

## CMSC 754: Lecture 17

### Geometric Sampling, VC-Dimension, and Applications

**Reading:** This material is not covered in our text. My presentation is taken from the book “Geometric Approximation Algorithms” by S. Har-Peled.

**Geometric Set Systems:** Many problems in computational geometry involve an interaction between points and subsets of points defined by geometric objects. For example, suppose that a point set  $P$  represents a set of  $n$  locations on campus where students tend to congregate (see Fig. 1(a)). An internet wireless service provider wants to place a set of towers around the campus equipt with wireless routers to provide high-capacity data service to these locations. Due to power considerations, each wireless user needs to be within a certain distance  $\delta$  of one of these towers in order to benefit from the special service. The service provider would like to determine the smallest number of locations such that each of the congregation points is within distance  $\delta$  of one of these towers (see Fig. 1(b)). This is equivalent to a set-cover problem, where we want to cover a set of  $n$  points with set of circular disks of radius  $\delta$ . In general, set cover is a hard problem, but the constraint of having geometric sets can help ameliorate the situation. We begin with a discussion of the concept of geometric range spaces.

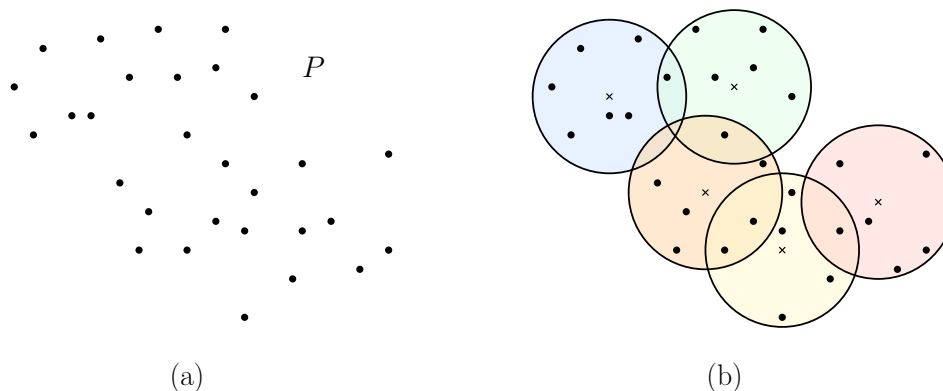


Fig. 1: Set cover by circular disks.

**Range Spaces:** Given a set  $P$  of  $n$  points in  $\mathbb{R}^d$ , its *power set*, denoted  $2^P$ , is the set of all subsets of  $P$ , including  $P$  and the empty set. The power set has  $2^n$  elements. If we constrain ourselves to subsets formed by some geometric property (e.g., the subsets of  $P$  lying within a circular disk, a halfplane, or a rectangle), this severely limits the types of subsets that can be formed.

We can characterize such geometric set systems abstractly as follows. A *range space* is defined to be a pair  $(X, \mathcal{R})$ , where  $X$  is an arbitrary set (which might be finite or infinite) and  $\mathcal{R}$  is a subset of the power set of  $X$ . We will usually apply range spaces to finite point sets. Given a set  $P \subseteq X$ , define the *restriction* (sometimes called the *projection*) of  $\mathcal{R}$  to  $P$  as

$$\mathcal{R}|_P = \{P \cap Q \mid Q \in \mathcal{R}\}.$$

For example, if  $X = \mathbb{R}^d$ ,  $P$  is a set of  $n$  points in  $\mathbb{R}^d$ , and  $\mathcal{R}$  consists of the subsets of real space contained within axis-parallel rectangles, then  $\mathcal{R}|_P$  consists of the subsets of  $P$

contained within axis-parallel rectangles (see Fig. 2). Note that not all subsets of  $P$  may be in  $\mathcal{R}|_P$ . For example, the sets  $\{1, 4\}$  and  $\{1, 2, 4\}$  cannot be formed by intersecting  $P$  with axis-parallel rectangular ranges.

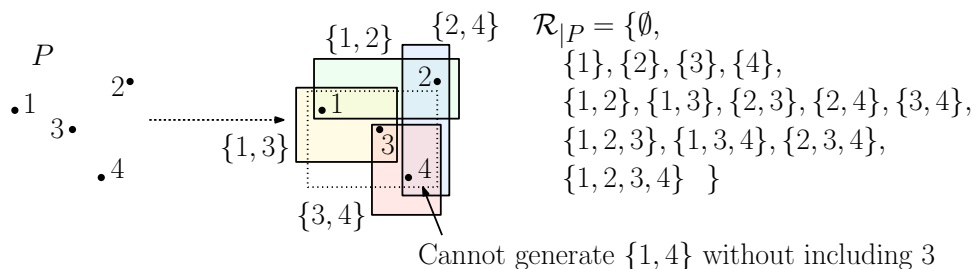


Fig. 2: A 4-point set and the range space of axis-parallel rectangles. Note that sets  $\{1, 4\}$  and  $\{1, 2, 4\}$  cannot be generated.

**Measures, Samples, and Nets:** When dealing with range spaces over very large point sets, it is often desirable to approximate the set with a much smaller sample of the set that does a good job of representing the set. What does it mean for a sample to be “good”? (We have already seen one example in our coverage of coresets.) The concept of a range space provides one way of making this precise.

