

# Distribution-Sensitive Bounds on Relative Approximations of Geometric Ranges

Yufei Tao

Chinese University of Hong Kong, Hong Kong  
taoyf@cse.cuhk.edu.hk

Yu Wang

Chinese University of Hong Kong, Hong Kong  
yuwang@cse.cuhk.edu.hk

---

## Abstract

---

A family  $\mathcal{R}$  of ranges and a set  $X$  of points, all in  $\mathbb{R}^d$ , together define a range space  $(X, \mathcal{R}|_X)$ , where  $\mathcal{R}|_X = \{X \cap h \mid h \in \mathcal{R}\}$ . We want to find a structure to estimate the quantity  $|X \cap h|/|X|$  for any range  $h \in \mathcal{R}$  with the  $(\rho, \epsilon)$ -guarantee: (i) if  $|X \cap h|/|X| > \rho$ , the estimate must have a relative error  $\epsilon$ ; (ii) otherwise, the estimate must have an absolute error  $\rho\epsilon$ . The objective is to minimize the size of the structure. Currently, the dominant solution is to compute a relative  $(\rho, \epsilon)$ -approximation, which is a subset of  $X$  with  $\tilde{O}(\lambda/(\rho\epsilon^2))$  points, where  $\lambda$  is the VC-dimension of  $(X, \mathcal{R}|_X)$ , and  $\tilde{O}$  hides polylog factors.

This paper shows a more general bound sensitive to the content of  $X$ . We give a structure that stores  $O(\log(1/\rho))$  integers plus  $\tilde{O}(\theta \cdot (\lambda/\epsilon^2))$  points of  $X$ , where  $\theta$  – called the *disagreement coefficient* – measures how much the ranges differ from each other in their intersections with  $X$ . The value of  $\theta$  is between 1 and  $1/\rho$ , such that our space bound is never worse than that of relative  $(\rho, \epsilon)$ -approximations, but we improve the latter’s  $1/\rho$  term whenever  $\theta = o(\frac{1}{\rho \log(1/\rho)})$ . We also prove that, in the worst case, summaries with the  $(\rho, 1/2)$ -guarantee must consume  $\Omega(\theta)$  words even for  $d = 2$  and  $\lambda \leq 3$ .

We then constrain  $\mathcal{R}$  to be the set of halfspaces in  $\mathbb{R}^d$  for a constant  $d$ , and prove the existence of structures with  $o(1/(\rho\epsilon^2))$  size offering  $(\rho, \epsilon)$ -guarantees, when  $X$  is generated from various stochastic distributions. This is the first formal justification on why the term  $1/\rho$  is not compulsory for “realistic” inputs.

**2012 ACM Subject Classification** Theory of computation  $\rightarrow$  Computational geometry

**Keywords and phrases** Relative Approximation, Disagreement Coefficient, Data Summary

**Digital Object Identifier** 10.4230/LIPIcs.SoCG.2019.57

**Related Version** A full version of this paper is available at <http://arxiv.org/abs/1903.06617>.

**Funding** This work was partially supported by a direct grant (Project Number: 4055079) from CUHK and by a Faculty Research Award from Google.

## 1 Introduction

A (data) *summary*, in general, refers to a structure that captures certain information up to a specified precision about a set of objects, but using space significantly smaller than the size of the set. These summaries have become important tools in algorithm design, especially in distributed/parallel computing where the main performance goal is to minimize the communication across different servers.

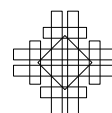
In this paper, we revisit the problem of finding a small-space summary to perform range estimation in  $\mathbb{R}^d$  with relative-error guarantees. Let  $\mathcal{R}$  be a family of geometric ranges in  $\mathbb{R}^d$  (e.g., a “halfspace family”  $\mathcal{R}$  is the set of all halfspaces in  $\mathbb{R}^d$ ), and  $X$  be a set of points in  $\mathbb{R}^d$ .  $\mathcal{R}$  and  $X$  together define a range space  $(X, \mathcal{R}|_X)$ , where  $\mathcal{R}|_X = \{X \cap h \mid h \in \mathcal{R}\}$ . Denote by  $\lambda$  the VC-dimension of  $(X, \mathcal{R}|_X)$ .



