteq \mathsf{P}$, such that the maximum distance of a point in P to its closest point in $\mathbf{C}$ is minimized.

As a concrete example, consider the set of points to be a set of cities. Distances between points represent the time it takes to travel between the corresponding cities. We would like to build k hospitals and minimize the maximum time it takes a patient to arrive at her closest hospital. Naturally, we want to build the hospitals in the cities and not in the middle of nowhere.[①]

Formally, given a set of centers $\mathbf{C}$, the k-center clustering ***price*** of P by $\mathbf{C}$ is denoted by

$$\|\mathsf{P}_{\mathbf{C}}\|_\infty = \max_{\mathsf{p} \in \mathsf{P}} \mathbf{d}(\mathsf{p}, \mathbf{C}).$$

Note that every point in a cluster is within a distance at most $\|\mathsf{P}_{\mathbf{C}}\|_\infty$ from its respective center.

Formally, the ***k-center problem*** is to find a set $\mathbf{C}$ of k points, such that $\|\mathsf{P}_{\mathbf{C}}\|_\infty$ is minimized; namely,

$$\mathrm{opt}_\infty(\mathsf{P}, k) = \min_{\mathbf{C} \subseteq \mathsf{P}, |\mathbf{C}| = k} \|\mathsf{P}_{\mathbf{C}}\|_\infty .$$

We will denote the set of centers realizing the optimal clustering by C_{opt}. A more explicit definition (and somewhat more confusing) of the k-center clustering is to compute the set $\mathbf{C}$ of size k realizing $\min_{\mathbf{C} \subseteq \mathsf{P}} \max_{\mathsf{p} \in \mathsf{P}} \min_{\overline{\mathsf{c}} \in \mathbf{C}} \mathbf{d}_{\mathcal{M}}(\mathsf{p}, \overline{\mathsf{c}})$.

It is known that k-center clustering is NP-Hard, and it is in fact hard to approximate within a factor of 1.86, even for a point set in the plane (under the Euclidean distance). Surprisingly, there is a simple and elegant algorithm that achieves a 2-approximation.

Discrete vs. continuous clustering. If the input is a point set in $\mathbb{R}^{\mathsf{d}}$, the centers of the clustering are not necessarily restricted to be a subset of the input point, as they might be placed anywhere in $\mathbb{R}^{\mathsf{d}}$. Allowing this flexibility might further reduce the price of the clustering (by a constant factor). The variant where one is restricted to use the input points as centers is the ***discrete clustering*** problem. The version where centers might be placed anywhere in the given metric space is the ***continuous clustering*** version.

4.2.1. The greedy clustering algorithm. The algorithm **GreedyKCenter** starts by picking an arbitrary point, $\overline{\mathsf{c}}_1$, and setting $\mathbf{C}_1 = \{\overline{\mathsf{c}}_1\}$. Next, we compute for every point $\mathsf{p} \in \mathsf{P}$ its distance $d_1[\mathsf{p}]$ from $\overline{\mathsf{c}}_1$. Now, consider the point worst served by $\mathbf{C}_1$; this is the point realizing $r_1 = \max_{\mathsf{p} \in \mathsf{P}} d_1[\mathsf{p}]$. Let $\overline{\mathsf{c}}_2$ denote this point, and add it to the set $\mathbf{C}_1$, resulting in the set $\mathbf{C}_2$.

Specifically, in the ith iteration, we compute for each point $\mathsf{p} \in \mathsf{P}$ the quantity $d_{i-1}[\mathsf{p}] = \min_{\overline{\mathsf{c}} \in \mathbf{C}_{i-1}} \mathbf{d}_{\mathcal{M}}(\mathsf{p}, \overline{\mathsf{c}})$. We also compute the radius of the clustering

$$r_{i-1} = \left\| \mathsf{P}_{\mathbf{C}_{i-1}} \right\|_\infty = \max_{\mathsf{p} \in \mathsf{P}} d_{i-1}[\mathsf{p}] = \max_{\mathsf{p} \in \mathsf{P}} \mathbf{d}(\mathsf{p}, \mathbf{C}_{i-1}) \tag{4.2}$$

[①] Although, there are recorded cases in history of building universities in the middle of nowhere.

and the bottleneck point $\overline{c}_i$ that realizes it. Next, we add $\overline{c}_i$ to $\mathbf{C}_{i-1}$ to form the new set $\mathbf{C}_i$. We repeat this process k times.

Namely, the algorithm repeatedly picks the point furthest away from the current set of centers and adds it to this set.

To make this algorithm slightly faster, observe that

$$d_i[\mathsf{p}] = \mathbf{d}(\mathsf{p}, \mathbf{C}_i) = \min(\mathbf{d}(\mathsf{p}, \mathbf{C}_{i-1}), \mathbf{d}_{\mathcal{M}}(\mathsf{p}, \overline{c}_i)) = \min(d_{i-1}[\mathsf{p}], \mathbf{d}_{\mathcal{M}}(\mathsf{p}, \overline{c}_i)).$$

In particular, if we maintain for each point $\mathsf{p} \in \mathsf{P}$ a single variable $d[\mathsf{p}]$ with its current distance to its closest center in the current center set, then the above formula boils down to

$$d[\mathsf{p}] \leftarrow \min(d[\mathsf{p}], \mathbf{d}_{\mathcal{M}}(\mathsf{p}, \overline{c}_i)).$$

Namely, the above algorithm can be implemented using $O(n)$ space, where $n = |\mathsf{P}|$. The ith iteration of choosing the ith center takes $O(n)$ time. Thus, overall, this approximation algorithm takes $O(nk)$ time.

A ***ball*** of radius r around a point $\mathsf{p} \in \mathsf{P}$ is the set of points in P with distance at most r from p; namely, $\mathbf{b}(\mathsf{p}, r) = \left\{ \mathsf{q} \in \mathsf{P} \;\middle|\; \mathbf{d}_{\mathcal{M}}(\mathsf{p}, \mathsf{q}) \le r \right\}$. Thus, the k-center problem can be interpreted as the problem of covering the points of P using k balls of minimum (maximum) radius.

Theorem 4.3. *Given a set of n points P in a metric space $(\mathcal{X}, \mathbf{d})$, the algorithm* **GreedyK-Center** *computes a set $\mathbf{K}$ of k centers, such that $\mathbf{K}$ is a 2-approximation to the optimal k-center clustering of P; namely, $\|\mathsf{P}_{\mathbf{K}}\|_\infty \le 2\mathrm{opt}_\infty$, where $\mathrm{opt}_\infty = \mathrm{opt}_\infty(\mathsf{P}, k)$ is the price of the optimal clustering. The algorithm takes $O(nk)$ time.*

Proof. The running time follows by the above description, so we concern ourselves only with the approximation quality.

By definition, we have $r_k = \|\mathsf{P}_{\mathbf{K}}\|_\infty$, and let $\overline{c}_{k+1}$ be the point in P realizing $r_k = \max_{\mathsf{p} \in \mathsf{P}} \mathbf{d}(\mathsf{p}, \mathbf{K})$. Let $\mathbf{C} = \mathbf{K} \cup \{\overline{c}_{k+1}\}$. Observe that by the definition of r_i (see (4.2)), we have that $r_1 \ge r_2 \ge \ldots \ge r_k$. Furthermore, for $i < j \le k+1$ we have that

$$\mathbf{d}_{\mathcal{M}}\left(\overline{c}_i, \overline{c}_j\right) \ge \mathbf{d}_{\mathcal{M}}\left(\overline{c}_j, \mathbf{C}_{j-1}\right) = r_{j-1} \ge r_k.$$

Namely, the distance between any pair of points in $\mathbf{C}$ is at least r_k. Now, assume for the sake of contradiction that $r_k > 2\mathrm{opt}_\infty(\mathsf{P}, k)$. Consider the optimal solution that covers P with k balls of radius opt_∞. By the triangle inequality, any two points inside such a ball are within a distance at most $2\mathrm{opt}_\infty$ from each other. Thus, none of these balls can cover two points of $\mathbf{C} \subseteq \mathsf{P}$, since the minimum distance between members of $\mathbf{C}$ is $> 2\mathrm{opt}_\infty$. As such, the optimal cover by k balls of radius opt_∞ cannot cover $\mathbf{C}$ (and thus P), as $|\mathbf{C}| = k+1$, a contradiction. ■

In the spirit of never trusting a claim that has only a single proof, we provide an alternative proof.[②]

Alternative proof. If every cluster of C_{opt} contains exactly one point of $\mathbf{K}$, then the claim follows. Indeed, consider any point $\mathsf{p} \in \mathsf{P}$, and let $\overline{c}$ be the center it belongs to in C_{opt}. Also, let $\overline{g}$ be the center of $\mathbf{K}$ that is in cluster$\left(C_{\mathrm{opt}}, \overline{c}\right)$. We have that $\mathbf{d}_{\mathcal{M}}(\mathsf{p}, \overline{c}) = \mathbf{d}\left(\mathsf{p}, C_{\mathrm{opt}}\right) \le \mathrm{opt}_\infty = \mathrm{opt}_\infty(\mathsf{P}, k)$. Similarly, observe that $\mathbf{d}_{\mathcal{M}}(\overline{g}, \overline{c}) = \mathbf{d}\left(\overline{g}, C_{\mathrm{opt}}\right) \le \mathrm{opt}_\infty$. As such, by the triangle inequality, we have that $\mathbf{d}_{\mathcal{M}}(\mathsf{p}, \overline{g}) \le \mathbf{d}_{\mathcal{M}}(\mathsf{p}, \overline{c}) + \mathbf{d}_{\mathcal{M}}(\overline{c}, \overline{g}) \le 2\mathrm{opt}_\infty$.

[②] Mark Twain is credited with saying that "I don't give a damn for a man that can only spell a word one way." However, there seems to be some doubt if he really said that, which brings us to the conclusion of never trusting a quote if it is credited only to a single person.

By the pigeon hole principle, the only other possibility is that there are at least two centers $\overline{\mathsf{g}}$ and $\overline{\mathsf{h}}$ of $\mathbf{K}$ that are both in cluster$\left(C_{\mathrm{opt}}, \overline{\mathsf{c}}\right)$, for some $\overline{\mathsf{c}} \in C_{\mathrm{opt}}$. Assume, without loss of generality, that $\overline{\mathsf{h}}$ was added later than $\overline{\mathsf{g}}$ to the center set $\mathbf{K}$ by the algorithm **GreedyKCenter**, say in the ith iteration. But then, since **GreedyKCenter** always chooses the point furthest away from the current set of centers, we have that

$$\|\mathsf{P}_{\mathbf{K}}\|_\infty \leq \left\|\mathsf{P}_{\mathbf{C}_{i-1}}\right\|_\infty = \mathbf{d}\left(\overline{\mathsf{h}}, \mathbf{C}_{i-1}\right) \leq \mathbf{d}_{\mathcal{M}}\left(\overline{\mathsf{h}}, \overline{\mathsf{g}}\right) \leq \mathbf{d}_{\mathcal{M}}\left(\overline{\mathsf{h}}, \overline{\mathsf{c}}\right) + \mathbf{d}_{\mathcal{M}}(\overline{\mathsf{c}}, \overline{\mathsf{g}}) \leq 2\mathrm{opt}_\infty. \qquad \blacksquare$$

4.2.2. The greedy permutation. There is an interesting phenomena associated with **GreedyKCenter**. If we run it till it exhausts all the points of P (i.e., $k = n$), then this algorithm generates a permutation of P; that is, $\langle \mathsf{P} \rangle = \langle \overline{\mathsf{c}}_1, \overline{\mathsf{c}}_2, \ldots, \overline{\mathsf{c}}_n \rangle$. We will refer to $\langle \mathsf{P} \rangle$ as the ***greedy permutation*** of P. There is also an associated sequence of radii $\langle r_1, r_2, \ldots, r_n \rangle$, where all the points of P are within a distance at most r_i from the points of $\mathbf{C}_i = \langle \overline{\mathsf{c}}_1, \ldots, \overline{\mathsf{c}}_i \rangle$.

Definition 4.4. A set $S \subseteq \mathsf{P}$ is an ***r-packing*** for P if the following two properties hold.

(i) ***Covering property***: All the points of P are within a distance at most r from the points of S.

(ii) ***Separation property***: For any pair of points $\mathsf{p}, \mathsf{q} \in S$, we have that $\mathbf{d}_{\mathcal{M}}(\mathsf{p}, \mathsf{q}) \geq r$.

(One can relax the separation property by requiring that the points of S be at a distance $\Omega(r)$ apart.)

Intuitively, an r-packing of a point set P is a compact representation of P in the resolution r. Surprisingly, the greedy permutation of P provides us with such a representation for all resolutions.

Theorem 4.5. *Let P be a set of n points in a finite metric space, and let its greedy permutation be $\langle \overline{\mathsf{c}}_1, \overline{\mathsf{c}}_2, \ldots, \overline{\mathsf{c}}_n \rangle$ with the associated sequence of radii $\langle r_1, r_2, \ldots, r_n \rangle$. For any i, we have that $\mathbf{C}_i = \langle \overline{\mathsf{c}}_1, \ldots, \overline{\mathsf{c}}_i \rangle$ is an r_i-packing of P.*

Proof. Note that by construction $r_{k-1} = \mathbf{d}(\overline{\mathsf{c}}_k, \mathbf{C}_{k-1})$, for all $k = 2, \ldots, n$. As such, for $j < k \leq i \leq n$, we have that $\mathbf{d}_{\mathcal{M}}\left(\overline{\mathsf{c}}_j, \overline{\mathsf{c}}_k\right) \geq \mathbf{d}(\overline{\mathsf{c}}_k, \mathbf{C}_{k-1}) = r_{k-1} \geq r_i$, since $r_1, r_2, \ldots, r_n$ is a monotonically non-increasing sequence. This implies the required separation property.

The covering property follows by definition; see $(4.2)_{\mathrm{p49}}$. $\blacksquare$

4.3. On k-median clustering

In the ***k-median clustering problem***, a set $\mathsf{P} \subseteq \mathcal{X}$ is provided together with a parameter k. We would like to find a set of k points, $\mathbf{C} \subseteq \mathsf{P}$, such that the sum of the distances of points of P to their closest point in $\mathbf{C}$ is minimized.

Formally, given a set of centers $\mathbf{C}$, the k-median clustering ***price*** of clustering P by $\mathbf{C}$ is denoted by

$$\|\mathsf{P}_{\mathbf{C}}\|_1 = \sum_{\mathsf{p} \in \mathsf{P}} \mathbf{d}(\mathsf{p}, \mathbf{C}) .$$

Formally, the ***k-median problem*** is to find a set $\mathbf{C}$ of k points, such that $\|\mathsf{P}_{\mathbf{C}}\|_1$ is minimized; namely,

$$\mathrm{opt}_1(\mathsf{P}, k) = \min_{\mathbf{C} \subseteq \mathsf{P}, |\mathbf{C}| = k} \|\mathsf{P}_{\mathbf{C}}\|_1 .$$

We will denote the set of centers realizing the optimal clustering by C_{opt}.

There is a simple and elegant constant factor approximation algorithm for k-median clustering using ***local search*** (its analysis however is painful).

A note on notation. Consider the set $U = \left\{ \mathsf{P}_{\mathbf{C}} \;\middle|\; \mathbf{C} \in \mathsf{P}^k \right\}$. Clearly, we have that $\mathrm{opt}_\infty(\mathsf{P}, k) = \min_{\mathsf{q} \in U} \|\mathsf{q}\|_\infty$ and $\mathrm{opt}_1(\mathsf{P}, k) = \min_{\mathsf{q} \in U} \|\mathsf{q}\|_1$.

Namely, k-center clustering under this interpretation is just finding the point minimizing the L_∞-norm in a set U of points in n dimensions. Similarly, the k-median problem is to find the point minimizing the L_1-norm in the set U.

Claim 4.6. *For any point set* P *of* n *points and a parameter* k*, we have that* $\mathrm{opt}_\infty(\mathsf{P}, k) \le \mathrm{opt}_1(\mathsf{P}, k) \le n \,\mathrm{opt}_\infty(\mathsf{P}, k)$.

Proof. For any point $\mathsf{p} \in \mathbb{R}^n$, we have that $\|\mathsf{p}\|_\infty = \max_{i=1}^n |\mathsf{p}_i| \le \sum_{i=1}^n |\mathsf{p}_i| = \|\mathsf{p}\|_1$ and $\|\mathsf{p}\|_1 = \sum_{i=1}^n |\mathsf{p}_i| \le \sum_{i=1}^n \max_{j=1}^n |\mathsf{p}_j| \le n \|\mathsf{p}\|_\infty$.

Let $\mathbf{C}$ be the set of k points realizing $\mathrm{opt}_1(\mathsf{P}, k)$; that is, $\mathrm{opt}_1(\mathsf{P}, k) = \|\mathsf{P}_{\mathbf{C}}\|_1$. We have that $\mathrm{opt}_\infty(\mathsf{P}, k) \le \|\mathsf{P}_{\mathbf{C}}\|_\infty \le \|\mathsf{P}_{\mathbf{C}}\|_1 = \mathrm{opt}_1(\mathsf{P}, k)$. Similarly, if $\mathbf{K}$ is the set realizing $\mathrm{opt}_\infty(\mathsf{P}, k)$, then $\mathrm{opt}_1(\mathsf{P}, k) = \|\mathsf{P}_{\mathbf{C}}\|_1 \le \|\mathsf{P}_{\mathbf{K}}\|_1 \le n \|\mathsf{P}_{\mathbf{K}}\|_\infty = n \cdot \mathrm{opt}_\infty(\mathsf{P}, k)$. ■

4.3.1. Approximation algorithm – local search. We are given a set P of n points and a parameter k. In the following, let C_{opt} denote the set of centers realizing the optimal solution, and let $\mathrm{opt}_1 = \mathrm{opt}_1(\mathsf{P}, k)$.

4.3.1.1. *The algorithm.*

A $2n$-approximation. The algorithm starts by computing a set of k centers L using Theorem 4.3. Claim 4.6 implies that

$$\|\mathsf{P}_L\|_1 / 2n \le \|\mathsf{P}_L\|_\infty / 2 \le \mathrm{opt}_\infty(\mathsf{P}, k) \le \mathrm{opt}_1 \le \|\mathsf{P}_L\|_1 \implies \mathrm{opt}_1 \le \|\mathsf{P}_L\|_1 \le 2n \,\mathrm{opt}_1. \tag{4.3}$$

Namely, L is a $2n$-approximation to the optimal solution.

Improving it. Let $0 < \tau < 1$ be a parameter to be determined shortly. The local search algorithm **algLocalSearchKMed** initially sets the current set of centers L_{curr} to be L, the set of centers computed above. Next, at each iteration it checks if the current solution L_{curr} can be improved by replacing one of the centers in it by a center from the outside. We will refer to such an operation as a ***swap***. There are at most $|\mathsf{P}|\,|L_{\mathrm{curr}}| = nk$ choices to consider, as we pick a center $\bar{\mathsf{c}} \in L_{\mathrm{curr}}$ to throw away and a new center to replace it by $\bar{\mathsf{o}} \in (\mathsf{P} \setminus L_{\mathrm{curr}})$. We consider the new candidate set of centers $\mathbf{K} \leftarrow (L_{\mathrm{curr}} \setminus \{\bar{\mathsf{c}}\}) \cup \{\bar{\mathsf{o}}\}$. If $\|\mathsf{P}_{\mathbf{K}}\|_1 \le (1-\tau) \left\|\mathsf{P}_{L_{\mathrm{curr}}}\right\|_1$, then the algorithm sets $L_{\mathrm{curr}} \leftarrow \mathbf{K}$. The algorithm continues iterating in this fashion over all possible swaps.

The algorithm **algLocalSearchKMed** stops when there is no swap that would improve the current solution by a factor of (at least) $(1-\tau)$. The final content of the set L_{curr} is the required constant factor approximation.

4.3.1.2. *Running time.* An iteration requires checking $O(nk)$ swaps (i.e., $n-k$ candidates to be swapped in and k candidates to be swapped out). Computing the price of every such swap, done naively, requires computing the distance of every point to its nearest center, and that takes $O(nk)$ time per swap. As such, overall, each iteration takes $O\left((nk)^2\right)$ time.

Since $1/(1-\tau) \ge 1+\tau$, the running time of the algorithm is

$$O\left((nk)^2 \log_{1/(1-\tau)} \frac{\|\mathsf{P}_L\|_1}{\mathrm{opt}_1}\right) = O\left((nk)^2 \log_{1+\tau} 2n\right) = O\left((nk)^2 \frac{\log n}{\tau}\right),$$

by (4.3) and Lemma 28.10$_{\mathrm{p348}}$. Thus, if τ is polynomially small, then the running time would be polynomial.

4.3.2. Proof of quality of approximation. We claim that the above algorithm provides a constant factor approximation for the optimal k-median clustering.