Given a range space  $(P, \mathcal{R})$ , where  $P$  is finite, and given a range  $Q \in \mathcal{R}$ , we define  $Q$ 's *measure* to be the fraction of points of  $P$  that it contains, that is

$$\mu(Q) = \frac{|Q \cap P|}{|P|}.$$

Given a subset  $S \subseteq P$  (which we want to think of as being our sample, so that  $|S| \ll |P|$ ) it provides an *estimate* on the measure of a range. Define

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

(When  $S$  is clear, we will omit it from the subscript and just write  $\hat{\mu}(Q)$ .) A set  $S$  is a good sample of  $P$  if the estimate is close to the actual measure. That is, we would like to select  $S$  so that for all  $Q \in \mathcal{R}$ ,  $\hat{\mu}(Q) \approx \mu(Q)$ .

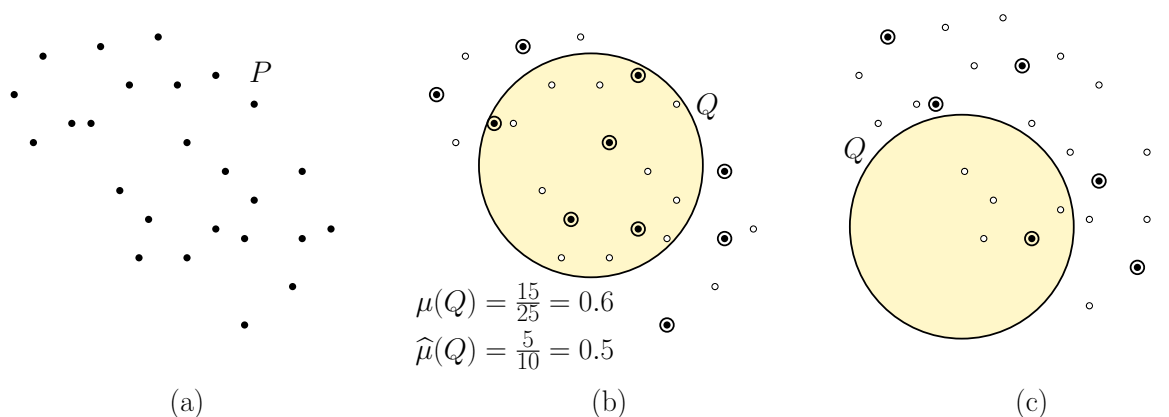
There are two common ways of characterizing good sample sets:  $\epsilon$ -samples and  $\epsilon$ -nets.

**$\epsilon$ -sample:** Given a range space  $(P, \mathcal{R})$  and any  $\epsilon > 0$ , a subset  $S \subseteq P$  is an  $\epsilon$ -sample if for any range  $Q \in \mathcal{R}$  we have

$$|\mu(Q) - \hat{\mu}(Q)| \leq \epsilon.$$

For example, suppose that  $\epsilon = 0.1$  and  $Q$  encloses 60% of the points of  $P$  ( $\mu(Q) = 0.6$ ) then  $Q$  should enclose a fraction of  $60 \pm 10\%$  (50–70%) of the points of  $S$  (see Fig. 3(b)). If this is true for every possible choice of  $Q$ , then  $S$  is a 0.1-sample for  $P$ .

**$\epsilon$ -net:** While  $\epsilon$ -samples intuitively correspond a desirable standard for good samples, it is often the case that we are can be satisfied with something weaker, where we just require

Fig. 3:  $\varepsilon$ -samples and  $\varepsilon$ -nets.

the sample to provide us with a single point. Given a range space  $(P, \mathcal{R})$  and any  $\varepsilon > 0$ , a subset  $S \subseteq P$  is an  $\varepsilon$ -net if for any range  $Q \in \mathcal{R}$ , if  $\mu(Q) \geq \varepsilon$  then  $Q$  contains at least one point of  $S$ . For example, if  $\varepsilon = 0.2$  and  $|P| = 25$ , then any range  $Q$  that contains at least  $0.2 \cdot 25 = 5$  points of  $P$  must contain at least one point of the  $\varepsilon$ -net (see Fig. 3(c)).

Observe that if  $S$  is an  $\varepsilon$ -sample, then it is surely an  $\varepsilon$ -net. The reason that  $\varepsilon$ -nets are of interest is that they are usually much smaller than  $\varepsilon$ -samples, and so it is more economical to use  $\varepsilon$ -nets whenever they are applicable.

**VC Dimension:** In looking at the definitions, it is clear that any sufficiently large random sample is likely to be an  $\varepsilon$ -sample (or  $\varepsilon$ -net) for an arbitrary range  $Q$ . But we want it to be the case that the sampling/netting condition applies to *all* ranges. Generally, this cannot be achieved unless we impose additional constraints on the range space, but these constraints are usually satisfied by the sorts of ranges that arise in geometric contexts.

Suppose that we are given a set  $P$  of  $n$  points in the plane and  $\mathcal{R}$  consists of axis parallel rectangles. How large might  $\mathcal{R}_{|P}$  be? If we take any axis-parallel rectangle that encloses some subset of  $P$ , and we shrink it as much as possible without altering the points contained within, we see that such a rectangle is generally determined by at most four points of  $P$ , that is, the points that lie on the rectangle's top, bottom, left, and right sides (see Fig. 4(b)). (It might be fewer if a point lies in the corner of the range.) Since there are essentially at most  $\binom{n}{4} = O(n^4)$  distinct ranges, we have  $|\mathcal{R}_{|P}| = O(n^4)$ .