© Yufei Tao and Yu Wang;  
licensed under Creative Commons License CC-BY  
35th International Symposium on Computational Geometry (SoCG 2019).  
Editors: Gill Barequet and Yusu Wang; Article No. 57; pp. 57:1–57:14



Leibniz International Proceedings in Informatics  
LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany



Following the notations of [6, 11], define

$$\overline{X}(h) = |X \cap h|/|X|$$

for each  $h \in \mathcal{R}$ , namely,  $\overline{X}(h)$  is the fraction of points in  $X$  that are covered by  $h$ . Given real-valued parameters  $0 < \rho, \epsilon < 1$ , we need to produce a structure – called a  $(\rho, \epsilon)$ -summary henceforth – that allows us to obtain, for every range  $h \in \mathcal{R}$ , a real-valued estimate  $\tau$  satisfying the following  $(\rho, \epsilon)$ -guarantee:

$$|\overline{X}(h) - \tau| \leq \epsilon \cdot \max\{\rho, \overline{X}(h)\}. \quad (1)$$

Phrased differently, the guarantee says that (i) if  $\overline{X}(h) > \rho$ ,  $\tau$  must have a relative error at most  $\epsilon$ ; (ii) otherwise,  $\tau$  must have an absolute error at most  $\rho\epsilon$ . The main challenge is to minimize the size of the structure.

## 1.1 Previous results

Throughout the paper, all logarithms have base 2 by default.

### 1.1.1 Sample-Based $(\rho, \epsilon)$ -Summaries

We say that a  $(\rho, \epsilon)$ -summary of  $(X, \mathcal{R}|_X)$  is *sample-based* if it meets the requirements below: it stores a subset  $Z \subseteq X$  such that, for any range  $h \in \mathcal{R}$  with  $Z \cap h = \emptyset$ , it returns an estimate 0 for  $\overline{X}(h)$ .

A *relative  $(\rho, \epsilon)$ -approximation* [11, 17] is a subset  $Z \subseteq X$  such that  $|\overline{X}(h) - \overline{Z}(h)| \leq \epsilon \cdot \max\{\rho, \overline{X}(h)\}$  holds for all ranges  $h \in \mathcal{R}$ . Hence, the  $(\rho, \epsilon)$ -guarantee can be fulfilled by simply setting  $\tau$  to  $\overline{Z}(h)$ , rendering  $Z$  a legal (sample-based)  $(\rho, \epsilon)$ -summary. Strengthening earlier results [3, 12, 21], Li et al. [17] proved that a random sample of  $X$  with size  $O(\frac{1}{\rho} \cdot \frac{1}{\epsilon^2} (\lambda \log \frac{1}{\rho} + \log \frac{1}{\delta}))$  is a relative  $(\rho, \epsilon)$ -approximation with probability at least  $1 - \delta$ . This implies the existence of a  $(\rho, \epsilon)$ -summary of size  $O(\frac{1}{\rho} \cdot \frac{\lambda}{\epsilon^2} \log \frac{1}{\rho})$ .