4.3.2.1. *Definitions and intuition.* Intuitively, since the local search got stuck in a locally optimal solution, it cannot be too far from the true optimal solution.

For the sake of simplicity of exposition, let us assume (for now) that the solution returned by the algorithm cannot be improved (at all) by any swap, and let L be this set of centers. For a center $\overline{c} \in L$ and a point $\overline{o} \in \mathsf{P} \setminus L$, let $L - \overline{c} + \overline{o} = (L \setminus \{\overline{c}\}) \cup \{\overline{o}\}$ denote the set of centers resulting from applying the swap $\overline{c} \to \overline{o}$ to L. We are assuming that there is no beneficial swap; that is,

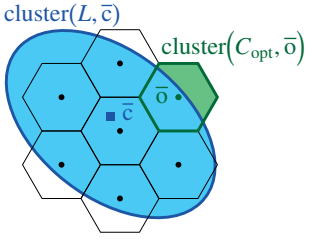
$$\forall \overline{c} \in L, \overline{o} \in \mathsf{P} \setminus L \qquad 0 \leq \Delta(\overline{c}, \overline{o}) = \nu_1(L - \overline{c} + \overline{o}) - \nu_1(L), \tag{4.4}$$

where $\nu_1(X) = \|\mathsf{P}_X\|_1$.

Equation (4.4) provides us with a large family of inequalities that all hold together. Each inequality is represented by a swap $\overline{c} \to \overline{o}$. We would like to combine these inequalities such that they will imply that $5 \left\|\mathsf{P}_{C_{\mathrm{opt}}}\right\|_1 \geq \|\mathsf{P}_L\|_1$, namely, that the local search algorithm provides a constant factor approximation to optimal clustering. This idea seems to be somewhat mysterious (or even impossible), but hopefully it will become clearer shortly.

From local clustering to local clustering complying with the optimal clustering. The first hurdle in the analysis is that a cluster of the optimal solution $\mathrm{cluster}\left(C_{\mathrm{opt}}, \overline{o}\right)$, for $\overline{o} \in C_{\mathrm{opt}}$, might intersect a large number of clusters in the local clustering (i.e., clusters of the form $\mathrm{cluster}(L, \overline{c})$ for $\overline{c} \in L$).

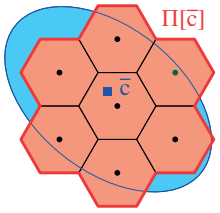
Fortunately, one can modify the assignment of points to clusters in the locally optimal clustering so that the resulting clustering of P complies with the optimal partition and the price of the clustering increases only moderately; that is, every cluster in the optimal clustering would be contained in a single cluster of the modified local solution. In particular, now an optimal cluster would intersect only a single cluster in the modified local solution.



Furthermore, this modified local solution Π is not much more expensive. Now, in this modified partition there are many beneficial swaps (by making it into the optimal clustering). But these swaps cannot be too profitable, since then they would have been profitable for the original local solution. This would imply that the local solution cannot be too expensive. The picture on the right depicts a local cluster and the optimal clusters in its vicinity such that their centers are contained inside it.

In the following, we denote by $\mathsf{nn}(\mathsf{p}, X)$ the nearest neighbor to p in the set X.

For a point $\mathsf{p} \in \mathsf{P}$, let $\overline{o}_{\mathsf{p}} = \mathsf{nn}\left(\mathsf{p}, C_{\mathrm{opt}}\right)$ be its optimal center, and let $\alpha(\mathsf{p}) = \mathsf{nn}\left(\overline{o}_{\mathsf{p}}, L\right)$ be the center p should use if p follows its optimal center's assignment. Let Π be the modified partition of P by the function $\alpha(\cdot)$.



That is, for $\overline{c} \in L$, its cluster in Π, denoted by $\Pi[\overline{c}]$, is the set of all points $\mathsf{p} \in \mathsf{P}$ such that $\alpha(\mathsf{p}) = \overline{c}$.

Now, for any center $\overline{o} \in C_{\mathrm{opt}}$, let $\mathsf{nn}(\overline{o}, L)$ be its nearest neighbor in L, and observe that $\mathrm{cluster}\left(C_{\mathrm{opt}}, \overline{o}\right) \subseteq \Pi\left[\mathsf{nn}(\overline{o}, L)\right]$ (see (4.1)$_{\mathrm{p48}}$). The picture on the right shows the resulting modified cluster for the above example.

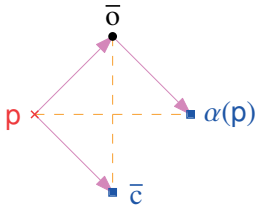
Let δ_{p} denote the price of this reassignment for the point p; that is, $\delta_{\mathsf{p}} = \mathbf{d}_{\mathcal{M}}(\mathsf{p}, \alpha(\mathsf{p})) - \mathbf{d}(\mathsf{p}, L)$. Note that if p does not get reassigned, then $\delta_{\mathsf{p}} = 0$ and otherwise $\delta_{\mathsf{p}} \geq 0$, since $\alpha(\mathsf{p}) \in L$ and $\mathbf{d}(\mathsf{p}, L) = \min_{\overline{\mathsf{c}} \in L} \mathbf{d}_{\mathcal{M}}(\mathsf{p}, \overline{\mathsf{c}})$.

Lemma 4.7. *The increase in cost from moving from the clustering induced by L to the clustering of Π is bounded by $2\left\|\mathsf{P}_{C_{\mathrm{opt}}}\right\|_1$. That is, $\sum_{\mathsf{p}\in\mathsf{P}} \delta_{\mathsf{p}} \leq 2\left\|\mathsf{P}_{C_{\mathrm{opt}}}\right\|_1$.*

Proof. For a point $\mathsf{p} \in \mathsf{P}$, let $\overline{\mathsf{c}} = \mathsf{nn}(\mathsf{p}, L)$ be its local center, let $\overline{\mathsf{o}} = \mathsf{nn}\left(\mathsf{p}, C_{\mathrm{opt}}\right)$ be its optimal center, and let $\alpha(\mathsf{p}) = \mathsf{nn}(\overline{\mathsf{o}}, L)$ be its new assigned center in Π. Observe that $\mathbf{d}_{\mathcal{M}}(\overline{\mathsf{o}}, \alpha(\mathsf{p})) = \mathbf{d}_{\mathcal{M}}(\overline{\mathsf{o}}, \mathsf{nn}(\overline{\mathsf{o}}, L)) \leq \mathbf{d}_{\mathcal{M}}(\overline{\mathsf{o}}, \overline{\mathsf{c}})$.

As such, by the triangle inequality, we have that

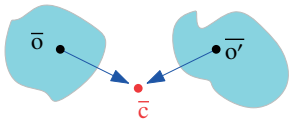
$$\begin{aligned}\mathbf{d}_{\mathcal{M}}(\mathsf{p}, \alpha(\mathsf{p})) &\leq \mathbf{d}_{\mathcal{M}}(\mathsf{p}, \overline{\mathsf{o}}) + \mathbf{d}_{\mathcal{M}}(\overline{\mathsf{o}}, \alpha(\mathsf{p})) \leq \mathbf{d}_{\mathcal{M}}(\mathsf{p}, \overline{\mathsf{o}}) + \mathbf{d}_{\mathcal{M}}(\overline{\mathsf{o}}, \overline{\mathsf{c}}) \\ &\leq \mathbf{d}_{\mathcal{M}}(\mathsf{p}, \overline{\mathsf{o}}) + (\mathbf{d}_{\mathcal{M}}(\overline{\mathsf{o}}, \mathsf{p}) + \mathbf{d}_{\mathcal{M}}(\mathsf{p}, \overline{\mathsf{c}})) \\ &= 2\mathbf{d}_{\mathcal{M}}(\mathsf{p}, \overline{\mathsf{o}}) + \mathbf{d}_{\mathcal{M}}(\mathsf{p}, \overline{\mathsf{c}}).\end{aligned}$$



Finally, $\delta_{\mathsf{p}} = \mathbf{d}_{\mathcal{M}}(\mathsf{p}, \alpha(\mathsf{p})) - \mathbf{d}(\mathsf{p}, L) \leq 2\mathbf{d}_{\mathcal{M}}(\mathsf{p}, \overline{\mathsf{o}}) + \mathbf{d}_{\mathcal{M}}(\mathsf{p}, \overline{\mathsf{c}}) - \mathbf{d}_{\mathcal{M}}(\mathsf{p}, \overline{\mathsf{c}}) = 2\mathbf{d}_{\mathcal{M}}(\mathsf{p}, \overline{\mathsf{o}}) = 2\mathbf{d}\left(\mathsf{p}, C_{\mathrm{opt}}\right)$. As such, $\sum_{\mathsf{p}\in\mathsf{P}} \delta_{\mathsf{p}} \leq \sum_{\mathsf{p}\in\mathsf{P}} 2\mathbf{d}\left(\mathsf{p}, C_{\mathrm{opt}}\right) = 2\left\|\mathsf{P}_{C_{\mathrm{opt}}}\right\|_1$. ■

Drifters, anchors, and tyrants. A center of L that does not serve any center of C_{opt} (i.e., its cluster in Π is empty) is a ***drifter***. Formally, we map each center of C_{opt} to its nearest neighbor in L, and for a center $\overline{\mathsf{c}} \in L$ its ***degree***, denoted by $\deg(\overline{\mathsf{c}})$, is the number of points of C_{opt} mapped to it by this nearest neighbor mapping.

As such, a center $\overline{\mathsf{c}} \in L$ is a ***drifter*** if $\deg(\overline{\mathsf{c}}) = 0$, an ***anchor*** if $\deg(\overline{\mathsf{c}}) = 1$, and a ***tyrant*** if $\deg(\overline{\mathsf{c}}) > 1$. Observe that if $\overline{\mathsf{c}}$ is a drifter, then $\Pi[\overline{\mathsf{c}}] = \emptyset$.



The reader should not take these names too seriously, but observe that centers that are tyrants cannot easily move around and are bad candidates for swaps. Indeed, consider the situation depicted in the figure on the right. Here the center $\overline{\mathsf{c}}$ serves points of P that belong to two optimal clusters $\overline{\mathsf{o}}$ and $\overline{\mathsf{o}'}$, such that $\overline{\mathsf{c}} = \mathsf{nn}(\overline{\mathsf{o}}, L) = \mathsf{nn}\left(\overline{\mathsf{o}'}, L\right)$. If we swap $\overline{\mathsf{c}} \to \overline{\mathsf{o}}$, then the points in the cluster $\mathrm{cluster}\left(C_{\mathrm{opt}}, \overline{\mathsf{o}'}\right)$ might find themselves very far from any center in $L - \overline{\mathsf{c}} + \overline{\mathsf{o}}$. Similarly, the points of $\mathrm{cluster}\left(C_{\mathrm{opt}}, \overline{\mathsf{o}}\right)$ might be in trouble if we swap $\overline{\mathsf{c}} \to \overline{\mathsf{o}'}$.

Intuitively, since we shifted our thinking from the local solution to the partition Π, a drifter center is not being used by the clustering, and we can reassign it so that it decreases the price of the clustering.

That is, since moving from the local clustering of L to Π is relatively cheap, we can free a drifter $\overline{\mathsf{c}}$ from all its clients in the local partition. Formally, the ***ransom*** of a drifter center $\overline{\mathsf{c}}$ is $\mathrm{ransom}(\overline{\mathsf{c}}) = \sum_{\mathsf{p} \in \mathrm{cluster}\left(L, \overline{\mathsf{c}}\right)} \delta_{\mathsf{p}}$. This is the price of reassigning all the points that are currently served by the drifter $\overline{\mathsf{c}}$ to the center in L serving their optimal center. Once this ransom is paid, $\overline{\mathsf{c}}$ serves nobody and can be moved with no further charge.

More generally, the ***ransom*** of any center $\overline{\mathsf{c}} \in L$ is

$$\mathrm{ransom}(\overline{\mathsf{c}}) = \sum_{\mathsf{p} \in \mathrm{cluster}\left(L, \overline{\mathsf{c}}\right) \setminus \Pi[\overline{\mathsf{c}}]} \delta_{\mathsf{p}}.$$

Note that for a drifter $\Pi[\overline{\mathsf{c}}] = \emptyset$ and $\mathrm{cluster}(L, \overline{\mathsf{c}}) = \mathrm{cluster}(L, \overline{\mathsf{c}}) \setminus \Pi[\overline{\mathsf{c}}]$, and in general, the points of $\mathrm{cluster}(L, \overline{\mathsf{c}}) \setminus \Pi[\overline{\mathsf{c}}]$ are exactly the points of $\mathrm{cluster}(L, \overline{\mathsf{c}})$ being reassigned.

Hence, ransom($\overline{\mathsf{c}}$) is the increase in cost of reassigning the points of cluster $(L, \overline{\mathsf{c}})$ when moving from the local clustering of L to the clustering of Π.

Observe that, by Lemma 4.7, we have that

$$\sum_{\overline{\mathsf{c}} \in L} \text{ransom}(\overline{\mathsf{c}}) \leq 2 \left\| \mathsf{P}_{C_{\text{opt}}} \right\|_1 . \tag{4.5}$$

For $\overline{\mathsf{o}} \in C_{\text{opt}}$, the ***optimal price*** and ***local price*** of cluster $\left(C_{\text{opt}}, \overline{\mathsf{o}}\right)$ are

$$\text{opt}\left(\overline{\mathsf{o}}\right) = \sum_{\mathsf{p} \in \text{cluster}\left(C_{\text{opt}}, \overline{\mathsf{o}}\right)} \mathbf{d}\left(\mathsf{p}, C_{\text{opt}}\right) \qquad \text{and} \qquad \text{local}(\overline{\mathsf{o}}) = \sum_{\mathsf{p} \in \text{cluster}\left(C_{\text{opt}}, \overline{\mathsf{o}}\right)} \mathbf{d}(\mathsf{p}, L) ,$$

respectively.

Lemma 4.8. *If $\overline{\mathsf{c}} \in L$ is a drifter and $\overline{\mathsf{o}}$ is any center of C_{opt}, then* $\text{local}(\overline{\mathsf{o}}) \leq \text{ransom}(\overline{\mathsf{c}}) + \text{opt}(\overline{\mathsf{o}})$.

Proof. Since $\overline{\mathsf{c}} \in L$ is a drifter, we can swap it with any center in $\overline{\mathsf{o}} \in C_{\text{opt}}$. Since L is a locally optimal solution, we have that the change in the cost caused by the swap $\overline{\mathsf{c}} \to \overline{\mathsf{o}}$ is

$$\begin{aligned} 0 \leq \Delta(\overline{\mathsf{c}}, \overline{\mathsf{o}}) &\leq \text{ransom}(\overline{\mathsf{c}}) - \text{local}(\overline{\mathsf{o}}) + \text{opt}(\overline{\mathsf{o}}) \\ &\implies \quad \text{local}(\overline{\mathsf{o}}) \leq \text{ransom}(\overline{\mathsf{c}}) + \text{opt}(\overline{\mathsf{o}}) . \end{aligned} \tag{4.6}$$

Indeed, $\overline{\mathsf{c}}$ pays its ransom so that all the clients using it are now assigned to some other centers of L. Now, all the points of cluster $\left(C_{\text{opt}}, \overline{\mathsf{o}}\right)$ instead of paying $\text{local}(\overline{\mathsf{o}})$ are now paying (at most) $\text{opt}(\overline{\mathsf{o}})$. (We might pay less for a point $\mathsf{p} \in \text{cluster}\left(C_{\text{opt}}, \overline{\mathsf{o}}\right)$ if it is closer to $L - \overline{\mathsf{c}} + \overline{\mathsf{o}}$ than to $\overline{\mathsf{o}}$.) ■

Equation (4.6) provides us with a glimmer of hope that we can bound the price of the local clustering. We next argue that if there are many tyrants, then there must also be many drifters. In particular, with these drifters we can bound the price of the local clustering cost of the optimal clusters assigned to tyrants. Also, we argue that an anchor and its associated optimal center define a natural swap which is relatively cheap. Putting all of these together will imply the desired claim.

There are many drifters. Let S_{opt} (resp. A_{opt}) be the set of all the centers of C_{opt} that are assigned to tyrants (resp. anchors) by $\mathsf{nn}(\cdot, L)$. Observe that $S_{\text{opt}} \cup \mathrm{A}_{\text{opt}} = C_{\text{opt}}$. Let $\mathcal{D}$ be the set of drifters in L.

Observe that every tyrant has at least two followers in C_{opt}; that is, $\left|S_{\text{opt}}\right| \geq 2\#_{\text{tyrants}}$. Also, $k = \left|C_{\text{opt}}\right| = |L|$ and $\#_{\text{anchors}} = \left|\mathrm{A}_{\text{opt}}\right|$. As such, we have that

$$\begin{aligned} &\#_{\text{tyrants}} + \#_{\text{anchors}} + \#_{\text{drifters}} = |L| = \left|C_{\text{opt}}\right| = \left|S_{\text{opt}}\right| + \left|\mathrm{A}_{\text{opt}}\right| \\ &\implies \quad \#_{\text{drifters}} = \left|S_{\text{opt}}\right| + \left|\mathrm{A}_{\text{opt}}\right| - \#_{\text{anchors}} - \#_{\text{tyrants}} = \left|S_{\text{opt}}\right| - \#_{\text{tyrants}} \geq \left|S_{\text{opt}}\right|/2. \end{aligned} \tag{4.7}$$

Namely, $2\#_{\text{drifters}} \geq \left|S_{\text{opt}}\right|$.

Lemma 4.9. *We have that* $\displaystyle\sum_{\overline{\mathsf{o}} \in S_{\text{opt}}} \text{local}(\overline{\mathsf{o}}) \leq 2 \sum_{\overline{\mathsf{c}} \in \mathcal{D}} \text{ransom}(\overline{\mathsf{c}}) + \sum_{\overline{\mathsf{o}} \in S_{\text{opt}}} \text{opt}(\overline{\mathsf{o}})$.

Proof. If $\left|S_{\text{opt}}\right| = 0$, then the statement holds trivially.

So assume $\left|S_{\text{opt}}\right| > 0$ and let $\overline{\mathsf{c}}$ be the drifter with the lowest ransom($\overline{\mathsf{c}}$). For any $\overline{\mathsf{o}} \in S_{\text{opt}}$, we have that $\text{local}(\overline{\mathsf{o}}) \leq \text{ransom}(\overline{\mathsf{c}}) + \text{opt}(\overline{\mathsf{o}})$, by (4.6). Summing over all such

centers, we have that

$$\sum_{\overline{o}\in S_{\text{opt}}} \text{local}(\overline{o}) \leq \left|S_{\text{opt}}\right| \text{ransom}(\overline{c}) + \sum_{\overline{o}\in S_{\text{opt}}} \text{opt}(\overline{o}),$$

which is definitely smaller than the stated bound, since $\left|S_{\text{opt}}\right| \leq 2\,|\mathcal{D}|$, by (4.7). ■

Lemma 4.10. *We have that* $\sum_{\overline{o}\in \mathrm{A}_{\text{opt}}} \text{local}(\overline{o}) \leq \sum_{\overline{o}\in \mathrm{A}_{\text{opt}}} \text{ransom}(\mathsf{nn}(\overline{o}, L)) + \sum_{\overline{o}\in \mathrm{A}_{\text{opt}}} \text{opt}(\overline{o})$.

Proof. For a center $\overline{o} \in \mathrm{A}_{\text{opt}}$, its anchor is $\overline{c} = \mathsf{nn}(\overline{o}, L)$. Consider the swap $\overline{c} \to \overline{o}$, and the increase in clustering cost as we move from L to $L - \overline{c} + \overline{o}$.

We claim that $\text{local}(\overline{o}) \leq \text{ransom}(\overline{c}) + \text{opt}(\overline{o})$ (i.e., (4.6)) holds in this setting. The points for which their clustering is negatively affected (i.e., their clustering price might increase) by the swap are in the set $\text{cluster}(L, \overline{c}) \cup \text{cluster}\left(C_{\text{opt}}, \overline{o}\right)$, and we split this set into two disjoint sets $X = \text{cluster}(L, \overline{c}) \setminus \text{cluster}\left(C_{\text{opt}}, \overline{o}\right)$ and $Y = \text{cluster}\left(C_{\text{opt}}, \overline{o}\right)$.

The increase in price by reassigning the points of X to some other center in L is exactly the ransom of $\overline{c}$. Now, the points of Y might get reassigned to $\overline{o}$, and the change in price of the points of Y can now be bounded by $-\text{local}(\overline{o}) + \text{opt}(\overline{o})$, as was argued in the proof of Lemma 4.8.