How can we generalize this to arbitrary range spaces? (Including those where ranges are not defined by geometric shapes.) Remarkably, the definition makes no mention of geometry at all! This is the notion of *VC-dimension*, which is short for *Vapnik-Chervonenkis dimension*.<sup>1</sup>

Given an arbitrary range space  $(X, \mathcal{R})$  and finite point set  $P$ , we say that  $\mathcal{R}$  *shatters*  $P$  if  $\mathcal{R}_{|P}$  is equal to the power set of  $P$ , that is, we can form any of the  $2^{|P|}$  subsets of  $P$  by taking intersections with the ranges of  $\mathcal{R}$ . For example, the point set shown in Fig. 2 is *not* shattered

<sup>1</sup>The concept of VC-dimension was first developed in the field of probability theory in the 1970's. The topic was discovered to be very relevant to the fields of machine learning and computational geometry in late 1980's.

by the range space of axis-parallel rectangles. However, the four-element point set  $P$  shown in Fig. 4 is shattered by this range space, because we can form all  $2^4 = 16$  subsets of this set.

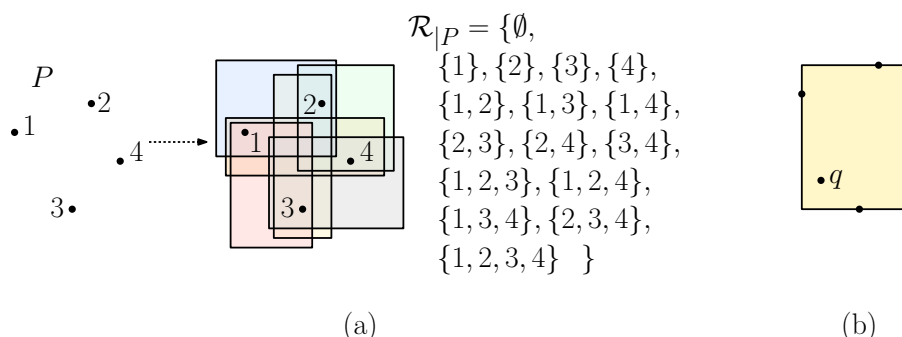


Fig. 4: (a) a 4-element point set that is shattered by the range space of axis-parallel rectangles (showing only the 2-element subsets in the drawing), and (b) the proof that no 5-element point set is shattered.

**Definition:** The *VC-dimension* of a range space  $(X, \mathcal{R})$ , is defined to be the size of the *largest* point set that is shattered by the range space.

We will denote the VC-dimension as  $\text{dim}_{\text{VC}}(X, \mathcal{R})$ , or simply  $\text{dim}_{\text{VC}}(\mathcal{R})$  when  $X$  is clear. Here are a couple of examples:

**Axis-parallel rectangles:** Axis-parallel rectangles have VC-dimension four. In Fig. 4(a) we gave a 4-element point set that can be shattered. We assert that no five points can be shattered. Consider any set  $P$  of five points in the plane, and assume the points are in general position. Because of general position, at least one of the points of  $P$ , call it  $q$ , does not lie on the boundary of  $P$ 's smallest enclosing axis-parallel rectangle (see Fig. 4(b)). It is easy to see that it is not possible to form the subset  $P \setminus \{q\}$ , since any axis-parallel rectangle containing the points that define the minimum bounding rectangle must contain all the points of  $P$ .

**Euclidean disks in the plane:** Planar Euclidean disks have VC-dimension three. A 3-element point set that is shattered is shown Fig. 5(a).

To see that no four-element set can be shattered, consider any set of four points  $\{a, b, c, d\}$  in general position (see Fig. 5(b)). If any point lies in the convex hull of the other three, then clearly it is not possible to form the subset that excludes this one point and contains all the others. Otherwise, all the points are on the convex hull. Consider their Delaunay triangulation. Let  $a$  and  $b$  denote the two points of the group that are *not* connected by an edge of the triangulation (see Fig. 5(b)). Because  $\overline{ab}$  is *not* an edge of the Delaunay triangulation, by the empty-circle property, any circle that contains  $a$  and  $b$ , must contain at least one other point of the set. Therefore, the subset  $\{a, b\}$  cannot be generated.

**Sauer's Lemma:** We have seen (1) that the range space of axis-parallel rectangles over an  $n$  element point set contains  $O(n^4)$  ranges and (2) that such a range space has VC-dimension

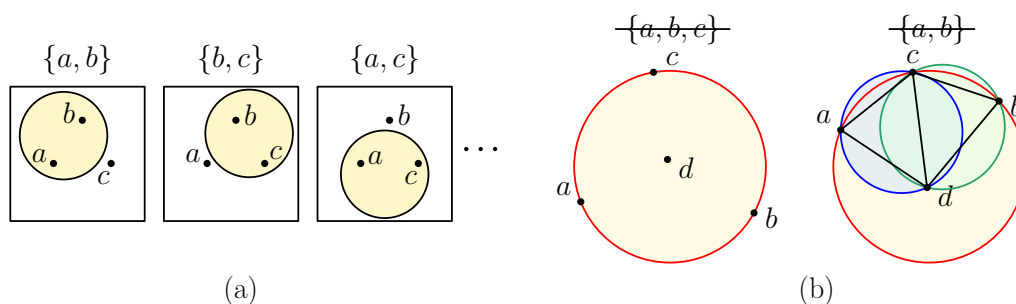


Fig. 5: (a) a 3-element point set that is shattered by the range space of Euclidean disks (showing just the 2-element subsets), and (b) the proof that no 4-element point set is shattered.

four. This raises the interesting conjecture that the size of *any* range space is related to its VC-dimension. Indeed, this is the case, and it is proved by a useful result called *Sauer's Lemma* (also called the Sauer-Shelah Lemma).