A range space  $(X, \mathcal{R}|_X)$  of a constant VC-dimension is said to be *well-behaved*, if  $\mathcal{R}|_X$  contains at most  $O(|X|) \cdot k^{O(1)}$  sets of size not exceeding  $k$ , for any integer  $k$  from 1 to  $|X|$ . Ezra [6] showed that such a range space admits a sample-based  $(\rho, \epsilon)$ -summary of size  $O(\frac{1}{\rho} \cdot \frac{1}{\epsilon^2} (\log \frac{1}{\epsilon} + \log \log \frac{1}{\rho}))$ ; note that this is smaller than the corresponding result  $O(\frac{1}{\rho} \cdot \frac{1}{\epsilon^2} \log \frac{1}{\rho})$  of [17] when  $\rho \ll \epsilon$ . It is worth mentioning that, when  $d \leq 3$  and  $\mathcal{R}$  is the halfspace family, any  $(X, \mathcal{R}|_X)$  is well-behaved; this, however, is not true when  $d \geq 4$ .

As opposed to the above “generic” bounds, Har-Peled and Sharir [11] proved specific bounds on the halfspace family  $\mathcal{R}$ . For  $d = 2$ , they showed that any  $(X, \mathcal{R})$  has a relative  $(\rho, \epsilon)$ -approximation of size  $O(\frac{1}{\rho} \cdot \frac{1}{\epsilon^{4/3}} \log^{4/3} \frac{1}{\rho\epsilon})$ ; similarly, for  $d = 3$ , the bound becomes  $O(\frac{1}{\rho} \cdot \frac{1}{\epsilon^{3/2}} \log^{3/2} \frac{1}{\rho\epsilon})$ . Combining these results and those of [6] gives the currently best bounds for these range spaces.

### 1.1.2 A Lower Bound of $\Omega(1/\rho)$

Notice that all the above bounds contain a term  $1/\rho$ . This is not a coincidence, but instead is due to a connection to “ $\epsilon$ -nets”. Given a range space  $(X, \mathcal{R}|_X)$ , an  $\epsilon$ -net [13] is a subset  $Z \subseteq X$  such that  $\overline{Z}(h) > 0$  holds for any range  $h \in \mathcal{R}$  satisfying  $\overline{X}(h) \geq \epsilon$ . As can be verified easily, any sample-based  $(\rho, 1/2)$ -summary of  $(X, \mathcal{R}|_X)$  must also be a  $\rho$ -net. This implies that the smallest size of sample-based  $(\rho, 1/2)$ -summaries must be at least that of  $\rho$ -nets.

Regarding the sizes of  $\epsilon$ -nets, a lower bound of  $\Omega(\frac{1}{\epsilon} \log \frac{1}{\epsilon})$  is known for many range families  $\mathcal{R}$  (see [14, 15, 20] and the references therein). More precisely, this means that, for each such family  $\mathcal{R}$ , one cannot hope to obtain an  $\epsilon$ -net of size  $o(\frac{1}{\epsilon} \log \frac{1}{\epsilon})$  for *every* possible  $X$ . It thus

follows that, for these families  $\mathcal{R}$ ,  $\Omega(\frac{1}{\rho} \log \frac{1}{\rho})$  is a lower bound on the sizes of sample-based  $(\rho, 1/2)$ -summaries. This, in turn, indicates that range spaces  $(X, \mathcal{R}|_X)$  defined by such an  $\mathcal{R}$  cannot always be well-behaved.

Coming back to the sizes of  $\epsilon$ -nets, a weaker lower bound of  $\Omega(1/\epsilon)$  holds quite commonly even on the range families that evade the  $\Omega(\frac{1}{\epsilon} \log \frac{1}{\epsilon})$  bound. Consider, for example, the halfspace family  $\mathcal{R}$  in  $\mathbb{R}^2$ . For any  $X$ , the range space  $(X, \mathcal{R}|_X)$  definitely has an  $\epsilon$ -net of size  $O(1/\epsilon)$  [10, 19]. This is tight: place a set  $X$  of points on the boundary of a circle; and it is easy to show that any  $\epsilon$ -net of  $(X, \mathcal{R}|_X)$  must have a size of at least  $1/\epsilon$ . This means that the size of any sample-based  $(\rho, 1/2)$ -summary of  $(X, \mathcal{R}|_X)$  must be  $\Omega(1/\rho)$ .