Note that it might be that points outside $X \cup Y$ get reassigned to $\overline{o}$ in the clustering induced by $L - \overline{c} + \overline{o}$. However, such reassignment only further reduce the price of the swap. As such, we have that $0 \leq \Delta(\overline{c}, \overline{o}) \leq \text{ransom}(\overline{c}) - \text{local}(\overline{o}) + \text{opt}(\overline{o})$. As such, summing up the inequality $\text{local}(\overline{o}) \leq \text{ransom}(\overline{c}) + \text{opt}(\overline{o})$ over all the centers in A_{opt} implies the claim. ■

Lemma 4.11. *Let L be the set of k centers computed by the local search algorithm. We have that* $\|\mathsf{P}_L\|_1 \leq 5\text{opt}_1(\mathsf{P}, k)$.

Proof. From the above two lemmas, we have that

$$\begin{aligned}
\|\mathsf{P}_L\|_1 &= \sum_{\overline{o}\in C_{\text{opt}}} \text{local}(\overline{o}) = \sum_{\overline{o}\in S_{\text{opt}}} \text{local}(\overline{o}) + \sum_{\overline{o}\in \mathrm{A}_{\text{opt}}} \text{local}(\overline{o}) \\
&\leq 2 \sum_{\overline{c}\in\mathcal{D}} \text{ransom}(\overline{c}) + \sum_{\overline{o}\in S_{\text{opt}}} \text{opt}(\overline{o}) + \sum_{\overline{o}\in \mathrm{A}_{\text{opt}}} \text{ransom}(\mathsf{nn}(\overline{o}, L)) + \sum_{\overline{o}\in \mathrm{A}_{\text{opt}}} \text{opt}(\overline{o}) \\
&\leq 2 \sum_{\overline{c}\in L} \text{ransom}(\overline{c}) + \sum_{\overline{o}\in C_{\text{opt}}} \text{opt}(\overline{o}) \leq 4 \left\|\mathsf{P}_{C_{\text{opt}}}\right\|_1 + \left\|\mathsf{P}_{C_{\text{opt}}}\right\|_1 = 5\text{opt}_1(\mathsf{P}, k),
\end{aligned}$$

by (4.5). ■

4.3.2.2. *Removing the strict improvement assumption.* In the above proof, we assumed that the current local minimum cannot be improved by a swap. Of course, this might not hold for the **algLocalSearchKMed** solution, since the algorithm allows a swap only if it makes "significant" progress. In particular, (4.4) is in fact

$$\forall \overline{c} \in L, \overline{o} \in \mathsf{P} \setminus L, \qquad -\tau\, \|\mathsf{P}_L\|_1 \leq \|\mathsf{P}_{L-\overline{c}+\overline{o}}\|_1 - \|\mathsf{P}_L\|_1 . \tag{4.8}$$

To adapt the proof to use this modified inequality, observe that the proof worked by adding up k inequalities defined by (4.4) and getting the inequality $0 \leq 5\left\|\mathsf{P}_{C_{\text{opt}}}\right\|_1 - \|\mathsf{P}_L\|_1$. Repeating the same argumentation on the modified inequalities, which is tedious but straightforward, yields

$$-\tau k\, \|\mathsf{P}_L\|_1 \leq 5\left\|\mathsf{P}_{C_{\text{opt}}}\right\|_1 - \|\mathsf{P}_L\|_1 .$$

This implies $\|\mathsf{P}_L\|_1 \leq 5\left\|\mathsf{P}_{C_{\text{opt}}}\right\|_1/(1-\tau k)$. For arbitrary $0 < \varepsilon < 1$, setting $\tau = \varepsilon/10k$, we have that $\|\mathsf{P}_L\|_1 \leq 5(1+\varepsilon/5)\text{opt}_1$, since $1/(1-\tau k) \leq 1 + 2\tau k = 1 + \varepsilon/5$, for $\tau \leq 1/10k$. We summarize:

Theorem 4.12. *Let P be a set of n points in a metric space. For $0 < \varepsilon < 1$, one can compute a $(5+\varepsilon)$-approximation to the optimal k-median clustering of P. The running time of the algorithm is $O\left(n^2 k^3 \frac{\log n}{\varepsilon}\right)$.*

4.4. On k-means clustering

In the ***k-means clustering problem***, a set $\mathsf{P} \subseteq \mathcal{X}$ is provided together with a parameter k. We would like to find a set of k points $\mathbf{C} \subseteq \mathsf{P}$, such that the sum of squared distances of all the points of P to their closest point in $\mathbf{C}$ is minimized.

Formally, given a set of centers $\mathbf{C}$, the k-center clustering ***price*** of clustering P by $\mathbf{C}$ is denoted by

$$\|\mathsf{P}_{\mathbf{C}}\|_2^2 = \sum_{\mathsf{p}\in\mathsf{P}} (\mathbf{d}_{\mathcal{M}}(\mathsf{p}, \mathbf{C}))^2 ,$$

and the ***k-means problem*** is to find a set $\mathbf{C}$ of k points, such that $\|\mathsf{P}_{\mathbf{C}}\|_2^2$ is minimized; namely,

$$\text{opt}_2(\mathsf{P}, k) = \min_{\mathbf{C},|\mathbf{C}|=k} \|\mathsf{P}_{\mathbf{C}}\|_2^2 .$$

Local search also works for k-means and yields a constant factor approximation. We leave the proof of the following theorem to Exercise 4.4.

Theorem 4.13. *Let P be a set of n points in a metric space. For $0 < \varepsilon < 1$, one can compute a $(25+\varepsilon)$-approximation to the optimal k-means clustering of P. The running time of the algorithm is $O\left(n^2 k^3 \frac{\log n}{\varepsilon}\right)$.*

4.5. Bibliographical notes

In this chapter we introduced the problem of clustering and showed some algorithms that achieve constant factor approximations. A lot more is known about these problems including faster and better clustering algorithms, but to discuss them, we need more advanced tools than what we currently have at hand.

Clustering is widely researched. Unfortunately, a large fraction of the work on this topic relies on heuristics or experimental studies. The inherent problem seems to be the lack of a universal definition of what is a good clustering. This depends on the application at hand, which is rarely clearly defined. In particular, no clustering algorithm can achieve all desired properties together; see the work by Kleinberg [**Kle02**] (although it is unclear if all these desired properties are indeed natural or even really desired).

***k*-center clustering.** The algorithm **GreedyKCenter** is by Gonzalez [**Gon85**], but it was probably known before, as the notion of r-packing is much older. The hardness of approximating k-center clustering was shown by Feder and Greene [**FG88**].

***k*-median/means clustering.** The analysis of the local search algorithm is due to Arya et al. [**AGK+01**]. Our presentation however follows the simpler proof of Gupta and Tangwongsan [**GT08**]. The extension to k-means is due to Kanungo et al. [**KMN+04**]. The extension is not completely trivial since the triangle inequality no longer holds. However, some approximate version of the triangle inequality does hold. Instead of performing a single swap, one can decide to do p swaps simultaneously. Thus, the running time deteriorates since there are more possibilities to check. This improves the approximation constant

for the k-median (resp., k-means) to $(3 + 2/p)$ (resp. $(3 + 2/p)^2$). Unfortunately, this is (essentially) tight in the worst case. See [**AGK+01, KMN+04**] for details.

The k-median and k-means clustering are more interesting in Euclidean settings where there is considerably more structure, and one can compute a $(1+\varepsilon)$-approximation in linear time for fixed ε and k and d; see [**HM04**].

Since k-median and k-means clustering can be used to solve the dominating set in a graph, this implies that both clustering problems are NP-Hard to solve exactly.

One can also compute a permutation similar to the greedy permutation (for k-center clustering) for k-median clustering. See the work by Mettu and Plaxton [**MP03**].

Handling outliers. The problem of handling outliers is still not well understood. See the work of Charikar et al. [**CKMN01**] for some relevant results. In particular, for k-center clustering they get a constant factor approximation, and Exercise 4.3 is taken from there. For k-median clustering they present a constant factor approximation using a linear programming relaxation that also approximates the number of outliers. Recently, Chen [**Che08**] provided a constant factor approximation algorithm by extending the work of Charikar et al. The problem of finding a simple algorithm with simple analysis for k-median clustering with outliers is still open, as Chen's work is quite involved.

Open Problem 4.14. Get a *simple* constant factor k-median clustering algorithm that runs in polynomial time and uses exactly m outliers. Alternatively, solve this problem in the case where P is a set of n points in the plane. (The emphasize here is that the analysis of the algorithm should be simple.)

Bi-criteria approximation. All clustering algorithms tend to become considerably easier if one allows trade-off in the number of clusters. In particular, one can compute a constant factor approximation to the optimal k-median/means clustering using $O(k)$ centers in $O(nk)$ time. The algorithm succeeds with constant probability. See the work by Indyk [**Ind99**] and Chen [**Che06**] and references therein.

Facility location. All the problems mentioned here fall into the family of facility location problems. There are numerous variants. The more specific ***facility location*** problem is a variant of k-median clustering where the number of clusters is not specified, but instead one has to pay to open a facility in a certain location. Local search also works for this variant.

Local search. As mentioned above, ***local search*** also works for k-means clustering [**AGK+01**]. A collection of some basic problems for which local search works is described in the book by Kleinberg and Tardos [**KT06**]. Local search is a widely used heuristic for attacking NP-Hard problems. The idea is usually to start from a solution and try to locally improve it. Here, one defines a neighborhood of the current solution, and one tries to move to the best solution in this neighborhood. In this sense, local search can be thought of as a hill-climbing/EM (expectation maximization) algorithm. Problems for which local search was used include vertex cover, traveling salesperson, and satisfiability, and probably many other problems.

Provable cases where local search generates a guaranteed solution are less common and include facility location, k-median clustering [**AGK+01**], weighted max cut, k-means [**KMN+04**], the metric labeling problem with the truncated linear metric [**GT00**], and image segmentation [**BVZ01**]. See [**KT06**] for more references and a nice discussion of the connection of local search to the ***Metropolis algorithm*** and ***simulated annealing***.

4.6. Exercises

Exercise 4.1 (Another algorithm for k-center clustering)**.** Consider the algorithm that, given a point set P and a parameter k, initially picks an arbitrary set $\mathbf{C} \subseteq \mathsf{P}$ of k points. Next, it computes the closest pair of points $\overline{\mathsf{c}}, \overline{\mathsf{f}} \in \mathbf{C}$ and the point s realizing $\|\mathsf{P}_{\mathbf{C}}\|_\infty$. If $\mathbf{d}(\mathsf{s}, \mathbf{C}) > \mathbf{d}_{\mathcal{M}}\left(\overline{\mathsf{c}}, \overline{\mathsf{f}}\right)$, then the algorithm sets $\mathbf{C} \leftarrow \mathbf{C} - \overline{\mathsf{c}} + \mathsf{s}$ and repeats this process till the condition no longer holds.

(A) Prove that this algorithm outputs a k-center clustering of radius $\leq 2\mathrm{opt}_\infty(\mathsf{P}, k)$.
(B) What is the running time of this algorithm?
(C) If one is willing to trade off the approximation quality of this algorithm, it can be made faster. In particular, suggest a variant of this algorithm that in $O(k)$ iterations computes an $O(1)$-approximation to the optimal k-center clustering.

Exercise 4.2 (Handling outliers)**.** Given a point set P, we would like to perform a k-median clustering of it, where we are allowed to ignore m of the points. These m points are ***outliers*** which we would like to ignore since they represent irrelevant data. Unfortunately, we do not know the m outliers in advance. It is natural to conjecture that one can perform a local search for the optimal solution. Here one maintains a set of k centers and a set of m outliers. At every point in time the algorithm moves one of the centers or the outliers if it improves the solution.

Show that local search does not work for this problem; namely, the approximation factor is not a constant.

Exercise 4.3 (Handling outliers for k-center clustering)**.** Given P, k, and m, present a polynomial time algorithm that computes a constant factor approximation to the optimal k-center clustering of P with m outliers. (Hint: Assume first that you know the radius of the optimal solution.)

Exercise 4.4 (Local search for k-means clustering)**.** Prove Theorem 4.13.

CHAPTER 5

On Complexity, Sampling, and ε-Nets and ε-Samples

In this chapter we will try to quantify the notion of geometric complexity. It is intuitively clear that a ● (i.e., disk) is a simpler shape than an ⬬ (i.e., ellipse), which is in turn simpler than a ☻ (i.e., smiley). This becomes even more important when we consider several such shapes and how they interact with each other. As these examples might demonstrate, this notion of complexity is somewhat elusive.

To this end, we show that one can capture the structure of a distribution/point set by a small subset. The size here would depend on the complexity of the shapes/ranges we care about, but surprisingly it would be independent of the size of the point set.

5.1. VC dimension

Definition 5.1. A ***range space*** S is a pair $(\mathsf{X}, \mathcal{R})$, where X is a ***ground set*** (finite or infinite) and $\mathcal{R}$ is a (finite or infinite) family of subsets of X. The elements of X are ***points*** and the elements of $\mathcal{R}$ are ***ranges***.

Our interest is in the size/weight of the ranges in the range space. For technical reasons, it will be easier to consider a finite subset x as the underlining ground set.

Definition 5.2. Let $\mathsf{S} = (\mathsf{X}, \mathcal{R})$ be a range space, and let x be a finite (fixed) subset of X. For a range $\mathbf{r} \in \mathcal{R}$, its ***measure*** is the quantity

$$\overline{m}(\mathbf{r}) = \frac{|\mathbf{r} \cap \mathsf{x}|}{|\mathsf{x}|}.$$

While x is finite, it might be very large. As such, we are interested in getting a good estimate to $\overline{m}(\mathbf{r})$ by using a more compact set to represent the range space.

Definition 5.3. Let $\mathsf{S} = (\mathsf{X}, \mathcal{R})$ be a range space. For a subset N (which might be a multi-set) of x, its ***estimate*** of the measure of $\overline{m}(\mathbf{r})$, for $\mathbf{r} \in \mathcal{R}$, is the quantity

$$\overline{s}(\mathbf{r}) = \frac{|\mathbf{r} \cap N|}{|N|}.$$

The main purpose of this chapter is to come up with methods to generate a sample N, such that $\overline{m}(\mathbf{r}) \approx \overline{s}(\mathbf{r})$, for all the ranges $\mathbf{r} \in \mathcal{R}$.

It is easy to see that in the worst case, no sample can capture the measure of all ranges. Indeed, given a sample N, consider the range $\mathsf{x} \setminus N$ that is being completely missed by N. As such, we need to concentrate on range spaces that are "low dimensional", where not all subsets are allowable ranges. The notion of VC dimension (named after Vapnik and Chervonenkis [**VC71**]) is one way to limit the complexity of a range space.

Definition 5.4. Let $\mathsf{S} = (\mathsf{X}, \mathcal{R})$ be a range space. For $Y \subseteq \mathsf{X}$, let

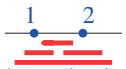
$$\mathcal{R}_{|Y} = \left\{ \mathbf{r} \cap Y \;\middle|\; \mathbf{r} \in \mathcal{R} \right\} \tag{5.1}$$

denote the ***projection*** of $\mathcal{R}$ on Y. The range space S projected to Y is $\mathsf{S}_{|Y} = (Y, \mathcal{R}_{|Y})$.

If $\mathcal{R}_{|Y}$ contains all subsets of Y (i.e., if Y is finite, we have $\left|\mathcal{R}_{|Y}\right| = 2^{|Y|}$), then Y is ***shattered*** by $\mathcal{R}$ (or equivalently Y is shattered by S).

The ***Vapnik-Chervonenkis*** dimension (or ***VC dimension***) of S, denoted by $\dim_{\mathsf{VC}}(S)$, is the maximum cardinality of a shattered subset of X. If there are arbitrarily large shattered subsets, then $\dim_{\mathsf{VC}}(S) = \infty$.

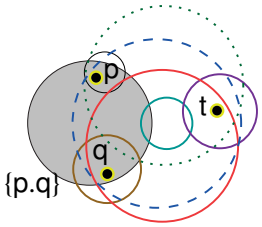
5.1.1. Examples.



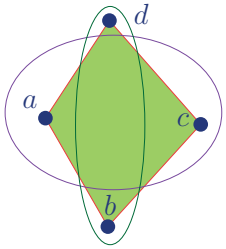
Intervals. Consider the set X to be the real line, and consider $\mathcal{R}$ to be the set of all intervals on the real line. Consider the set $Y = \{1, 2\}$. Clearly, one can find four intervals that contain all possible subsets of Y. Formally, the projection $\mathcal{R}_{|Y} = \left\{ \{\}, \{1\}, \{2\}, \{1, 2\} \right\}$. The intervals realizing each of these subsets are depicted on the right.

However, this is false for a set of three points $B = \{\mathsf{p}, \mathsf{q}, \mathsf{s}\}$, since there is no interval that can contain the two extreme points p and s without also containing q. Namely, the subset $\{\mathsf{p}, \mathsf{s}\}$ is not realizable for intervals, implying that the largest shattered set by the range space (real line, intervals) is of size two. We conclude that the VC dimension of this space is two.

Disks. Let $\mathsf{X} = \mathbb{R}^2$, and let $\mathcal{R}$ be the set of disks in the plane. Clearly, for any three points in the plane (in general position), denoted by p, q, and s, one can find eight disks that realize all possible 2^3 different subsets. See the figure on the right.

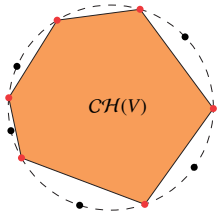


But can disks shatter a set with four points? Consider such a set P of four points. If the convex hull of P has only three points on its boundary, then the subset X having only those three vertices (i.e., it does not include the middle point) is impossible, by convexity. Namely, there is no disk that contains only the points of X without the middle point.



Alternatively, if all four points are vertices of the convex hull and they are a, b, c, d along the boundary of the convex hull, either the set $\{a, c\}$ or the set $\{b, d\}$ is not realizable. Indeed, if both options are realizable, then consider the two disks D_1 and D_2 that realize those assignments. Clearly, ∂D_1 and ∂D_2 must intersect in four points, but this is not possible, since two circles have at most two intersection points. See the figure on the left. Hence the VC dimension of this range space is 3.

Convex sets. Consider the range space $\mathsf{S} = (\mathbb{R}^2, \mathcal{R})$, where $\mathcal{R}$ is the set of all (closed) convex sets in the plane. We claim that $\dim_{\mathsf{VC}}(\mathsf{S}) = \infty$. Indeed, consider a set U of n points $p_1, \ldots, p_n$ all lying on the boundary of the unit circle in the plane. Let V be any subset of U, and consider the convex hull $\mathcal{CH}(V)$. Clearly, $\mathcal{CH}(V) \in \mathcal{R}$, and furthermore, $\mathcal{CH}(V) \cap U = V$. Namely, any subset of U is realizable by S. Thus, S can shatter sets of arbitrary size, and its VC dimension is unbounded.



Complement. Consider the range space $\mathsf{S} = (\mathsf{X}, \mathcal{R})$ with $\delta = \dim_{\mathsf{VC}}(\mathsf{S})$. Next, consider the complement space, $\overline{\mathsf{S}} = \left(\mathsf{X}, \overline{\mathcal{R}}\right)$, where

$$\overline{\mathcal{R}} = \left\{ \mathsf{X} \setminus \mathbf{r} \;\middle|\; \mathbf{r} \in \mathcal{R} \right\};$$

namely, the ranges of $\overline{\mathsf{S}}$ are the complement of the ranges in S. What is the VC dimension of $\overline{\mathsf{S}}$? Well, a set $B \subseteq \mathsf{X}$ is shattered by $\overline{\mathsf{S}}$ if and only if it is shattered by S. Indeed, if S shatters B, then for any $Z \subseteq B$, we have that $(B \setminus Z) \in \mathcal{R}_{|B}$, which implies that $Z = B \setminus (B \setminus Z) \in \overline{\mathcal{R}}_{|B}$. Namely, $\overline{\mathcal{R}}_{|B}$ contains all the subsets of B, and $\overline{\mathsf{S}}$ shatters B. Thus, $\dim_{\mathsf{VC}}\left(\overline{\mathsf{S}}\right) = \dim_{\mathsf{VC}}(\mathsf{S})$.

Lemma 5.5. *For a range space $\mathsf{S} = (\mathsf{X}, \mathcal{R})$ we have that $\dim_{\mathsf{VC}}(\mathsf{S}) = \dim_{\mathsf{VC}}\left(\overline{\mathsf{S}}\right)$, where $\overline{\mathsf{S}}$ is the complement range space.*

5.1.1.1. *Halfspaces.* Let $\mathsf{S} = (\mathsf{X}, \mathcal{R})$, where $\mathsf{X} = \mathbb{R}^{\mathsf{d}}$ and $\mathcal{R}$ is the set of all (closed) halfspaces in $\mathbb{R}^{\mathsf{d}}$. We need the following technical claim.