Before giving this lemma, let us first define a useful function. Given  $0 \leq d \leq n$ , define  $\Phi_d(n)$  to be the number of subsets of size at most  $d$  over a ground set of size  $n$ , that is,

$$\Phi_d(n) = \binom{n}{0} + \binom{n}{1} + \dots + \binom{n}{d} = \sum_{i=0}^d \binom{n}{i}.$$

An important fact about this function is that it satisfies the following recurrence

$$\Phi_d(n) = \Phi_d(n-1) + \Phi_{d-1}(n-1).$$

An intuitive way to justify the recurrence is to fix one element  $x_0$  of the  $n$ -element set. The number of sets of size at most  $d$  that do not contain  $x_0$  is  $\Phi_d(n-1)$  (since the element itself is not available from the  $n$  elements) and the number of sets that do contain  $x_0$  is  $\Phi_{d-1}(n-1)$  (since once  $x_0$  is removed from each of these sets, we have  $d-1$  remaining elements to pick from).

**Sauer's Lemma:** If  $(X, \mathcal{R})$  is a range space with VC-dimension  $d$  and  $|X| = n$ , then  $|\mathcal{R}| \leq \Phi_d(n)$ .

**Proof:** The proof is by induction on  $d$  and  $n$ . It is trivially true if  $d = 0$  or  $n = 0$ . (Note that  $\Phi_0(n) = \Phi_d(0) = 1$ , and the range space consists of just the empty set.) Assuming  $n \geq 1$ , fix any one element  $x \in X$ . Consider the following two range sets:

$$\begin{aligned} \mathcal{R}_x &= \{Q \setminus \{x\} : Q \cup \{x\} \in \mathcal{R} \text{ and } Q \setminus \{x\} \in \mathcal{R}\} \\ \mathcal{R} \setminus \{x\} &= \{Q \setminus \{x\} : Q \in \mathcal{R}\} \end{aligned}$$

Intuitively,  $\mathcal{R}_x$  is formed from pairs of ranges from  $\mathcal{R}$  that are identical except that one contains  $x$  and the other does not. (For example, if  $x$  is along the side of some axis-parallel rectangle, then there is a range that includes  $x$  and a slightly smaller one that does not. We put the range that does not contain  $x$  into  $\mathcal{R}_x$ .) The set  $\mathcal{R} \setminus \{x\}$  is the result of throwing  $x$  entirely out of the point set and considering the remaining ranges.

We assert that  $|\mathcal{R}| = |\mathcal{R}_x| + |\mathcal{R} \setminus \{x\}|$ . To see why, suppose that we charge each range of  $\mathcal{R}$  to its corresponding range in  $\mathcal{R} \setminus \{x\}$ . Every range of  $\mathcal{R} \setminus \{x\}$  receives at least one charge, but it receives two charges if there exist two ranges that are identical except that one contains  $x$  and one doesn't. The elements of  $\mathcal{R}_x$  account for these extra charges.

Now, let us apply induction. Observe that the range space  $(X \setminus \{x\}, \mathcal{R}_x)$  has VC-dimension  $d - 1$ . In particular, we claim that no set  $P'$  of size  $d$  can be shattered. To see why, suppose that we were to throw  $x$  back into the mix. The pairs of sets of  $\mathcal{R}$  that gave rise to the ranges of  $\mathcal{R}_x$  would then shatter the  $d + 1$  element set  $P' \cup \{x\}$ . (This is the critical step of the proof, so you should take a little time to convince yourself of it!) Clearly, the VC-dimension of  $\mathcal{R} \setminus \{x\}$  cannot be larger than the original, so its VC-dimension is at most  $d$ . Since both range sets involve one fewer element ( $n - 1$ ), by applying the induction hypothesis and our earlier recurrence for  $\Phi_d(n)$ , we have

$$|\mathcal{R}| = |\mathcal{R}_x| + |\mathcal{R} \setminus \{x\}| \leq \Phi_{d-1}(n-1) + \Phi_d(n-1) = \Phi_d(n).$$

And this completes the proof.

Clearly,  $\Phi_d(n) = \Theta(n^d)$ , so Sauer's Lemma implies that an range space of VC-dimension  $d$  over a point set of size  $n$  contains at most  $O(n^d)$  ranges. It can be shown that this bound is tight.

**On the Sizes of  $\varepsilon$ -nets and  $\varepsilon$ -samples:** One of the important features of range spaces of low VC-dimension is that there exist good samples of small size. Intuitively, by restricting ourselves to simple geometric ranges, we do not have the power to construct arbitrarily complicated sets. Observe that if sets of arbitrary complexity are allowed, then it would be hopeless to try to construct  $\varepsilon$ -samples or  $\varepsilon$ -nets, because given any sample, we could find some nasty range  $Q$  that manages to exclude every point of the sample and include all the remaining points of  $P$  (see Fig. 6).