## 1.2 Our results

### 1.2.1 On One Input: Moving Beyond $\Omega(1/\rho)$

The  $\Omega(1/\rho)$  lower bound discussed earlier holds only in the *worst case*, i.e., it is determined by the “hardest”  $X$ . For other  $X$ , the range space  $(X, \mathcal{R}|_X)$  may admit much smaller  $(\rho, \epsilon)$ -summaries. For example, let  $\mathcal{R}$  be again the set of halfplanes in  $\mathbb{R}^2$ . When all the points of  $X$  lie on a line,  $(X, \mathcal{R})$  has a sample-based  $(\rho, \epsilon)$ -approximation of size only  $O(\frac{1}{\epsilon} \log \frac{1}{\rho})$ ; also, it would be interesting to note that  $(X, \mathcal{R})$  has an  $\epsilon$ -net that contains only 2 points! In general, the existing bounds on  $(\rho, \epsilon)$ -summaries can be excessively loose on individual inputs  $X$ . This calls for an alternative, *distribution-sensitive*, analytical framework that is able to prove tighter bounds using extra complexity parameters that depend on the *content* of  $X$ .

The first contribution of this paper is to establish such a framework by resorting to the concept of *disagreement coefficient* [8] from active learning. This notion was originally defined in a context different from ours; and we will adapt it to range spaces in the next section. At this moment, it suffices to understand that the disagreement coefficient  $\theta$  is a real value satisfying:  $1 \leq \theta \leq 1/\rho$ . The coefficient quantifies the differences among the sets in  $\mathcal{R}|_X$  (a larger  $\theta$  indicates greater differences). Even under the same  $\mathcal{R}$ ,  $\theta$  may vary considerably depending on  $X$ .

We will show that, for any range space  $(X, \mathcal{R}|_X)$  of VC-dimension  $\lambda$ , there is a  $(\rho, \epsilon)$ -summary that keeps  $O(\log(1/\rho))$  integers plus

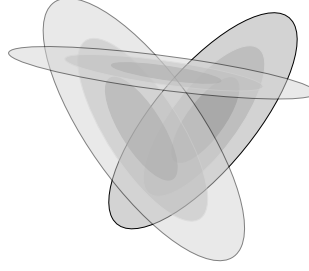
$$O\left(\min\left\{\frac{1}{\rho}, \theta \log \frac{1}{\rho}\right\} \cdot \frac{\lambda}{\epsilon^2} \log \frac{1}{\rho}\right) \quad (2)$$

points of  $X$ . The above is never worse than the general bound  $O(\frac{1}{\rho} \cdot \frac{\lambda}{\epsilon^2} \log \frac{1}{\rho})$  of relative  $(\rho, \epsilon)$ -approximations.

We will also prove that  $\Omega(\theta)$  is a lower bound on the number of words needed to encode a  $(\rho, 1/2)$ -summary even when  $d = 2$  and  $\lambda \leq 3$ . This generalizes the  $\Omega(1/\rho)$  lower bound in Section 1.1 because  $\theta$  is at most, but can reach,  $1/\rho$ . Thus, our result in (2) reflects the hardness of the input, and is tight within polylog factors for constant  $\epsilon$ . Our lower bound is information-theoretic, and does not require the summary to be sample-based.

### 1.2.2 On a Distribution of Inputs: Small Summaries for Halfspaces

Our framework allows us to explain – for the first time we believe – why  $\Omega(1/\rho) \cdot \text{poly}(1/\epsilon)$  is too pessimistic a bound on the sizes of  $(\rho, \epsilon)$ -summaries for inputs encountered in practice. For this purpose, we *must not* allow arbitrary inputs because of the prevalent  $\Omega(1/\rho)$  lower bound; instead, we will examine a class of inputs following a certain distribution.



■ **Figure 1** Mixture of 3 truncated Gaussian distributions in 2D space.