Claim 5.6. *Let $\mathsf{P} = \{\mathsf{p}_1, \ldots, \mathsf{p}_{\mathsf{d}+2}\}$ be a set of $\mathsf{d}+2$ points in $\mathbb{R}^{\mathsf{d}}$. There are real numbers $\beta_1, \ldots, \beta_{\mathsf{d}+2}$, not all of them zero, such that $\sum_i \beta_i \mathsf{p}_i = 0$ and $\sum_i \beta_i = 0$.*

Proof. Indeed, set $\mathsf{q}_i = (\mathsf{p}_i, 1)$, for $i = 1, \ldots, d+2$. Now, the points $\mathsf{q}_1, \ldots, \mathsf{q}_{\mathsf{d}+2} \in \mathbb{R}^{\mathsf{d}+1}$ are linearly dependent, and there are coefficients $\beta_1, \ldots, \beta_{\mathsf{d}+2}$, not all of them zero, such that $\sum_{i=1}^{\mathsf{d}+2} \beta_i \mathsf{q}_i = 0$. Considering only the first d coordinates of these points implies that $\sum_{i=1}^{\mathsf{d}+2} \beta_i \mathsf{p}_i = 0$. Similarly, by considering only the $(d+1)$st coordinate of these points, we have that $\sum_{i=1}^{\mathsf{d}+2} \beta_i = 0$. ■

To see what the VC dimension of halfspaces in $\mathbb{R}^{\mathsf{d}}$ is, we need the following result of Radon. (For a reminder of the formal definition of convex hulls, see Definition 28.1$_{\text{p347}}$.)

Theorem 5.7 (Radon's theorem)**.** *Let $\mathsf{P} = \{\mathsf{p}_1, \ldots, \mathsf{p}_{\mathsf{d}+2}\}$ be a set of $\mathsf{d}+2$ points in $\mathbb{R}^{\mathsf{d}}$. Then, there exist two disjoint subsets C and D of P, such that $\mathcal{CH}(C) \cap \mathcal{CH}(D) \neq \emptyset$ and $C \cup D = \mathsf{P}$.*

Proof. By Claim 5.6 there are real numbers $\beta_1, \ldots, \beta_{\mathsf{d}+2}$, not all of them zero, such that $\sum_i \beta_i \mathsf{p}_i = 0$ and $\sum_i \beta_i = 0$.

Assume, for the sake of simplicity of exposition, that $\beta_1, \ldots, \beta_k \geq 0$ and $\beta_{k+1}, \ldots, \beta_{\mathsf{d}+2} < 0$. Furthermore, let $\mu = \sum_{i=1}^{k} \beta_i = -\sum_{i=k+1}^{\mathsf{d}+2} \beta_i$. We have that

$$\sum_{i=1}^{k} \beta_i \mathsf{p}_i = -\sum_{i=k+1}^{\mathsf{d}+2} \beta_i \mathsf{p}_i.$$

In particular, $v = \sum_{i=1}^{k} (\beta_i/\mu)\mathsf{p}_i$ is a point in $\mathcal{CH}(\{\mathsf{p}_1, \ldots, \mathsf{p}_k\})$. Furthermore, for the same point v we have $v = \sum_{i=k+1}^{\mathsf{d}+2} -(\beta_i/\mu)\mathsf{p}_i \in \mathcal{CH}(\{\mathsf{p}_{k+1}, \ldots, \mathsf{p}_{\mathsf{d}+2}\})$. We conclude that v is in the intersection of the two convex hulls, as required. ■

The following is a trivial observation, and yet we provide a proof to demonstrate it is true.

Lemma 5.8. *Let $\mathsf{P} \subseteq \mathbb{R}^{\mathsf{d}}$ be a finite set, let s be any point in $\mathcal{CH}(\mathsf{P})$, and let h^+ be a halfspace of $\mathbb{R}^{\mathsf{d}}$ containing s. Then there exists a point of P contained inside h^+.*

PROOF. The halfspace h^+ can be written as $h^+ = \left\{ \mathsf{t} \in \mathbb{R}^{\mathsf{d}} \;\middle|\; \langle \mathsf{t}, v \rangle \le c \right\}$. Now $\mathsf{s} \in \mathcal{CH}(\mathsf{P}) \cap h^+$, and as such there are numbers $\alpha_1, \ldots, \alpha_m \ge 0$ and points $\mathsf{p}_1, \ldots, \mathsf{p}_m \in \mathsf{P}$, such that $\sum_i \alpha_i = 1$ and $\sum_i \alpha_i \mathsf{p}_i = \mathsf{s}$. By the linearity of the dot product, we have that

$$\mathsf{s} \in h^+ \implies \langle \mathsf{s}, v \rangle \le c \implies \left\langle \sum_{i=1}^{m} \alpha_i \mathsf{p}_i, v \right\rangle \le c \implies \beta = \sum_{i=1}^{m} \alpha_i \langle \mathsf{p}_i, v \rangle \le c.$$

Setting $\beta_i = \langle \mathsf{p}_i, v \rangle$, for $i = 1, \ldots, m$, the above implies that β is a weighted average of $\beta_1, \ldots, \beta_m$. In particular, there must be a β_i that is no larger than the average. That is $\beta_i \le c$. This implies that $\langle \mathsf{p}_i, v \rangle \le c$. Namely, $\mathsf{p}_i \in h^+$ as claimed. ■

Let S be the range space having $\mathbb{R}^{\mathsf{d}}$ as the ground set and all the close halfspaces as ranges. Radon's theorem implies that if a set Q of $\mathsf{d}+2$ points is being shattered by S, then we can partition this set Q into two disjoint sets Y and Z such that $\mathcal{CH}(Y) \cap \mathcal{CH}(Z) \neq \emptyset$. In particular, let s be a point in $\mathcal{CH}(Y) \cap \mathcal{CH}(Z)$. If a halfspace h^+ contains all the points of Y, then $\mathcal{CH}(Y) \subseteq h^+$, since a halfspace is a convex set. Thus, any halfspace h^+ containing all the points of Y will contain the point $\mathsf{s} \in \mathcal{CH}(Y)$. But $\mathsf{s} \in \mathcal{CH}(Z) \cap h^+$, and this implies that a point of Z must lie in h^+, by Lemma 5.8. Namely, the subset $Y \subseteq \mathsf{Q}$ cannot be realized by a halfspace, which implies that Q cannot be shattered. Thus $\dim_{\mathsf{VC}}(S) < d+2$. It is also easy to verify that the regular simplex with $\mathsf{d}+1$ vertices is shattered by S. Thus, $\dim_{\mathsf{VC}}(S) = \mathsf{d}+1$.

5.2. Shattering dimension and the dual shattering dimension

The main property of a range space with bounded VC dimension is that the number of ranges for a set of n elements grows polynomially in n (with the power being the dimension) instead of exponentially. Formally, let the ***growth function*** be

$$\mathcal{G}_\delta(n) = \sum_{i=0}^{\delta} \binom{n}{i} \le \sum_{i=0}^{\delta} \frac{n^i}{i!} \le n^\delta, \tag{5.2}$$

for $\delta > 1$ (the cases where $\delta = 0$ or $\delta = 1$ are not interesting and we will just ignore them). Note that for all $n, \delta \ge 1$, we have $\mathcal{G}_\delta(n) = \mathcal{G}_\delta(n-1) + \mathcal{G}_{\delta-1}(n-1)$[①].

Lemma 5.9 (Sauer's lemma)**.** *If $(X, \mathcal{R})$ is a range space of VC dimension δ with $|X| = n$, then $|\mathcal{R}| \le \mathcal{G}_\delta(n)$.*

PROOF. The claim trivially holds for $\delta = 0$ or $n = 0$.

Let x be any element of X, and consider the sets

$$\mathcal{R}_x = \left\{ \mathbf{r} \setminus \{x\} \;\middle|\; \mathbf{r} \cup \{x\} \in \mathcal{R} \text{ and } \mathbf{r} \setminus \{x\} \in \mathcal{R} \right\} \quad \text{and} \quad \mathcal{R} \setminus x = \left\{ \mathbf{r} \setminus \{x\} \;\middle|\; \mathbf{r} \in \mathcal{R} \right\}.$$

Observe that $|\mathcal{R}| = |\mathcal{R}_x| + |\mathcal{R} \setminus x|$. Indeed, we charge the elements of $\mathcal{R}$ to their corresponding element in $\mathcal{R} \setminus x$. The only bad case is when there is a range $\mathbf{r}$ such that both $\mathbf{r} \cup \{x\} \in \mathcal{R}$ and $\mathbf{r} \setminus \{x\} \in \mathcal{R}$, because then these two distinct ranges get mapped to the same range in $\mathcal{R} \setminus x$. But such ranges contribute exactly one element to $\mathcal{R}_x$. Similarly, every element of $\mathcal{R}_x$ corresponds to two such "twin" ranges in $\mathcal{R}$.

[①] Here is a cute (and standard) counting argument: $\mathcal{G}_\delta(n)$ is just the number of different subsets of size at most δ out of n elements. Now, we either decide to not include the first element in these subsets (i.e., $\mathcal{G}_\delta(n-1)$) or, alternatively, we include the first element in these subsets, but then there are only $\delta - 1$ elements left to pick (i.e., $\mathcal{G}_{\delta-1}(n-1)$).

Observe that $(X\setminus\{x\}, \mathcal{R}_x)$ has VC dimension $\delta-1$, as the largest set that can be shattered is of size $\delta-1$. Indeed, any set $B \subset \mathsf{X}\setminus\{x\}$ shattered by $\mathcal{R}_x$ implies that $B\cup\{x\}$ is shattered in $\mathcal{R}$.

Thus, we have

$$|\mathcal{R}| = |\mathcal{R}_x| + |\mathcal{R}\setminus x| \leq \mathcal{G}_{\delta-1}(n-1) + \mathcal{G}_\delta(n-1) = \mathcal{G}_\delta(n),$$

by induction. ■

Interestingly, Lemma 5.9 is tight. See Exercise 5.4.

Next, we show pretty tight bounds on $\mathcal{G}_\delta(n)$. The proof is technical and not very interesting, and it is delegated to Section 5.6.

Lemma 5.10. *For $n \geq 2\delta$ and $\delta \geq 1$, we have $\left(\frac{n}{\delta}\right)^\delta \leq \mathcal{G}_\delta(n) \leq 2\left(\frac{ne}{\delta}\right)^\delta$, where $\mathcal{G}_\delta(n) = \sum_{i=0}^{\delta}\binom{n}{i}$.*

Definition 5.11 (Shatter function)**.** Given a range space $\mathsf{S} = (\mathsf{X}, \mathcal{R})$, its ***shatter function*** $\pi_{\mathsf{S}}(m)$ is the maximum number of sets that might be created by S when restricted to subsets of size m. Formally,

$$\pi_{\mathsf{S}}(m) = \max_{\substack{B\subset\mathsf{X}\\ |B|=m}} \left|\mathcal{R}_{|B}\right|;$$

see (5.1).

The ***shattering dimension*** of S is the smallest d such that $\pi_{\mathsf{S}}(m) = O(m^{\mathsf{d}})$, for all m.

By applying Lemma 5.9 to a finite subset of X, we get:

Corollary 5.12. *If $\mathsf{S} = (\mathsf{X}, \mathcal{R})$ is a range space of VC dimension δ, then for every finite subset B of X, we have $\left|\mathcal{R}_{|B}\right| \leq \pi_{\mathsf{S}}(|B|) \leq \mathcal{G}_\delta(|B|)$. That is, the VC dimension of a range space always bounds its shattering dimension.*

Proof. Let $n = |B|$, and observe that $\left|\mathcal{R}_{|B}\right| \leq \mathcal{G}_\delta(n) \leq n^\delta$, by (5.2). As such, $\left|\mathcal{R}_{|B}\right| \leq n^\delta$, and, by definition, the shattering dimension of S is at most δ; namely, the shattering dimension is bounded by the VC dimension. ■

Our arch-nemesis in the following is the function $x/\ln x$. The following lemma states some properties of this function, and its proof is delegated to Exercise 5.2.

Lemma 5.13. *For the function $f(x) = x/\ln x$ the following hold.*

(A) $f(x)$ is monotonically increasing for $x \geq e$.
(B) $f(x) \geq e$, for $x > 1$.
(C) For $u \geq \sqrt{e}$, if $f(x) \leq u$, then $x \leq 2u\ln u$.
(D) For $u \geq \sqrt{e}$, if $x > 2u\ln u$, then $f(x) > u$.
(E) For $u \geq e$, if $f(x) \geq u$, then $x \geq u\ln u$.

The next lemma introduces a standard argument which is useful in bounding the VC dimension of a range space by its shattering dimension. It is easy to see that the bound is tight in the worst case.

Lemma 5.14. *If $\mathsf{S} = (\mathsf{X}, \mathcal{R})$ is a range space with shattering dimension d, then its VC dimension is bounded by $O(\mathsf{d}\log\mathsf{d})$.*

PROOF. Let $N \subseteq \mathsf{X}$ be the largest set shattered by S, and let δ denote its cardinality. We have that $2^\delta = \left|\mathcal{R}_{|N}\right| \leq \pi_{\mathsf{S}}(|N|) \leq c\delta^{\mathsf{d}}$, where c is a fixed constant. As such, we have that $\delta \leq \lg c + \mathsf{d}\lg\delta$, which in turn implies that $\dfrac{\delta - \lg c}{\lg \delta} \leq \mathsf{d}$.[②] Assuming $\delta \geq \max(2, 2\lg c)$, we have that

$$\frac{\delta}{2\lg\delta} \leq \mathsf{d} \implies \frac{\delta}{\ln\delta} \leq \frac{2\mathsf{d}}{\ln 2} \leq 6\mathsf{d} \implies \delta \leq 2(6\mathsf{d})\ln(6\mathsf{d}),$$

by Lemma 5.13(C). ■

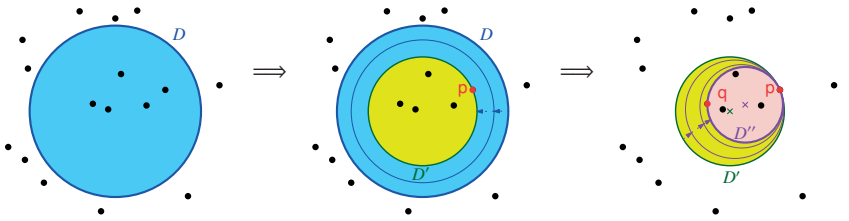
Disks revisited. To see why the shattering dimension is more convenient to work with than the VC dimension, consider the range space $\mathsf{S} = (\mathsf{X}, \mathcal{R})$, where $\mathsf{X} = \mathbb{R}^2$ and $\mathcal{R}$ is the set of disks in the plane. We know that the VC dimension of S is 3 (see Section 5.1.1).

We next use a standard continuous deformation argument to argue that the shattering dimension of this range space is also 3.

Lemma 5.15. *Consider the range space $\mathsf{S} = (\mathsf{X}, \mathcal{R})$, where $\mathsf{X} = \mathbb{R}^2$ and $\mathcal{R}$ is the set of disks in the plane. The shattering dimension of S is 3.*

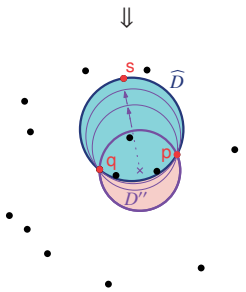
PROOF. Consider any set P of n points in the plane, and consider the set $\mathcal{F} = \mathcal{R}_{|\mathsf{P}}$. We claim that $|\mathcal{F}| \leq 4n^3$.

The set $\mathcal{F}$ contains only n sets with a single point in them and only $\binom{n}{2}$ sets with two points in them. So, fix $\mathsf{Q} \in \mathcal{F}$ such that $|\mathsf{Q}| \geq 3$.



There is a disk D that realizes this subset; that is, $\mathsf{P} \cap D = \mathsf{Q}$. For the sake of simplicity of exposition, assume that P is in general position. Shrink D till its boundary passes through a point p of P.

Now, continue shrinking the new disk D' in such a way that its boundary passes through the point p (this can be done by moving the center of D' towards p). Continue in this continuous deformation till the new boundary hits another point q of P. Let D'' denote this disk.



Next, we continuously deform D'' so that it has both $\mathsf{p} \in \mathsf{Q}$ and $\mathsf{q} \in \mathsf{Q}$ on its boundary. This can be done by moving the center of D'' along the bisector linear between p and q. Stop as soon as the boundary of the disk hits a third point $\mathsf{s} \in \mathsf{P}$. (We have freedom in choosing in which direction to move the center. As such, move in the direction that causes the disk boundary to hit a new point s.) Let $\widehat{D}$ be the resulting disk. The boundary of $\widehat{D}$ is the unique circle passing through p, q, and s. Furthermore, observe that

$$D \cap \big(\mathsf{P} \setminus \{\mathsf{s}\}\big) = \widehat{D} \cap \big(\mathsf{P} \setminus \{\mathsf{s}\}\big).$$

[②] We remind the reader that $\lg = \log_2$.

That is, we can specify the point set $\mathsf{P} \cap D$ by specifying the three points $\mathsf{p}, \mathsf{q}, \mathsf{s}$ (and thus specifying the disk $\widehat{D}$) and the status of the three special points; that is, we specify for each point $\mathsf{p}, \mathsf{q}, \mathsf{s}$ whether or not it is inside the generated subset.

As such, there are at most $8\binom{n}{3}$ different subsets in $\mathcal{F}$ containing more than three points, as each such subset maps to a "canonical" disk, there are at most $\binom{n}{3}$ different such disks, and each such disk defines at most eight different subsets.

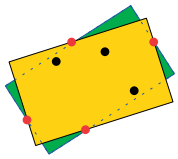
Similar argumentation implies that there are at most $4\binom{n}{2}$ subsets that are defined by a pair of points that realizes the diameter of the resulting disk. Overall, we have that

$$|\mathcal{F}| = 1 + n + 4\binom{n}{2} + 8\binom{n}{3} \leq 4n^3,$$

since there is one empty set in $\mathcal{F}$, n sets of size 1, and the rest of the sets are counted as described above. ■

The proof of Lemma 5.15 might not seem like a great simplification over the same bound we got by arguing about the VC dimension. However, the above argumentation gives us a very powerful tool – the shattering dimension of a range space defined by a family of shapes is always bounded by the number of points that determine a shape in the family.

Thus, the shattering dimension of, say, arbitrarily oriented rectangles in the plane is bounded by (and in this case, equal to) five, since such a rectangle is uniquely determined by five points. To see that, observe that if a rectangle has only four points on its boundary, then there is one degree of freedom left, since we can rotate the rectangle "around" these points; see the figure on the right.



5.2.1. The dual shattering dimension. Given a range space $\mathsf{S} = (\mathsf{X}, \mathcal{R})$, consider a point $\mathsf{p} \in \mathsf{X}$. There is a set of ranges of $\mathcal{R}$ associated with p, namely, the set of all ranges of $\mathcal{R}$ that contains p which we denote by

$$\mathcal{R}_{\mathsf{p}} = \left\{ \mathbf{r} \;\middle|\; \mathbf{r} \in \mathcal{R},\ \text{the range } \mathbf{r} \text{ contains } \mathsf{p} \right\}.$$

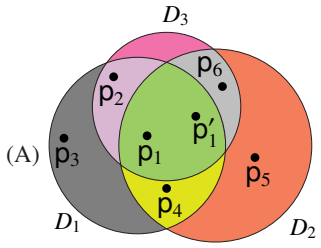
This gives rise to a natural dual range space to S.

Definition 5.16. The ***dual range space*** to a range space $\mathsf{S} = (\mathsf{X}, \mathcal{R})$ is the space $\mathsf{S}^\star = (\mathcal{R}, \mathsf{X}^\star)$, where $\mathsf{X}^\star = \left\{ \mathcal{R}_{\mathsf{p}} \;\middle|\; \mathsf{p} \in \mathsf{X} \right\}$.

Naturally, the dual range space to $\mathsf{S}^\star$ is the original S, which is thus sometimes referred to as the ***primal range space***. (In other words, the dual to the dual is the primal.) The easiest way to see this, is to think about it as an abstract set system realized as an incidence matrix, where each point is a column and a set is a row in the matrix having 1 in an entry if and only if it contains the corresponding point; see Figure 5.1. Now, it is easy to verify that the dual range space is the transposed matrix.

To understand what the dual space is, consider X to be the plane and $\mathcal{R}$ to be a set of m disks. Then, in the dual range space $\mathsf{S}^\star = (\mathcal{R}, \mathsf{X}^\star)$, every point p in the plane has a set associated with it in $\mathsf{X}^\star$, which is the set of disks of $\mathcal{R}$ that contains p. In particular, if we consider the arrangement formed by the m disks of $\mathcal{R}$, then all the points lying inside a single face of this arrangement correspond to the same set of $\mathsf{X}^\star$. The number of ranges in $\mathsf{X}^\star$ is bounded by the complexity of the arrangement of these disks, which is $O(m^2)$; see Figure 5.1.