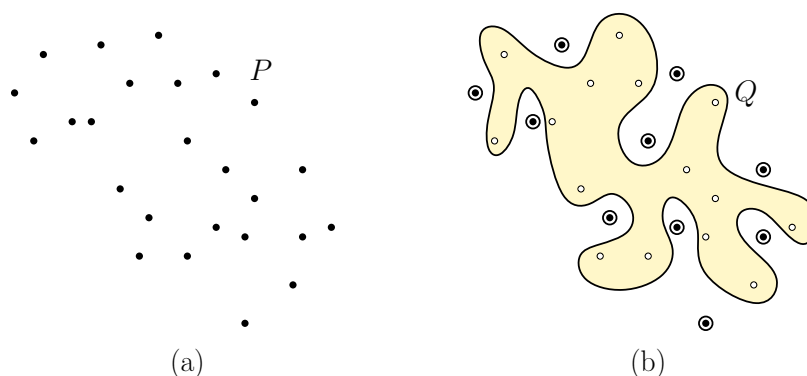


Fig. 6: Why VC-dimension matters.

If a range space has VC-dimension  $d$ , we will show that there exist  $\varepsilon$ -samples and  $\varepsilon$ -nets whose sizes depend on  $\varepsilon$  and  $d$  alone, independent of the original point set  $n$ ! This is very important in geometric approximation algorithms, because it allows us to extract a tiny set from a huge one, with the knowledge that the tiny set is guaranteed to do a good job of representing the huge one.

**Theorem:** ( $\varepsilon$ -Sample Theorem) Let  $(X, \mathcal{R})$  be a range space of VC-dimension  $d$ , and let  $P$  be any finite subset of  $X$ . There exists a positive constant  $c$  (independent of the range space) such that with probability at least  $1 - \varphi$  any random sample  $S$  of  $P$  of size at least

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

is an  $\varepsilon$ -sample for  $(P, \mathcal{R})$ . Assuming that  $d$  and  $\varphi$  are constants, this is  $O((1/\varepsilon^2) \log(1/\varepsilon))$ .

**Theorem:** ( $\varepsilon$ -Net Theorem) Let  $(X, \mathcal{R})$  be a range space of VC-dimension  $d$ , and let  $P$  be any finite subset of  $X$ . There exists a positive constant  $c$  (independent of the range space) such that with probability at least  $1 - \varphi$  any random sample  $S$  of  $P$  of size at least

$$\frac{c}{\varepsilon} \left( d \log \frac{1}{\varepsilon} + \log \frac{1}{\varphi} \right)$$

is an  $\varepsilon$ -sample for  $(P, \mathcal{R})$ . Assuming that  $d$  and  $\varphi$  are constants, this is  $O((1/\varepsilon) \log(1/\varepsilon))$ .

We will not prove these theorems. Both involve fairly standard applications of techniques from probability theory (particularly the Chernoff bounds), but there are quite a few non-trivial technical details involved.

**Application — Geometric Set Cover:** Nets and samples have applications in many areas of computational geometry. We will discuss one such application involving geometric set cover. Given an  $n$ -element ground set  $X$  and a collection of subsets  $\mathcal{R}$  over  $X$ , the *set cover problem* is that of computing a subset of  $\mathcal{R}$  of minimum size whose union contains all the elements of  $X$ . It is well known that this problem is NP-hard, and assuming that  $P \neq NP$ , it is hard to approximate to within a factor of  $\Omega(\log n)$ .

There is a well-known greedy approximation algorithm for set cover that achieves an approximation ratio of  $\ln n$ . This algorithm repeatedly selects the set of  $\mathcal{R}$  that contains the largest number of elements of  $X$  that have not yet been covered. This algorithm can be applied to arbitrary set systems, but we will show that if the range space  $(X, \mathcal{R})$  has constant VC-dimension then there exists an approximation algorithm that achieves an approximation ratio of  $O(\log k^*)$ , where  $k^*$  is the number of sets in the optimal solution. If  $k^* \ll n$ , then this algorithm provides a significant theoretical improvement over the greedy algorithm. (In practice, the greedy heuristic is very good.)

For the sake of simplicity, we will present this algorithm in a slightly simpler context, but it readily generalizes to any range space of constant VC-dimension. Rather than solve the set cover problem, we will present an approximation for the dual problem, called *hitting set*. Here, we are given a set system  $(X, \mathcal{R})$ , and the problem is to compute a subset of  $H \subseteq X$  (as small as possible) so that every set  $Q$  in  $\mathcal{R}$  contains at least one element (that is, it is “hit”) by  $H$ .

The set cover and hitting set problems are actually dual equivalents. Given a set system  $(X, \mathcal{R})$  for the set cover problem (see Fig. 7(a)), we can form an equivalent instance of hitting set as follows. We create a new set system  $(X', \mathcal{R}')$ , where each element of  $X'$  corresponds to a set in  $\mathcal{R}$ , and each set in  $\mathcal{R}'$  corresponds to an element in  $X$ . For each  $x \in X$ , let  $R_{i_1}, R_{i_2}, \dots, R_{i_j}$  denote the sets of  $\mathcal{R}$  that contain  $x$ . We create the set  $\{i_1, \dots, i_j\}$  and add

it to  $\mathcal{R}'$ . (See Fig. 7(c), where the point  $x$  is mapped to the disk  $x^*$  and ranges  $R_1, R_2, R_3$  are mapped to points  $R_1^*, R_2^*, R_3^*$ . Observe that  $x \in R_i$  and  $R_i^* \in x^*$ .) It is easy to see that any set cover for  $(X, \mathcal{R})$  (see Fig. 7(b)) corresponds to an equivalent hitting set for  $(X', \mathcal{R}')$  (see Fig. 7(d)). (We'll leave the formal proof as an exercise.)