(A)



(B)

	p_1	p'_1	p_2	p_3	p_4	p_5	p_6
D_1	1	1	1	1	1	0	0
D_2	1	1	0	0	1	1	1
D_3	1	1	1	0	0	0	1

(C)

	D_1	D_2	D_3
p_1	1	1	1
p'_1	1	1	1
p_2	1	0	1
p_3	1	0	0
p_4	1	1	0
p_5	0	1	0
p_6	0	1	1

FIGURE 5.1. (A) $\mathcal{R}_{p_1} = \mathcal{R}_{p'_1}$. (B) Writing the set system as an incidence matrix where a point is a column and a set is a row. For example, D_2 contains p_4, and as such the column of p_4 has a 1 in the row corresponding to D_2. (C) The dual set system is represented by a matrix which is the transpose of the original incidence matrix.

Let the ***dual shatter function*** of the range space S be $\pi^{\star}_{\mathsf{S}}(m) = \pi_{\mathsf{S}^{\star}}(m)$, where $\mathsf{S}^{\star}$ is the dual range space to S.

Definition 5.17. The ***dual shattering dimension*** of S is the shattering dimension of the dual range space $\mathsf{S}^{\star}$.

Note that the dual shattering dimension might be smaller than the shattering dimension and hence also smaller than the VC dimension of the range space. Indeed, in the case of disks in the plane, the dual shattering dimension is just 2, while the VC dimension and the shattering dimension of this range space is 3. Note, also, that in geometric settings bounding the dual shattering dimension is relatively easy, as all you have to do is bound the complexity of the arrangement of m ranges of this space.

The following lemma shows a connection between the VC dimension of a space and its dual. The interested reader[③] might find the proof amusing.

Lemma 5.18. *Consider a range space $\mathsf{S} = (X, \mathcal{R})$ with VC dimension δ. The dual range space $\mathsf{S}^{\star} = (\mathcal{R}, X^{\star})$ has VC dimension bounded by $2^{\delta+1}$.*

PROOF. Assume that $\mathsf{S}^{\star}$ shatters a set $\mathcal{F} = \{\mathbf{r}_1, \ldots, \mathbf{r}_k\} \subseteq \mathcal{R}$ of k ranges. Then, there is a set $\mathsf{P} \subseteq X$ of $m = 2^k$ points that shatters $\mathcal{F}$. Formally, for every subset $V \subseteq \mathcal{F}$, there exists a point $\mathsf{p} \in \mathsf{P}$, such that $\mathcal{F}_{\mathsf{p}} = V$.

So, consider the matrix $\mathbf{M}$ (of dimensions $k \times 2^k$) having the points $\mathsf{p}_1, \ldots, \mathsf{p}_{2^k}$ of P as the columns, and every row is a set of $\mathcal{F}$, where the entry in the matrix corresponding to a point $\mathsf{p} \in \mathsf{P}$ and a range $\mathbf{r} \in \mathcal{F}$ is 1 if and only if $\mathsf{p} \in \mathbf{r}$ and zero otherwise. Since P shatters $\mathcal{F}$, we know that this matrix has all possible 2^k binary vectors as columns.

[③]The author is quite aware that the interest of the reader in this issue might not be the result of free choice. Nevertheless, one might draw some comfort from the realization that the existence of the interested reader is as much an illusion as the existence of free choice. Both are convenient to assume, and both are probably false. Or maybe not.

Next, let $\kappa' = 2^{\lfloor \lg k \rfloor} \leq k$, and consider the matrix $\mathbf{M}'$ of size $\kappa' \times \lg \kappa'$, where the ith row is the binary representation of the number $i-1$ (formally, the jth entry in the ith row is 1 if the jth bit in the binary representation of $i-1$ is 1), where $i = 1, \ldots, \kappa'$. See the figure on the right. Clearly, the $\lg \kappa'$ columns of $\mathbf{M}'$ are all different, and we can find $\lg \kappa'$ columns of $\mathbf{M}$ that are identical to the columns of $\mathbf{M}'$ (in the first κ' entries starting from the top of the columns).

$\mathbf{M}$:

	p_1	p_2	...	p_{2^k}
$\mathbf{r}_1$	0	1		0
$\mathbf{r}_2$	1	1		1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathbf{r}_{k-2}$	1	1	...	0
$\mathbf{r}_{k-1}$	0	0	...	1
$\mathbf{r}_k$	1	0	...	1

$\mathbf{M}'$:

0	0	0
0	0	1
0	1	0
0	1	1
1	0	0
1	0	1
1	1	0
1	1	1

Each such column corresponds to a point $\mathsf{p} \in \mathsf{P}$, and let $\mathsf{Q} \subset \mathsf{P}$ be this set of $\lg \kappa'$ points. Note that for any subset $Z \subseteq \mathsf{Q}$, there is a row t in $\mathbf{M}'$ that encodes this subset. Consider the corresponding row in $\mathbf{M}$; that is, the range $\mathbf{r}_t \in \mathcal{F}$. Since $\mathbf{M}$ and $\mathbf{M}'$ are identical (in the relevant $\lg \kappa'$ columns of $\mathbf{M}$) on the first κ', we have that $\mathbf{r}_t \cap \mathsf{Q} = Z$. Namely, the set of ranges $\mathcal{F}$ shatters Q. But since the original range space has VC dimension δ, it follows that $|\mathsf{Q}| \leq \delta$. Namely, $|\mathsf{Q}| = \lg \kappa' = \lfloor \lg k \rfloor \leq \delta$, which implies that $\lg k \leq \delta + 1$, which in turn implies that $k \leq 2^{\delta+1}$. ■

Lemma 5.19. *If a range space $\mathsf{S} = (X, \mathcal{R})$ has dual shattering dimension δ, then its* VC *dimension is bounded by $\delta^{O(\delta)}$.*

Proof. The shattering dimension of the dual range space $\mathsf{S}^\star$ is bounded by δ, and as such, by Lemma 5.14, its VC dimension is bounded by $\delta' = O(\delta \log \delta)$. Since the dual range space to $\mathsf{S}^\star$ is S, we have by Lemma 5.18 that the VC dimension of S is bounded by $2^{\delta'+1} = \delta^{O(\delta)}$. ■

The bound of Lemma 5.19 might not be pretty, but it is sufficient in a lot of cases to bound the VC dimension when the shapes involved are simple.

Example 5.20. Consider the range space $\mathsf{S} = \left(\mathbb{R}^2, \mathcal{R}\right)$, where $\mathcal{R}$ is a set of shapes in the plane, so that the boundary of any pair of them intersects at most s times. Then, the VC dimension of S is $O(1)$. Indeed, the dual shattering dimension of S is $O(1)$, since the complexity of the arrangement of n such shapes is $O(sn^2)$. As such, by Lemma 5.19, the VC dimension of S is $O(1)$.

5.2.1.1. *Mixing range spaces.*

Lemma 5.21. *Let $\mathsf{S} = (X, \mathcal{R})$ and $\mathsf{T} = (X, \mathcal{R}')$ be two range spaces of* VC *dimension δ and δ', respectively, where $\delta, \delta' > 1$. Let $\widehat{\mathcal{R}} = \left\{ \mathbf{r} \cup \mathbf{r}' \;\middle|\; \mathbf{r} \in \mathcal{R}, \mathbf{r}' \in \mathcal{R}' \right\}$. Then, for the range space $\widehat{\mathsf{S}} = \left(X, \widehat{\mathcal{R}}\right)$, we have that $\dim_{\mathsf{VC}}\left(\widehat{\mathsf{S}}\right) = O(\delta + \delta')$.*

Proof. As a warm-up exercise, we prove a somewhat weaker bound here of $O((\delta + \delta') \log(\delta + \delta'))$. The stronger bound follows from Theorem 5.22 below. Let B be a set of n points in X that are shattered by $\widehat{\mathsf{S}}$. There are at most $\mathcal{G}_\delta(n)$ and $\mathcal{G}_{\delta'}(n)$ different ranges of B in the range sets $\mathcal{R}_{|B}$ and $\mathcal{R}'_{|B}$, respectively, by Lemma 5.9. Every subset C of B realized by $\widehat{r} \in \widehat{\mathcal{R}}$ is a union of two subsets $B \cap \mathbf{r}$ and $B \cap \mathbf{r}'$, where $\mathbf{r} \in R$ and $\mathbf{r}' \in \mathcal{R}'$, respectively. Thus, the number of different subsets of B realized by $\widehat{\mathsf{S}}$ is bounded by $\mathcal{G}_\delta(n)\mathcal{G}_{\delta'}(n)$. Thus, $2^n \leq n^\delta n^{\delta'}$, for $\delta, \delta' > 1$. We conclude that $n \leq (\delta + \delta') \lg n$, which implies that $n = O((\delta + \delta') \log(\delta + \delta'))$, by Lemma 5.13(C). ■

Interestingly, one can prove a considerably more general result with tighter bounds. The required computations are somewhat more painful.

Theorem 5.22. *Let* $\mathsf{S}_1 = \left(X, \mathcal{R}^1\right), \ldots, \mathsf{S}_k = \left(X, \mathcal{R}^k\right)$ *be range spaces with* VC *dimension* $\delta_1, \ldots, \delta_k$*, respectively. Next, let* $f(\mathbf{r}_1, \ldots, \mathbf{r}_k)$ *be a function that maps any k-tuple of sets* $\mathbf{r}_1 \in \mathcal{R}^1, \ldots, \mathbf{r}_k \in \mathcal{R}^k$ *into a subset of* X*. Consider the range set*

$$\mathcal{R}' = \left\{ f(\mathbf{r}_1, \ldots, \mathbf{r}_k) \;\middle|\; \mathbf{r}_1 \in \mathcal{R}_1, \ldots, \mathbf{r}_k \in \mathcal{R}_k \right\}$$

and the associated range space $\mathsf{T} = (X, \mathcal{R}')$*. Then, the* VC *dimension of* T *is bounded by* $O(k\delta \lg k)$*, where* $\delta = \max_i \delta_i$*.*

Proof. Assume a set $Y \subseteq \mathsf{X}$ of size t is being shattered by $\mathcal{R}'$, and observe that

$$\left|\mathcal{R}'_{|Y}\right| \leq \left| \left\{ (\mathbf{r}_1, \ldots, \mathbf{r}_k) \;\middle|\; \mathbf{r}_1 \in \mathcal{R}^1_{|Y}, \ldots, \mathbf{r}_k \in \mathcal{R}^k_{|Y} \right\} \right| \leq \left|\mathcal{R}^1_{|Y}\right| \cdots \left|\mathcal{R}^k_{|Y}\right| \leq \mathcal{G}_{\delta_1}(t) \cdot \mathcal{G}_{\delta_2}(t) \cdots \mathcal{G}_{\delta_k}(t)$$

$$\leq \left(\mathcal{G}_{\delta}(t)\right)^k \leq \left(2\left(\frac{te}{\delta}\right)^{\delta}\right)^k,$$

by Lemma 5.9 and Lemma 5.10. On the other hand, since Y is being shattered by $\mathcal{R}'$, this implies that $\left|\mathcal{R}'_{|Y}\right| = 2^t$. Thus, we have the inequality $2^t \leq \left(2(te/\delta)^{\delta}\right)^k$, which implies $t \leq k(1 + \delta \lg(te/\delta))$. Assume that $t \geq e$ and $\delta \lg(te/\delta) \geq 1$ since otherwise the claim is trivial, and observe that $t \leq k(1 + \delta \lg(te/\delta)) \leq 3k\delta \lg(t/\delta)$. Setting $x = t/\delta$, we have

$$\frac{t}{\delta} \leq 3k \frac{\ln(t/\delta)}{\ln 2} \leq 6k \ln \frac{t}{\delta} \implies \frac{x}{\ln x} \leq 6k \implies x \leq 2 \cdot 6k \ln(6k) \implies x \leq 12k \ln(6k),$$

by Lemma 5.13(C). We conclude that $t \leq 12\delta k \ln(6k)$, as claimed. ∎

Corollary 5.23. *Let* $\mathsf{S} = (X, \mathcal{R})$ *and* $\mathsf{T} = (X, \mathcal{R}')$ *be two range spaces of* VC *dimension* δ *and* δ'*, respectively, where* $\delta, \delta' > 1$*. Let* $\widehat{\mathcal{R}} = \left\{ \mathbf{r} \cap \mathbf{r}' \;\middle|\; \mathbf{r} \in \mathcal{R}, \mathbf{r}' \in \mathcal{R}' \right\}$*. Then, for the range space* $\widehat{\mathsf{S}} = (X, \widehat{\mathcal{R}})$*, we have that* $\dim_{\mathsf{VC}}(\widehat{\mathsf{S}}) = O(\delta + \delta')$*.*

Corollary 5.24. *Any finite sequence of combining range spaces with finite* VC *dimension (by intersecting, complementing, or taking their union) results in a range space with a finite* VC *dimension.*

5.3. On ε-nets and ε-sampling

5.3.1. ε-nets and ε-samples.

Definition 5.25 (ε-sample)**.** Let $\mathsf{S} = (\mathsf{X}, \mathcal{R})$ be a range space, and let x be a finite subset of X. For $0 \leq \varepsilon \leq 1$, a subset $C \subseteq \mathsf{x}$ is an ***ε-sample*** for x if for any range $\mathbf{r} \in \mathcal{R}$, we have

$$\left| \overline{m}(\mathbf{r}) - \overline{s}(\mathbf{r}) \right| \leq \varepsilon,$$

where $\overline{m}(\mathbf{r}) = |\mathsf{x} \cap \mathbf{r}| \,/\, |\mathsf{x}|$ is the measure of $\mathbf{r}$ (see Definition 5.2) and $\overline{s}(\mathbf{r}) = |C \cap \mathbf{r}| \,/\, |C|$ is the estimate of $\mathbf{r}$ (see Definition 5.3). (Here C might be a multi-set, and as such $|C \cap \mathbf{r}|$ is counted with multiplicity.)

As such, an ε-sample is a subset of the ground set x that "captures" the range space up to an error of ε. Specifically, to estimate the fraction of the ground set covered by a range $\mathbf{r}$, it is sufficient to count the points of C that fall inside $\mathbf{r}$.

If X is a finite set, we will abuse notation slightly and refer to C as an ***ε-sample*** for S.

To see the usage of such a sample, consider $\mathsf{x} = \mathsf{X}$ to be, say, the population of a country (i.e., an element of X is a citizen). A range in $\mathcal{R}$ is the set of all people in the country that answer yes to a question (i.e., would you vote for party Y?, would you buy a bridge from me?, questions like that). An ε-sample of this range space enables us to estimate reliably (up to an error of ε) the answers for all these questions, by just asking the people in the sample.

The natural question of course is how to find such a subset of small (or minimal) size.

Theorem 5.26 (ε-sample theorem, [**VC71**])**.** *There is a positive constant c such that if $(X, \mathcal{R})$ is any range space with VC dimension at most δ, $\mathsf{x} \subseteq X$ is a finite subset and $\varepsilon, \varphi > 0$, then a random subset $C \subseteq \mathsf{x}$ of cardinality*

$$s = \frac{c}{\varepsilon^2}\left(\delta \log \frac{\delta}{\varepsilon} + \log \frac{1}{\varphi}\right)$$

is an ε-sample for x with probability at least $1 - \varphi$.

(In the above theorem, if $s > |\mathsf{x}|$, then we can just take all of x to be the ε-sample.)

For a strengthened version of the above theorem with slightly better bounds, see Theorem 7.13_{p107}.

Sometimes it is sufficient to have (hopefully smaller) samples with a weaker property – if a range is "heavy", then there is an element in our sample that is in this range.

Definition 5.27 (ε-net)**.** A set $N \subseteq \mathsf{x}$ is an ***ε-net*** for x if for any range $\mathbf{r} \in \mathcal{R}$, if $\overline{m}(\mathbf{r}) \geq \varepsilon$ (i.e., $|\mathbf{r} \cap \mathsf{x}| \geq \varepsilon |\mathsf{x}|$), then $\mathbf{r}$ contains at least one point of N (i.e., $\mathbf{r} \cap N \neq \emptyset$).

Theorem 5.28 (ε-net theorem, [**HW87**])**.** *Let $(X, \mathcal{R})$ be a range space of VC dimension δ, let x be a finite subset of X, and suppose that $0 < \varepsilon \leq 1$ and $\varphi < 1$. Let N be a set obtained by m random independent draws from x, where*

$$m \geq \max\left(\frac{4}{\varepsilon} \lg \frac{4}{\varphi}, \frac{8\delta}{\varepsilon} \lg \frac{16}{\varepsilon}\right). \tag{5.3}$$

Then N is an ε-net for x with probability at least $1 - \varphi$.

(We remind the reader that $\lg = \log_2$.)

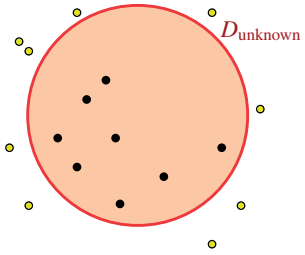
The proofs of the above theorems are somewhat involved and we first turn our attention to some applications before presenting the proofs.

Remark 5.29. The above two theorems also hold for spaces with shattering dimension at most δ, in which case the sample size is slightly larger. Specifically, for Theorem 5.28, the sample size needed is $O\left(\frac{1}{\varepsilon} \lg \frac{1}{\varphi} + \frac{\delta}{\varepsilon} \lg \frac{\delta}{\varepsilon}\right)$.

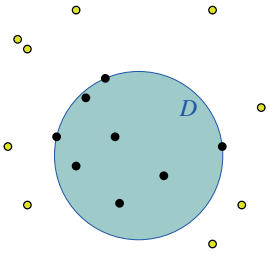
5.3.2. Some applications. We mention two (easy) applications of these theorems, which (hopefully) demonstrate their power.

5.3.2.1. *Range searching.* So, consider a (very large) set of points P in the plane. We would like to be able to quickly decide how many points are included inside a query rectangle. Let us assume that we allow ourselves 1% error. What Theorem 5.26 tells us is that there is a subset of *constant size* (that depends only on ε) that can be used to perform this estimation, and it works for *all* query rectangles (we used here the fact that rectangles in the plane have finite VC dimension). In fact, a random sample of this size works with constant probability.

5.3.2.2. *Learning a concept.* Assume that we have a function f defined in the plane that returns '1' inside an (unknown) disk D_{unknown} and '0' outside it. There is some distribution $\mathcal{D}$ defined over the plane, and we pick points from this distribution. Furthermore, we can compute the function for these labels (i.e., we can compute f for certain values, but it is expensive). For a mystery value $\varepsilon > 0$, to be explained shortly, Theorem 5.28 tells us to pick (roughly) $O((1/\varepsilon)\log(1/\varepsilon))$ random points in a sample R from this distribution and to compute the labels for the samples. This is demonstrated in the figure on the right, where black dots are the sample points for which $f(\cdot)$ returned 1.

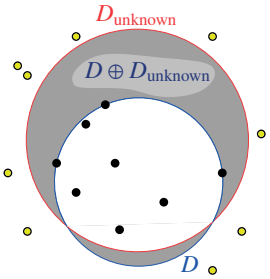


So, now we have positive examples and negative examples. We would like to find a hypothesis that agrees with all the samples we have and that hopefully is close to the true unknown disk underlying the function f. To this end, compute the smallest disk D that contains the sample labeled by '1' and does not contain any of the '0' points, and let $g : \mathbb{R}^2 \to \{0, 1\}$ be the function g that returns '1' inside the disk and '0' otherwise. We claim that g classifies correctly all but an ε-fraction of the points (i.e., the probability of misclassifying a point picked according to the given distribution is smaller than ε); that is, $\mathbf{Pr}_{\mathsf{p}\in\mathcal{D}}\left[f(\mathsf{p}) \neq g(\mathsf{p})\right] \leq \varepsilon$.



Geometrically, the region where g and f disagree is all the points in the symmetric difference between the two disks. That is, $\mathcal{E} = D \oplus D_{\text{unknown}}$; see the figure on the right.

Thus, consider the range space S having the plane as the ground set and the symmetric difference between any two disks as its ranges. By Corollary 5.24, this range space has finite VC dimension. Now, consider the (unknown) disk D' that induces f and the region $\mathbf{r} = D_{\text{unknown}} \oplus D$. Clearly, the learned classifier g returns incorrect answers only for points picked inside $\mathbf{r}$.



Thus, the probability of a mistake in the classification is the measure of $\mathbf{r}$ under the distribution $\mathcal{D}$. So, if $\mathbf{Pr}_{\mathcal{D}}\left[\mathbf{r}\right] > \varepsilon$ (i.e., the probability that a sample point falls inside $\mathbf{r}$), then by the ε-net theorem (i.e., Theorem 5.28) the set R is an ε-net for S (ignore for the time being the possibility that the random sample fails to be an ε-net) and as such, R contains a point q inside $\mathbf{r}$. But, it is not possible for g (which classifies correctly all the sampled points of R) to make a mistake on q, a contradiction, because by construction, the range $\mathbf{r}$ is where g misclassifies points. We conclude that $\mathbf{Pr}_{\mathcal{D}}\left[\mathbf{r}\right] \leq \varepsilon$, as desired.

Little lies. The careful reader might be tearing his or her hair out because of the above description. First, Theorem 5.28 might fail, and the above conclusion might not hold. This is of course true, and in real applications one might use a much larger sample to guarantee that the probability of failure is so small that it can be practically ignored. A more serious issue is that Theorem 5.28 is defined only for finite sets. Nowhere does it speak about a continuous distribution. Intuitively, one can approximate a continuous distribution to an

arbitrary precision using a huge sample and apply the theorem to this sample as our ground set. A formal proof is more tedious and requires extending the proof of Theorem 5.28 to continuous distributions. This is straightforward and we will ignore this topic altogether.

5.3.2.3. *A naive proof of the ε-sample theorem.* To demonstrate why the ε-sample/net theorems are interesting, let us try to prove the ε-sample theorem in the natural naive way. Thus, consider a finite range space $\mathsf{S} = (\mathsf{x}, \mathcal{R})$ with shattering dimension δ. Also, consider a range $\mathbf{r}$ that contains, say, a p fraction of the points of x, where $p \geq \varepsilon$. Consider a random sample R of r points from x, picked with replacement.

Let p_i be the ith sample point, and let X_i be an indicator variable which is one if and only if $\mathsf{p}_i \in \mathbf{r}$. Clearly, $(\sum_i X_i)/r$ is an estimate for $p = |\mathbf{r} \cap \mathsf{x}| / |\mathsf{x}|$. We would like this estimate to be within $\pm\varepsilon$ of p and with confidence $\geq 1 - \varphi$.

As such, the sample failed if $\left|\sum_{i=1}^{r} X_i - pr\right| \geq \varepsilon r = (\varepsilon/p)pr$. Set $\phi = \varepsilon/p$ and $\mu = \mathbf{E}[\sum_i X_i] = pr$. Using Chernoff's inequality (Theorem 27.17_{p340} and Theorem 27.18_{p341}), we have

$$\mathbf{Pr}\left[\left|\sum_{i=1}^{r} X_i - pr\right| \geq (\varepsilon/p)pr\right] = \mathbf{Pr}\left[\left|\sum_{i=1}^{r} X_i - \mu\right| \geq \phi\mu\right] \leq \exp\left(-\mu\phi^2/2\right) + \exp\left(-\mu\phi^2/4\right)$$
$$\leq 2\exp\left(-\mu\phi^2/4\right) = 2\exp\left(-\frac{\varepsilon^2}{4p}r\right) \leq \varphi,$$

for $r \geq \left\lceil \frac{4}{\varepsilon^2} \ln \frac{2}{\varphi} \right\rceil \geq \left\lceil \frac{4p}{\varepsilon^2} \ln \frac{2}{\varphi} \right\rceil$.

Viola! We proved the ε-sample theorem. Well, not quite. We proved that the sample works correctly for a single range. Namely, we proved that for a specific range $\mathbf{r} \in \mathcal{R}$, we have that $\mathbf{Pr}\left[\left| \overline{m}(\mathbf{r}) - \overline{s}(\mathbf{r}) \right| > \varepsilon\right] \leq \varphi$. However, we need to prove that $\forall \mathbf{r} \in \mathcal{R}$, $\mathbf{Pr}\left[\left| \overline{m}(\mathbf{r}) - \overline{s}(\mathbf{r}) \right| > \varepsilon\right] \leq \varphi$.

Now, naively, we can overcome this by using a union bound on the bad probability. Indeed, if there are k different ranges under consideration, then we can use a sample that is large enough such that the probability of it to fail for each range is at most φ/k. In particular, let $\mathcal{E}_i$ be the bad event that the sample fails for the ith range. We have that $\mathbf{Pr}[\mathcal{E}_i] \leq \varphi/k$, which implies that

$$\mathbf{Pr}\left[\text{sample fails for any range}\right] \leq \mathbf{Pr}\left[\bigcup_{i=1}^{k} \mathcal{E}_i\right] \leq \sum_{i=1}^{k} \mathbf{Pr}\left[\mathcal{E}_i\right] \leq k(\varphi/k) \leq \varphi,$$

by the union bound; that is, the sample works for all ranges with good probability.

However, the number of ranges that we need to prove the theorem for is $\pi_{\mathsf{S}}(|\mathsf{x}|)$ (see Definition 5.11). In particular, if we plug in confidence $\varphi/\pi_{\mathsf{S}}(|\mathsf{x}|)$ to the above analysis and use the union bound, we get that for

$$r \geq \left\lceil \frac{4}{\varepsilon^2} \ln \frac{\pi_{\mathsf{S}}(|\mathsf{x}|)}{\varphi} \right\rceil$$

the sample estimates correctly (up to $\pm\varepsilon$) the size of all ranges with confidence $\geq 1 - \varphi$. Bounding $\pi_{\mathsf{S}}(|\mathsf{x}|)$ by $O\left(|\mathsf{x}|^{\delta}\right)$ (using $(5.2)_{\text{p64}}$ for a space with VC dimension δ), we can bound the required size of r by $O\left(\delta\varepsilon^{-2} \log(|\mathsf{x}| / \varphi)\right)$. We summarize the result.

Lemma 5.30. *Let $(\mathsf{x}, \mathcal{R})$ be a finite range space with* VC *dimension at most δ, and let $\varepsilon, \varphi > 0$ be parameters. Then a random subset $C \subseteq \mathsf{x}$ of cardinality $O\left(\delta\varepsilon^{-2} \log(|\mathsf{x}| / \varphi)\right)$ is an ε-sample for x with probability at least $1 - \varphi$.*

Namely, the "naive" argumentation gives us a sample bound which depends on the underlying size of the ground set. However, the sample size in the ε-sample theorem (Theorem 5.26) is independent of the size of the ground set. This is the magical property of the ε-sample theorem[④].

Interestingly, using a chaining argument on Lemma 5.30, one can prove the ε-sample theorem for the finite case; see Exercise 5.3. We provide a similar proof when using discrepancy, in Section 5.4. However, the original proof uses a clever double sampling idea that is both interesting and insightful that makes the proof work for the infinite case also.

5.3.3. A quicky proof of the ε-net theorem (Theorem 5.28). Here we provide a sketchy proof of Theorem 5.28, which conveys the main ideas. The full proof in all its glory and details is provided in Section 5.5.

Let $N = (x_1, \ldots, x_m)$ be the sample obtained by m independent samples from x (observe that N might contain the same element several times, and as such it is a multi-set). Let $\mathcal{E}_1$ be the probability that N fails to be an ε-net. Namely, for $n = |\mathsf{x}|$, let

$$\mathcal{E}_1 = \left\{ \exists \mathbf{r} \in \mathcal{R} \;\middle|\; |\mathbf{r} \cap \mathsf{x}| \geq \varepsilon n \text{ and } \mathbf{r} \cap N = \emptyset \right\}.$$

To complete the proof, we must show that $\mathbf{Pr}[\mathcal{E}_1] \leq \varphi$.

Let $T = (y_1, \ldots, y_m)$ be another random sample generated in a similar fashion to N. It might be that N fails for a certain range $\mathbf{r}$, but then since T is an independent sample, we still expect that $|\mathbf{r} \cap T| = \varepsilon m$. In particular, the probability that $\mathbf{Pr}\left[|\mathbf{r} \cap T| \geq \frac{\varepsilon m}{2}\right]$ is a large constant close to 1, regardless of how N performs. Indeed, if m is sufficiently large, we expect the random variable $|\mathbf{r} \cap T|$ to concentrate around εm, and one can argue this formally using Chernoff's inequality. Namely, intuitively, for a heavy range $\mathbf{r}$ we have that

$$\mathbf{Pr}[\mathbf{r} \cap N = \emptyset] \approx \mathbf{Pr}\left[\mathbf{r} \cap N = \emptyset \text{ and } \left(|\mathbf{r} \cap T| \geq \frac{\varepsilon m}{2}\right)\right].$$

Inspired by this, let $\mathcal{E}_2$ be the event that N fails for some range $\mathbf{r}$ but T "works" for $\mathbf{r}$; formally

$$\mathcal{E}_2 = \left\{ \exists \mathbf{r} \in \mathcal{R} \;\middle|\; |\mathbf{r} \cap \mathsf{x}| \geq \varepsilon n,\ \mathbf{r} \cap N = \emptyset \text{ and } |\mathbf{r} \cap T| \geq \frac{\varepsilon m}{2} \right\}.$$

Intuitively, since $\mathbf{E}[|\mathbf{r} \cap T|] \geq \varepsilon m$, then for the range $\mathbf{r}$ that N fails for, we have with "good" probability that $|\mathbf{r} \cap T| \geq \varepsilon m/2$. Namely, $\mathbf{Pr}[\mathcal{E}_1] \approx \mathbf{Pr}[\mathcal{E}_2]$.

Next, let

$$\mathcal{E}_2' = \left\{ \exists \mathbf{r} \in R \;\middle|\; \mathbf{r} \cap N = \emptyset \text{ and } |\mathbf{r} \cap T| \geq \frac{\varepsilon m}{2} \right\}.$$

Clearly, $\mathcal{E}_2 \subseteq \mathcal{E}_2'$ and as such $\mathbf{Pr}[\mathcal{E}_2] \leq \mathbf{Pr}\left[\mathcal{E}_2'\right]$. Now, fix $Z = N \cup T$, and observe that $|Z| = 2m$. Next, fix a range $\mathbf{r}$, and observe that the bad probability of $\mathcal{E}_2'$ is maximized if $|\mathbf{r} \cap Z| = \varepsilon m/2$. Now, the probability that all the elements of $\mathbf{r} \cap Z$ fall only into the second half of the sample is at most $2^{-\varepsilon m/2}$ as a careful calculation shows. Now, there are at most $\left|Z_{|R}\right| \leq \mathcal{G}_d(2m)$ different ranges that one has to consider. As such, $\mathbf{Pr}[\mathcal{E}_1] \approx \mathbf{Pr}[\mathcal{E}_2] \leq \mathbf{Pr}\left[\mathcal{E}_2'\right] \leq \mathcal{G}_d(2m)2^{-\varepsilon m/2}$ and this is smaller than φ, as a careful calculation shows by just plugging the value of m into the right-hand side; see $(5.3)_{\text{p71}}$. ■

[④]The notion of magic is used here in the sense of Arthur C. Clarke's statement that "any sufficiently advanced technology is indistinguishable from magic."

5.4. Discrepancy

The proof of the ε-sample/net theorem is somewhat complicated. It turns out that one can get a somewhat similar result by attacking the problem from the other direction; namely, let us assume that we would like to take a truly large sample of a finite range space $\mathsf{S} = (\mathsf{X}, \mathcal{R})$ defined over n elements with m ranges. We would like this sample to be as representative as possible as far as S is concerned. In fact, let us decide that we would like to pick exactly half of the points of X in our sample (assume that $n = |\mathsf{X}|$ is even).

To this end, let us color half of the points of X by -1 (i.e., black) and the other half by 1 (i.e., white). If for every range, $\mathbf{r} \in \mathcal{R}$, the number of black points inside it is equal to the number of white points, then doubling the number of black points inside a range gives us the exact number of points inside the range. Of course, such a perfect coloring is unachievable in almost all situations. To see this, consider the complete graph K_3 – clearly, in any coloring (by two colors) of its vertices, there must be an edge with two endpoints having the same color (i.e., the edges are the ranges).

Formally, let $\chi : \mathsf{X} \to \{-1, 1\}$ be a coloring. The ***discrepancy*** of χ over a range $\mathbf{r}$ is the amount of imbalance in the coloring inside χ. Namely,

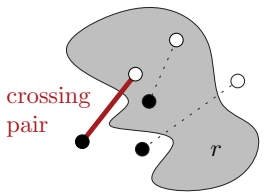
$$|\chi(\mathbf{r})| = \left| \sum_{\mathsf{p} \in \mathbf{r}} \chi(\mathsf{p}) \right|.$$

The overall ***discrepancy*** of χ is $\operatorname{disc}(\chi) = \max_{\mathbf{r} \in \mathcal{R}} |\chi(\mathbf{r})|$. The ***discrepancy*** of a (finite) range space $\mathsf{S} = (\mathsf{X}, \mathcal{R})$ is the discrepancy of the best possible coloring; namely,

$$\operatorname{disc}(\mathsf{S}) = \min_{\chi : \mathsf{X} \to \{-1, +1\}} \operatorname{disc}(\chi).$$

The natural question is, of course, how to compute the coloring χ of minimum discrepancy. This seems like a very challenging question, but when you do not know what to do, you might as well do something random. So, let us pick a random coloring χ of X. To this end, let Π be an arbitrary partition of X into pairs (i.e., a perfect matching). For a pair $\{\mathsf{p}, \mathsf{q}\} \in \Pi$, we will either color $\chi(\mathsf{p}) = -1$ and $\chi(\mathsf{q}) = 1$ or the other way around; namely, $\chi(\mathsf{p}) = 1$ and $\chi(\mathsf{q}) = -1$. We will decide how to color this pair using a single coin flip. Thus, our coloring would be induced by making such a decision for every pair of Π, and let χ be the resulting coloring. We will refer to χ as ***compatible*** with the partition Π if, for all $\{\mathsf{p}, \mathsf{q}\} \in \Pi$, we have that $\chi(\{\mathsf{p}, \mathsf{q}\}) = 0$; namely,

$$\forall \{\mathsf{p}, \mathsf{q}\} \in \Pi \quad \Big(\chi(\mathsf{p}) = +1 \text{ and } \chi(\mathsf{q}) = -1\Big) \\ \text{or} \quad \Big(\chi(\mathsf{p}) = -1 \text{ and } \chi(\mathsf{q}) = +1\Big).$$



.

Consider a range $\mathbf{r}$ and a coloring χ compatible with Π. If a pair $\{\mathsf{p}, \mathsf{q}\} \in \Pi$ falls completely inside $\mathbf{r}$ or completely outside $\mathbf{r}$, then it does not contribute anything to the discrepancy of $\mathbf{r}$. Thus, the only pairs that contribute to the discrepancy of $\mathbf{r}$ are the ones that ***cross*** it. Namely, $\{\mathsf{p}, \mathsf{q}\} \cap \mathbf{r} \neq \emptyset$ and $\{\mathsf{p}, \mathsf{q}\} \cap (\mathsf{X} \setminus \mathbf{r}) \neq \emptyset$.

As such, let $\#_{\mathbf{r}}$ denote the ***crossing number*** of $\mathbf{r}$, that is, the number of pairs that cross $\mathbf{r}$. Next, let $X_i \in \{-1, +1\}$ be the indicator variable which is the contribution of the ith crossing pair to the discrepancy of $\mathbf{r}$. For $\Delta_{\mathbf{r}} = \sqrt{2\#_{\mathbf{r}} \ln(4m)}$, we have by Chernoff's

inequality (Theorem 27.13$_{\text{p338}}$), that

$$\mathbf{Pr}\Big[\,|\chi(\mathbf{r})| \geq \Delta_{\mathbf{r}}\Big] = \mathbf{Pr}\Big[\chi(\mathbf{r}) \geq \Delta_{\mathbf{r}}\Big] + \mathbf{Pr}\Big[\chi(\mathbf{r}) \leq -\Delta_{\mathbf{r}}\Big] = 2\,\mathbf{Pr}\left[\sum_i X_i \geq \Delta_{\mathbf{r}}\right]$$
$$\leq 2\exp\left(-\frac{\Delta_{\mathbf{r}}^2}{2\#_{\mathbf{r}}}\right) = \frac{1}{2m}.$$

Since there are m ranges in $\mathcal{R}$, it follows that with good probability (i.e., at least half) for all $\mathbf{r} \in \mathcal{R}$ the discrepancy of $\mathbf{r}$ is at most $\Delta_{\mathbf{r}}$.

Theorem 5.31. *Let $\mathsf{S} = (\mathsf{X}, \mathcal{R})$ be a range space defined over $n = |\mathsf{X}|$ elements with $m = |\mathcal{R}|$ ranges. Consider any partition Π of the elements of X into pairs. Then, with probability $\geq 1/2$, for any range $\mathbf{r} \in \mathcal{R}$, a random coloring $\chi : \mathsf{X} \to \{-1, +1\}$ that is compatible with the partition Π has discrepancy at most*

$$|\chi(\mathbf{r})| < \Delta_{\mathbf{r}} = \sqrt{2\#_{\mathbf{r}} \ln(4m)},$$

where $\#_{\mathbf{r}}$ denotes the number of pairs of Π that cross $\mathbf{r}$. In particular, since $\#_{\mathbf{r}} \leq |\mathbf{r}|$, we have $|\chi(\mathbf{r})| \leq \sqrt{2\,|\mathbf{r}| \ln(4m)}$.

Observe that for every range $\mathbf{r}$ we have that $\#_{\mathbf{r}} \leq n/2$, since $2\#_{\mathbf{r}} \leq |\mathsf{X}|$. As such, we have:

Corollary 5.32. *Let $\mathsf{S} = (\mathsf{X}, \mathcal{R})$ be a range space defined over n elements with m ranges. Let Π be an arbitrary partition of X into pairs. Then a random coloring which is compatible with Π has $\operatorname{disc}(\chi) < \sqrt{n \ln(4m)}$, with probability $\geq 1/2$.*

One can easily amplify the probability of success of the coloring by increasing the threshold. In particular, for any constant $c \geq 1$, one has that

$$\forall \mathbf{r} \in \mathcal{R} \quad |\chi(\mathbf{r})| \leq \sqrt{2c\,\#_{\mathbf{r}} \ln(4m)},$$

with probability $\geq 1 - \dfrac{2}{(4m)^c}$.

5.4.1. Building ε-sample via discrepancy. Let $\mathsf{S} = (\mathsf{X}, \mathcal{R})$ be a range space with shattering dimension δ. Let $\mathsf{P} \subseteq \mathsf{X}$ be a set of n points, and consider the induced range space $\mathsf{S}_{|\mathsf{P}} = (\mathsf{P}, \mathcal{R}_{|\mathsf{P}})$; see Definition 5.4$_{\text{p62}}$. Here, by the definition of shattering dimension, we have that $m = \left|\mathcal{R}_{|\mathsf{P}}\right| = O\left(n^{\delta}\right)$. Without loss of generality, we assume that n is a power of 2. Consider a coloring χ of P with discrepancy bounded by Corollary 5.32. In particular, let Q be the points of P colored by, say, -1. We know that $|\mathsf{Q}| = n/2$, and for any range $\mathbf{r} \in \mathcal{R}$, we have that

$$\chi(\mathbf{r}) = \Big|\,|(\mathsf{P} \setminus \mathsf{Q}) \cap \mathbf{r}| - |\mathsf{Q} \cap \mathbf{r}|\Big| < \sqrt{n \ln(4m)} = \sqrt{n \ln O(n^{\delta})} \leq c\sqrt{n \ln(n^{\delta})},$$

for some absolute constant c. Observe that $|(\mathsf{P} \setminus \mathsf{Q}) \cap \mathbf{r}| = |\mathsf{P} \cap \mathbf{r}| - |\mathsf{Q} \cap \mathbf{r}|$. In particular, we have that for any range $\mathbf{r}$,

$$\Big|\,|\mathsf{P} \cap \mathbf{r}| - 2\,|\mathsf{Q} \cap \mathbf{r}|\Big| \leq c\sqrt{n \ln(n^{\delta})}. \tag{5.4}$$

Dividing both sides by $n = |\mathsf{P}| = 2\,|\mathsf{Q}|$, we have that

$$\left|\frac{|\mathsf{P} \cap \mathbf{r}|}{|\mathsf{P}|} - \frac{|\mathsf{Q} \cap \mathbf{r}|}{|\mathsf{Q}|}\right| \leq \tau(n) \;\text{ for }\; \tau(n) = c\sqrt{\frac{\delta \ln n}{n}}. \tag{5.5}$$

Namely, a coloring with discrepancy bounded by Corollary 5.32 yields a $\tau(n)$-sample. Intuitively, if n is very large, then Q provides a good approximation to P. However, we

want an ε-sample for a prespecified $\varepsilon > 0$. Conceptually, ε is a fixed constant while $\tau(n)$ is considerably smaller. Namely, Q is a sample which is too tight for our purposes (and thus too big). As such, we will coarsen (and shrink) Q till we get the desired ε-sample by repeated application of Corollary 5.32. Specifically, we can "chain" together several approximations generated by Corollary 5.32. This is sometime refereed to as the ***sketch*** property of samples. Informally, as testified by the following lemma, a sketch of a sketch is a sketch[5].