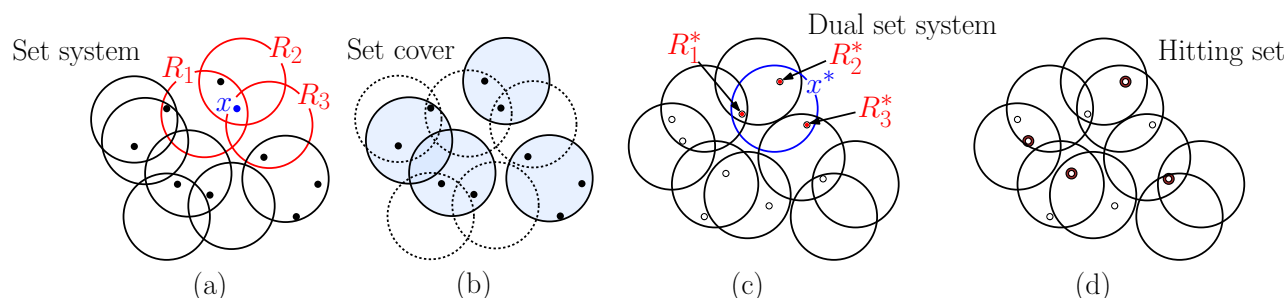


Fig. 7: (a) Set system  $(X, \mathcal{R})$ , (b) a set cover, (c) the dual set system  $(X', \mathcal{R}')$  (where points become disk centers and disk centers become points), and (d) the equivalent hitting set

**Iterative Reweighting Algorithm:** Given  $(X, \mathcal{R})$  and  $k$ , we present an algorithm which determine either that there exists a hitting set of size at most  $k' = ck \log k$  (where  $c$  is a suitably chosen constant), or it will report failure, indicating that there is no hitting set of size  $k$ . To approximate the hitting set problem, we simply run this algorithm with  $k = 1, 2, 4, \dots, 2^j$ , until we achieve success. Letting  $k^*$  denote the size of the optimal hitting set, it follows that when the algorithm first reports success, the hitting set can have size at most  $O(k^* \log k^*)$ , and therefore it is an  $O(\log k^*)$  factor approximation.

To start, we associate each point of  $X$  with a positive integer *weight*, which initially is 1. When computing measures, a point  $x \in X$  with weight  $w(x)$  will contribute  $w(x)$  to the weight. For a suitable value of  $\varepsilon$  (which will depend on  $k$ ) we compute a weighted  $\varepsilon$ -net  $S$  of size  $k'$  for  $X$ . This means that any disk of radius  $\delta$  whose weight is at least  $\varepsilon$  times the total weight of  $X$  must contain at least one point of  $S$ . If  $S$  is a hitting set, we output  $S$  and we are done. If not, we must have failed to hit some disk. We double the weights of the points within this disk (thus making them more likely to be sampled in the future). If we don't succeed after a sufficient number of iterations, we declare that no hitting set of size  $k$  exists. Here is a detailed description:

- (1) Let  $\varepsilon \leftarrow 1/(4k)$ . For a suitable value of  $c$  (depending on  $\varepsilon$ ) set  $k' \leftarrow ck \log k$ .
- (2) Repeat the following either  $2k \log(n/k)$  times or until returns "success"
  - (a) Compute a weighted  $\varepsilon$ -net  $S$  of  $X$  of size  $k'$  (see Fig. 8(a)). (By the  $\varepsilon$ -Net Theorem, this can be done by computing a random sample of  $X$  of size  $k'$ , where the probability that a point is sampled is proportional to its weight.)
  - (b) Enumerate the disks of  $\mathcal{R}$ , and determine whether there exists any disk that is *not* hit by any of the points of  $S$ .
    - (i) If there is no such disk, then  $S$  is a hitting set, and we return "success" (and output  $S$  as the answer).

- (ii) If there is some disk  $Q$  that is not hit, we double the weight of all the points lying within  $Q$  (see Fig. 8(b)).
- (3) If we exit the loop without success, we return “failure”. (This means that there is no hitting set of size  $k$  or smaller.)