Lemma 5.33. *Let $\mathsf{Q} \subseteq \mathsf{P}$ be a ρ-sample for P (in some underlying range space S), and let $\mathsf{R} \subseteq \mathsf{Q}$ be a ρ'-sample for Q. Then R is a $(\rho + \rho')$-sample for P.*

Proof. By definition, we have that, for every $\mathbf{r} \in \mathcal{R}$,

$$\left|\frac{|\mathbf{r} \cap \mathsf{P}|}{|\mathsf{P}|} - \frac{|\mathbf{r} \cap \mathsf{Q}|}{|\mathsf{Q}|}\right| \leq \rho \quad \text{and} \quad \left|\frac{|\mathbf{r} \cap \mathsf{Q}|}{|\mathsf{Q}|} - \frac{|\mathbf{r} \cap \mathsf{R}|}{|\mathsf{R}|}\right| \leq \rho'.$$

By adding the two inequalities together, we get

$$\left|\frac{|\mathbf{r} \cap \mathsf{P}|}{|\mathsf{P}|} - \frac{|\mathbf{r} \cap \mathsf{R}|}{|\mathsf{R}|}\right| = \left|\frac{|\mathbf{r} \cap \mathsf{P}|}{|\mathsf{P}|} - \frac{|\mathbf{r} \cap \mathsf{Q}|}{|\mathsf{Q}|} + \frac{|\mathbf{r} \cap \mathsf{Q}|}{|\mathsf{Q}|} - \frac{|\mathbf{r} \cap \mathsf{R}|}{|\mathsf{R}|}\right| \leq \rho + \rho'. \qquad \blacksquare$$

Thus, let $\mathsf{P}_0 = \mathsf{P}$ and $\mathsf{P}_1 = \mathsf{Q}$. Now, in the ith iteration, we will compute a coloring χ_{i-1} of P_{i-1} with low discrepancy, as guaranteed by Corollary 5.32, and let P_i be the points of P_{i-1} colored white by χ_{i-1}. Let $\delta_i = \tau(n_{i-1})$, where $n_{i-1} = |\mathsf{P}_{i-1}| = n/2^{i-1}$. By Lemma 5.33, we have that P_k is a $(\sum_{i=1}^{k} \delta_i)$-sample for P. Since we would like the smallest set in the sequence $\mathsf{P}_1, \mathsf{P}_2, \ldots$ that is still an ε-sample, we would like to find the maximal k, such that $(\sum_{i=1}^{k} \delta_i) \leq \varepsilon$. Plugging in the value of δ_i and $\tau(\cdot)$, see (5.5), it is sufficient for our purposes that

$$\sum_{i=1}^{k} \delta_i = \sum_{i=1}^{k} \tau(n_{i-1}) = \sum_{i=1}^{k} c\sqrt{\frac{\delta \ln(n/2^{i-1})}{n/2^{i-1}}} \leq c_1 \sqrt{\frac{\delta \ln(n/2^{k-1})}{n/2^{k-1}}} = c_1 \sqrt{\frac{\delta \ln n_{k-1}}{n_{k-1}}} \leq \varepsilon,$$

since the above series behaves like a geometric series, and as such its total sum is proportional to its largest element[6], where c_1 is a sufficiently large constant. This holds for

$$c_1 \sqrt{\frac{\delta \ln n_{k-1}}{n_{k-1}}} \leq \varepsilon \iff c_1^2 \frac{\delta \ln n_{k-1}}{n_{k-1}} \leq \varepsilon^2 \iff \frac{c_1^2 \delta}{\varepsilon^2} \leq \frac{n_{k-1}}{\ln n_{k-1}}.$$

The last inequality holds for $n_{k-1} \geq 2\frac{c_1^2\delta}{\varepsilon^2} \ln \frac{c_1^2\delta}{\varepsilon^2}$, by Lemma 5.13(D). In particular, taking the largest k for which this holds results in a set P_k of size $O\Big((\delta/\varepsilon^2)\ln(\delta/\varepsilon)\Big)$ which is an ε-sample for P.

Theorem 5.34 (ε-sample via discrepancy)**.** *For a range space $(X, \mathcal{R})$ with shattering dimension at most δ and $B \subseteq X$ a finite subset and $\varepsilon > 0$, there exists a subset $C \subseteq B$, of cardinality $O\Big((\delta/\varepsilon^2)\ln(\delta/\varepsilon)\Big)$, such that C is an ε-sample for B.*

Note that it is not obvious how to turn Theorem 5.34 into an efficient construction algorithm of such an ε-sample. Nevertheless, this theorem can be turned into a relatively efficient deterministic algorithm using conditional probabilities. In particular, there is a

[5] Try saying this quickly 100 times.

[6] Formally, one needs to show that the ratio between two consecutive elements in the series is larger than some constant, say 1.1. This is easy but tedious, but the well-motivated reader (of little faith) might want to do this calculation.

deterministic $O\left(n^{\delta+1}\right)$ time algorithm for computing an ε-sample for a range space of VC dimension δ and with n points in its ground set using the above approach (see the bibliographical notes in Section 5.7 for details). Inherently, however, it is a far cry from the simplicity of Theorem 5.26 that just requires us to take a random sample. Interestingly, there are cases where using discrepancy leads to smaller ε-samples; again see bibliographical notes for details.

5.4.1.1. *Faster deterministic construction of ε-samples.* One can speed up the deterministic construction mentioned above by using a sketch-and-merge approach. To this end, we need the following ***merge*** property of ε-samples. (The proof of the following lemma is quite easy. Nevertheless, we provide the proof in excruciating detail for the sake of completeness.)

Lemma 5.35. *Consider the sets $\mathsf{R} \subseteq \mathsf{P}$ and $\mathsf{R}' \subseteq \mathsf{P}'$. Assume that P and P' are disjoint, $|\mathsf{P}| = |\mathsf{P}'|$, and $|\mathsf{R}| = |\mathsf{R}'|$. Then, if R is an ε-sample of P and R' is an ε-sample of P', then $\mathsf{R} \cup \mathsf{R}'$ is an ε-sample of $\mathsf{P} \cup \mathsf{P}'$.*

Proof. We have for any range $\mathbf{r}$ that

$$\begin{aligned}
\left|\frac{|\mathbf{r}\cap(\mathsf{P}\cup\mathsf{P}')|}{|\mathsf{P}\cup\mathsf{P}'|} - \frac{|\mathbf{r}\cap(\mathsf{R}\cup\mathsf{R}')|}{|\mathsf{R}\cup\mathsf{R}'|}\right| &= \left|\frac{|\mathbf{r}\cap\mathsf{P}|}{|\mathsf{P}\cup\mathsf{P}'|} + \frac{|\mathbf{r}\cap\mathsf{P}'|}{|\mathsf{P}\cup\mathsf{P}'|} - \frac{|\mathbf{r}\cap\mathsf{R}|}{|\mathsf{R}\cup\mathsf{R}'|} - \frac{|\mathbf{r}\cap\mathsf{R}'|}{|\mathsf{R}\cup\mathsf{R}'|}\right| \\
&= \left|\frac{|\mathbf{r}\cap\mathsf{P}|}{2\,|\mathsf{P}|} + \frac{|\mathbf{r}\cap\mathsf{P}'|}{2\,|\mathsf{P}'|} - \frac{|\mathbf{r}\cap\mathsf{R}|}{2\,|\mathsf{R}|} - \frac{|\mathbf{r}\cap\mathsf{R}'|}{2\,|\mathsf{R}'|}\right| \\
&= \frac{1}{2}\left|\left(\frac{|\mathbf{r}\cap\mathsf{P}|}{|\mathsf{P}|} - \frac{|\mathbf{r}\cap\mathsf{R}|}{|\mathsf{R}|}\right) + \left(\frac{|\mathbf{r}\cap\mathsf{P}'|}{|\mathsf{P}'|} - \frac{|\mathbf{r}\cap\mathsf{R}'|}{|\mathsf{R}'|}\right)\right| \\
&\le \frac{1}{2}\left|\frac{|\mathbf{r}\cap\mathsf{P}|}{|\mathsf{P}|} - \frac{|\mathbf{r}\cap\mathsf{R}|}{|\mathsf{R}|}\right| + \frac{1}{2}\left|\frac{|\mathbf{r}\cap\mathsf{P}'|}{|\mathsf{P}'|} - \frac{|\mathbf{r}\cap\mathsf{R}'|}{|\mathsf{R}'|}\right| \\
&\le \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.
\end{aligned}$$

■

Interestingly, by breaking the given ground sets into sets of equal size and building a balanced binary tree over these sets, one can speed up the deterministic algorithm for building ε-samples. The idea is to compute the sample bottom-up, where at every node we merge the samples provided by the children (i.e., using Lemma 5.35), and then we sketch the resulting set using Lemma 5.33. By carefully fine-tuning this construction, one can get an algorithm for computing ε-samples in time which is near linear in n (assuming ε and δ are small constants). We delegate the details of this construction to Exercise 5.6.

This algorithmic idea is quite useful and we will refer to it as ***sketch-and-merge***.

5.4.2. Building ε-net via discrepancy. We are given range space $(\mathsf{X}, \mathcal{R})$ with shattering dimension d and $\varepsilon > 0$ and the target is to compute an ε-net for this range space.

We need to be slightly more careful if we want to use discrepancy to build ε-nets, and we will use Theorem 5.31 instead of Corollary 5.32 in the analysis.

The construction is as before – we set $\mathsf{P}_0 = \mathsf{P}$, and P_i is all the points colored +1 in the coloring of P_{i-1} by Theorem 5.31. We repeat this till we get a set that is the required net.

To analyze this construction (and decide when it should stop), let $\mathbf{r}$ be a range in a given range space $(\mathsf{X}, \mathcal{R})$ with shattering dimension d, and let

$$\nu_i = |\mathsf{P}_i \cap \mathbf{r}|$$

denote the size of the range $\mathbf{r}$ in the ith set P_i and let $n_i = |\mathsf{P}_i|$, for $i \geq 0$. Observer that the number of points in $\mathbf{r}$ colored by $+1$ and -1 when coloring P_{i-1} is

$$\alpha_i = |\mathsf{P}_i \cap \mathbf{r}| = \nu_i \quad \text{and} \quad \beta_i = |\mathsf{P}_{i-1} \cap \mathbf{r}| - |\mathsf{P}_i \cap \mathbf{r}| = \nu_{i-1} - \nu_i,$$

respectively. As such, setting $m_i = \left|\mathcal{R}_{|\mathsf{P}_i}\right| = O\left(n_i^d\right)$, we have, by Theorem 5.31, that the discrepancy of $\mathbf{r}$ in this coloring of P_{i-1} is

$$|\alpha_i - \beta_i| = |\nu_i - 2\nu_{i-1}| \leq \sqrt{2\nu_{i-1} \ln 4m_{i-1}} \leq c\sqrt{d\nu_{i-1} \ln n_{i-1}}$$

for some constant c, since the crossing number $\#_{\mathbf{r}}$ of a range $\mathbf{r} \cap \mathsf{P}_{i-1}$ is always bounded by its size. This is equivalent to

$$\left|2^{i-1}\nu_{i-1} - 2^i\nu_i\right| \leq c2^{i-1}\sqrt{d\nu_{i-1} \ln n_{i-1}}. \tag{5.6}$$

We need the following technical claim that states that the size of ν_k behaves as we expect; as long as the set P_k is large enough, the size of ν_k is roughly $\nu_0/2^k$.

Claim 5.36. *There is a constant c_4 (independent of d), such that for all k with $\nu_0/2^k \geq c_4 d \ln n_k$, $(\nu_0/2^k)/2 \leq \nu_k \leq 2(\nu_0/2^k)$.*

Proof. The proof is by induction. For $k = 0$ the claim trivially holds. Assume that it holds for $i < k$. Adding up the inequalities of (5.6), for $i = 1, \ldots, k$, we have that

$$\left|\nu_0 - 2^k\nu_k\right| \leq \sum_{i=1}^{k} c2^{i-1}\sqrt{d\nu_{i-1} \ln n_{i-1}} \leq \sum_{i=1}^{k} c2^{i-1}\sqrt{2d\frac{\nu_0}{2^{i-1}} \ln n_{i-1}} \leq c_3\, 2^k\sqrt{d\frac{\nu_0}{2^k} \ln n_k},$$

for some constant c_3 since this summation behaves like an increasing geometric series and the last term dominates the summation. Thus,

$$\frac{\nu_0}{2^k} - c_3\sqrt{d\frac{\nu_0}{2^k} \ln n_k} \;\leq\; \nu_k \;\leq\; \frac{\nu_0}{2^k} + c_3\sqrt{d\frac{\nu_0}{2^k} \ln n_k}.$$

By assumption, we have that $\sqrt{\frac{\nu_0}{c_4 2^k}} \geq \sqrt{d \ln n_k}$. This implies that

$$\nu_k \leq \frac{\nu_0}{2^k} + c_3\sqrt{\frac{\nu_0}{2^k} \cdot \frac{\nu_0}{c_4 2^k}} = \frac{\nu_0}{2^k}\left(1 + \frac{c_3}{\sqrt{c_4}}\right) \leq 2\frac{\nu_0}{2^k},$$

by selecting $c_4 \geq 4c_3^2$. Similarly, we have

$$\nu_k \geq \frac{\nu_0}{2^k}\left(1 - \frac{c_3\sqrt{d \ln n_k}}{\sqrt{\nu_0/2^k}}\right) \geq \frac{\nu_0}{2^k}\left(1 - \frac{c_3\sqrt{\nu_0/c_4 2^k}}{\sqrt{\nu_0/2^k}}\right) = \frac{\nu_0}{2^k}\left(1 - \frac{c_3}{\sqrt{c_4}}\right) \geq \frac{\nu_0}{2^k}/2.$$

■

So consider a "heavy" range $\mathbf{r}$ that contains at least $\nu_0 \geq \varepsilon n$ points of P. To show that P_k is an ε-net, we need to show that $\mathsf{P}_k \cap \mathbf{r} \neq \emptyset$. To apply Claim 5.36, we need a k such that $\varepsilon n/2^k \geq c_4 d \ln n_{k-1}$, or equivalently, such that

$$\frac{2n_k}{\ln(2n_k)} \geq \frac{2c_4 d}{\varepsilon},$$

which holds for $n_k = \Omega\left(\frac{d}{\varepsilon} \ln \frac{d}{\varepsilon}\right)$, by Lemma 5.13(D). But then, by Claim 5.36, we have that

$$\nu_k = |\mathsf{P}_k \cap \mathbf{r}| \geq \frac{|\mathsf{P} \cap \mathbf{r}|}{2 \cdot 2^k} \geq \frac{1}{2} \cdot \frac{\varepsilon n}{2^k} = \frac{\varepsilon}{2} n_k = \Omega\left(d \ln \frac{d}{\varepsilon}\right) > 0.$$

We conclude that the set P_k, which is of size $\Omega\left(\frac{d}{\varepsilon} \ln \frac{d}{\varepsilon}\right)$, is an ε-net for P.

Theorem 5.37 (ε-net via discrepancy)**.** *For any range space $(\mathsf{X}, \mathcal{R})$ with shattering dimension at most d, a finite subset $B \subseteq \mathsf{X}$, and $\varepsilon > 0$, there exists a subset $C \subseteq B$, of cardinality $O((d/\varepsilon)\ln(d/\varepsilon))$, such that C is an ε-net for B.*

5.5. Proof of the ε-net theorem

In this section, we finally prove Theorem 5.28.

Let $(\mathsf{X}, \mathcal{R})$ be a range space of VC dimension δ, and let x be a subset of X of cardinality n. Suppose that m satisfies $(5.3)_{\text{p71}}$. Let $N = (x_1, \ldots, x_m)$ be the sample obtained by m independent samples from x (the elements of N are not necessarily distinct, and we treat N as an ordered set). Let $\mathcal{E}_1$ be the probability that N fails to be an ε-net. Namely,

$$\mathcal{E}_1 = \left\{ \exists \mathbf{r} \in \mathcal{R} \;\middle|\; |\mathbf{r} \cap \mathsf{x}| \geq \varepsilon n \text{ and } \mathbf{r} \cap N = \emptyset \right\}.$$

(Namely, there exists a "heavy" range $\mathbf{r}$ that does not contain any point of N.) To complete the proof, we must show that $\mathbf{Pr}[\mathcal{E}_1] \leq \varphi$. Let $T = (y_1, \ldots, y_m)$ be another random sample generated in a similar fashion to N. Let $\mathcal{E}_2$ be the event that N fails but T "works"; formally

$$\mathcal{E}_2 = \left\{ \exists \mathbf{r} \in R \;\middle|\; |\mathbf{r} \cap \mathsf{x}| \geq \varepsilon n,\ \mathbf{r} \cap N = \emptyset,\ \text{ and } |\mathbf{r} \cap T| \geq \frac{\varepsilon m}{2} \right\}.$$

Intuitively, since $\mathbf{E}[|\mathbf{r} \cap T|] \geq \varepsilon m$, we have that for the range $\mathbf{r}$ that N fails for, it follows with "good" probability that $|\mathbf{r} \cap T| \geq \varepsilon m/2$. Namely, $\mathcal{E}_1$ and $\mathcal{E}_2$ have more or less the same probability.

Claim 5.38. $\mathbf{Pr}[\mathcal{E}_2] \leq \mathbf{Pr}[\mathcal{E}_1] \leq 2\,\mathbf{Pr}[\mathcal{E}_2]$.

Proof. Clearly, $\mathcal{E}_2 \subseteq \mathcal{E}_1$, and thus $\mathbf{Pr}[\mathcal{E}_2] \leq \mathbf{Pr}[\mathcal{E}_1]$. As for the other part, note that by the definition of conditional probability, we have

$$\mathbf{Pr}\left[\mathcal{E}_2 \;\middle|\; \mathcal{E}_1\right] = \mathbf{Pr}[\mathcal{E}_2 \cap \mathcal{E}_1] \,/\, \mathbf{Pr}[\mathcal{E}_1] = \mathbf{Pr}[\mathcal{E}_2] \,/\, \mathbf{Pr}[\mathcal{E}_1].$$

It is thus enough to show that $\mathbf{Pr}\left[\mathcal{E}_2 \;\middle|\; \mathcal{E}_1\right] \geq 1/2$.

Assume that $\mathcal{E}_1$ occurs. There is $\mathbf{r} \in R$, such that $|\mathbf{r} \cap \mathsf{x}| > \varepsilon n$ and $\mathbf{r} \cap N = \emptyset$. The required probability is at least the probability that for this specific $\mathbf{r}$, we have $|\mathbf{r} \cap T| \geq \frac{\varepsilon n}{2}$. However, $X = |\mathbf{r} \cap T|$ is a binomial variable with expectation $\mathbf{E}[X] = pm$, and variance $\mathbf{V}[X] = p(1-p)m \leq pm$, where $p = |\mathbf{r} \cap \mathsf{x}|/n \geq \varepsilon$. Thus, by Chebychev's inequality (Theorem 27.3_{p335}),

$$\begin{aligned}
\mathbf{Pr}\left[X < \frac{\varepsilon m}{2}\right] &\leq \mathbf{Pr}\left[X < \frac{pm}{2}\right] \leq \mathbf{Pr}\left[|X - pm| > \frac{pm}{2}\right] \\
&= \mathbf{Pr}\left[|X - pm| > \frac{\sqrt{pm}}{2}\sqrt{pm}\right] \leq \mathbf{Pr}\left[|X - \mathbf{E}[X]| > \frac{\sqrt{pm}}{2}\sqrt{\mathbf{V}[X]}\right] \\
&\leq \left(\frac{2}{\sqrt{pm}}\right)^2 \leq \frac{1}{2},
\end{aligned}$$

since $m \geq 8/\varepsilon \geq 8/p$; see $(5.3)_{\text{p71}}$. Thus, for $\mathbf{r} \in \mathcal{E}_1$, we have

$$\frac{\mathbf{Pr}[\mathcal{E}_2]}{\mathbf{Pr}[\mathcal{E}_1]} \geq \mathbf{Pr}\left[|\mathbf{r} \cap T| \geq \tfrac{\varepsilon m}{2}\right] = 1 - \mathbf{Pr}\left[|\mathbf{r} \cap T| < \frac{\varepsilon m}{2}\right] \geq \frac{1}{2}. \qquad \blacksquare$$

Claim 5.38 implies that to bound the probability of $\mathcal{E}_1$, it is enough to bound the probability of $\mathcal{E}_2$. Let

$$\mathcal{E}_2' = \left\{ \exists \mathbf{r} \in R \;\middle|\; \mathbf{r} \cap N = \emptyset, |\mathbf{r} \cap T| \geq \frac{\varepsilon m}{2} \right\}.$$

Clearly, $\mathcal{E}_2 \subseteq \mathcal{E}'_2$. Thus, bounding the probability of $\mathcal{E}'_2$ is enough to prove Theorem 5.28. Note, however, that a shocking thing happened! We no longer have x participating in our event. Namely, we turned bounding an event that depends on a global quantity (i.e., the ground set x) into bounding a quantity that depends only on a local quantity/experiment (involving only N and T). This is the crucial idea in this proof.

Claim 5.39. $\mathbf{Pr}\Big[\mathcal{E}_2\Big] \leq \mathbf{Pr}\Big[\mathcal{E}'_2\Big] \leq \mathcal{G}_d(2m)2^{-\varepsilon m/2}$.

Proof. We imagine that we sample the elements of $N \cup T$ together, by picking $Z = (z_1, \ldots, z_{2m})$ independently from x. Next, we randomly decide the m elements of Z that go into N, and the remaining elements go into T. Clearly,

$$\begin{aligned}\mathbf{Pr}\Big[\mathcal{E}'_2\Big] &= \sum_{z \in \mathsf{X}^{2m}} \mathbf{Pr}\Big[\mathcal{E}'_2 \cap (Z = z)\Big] = \sum_{z \in \mathsf{X}^{2m}} \frac{\mathbf{Pr}\Big[\mathcal{E}'_2 \cap (Z = z)\Big]}{\mathbf{Pr}[Z = z]} \cdot \mathbf{Pr}\Big[Z = z\Big] \\ &= \sum_{z} \mathbf{Pr}\Big[\mathcal{E}'_2 \;\Big|\; Z = z\Big]\mathbf{Pr}\Big[Z = z\Big] = \mathbf{E}\Big[\mathbf{Pr}\Big[\mathcal{E}'_2 \;\Big|\; Z = z\Big]\Big].\end{aligned}$$

Thus, from this point on, we fix the set Z, and we bound $\mathbf{Pr}\Big[\mathcal{E}'_2 \;\Big|\; Z\Big]$. Note that $\mathbf{Pr}[\mathcal{E}'_2]$ is a weighted average of $\mathbf{Pr}[\mathcal{E}'_2 | Z = z]$, and as such a bound on this quantity would imply the same bound on $\mathbf{Pr}[\mathcal{E}'_2]$.

It is now enough to consider the ranges in the projection space $(Z, \mathcal{R}_{|Z})$ (which has VC dimension δ). By Lemma 5.9, we have $\big|\mathcal{R}_{|Z}\big| \leq \mathcal{G}_\delta(2m)$.

Let us fix any $\mathbf{r} \in \mathcal{R}_{|Z}$, and consider the event

$$\mathcal{E}_{\mathbf{r}} = \left\{\mathbf{r} \cap N = \emptyset \text{ and } |\mathbf{r} \cap T| > \frac{\varepsilon m}{2}\right\}.$$

We claim that $\mathbf{Pr}[\mathcal{E}_{\mathbf{r}}] \leq 2^{-\varepsilon m/2}$. Observe that if $k = |\mathbf{r} \cap (N \cup T)| \leq \varepsilon m/2$, then the event is empty, and this claim trivially holds. Otherwise, $\mathbf{Pr}[\mathcal{E}_{\mathbf{r}}] = \mathbf{Pr}[\mathbf{r} \cap N = \emptyset]$. To bound this probability, observe that we have the $2m$ elements of Z, and we can choose any m of them to be N, as long as none of them is one of the k "forbidden" elements of $\mathbf{r} \cap (N \cup T)$. The probability of that is $\binom{2m-k}{m}/\binom{2m}{m}$. We thus have

$$\begin{aligned}\mathbf{Pr}\Big[\mathcal{E}_{\mathbf{r}}\Big] \leq \mathbf{Pr}\Big[\mathbf{r} \cap N = \emptyset\Big] &= \frac{\binom{2m-k}{m}}{\binom{2m}{m}} = \frac{(2m-k)(2m-k-1)\cdots(m-k+1)}{2m(2m-1)\cdots(m+1)} \\ &= \frac{m(m-1)\cdots(m-k+1)}{2m(2m-1)\cdots(2m-k+1)} \leq 2^{-k} \leq 2^{-\varepsilon m/2}.\end{aligned}$$

Thus,

$$\mathbf{Pr}\Big[\mathcal{E}'_2 \;\Big|\; Z\Big] = \mathbf{Pr}\left[\bigcup_{\mathbf{r} \in \mathcal{R}_{|Z}} \mathcal{E}_{\mathbf{r}}\right] \leq \sum_{\mathbf{r} \in \mathcal{R}_{|Z}} \mathbf{Pr}[\mathcal{E}_{\mathbf{r}}] \leq \big|\mathcal{R}_{|Z}\big| 2^{-\varepsilon m/2} \leq \mathcal{G}_\delta(2m)2^{-\varepsilon m/2},$$

implying that $\mathbf{Pr}\Big[\mathcal{E}'_2\Big] \leq \mathcal{G}_\delta(2m)2^{-\varepsilon m/2}$. ∎

Proof of Theorem 5.28. By Claim 5.38 and Claim 5.39, we have that $\mathbf{Pr}[\mathcal{E}_1] \leq 2\mathcal{G}_\delta(2m)2^{-\varepsilon m/2}$. It thus remains to verify that if m satisfies (5.3), then $2\mathcal{G}_\delta(2m)2^{-\varepsilon m/2} \leq \varphi$.

Indeed, we know that $2m \geq 8\delta$ (by $(5.3)_{\text{p71}}$) and by Lemma 5.10, $\mathcal{G}_\delta(2m) \leq 2(2em/\delta)^\delta$, for $\delta \geq 1$. Thus, it is sufficient to show that the inequality $4(2em/\delta)^\delta 2^{-\varepsilon m/2} \leq \varphi$ holds. By

rearranging and taking lg of both sides, we have that this is equivalent to

$$2^{\varepsilon m/2} \geq \frac{4}{\varphi}\left(\frac{2em}{\delta}\right)^{\delta} \qquad \Longrightarrow \qquad \frac{\varepsilon m}{2} \geq \delta \lg \frac{2em}{\delta} + \lg \frac{4}{\varphi}.$$

By our choice of m (see (5.3)), we have that $\varepsilon m/4 \geq \lg(4/\varphi)$. Thus, we need to show that

$$\frac{\varepsilon m}{4} \geq \delta \lg \frac{2em}{\delta}.$$

We verify this inequality for $m = \dfrac{8\delta}{\varepsilon} \lg \dfrac{16}{\varepsilon}$ (this would also hold for bigger values, as can be easily verified). Indeed

$$2\delta \lg \frac{16}{\varepsilon} \geq \delta \lg\left(\frac{16e}{\varepsilon} \lg \frac{16}{\varepsilon}\right).$$

This is equivalent to $\left(\dfrac{16}{\varepsilon}\right)^2 \geq \dfrac{16e}{\varepsilon} \lg \dfrac{16}{\varepsilon}$, which is equivalent to $\dfrac{16}{e\,\varepsilon} \geq \lg \dfrac{16}{\varepsilon}$, which is certainly true for $0 < \varepsilon \leq 1$.

This completes the proof of the theorem. ∎

5.6. A better bound on the growth function

In this section, we prove Lemma 5.10$_{\text{p65}}$. Since the proof is straightforward but tedious, the reader can safely skip reading this section.

Lemma 5.40. *For any positive integer n, the following hold.*

(i) $(1 + 1/n)^n \leq e$. *(ii)* $(1 - 1/n)^{n-1} \geq e^{-1}$.

(iii) $n! \geq (n/e)^n$. *(iv) For any* $k \leq n$, *we have* $\left(\dfrac{n}{k}\right)^k \leq \binom{n}{k} \leq \left(\dfrac{ne}{k}\right)^k$.

Proof. (i) Indeed, $1 + 1/n \leq \exp(1/n)$, since $1 + x \leq e^x$, for $x \geq 0$. As such $(1 + 1/n)^n \leq \exp(n(1/n)) = e$.

(ii) Rewriting the inequality, we have that we need to prove $\left(\frac{n-1}{n}\right)^{n-1} \geq \frac{1}{e}$. This is equivalent to proving $e \geq \left(\frac{n}{n-1}\right)^{n-1} = \left(1 + \frac{1}{n-1}\right)^{n-1}$, which is our friend from (i).

(iii) Indeed,

$$\frac{n^n}{n!} \leq \sum_{i=0}^{\infty} \frac{n^i}{i!} = e^n,$$

by the Taylor expansion of $e^x = \sum_{i=0}^{\infty} \frac{x^i}{i!}$. This implies that $(n/e)^n \leq n!$, as required.

(iv) Indeed, for any $k \leq n$, we have $\frac{n}{k} \leq \frac{n-1}{k-1}$, as can be easily verified. As such, $\frac{n}{k} \leq \frac{n-i}{k-i}$, for $1 \leq i \leq k - 1$. As such,

$$\left(\frac{n}{k}\right)^k \leq \frac{n}{k} \cdot \frac{n-1}{k-1} \cdots \frac{n-k+1}{1} = \binom{n}{k}.$$

As for the other direction, by (iii), we have $\binom{n}{k} \leq \dfrac{n^k}{k!} \leq \dfrac{n^k}{\left(\frac{k}{e}\right)^k} = \left(\dfrac{ne}{k}\right)^k$. ∎

Lemma 5.10 restated. *For* $n \geq 2\delta$ *and* $\delta \geq 1$, *we have* $\left(\dfrac{n}{\delta}\right)^{\delta} \leq \mathcal{G}_\delta(n) \leq 2\left(\dfrac{ne}{\delta}\right)^{\delta}$, *where* $\mathcal{G}_\delta(n) = \displaystyle\sum_{i=0}^{\delta} \binom{n}{i}$.

PROOF. Note that by Lemma 5.40(iv), we have $\mathcal{G}_\delta(n) = \sum_{i=0}^{\delta} \binom{n}{i} \leq 1 + \sum_{i=1}^{\delta} \left(\frac{ne}{i}\right)^i$. This series behaves like a geometric series with constant larger than 2, since

$$\left(\frac{ne}{i}\right)^i \Big/ \left(\frac{ne}{i-1}\right)^{i-1} = \frac{ne}{i}\left(\frac{i-1}{i}\right)^{i-1} = \frac{ne}{i}\left(1-\frac{1}{i}\right)^{i-1} \geq \frac{ne}{i}\frac{1}{e} = \frac{n}{i} \geq \frac{n}{\delta} \geq 2,$$

by Lemma 5.40. As such, this series is bounded by twice the largest element in the series, implying the claim. ■

5.7. Bibliographical notes

The exposition of the ε-net and ε-sample theorems is roughly based on Alon and Spencer [**AS00**] and Komlós et al. [**KPW92**]. In fact, Komlós et al. proved a somewhat stronger bound; that is, a random sample of size $(\delta/\varepsilon)\ln(1/\varepsilon)$ is an ε-net with constant probability. For a proof that shows that in general ε-nets cannot be much smaller in the worst case, see [**PA95**]. The original proof of the ε-net theorem is due to Haussler and Welzl [**HW87**]. The proof of the ε-sample theorem is due to Vapnik and Chervonenkis [**VC71**]. The bound in Theorem 5.26 can be improved to $O\left(\frac{\delta}{\varepsilon^2} + \frac{1}{\varepsilon^2}\log\frac{1}{\varphi}\right)$ [**AB99**].

An alternative proof of the ε-net theorem proceeds by first computing an $(\varepsilon/4)$-sample of sufficient size, using the ε-sample theorem (Theorem 5.26_{p71}), and then computing and $\varepsilon/4$-net for this sample using a direct sample of the right size. It is easy to verify the resulting set is an ε-net. Furthermore, using the "naive" argument (see Section 5.3.2.3) then implies that this holds with the right probability, thus implying the ε-net theorem (the resulting constants might be slightly worse). Exercise 5.3 deploys similar ideas.

The beautiful alternative proof of both theorems via the usage of discrepancy is due to Chazelle and Matoušek [**CM96**]. The discrepancy method is a beautiful topic which is quite deep mathematically, and we have just skimmed the thin layer of melted water on top of the tip of the iceberg[⑦]. Two nice books on the topic are the books by Chazelle [**Cha01**] and Matoušek [**Mat99**]. The book by Chazelle [**Cha01**] is currently available online for free from Chazelle's webpage.

We will revisit discrepancy since in some geometric cases it yields better results than the ε-sample theorem. In particular, the random coloring of Theorem 5.31 can be derandomized using conditional probabilities. One can then use it to get an ε-sample/net by applying it repeatedly. A faster algorithm results from a careful implementation of the sketch-and-merge approach. The disappointing feature of all the deterministic constructions of ε-samples/nets is that their running time is exponential in the dimension δ, since the number of ranges is usually exponential in δ.

A similar result to the one derived by Haussler and Welzl [**HW87**], using a more geometric approach, was done independently by Clarkson at the same time [**Cla87**], exposing the fact that VC dimension is not necessary if we are interested only in geometric applications. This was later refined by Clarkson [**Cla88**], leading to a general technique that, in geometric settings, yields stronger results than the ε-net theorem. This technique has numerous applications in discrete and computational geometry and leads to several "proofs from the book" in discrete geometry.

Exercise 5.5 is from Anthony and Bartlett [**AB99**].

[⑦] The iceberg is melting because of global warming; so sorry, climate change.

5.7.1. Variants and extensions. A natural application of the ε-sample theorem is to use it to estimate the weights of ranges. In particular, given a finite range space $(\mathsf{X}, \mathcal{R})$, we would like to build a data-structure such that we can decide quickly, given a query range $\mathbf{r}$, what the number of points of X inside $\mathbf{r}$ is. We could always use a sample of size (roughly) $O(\varepsilon^{-2})$ to get an estimate of the weight of a range, using the ε-sample theorem. The error of the estimate of the size $|\mathbf{r} \cap \mathsf{X}|$ is $\leq \varepsilon n$, where $n = |\mathsf{X}|$; namely, the error is additive. The natural question is whether one can get a multiplicative estimate ρ, such that $|\mathbf{r} \cap \mathsf{X}| \leq \rho \leq (1+\varepsilon)|\mathbf{r} \cap \mathsf{X}|$, where $|\mathbf{r} \cap \mathsf{X}|$.

In particular, a subset $A \subset \mathsf{X}$ is a (relative) (ε, p)-***sample*** if for each $\mathbf{r} \in \mathcal{R}$ of weight $\geq pn$,

$$\left| \frac{|\mathbf{r} \cap A|}{|A|} - \frac{|\mathbf{r} \cap \mathsf{X}|}{|\mathsf{X}|} \right| \leq \varepsilon \frac{|\mathbf{r} \cap \mathsf{X}|}{|\mathsf{X}|}.$$

Of course, one can simply generate an εp-sample of size (roughly) $O(1/(\varepsilon p)^2)$ by the ε-sample theorem. This is not very interesting when $p = 1/\sqrt{n}$. Interestingly, the dependency on p can be improved.

Theorem 5.41 ([**LLS01**])**.** *Let $(X, \mathcal{R})$ be a range space with shattering dimension d, where $|X| = n$, and let $0 < \varepsilon < 1$ and $0 < p < 1$ be given parameters. Then, consider a random sample $A \subseteq X$ of size $\frac{c}{\varepsilon^2 p}\left(d \log \frac{1}{p} + \log \frac{1}{\varphi}\right)$, where c is a constant. Then, it holds that for each range $\mathbf{r} \in \mathcal{R}$ of at least pn points, we have*

$$\left| \frac{|\mathbf{r} \cap A|}{|A|} - \frac{|\mathbf{r} \cap X|}{|X|} \right| \leq \varepsilon \frac{|\mathbf{r} \cap X|}{|X|}.$$

In other words, A is a (p, ε)-sample for $(X, \mathcal{R})$. The probability of success is $\geq 1 - \varphi$.

A similar result is achievable by using discrepancy; see Exercise 5.7.

5.8. Exercises

Exercise 5.1 (Compute clustering radius)**.** Let C and P be two given sets of points in the plane, such that $k = |C|$ and $n = |P|$. Let $r = \max_{p \in \mathsf{P}} \min_{c \in C} \|c - p\|$ be the ***covering radius*** of P by C (i.e., if we place a disk of radius r centered at each point of C, all those disks cover the points of P).

(A) Give an $O(n + k \log n)$ expected time algorithm that outputs a number α, such that $r \leq \alpha \leq 10r$.

(B) For $\varepsilon > 0$ a prescribed parameter, give an $O(n + k\varepsilon^{-2} \log n)$ expected time algorithm that outputs a number α, such that $r \leq \alpha \leq (1+\varepsilon)r$.

Exercise 5.2 (Some calculus required)**.** Prove Lemma 5.13.

Exercise 5.3 (A direct proof of the ε-sample theorem)**.** For the case that the given range space is finite, one can prove the ε-sample theorem (Theorem 5.26_{p71}) directly. So, we are given a range space $\mathsf{S} = (\mathsf{x}, \mathcal{R})$ with VC dimension δ, where x is a finite set.

(A) Show that there exists an ε-sample of S of size $O\left(\delta\varepsilon^{-2} \log \frac{\log|\mathsf{x}|}{\varepsilon}\right)$ by extracting an $\varepsilon/3$-sample from an $\varepsilon/9$-sample of the original space (i.e., apply Lemma 5.30 twice and use Lemma 5.33).

(B) Show that for any k, there exists an ε-sample of S of size $O\left(\delta\varepsilon^{-2} \log \frac{\log^{(k)}|\mathsf{x}|}{\varepsilon}\right)$.

(C) Show that there exists an ε-sample of S of size $O\left(\delta\varepsilon^{-2} \log \frac{1}{\varepsilon}\right)$.

Exercise 5.4 (Sauer's lemma is tight)**.** Show that Sauer's lemma (Lemma 5.9) is tight. Specifically, provide a finite range space that has the number of ranges as claimed by Lemma 5.9.

Exercise 5.5 (Flip and flop)**.** (A) Let $b_1, \ldots, b_{2m}$ be m binary bits. Let Ψ be the set of all permutations of $1, \ldots, 2m$, such that for any $\sigma \in \Psi$, we have $\sigma(i) = i$ or $\sigma(i) = m + i$, for $1 \le i \le m$, and similarly, $\sigma(m + i) = i$ or $\sigma(m + i) = m + i$. Namely, $\sigma \in \Psi$ either leaves the pair $i, i + m$ in their positions or it exchanges them, for $1 \le i \le m$. As such $|\Psi| = 2^m$.

Prove that for a random $\sigma \in \Psi$, we have

$$\mathbf{Pr}\left[\left|\frac{\sum_{i=1}^{m} b_{\sigma(i)}}{m} - \frac{\sum_{i=1}^{m} b_{\sigma(i+m)}}{m}\right| \ge \varepsilon\right] \le 2e^{-\varepsilon^2 m/2}.$$

(B) Let Ψ' be the set of all permutations of $1, \ldots, 2m$. Prove that for a random $\sigma \in \Psi'$, we have

$$\mathbf{Pr}\left[\left|\frac{\sum_{i=1}^{m} b_{\sigma(i)}}{m} - \frac{\sum_{i=1}^{m} b_{\sigma(i+m)}}{m}\right| \ge \varepsilon\right] \le 2e^{-C\varepsilon^2 m/2},$$

where C is an appropriate constant. [**Hint:** Use (A), but be careful.**]**

(C) Prove Theorem 5.26 using (B).

Exercise 5.6 (Sketch and merge)**.** Assume that you are given a deterministic algorithm that can compute the discrepancy of Theorem 5.31 in $O(nm)$ time, where n is the size of the ground set and m is the number of induced ranges. We are assuming that the VC dimension δ of the given range space is small and that the algorithm input is only the ground set X (i.e., the algorithm can figure out on its own what the relevant ranges are).

(A) For a prespecified $\varepsilon > 0$, using the ideas described in Section 5.4.1.1, show how to compute a small ε-sample of X quickly. The running time of your algorithm should be (roughly) $O\big(n/\varepsilon^{O(\delta)}\mathrm{polylog}\big)$. What is the exact bound on the running time of your algorithm?

(B) One can slightly improve the running of the above algorithm by more aggressively sketching the sets used. That is, one can add additional sketch layers in the tree. Show how by using such an approach one can improve the running time of the above algorithm by a logarithmic factor.

Exercise 5.7 (Building relative approximations)**.** Prove the following theorem using discrepancy.

> **Theorem 5.42.** *Let $(X, \mathcal{R})$ be a range space with shattering dimension δ, where $|X| = n$, and let $0 < \varepsilon < 1$ and $0 < p < 1$ be given parameters. Then one can construct a set $N \subseteq X$ of size $O\big(\frac{\delta}{\varepsilon^2 p} \ln \frac{\delta}{\varepsilon p}\big)$, such that, for each range $\mathbf{r} \in \mathcal{R}$ of at least pn points, we have*
>
> $$\left|\frac{|\mathbf{r} \cap N|}{|N|} - \frac{|\mathbf{r} \cap X|}{|X|}\right| \le \varepsilon \frac{|\mathbf{r} \cap X|}{|X|}.$$
>
> *In other words, N is a relative (p, ε)-approximation for $(X, \mathcal{R})$.*