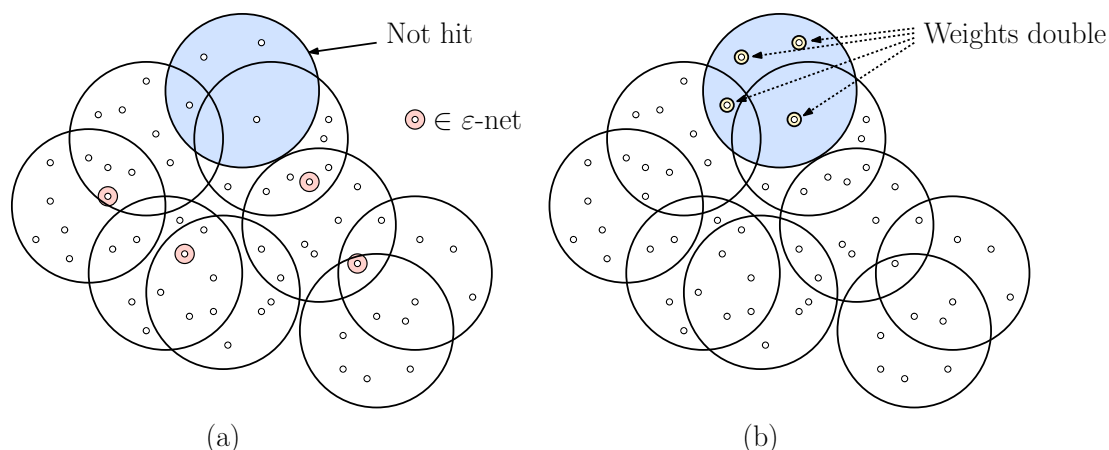


Fig. 8: The hitting-set approximation algorithm

**Analysis:** Before delving into the analysis, let’s see intuitively what the algorithm is doing. Clearly, if this algorithm terminates, then it has computed a hitting set of size  $k'$ . We want to argue that if such a hitting set exists, the algorithm will find it within  $2k \log(n/k)$  iterations. Observe that if an iteration is not successful, then some disk was not hit by our random sample. Because (by our assumption) the random sample is an  $\varepsilon$ -net, such a disk cannot contain more than an  $\varepsilon$  fraction of the total weight. All the points within this disk have their weights doubled. It follows that the total weight of the entire point set does not increase very much as a result, basically by a factor of at most  $(1 + \varepsilon)$ . Since the optimal hitting set must hit all disks, at least one of these doubled points is in the optimal hitting set. It follows that the total weight of the points in the optimal hitting set are increasing rapidly. Thus, the overall weight is growing slowly and the weight of the optimum set is growing rapidly. But since the optimum hitting set is a subset of the overall set, its weight can never be larger. Therefore, this process cannot go on forever. The analysis provides a formal bound on when it must end.

**Lemma:** If  $(X, \mathcal{R})$  has constant VC-dimension, and there exists a hitting set of size  $k$ , then after at most  $2k \log(n/k)$  iterations, the above algorithm will terminate successfully with a hitting set of size  $O(k \log k)$ . If not, it terminates with failure.

Let us assume that there exists a hitting set  $H$  of size  $k$  (which we’ll call the optimal hitting set). We will show that the algorithm terminates within  $2k \log(n/k)$  iterations. Let  $W_i(X)$  denote the total weight of all the points of  $X$  after the  $i$ th iteration. When the algorithm starts, each of the  $n$  points of  $X$  has weight 1, so  $W_0(X) = n$ .

Let’s consider the transition from the  $i - 1$ st to the  $i$ th iteration in detail. The set  $S$  is an  $\varepsilon$ -net, which means that any disk that is not hit by  $S$  must have a weight smaller than

$\varepsilon W_{i-1}(X)$ . If the iteration is not successful, then there is a disk  $Q$  that was not hit. All the points within  $Q$  have their weights doubled, which implies that the total weight has increased by at most  $W_{i-1}(Q) \leq \varepsilon W_{i-1}(X)$ . Therefore, after an unsuccessful iteration, we have

$$W_i(X) = W_{i-1}(X) + W_{i-1}(Q) \leq W_{i-1}(X) + \varepsilon W_{i-1}(X) = (1 + \varepsilon)W_{i-1}.$$

Since  $W_0(X) = n$ , we have  $W_i(X) \leq (1 + \varepsilon)^i n$ . Using the standard inequality  $1 + x \leq e^x$ , we have  $W_i(X) \leq n \cdot e^{\varepsilon i}$ .

Because any hitting set (including the optimal) must hit all the disks, we know that there is at least one point of the optimal hitting set that lies within the “unhit” disk, meaning that at least one of the  $k$  optimal points will have its weight doubled. For  $1 \leq j \leq k$ , let  $t_i(j)$  denote the number of times that the  $j$ th optimal point has been doubled during stage  $i$ . (It’s either once or not at all.) Since each of these points started with a weight of 1, the total weight of the optimal hitting set after  $i$  iterations, which we will denote by  $W_i(H)$  satisfies

$$W_i(H) = \sum_{j=1}^k 2^{t_i(j)}.$$

Because the function  $f(x) = 2^x$  is a convex function, it follows from standard combinatorics (in particular, Jensen’s inequality) that this sum is minimized when all the  $t_i(j)$ ’s are as nearly equal as possible. We know that at least one point must be doubled with each iteration, and therefore the minimum occurs when  $t_i(j) = i/k$ , for all  $j$ . (We’ll ignore the minor inconvenience that  $t_i(j)$  is an integer. It won’t affect the asymptotics.) Therefore:

$$W_i(H) \geq k 2^{i/k}.$$

Because  $H \subseteq X$ , we know that  $W_i(H) \leq W_i(X)$ . Therefore, we know that the number of iterations  $i$  must satisfy

$$k 2^{i/k} \leq n \cdot e^{\varepsilon i}.$$

Simplifying and recalling that  $\varepsilon = 1/(4k)$ , we obtain

$$\lg k + \frac{i}{k} \leq \lg n + \frac{i}{4k} \lg e \leq \lg n + \frac{i}{2k}.$$

(Here we have used the fact that  $\lg e \approx 1.45 \leq 2$ .) Therefore,  $i/(2k) \leq \lg n - \lg k$ , which implies that (assuming there is a hitting set of size  $k$ ) the number of iterations  $i$  satisfies

$$i \leq 2k \lg \frac{n}{k},$$

and therefore, if the algorithm runs for more than  $2 \lg(n/k)$  iterations, we know that there cannot be a hitting set of size  $k$ . (At this point, the algorithm will double the value of  $k$  and try again.)

The total running time is polynomial in  $n = |X|$  and  $m = |\mathcal{R}|$ . There are at most  $O(n \log(n/k)) = O(n \log n)$  iterations. Each iteration involves the generation of an  $\varepsilon$ -net, which takes  $O(k \log k)$  time. Also, for each of the  $m$  sets of  $\mathcal{R}$ , we take  $O(k \log k)$  time to determine whether any of the elements of  $S$  lie within this set. This takes  $O(km \log k) = O(nm \log n)$  time. Thus, the overall running time is (crudely)  $O(n^2 m \log^2 n)$ .

**(Naive) Proof of  $\varepsilon$ -Sample Theorem:** (Optional)

The proofs of the  $\varepsilon$ -Net and  $\varepsilon$ -Sample Theorems involve fairly straightforward applications of probability theory, but since we have not built up the necessary probability-theoretic background, it is not all that easy to present them. We will give a short (but not fully accurate) justification of the  $\varepsilon$ -Sample Theorem, which is due to Har-Peled (who presents the full proof in his book). This will give an idea of where the various expressions come from, even though the proof is not fully rigorous.

Before giving the proof, we will need to make use of a standard result from probability theory. This is well worth remembering, because it is used frequently in the analysis of randomized algorithms. To begin, let  $Y_1, \dots, Y_n$  be a set of  $n$  *Bernoulli trials*, that is, a sequence of independent “coin flips” having the value 1 with probability  $p_i$  and value 0 with probability  $q_i = 1 - p_i$ . Let  $Y = \sum_i Y_i$ , and let  $\mu = \mathbb{E}[Y] = \sum_i p_i$ , be its expected value. Clearly, we expect  $Y$  to be close to its expected value. What is the probability that it deviates from this by, say, a factor of  $1 \pm \delta$  (where  $\delta$  is a value that we will select based on our needs). The following result, called the *Chernoff bound*, credited to Herman Chernoff from 1952.

**Theorem:** (Chernoff Bound) For any  $\delta > 2e - 1$ :

$$\Pr[Y > (1 + \delta)\mu] < \exp(-\mu\delta^2/4) \quad \text{and} \quad \Pr[Y < (1 - \delta)\mu] < \exp(-\mu\delta^2/2).$$

Suppose we have a range space  $(X, \mathcal{R})$  with VC-dimension  $d$ , and consider any range  $Q$  that contains a fraction  $p$  of the points of  $X$ , where  $p \geq 2\varepsilon$ . Consider any random sample  $S$  of  $s$  points from  $X$  (sampling with replacement so that the probabilities of lying within  $r$  do not change).

Let  $q_i$  be the  $i$ th sample point, and let  $Y_i$  be an indicator variable whose value is 1 if  $q_i \in Q$  and 0 otherwise. Clearly,  $(\sum_i Y_i)/s$  is an estimate for  $p = |Q \cap X|/|X|$ . To prove the  $\varepsilon$ -Sample Theorem, we want to determine how many samples  $s$  we should take so that this estimate is within a factor  $1 \pm \varepsilon$  of  $p$  with confidence at least  $1 - \varphi$ .

Observe that the number of points of  $X$  we expect to see in  $Q$  is  $\mu = ps$ . A sample fails the sampling theorem if the actual count differs from the expected value by at least  $\varepsilon s$ , that is

$$\left| \sum_{i=1}^s Y_i - \mu \right| > \varepsilon s = \frac{\varepsilon}{p} ps = \delta \mu,$$

where  $\delta = \varepsilon/p$ . Letting  $Y = \sum_{i=1}^s Y_i$ , this is equivalent to saying that  $Y$  either exceeds its expected value  $\mu$  by a factor of  $1 + \delta$  or falls short by a factor of  $1 - \delta$ . We can therefore apply the Chernoff bound (combining both the upper and lower bounds), obtaining

$$\begin{aligned} \Pr \left[ \left| \sum_{i=1}^s Y_i - \mu \right| > \delta \mu \right] &\leq \Pr[Y < (1 - \delta)\mu] + \Pr[Y > (1 + \delta)\mu] \\ &\leq \exp(-\mu\delta^2/2) + \exp(-\mu\delta^2/4) \\ &\leq 2 \exp\left(-\frac{\mu\delta^2}{4}\right) = 2 \exp\left(-\frac{\varepsilon^2}{4p}s\right). \end{aligned}$$

We want this probability of a bad sample to be at most  $\varphi$ . If we set  $\varphi$  to this value and solve for  $s$ , we obtain

$$s \geq \frac{4}{\varepsilon^2} \ln \frac{2}{\varphi} \geq \frac{4p}{\varepsilon^2} \ln \frac{2}{\varphi}.$$

This is pretty close to the value given in the  $\varepsilon$ -Sample Theorem, but strangely we never made use of the VC-dimension  $d$ . (Something must be wrong!) The problem with this proof is that we only applied it to a single range  $Q$ . To generalize it to all ranges, we need work a bit harder, and in doing so, we will introduce the VC-dimension. (We refer you to Har-Peled's book for the details.)