In this paper, we demonstrate the above by concentrating on the family  $\mathcal{R}$  of halfspaces in  $\mathbb{R}^d$  where the dimensionality  $d$  is a constant; this is arguably the “most-studied” family in the literature of relative  $(\rho, \epsilon)$ -approximations. The core of our solutions concerns two stochastic distributions that have drastically different behavior:

- **(Box Uniform)** Suppose that  $X$  is obtained by drawing  $n$  points uniformly at random from the *unit box*  $[0, 1]^d$ . When  $\rho = \Omega(\frac{\log n}{n})$ , we will prove that  $\theta = O(\text{polylog } \frac{1}{\rho})$  with high probability (i.e., at least  $1 - 1/n^2$ ). Accordingly, (2) becomes  $O(\frac{1}{\epsilon^2} \text{polylog } \frac{1}{\rho})$ , improving the general bound of relative  $(\rho, \epsilon)$ -approximations by almost a factor of  $O(1/\rho)$ .
- **(Ball Uniform)** Consider instead that the  $n$  points are drawn uniformly at random from the *unit ball*:  $\{x \in \mathbb{R}^d \mid \sum_{i=1}^d x[i]^2 \leq 1\}$ , where  $x[i]$  represents the  $i$ -th coordinate of point  $x$ . This time, we will prove that  $\theta = O((\frac{1}{\rho})^{\frac{d-1}{d+1}})$  with high probability for  $\rho = \Omega(\frac{\log n}{n})$ . (2) indicates the existence of a  $(\rho, \epsilon)$ -summary with size  $\tilde{O}((\frac{1}{\rho})^{\frac{d-1}{d+1}} \cdot \frac{1}{\epsilon^2})$  for  $\rho = \Omega(\frac{\log n}{n})$ , again circumventing the  $\Omega(1/\rho)$  lower bound.

The very same bounds can also be obtained in *non-uniform* settings. Suppose that  $X$  is obtained by drawing  $n$  points in an *iid* manner, according to a distribution that can be described by a probabilistic density function (pdf)  $\pi(x)$  over  $\mathbb{R}^d$  where  $d = O(1)$ . Define the *support region* of  $\pi$  as  $\text{supp}(\pi) = \{x \in \mathbb{R}^d \mid \pi(x) > 0\}$ . When  $\pi$  satisfies:

- *C1*:  $\text{supp}(\pi)$  is the unit box (or unit ball, resp.);
- *C2*: for every point  $x \in \text{supp}(\pi)$ , it holds that  $\pi(x) = \Omega(1)$ ;

we will show that  $(X, \mathcal{R}|_X)$  has a  $(\rho, \epsilon)$ -summary whose size is asymptotically the same as the aforementioned bound for box uniform (or ball uniform, resp.). Conditions *C1* and *C2* are satisfied by many distributions encountered in practice (e.g., the truncated versions of the Gaussian, Elliptical, and Laplace distributions, etc.), suggesting that real-world datasets may have much smaller  $(\rho, \epsilon)$ -summaries than previously thought.

Even better, the linearity of halfspaces implies that, the same bounds still hold even when the shape of  $\text{supp}(\pi)$  in Condition *C1* is obtained from the unit box/ball by an affine transformation. Call a distribution *atomic* if it satisfies *C1* (perhaps after an affine transformation) and *C2*. Our results hold also on “composite distributions” synthesized from a constant  $z$  number of atomic distributions whose support regions may overlap *arbitrarily*. Specifically, let  $\pi_1, \pi_2, \dots, \pi_z$  be the pdfs of atomic distributions; and define  $\pi(x) = \sum_{i=1}^z \gamma_i \cdot \pi_i(x)$ , for arbitrary positive constants  $\gamma_1, \gamma_2, \dots, \gamma_z$  that sum up to 1; see Figure 1 for an example. Then, the  $(\rho, \epsilon)$ -summary bound on  $\pi$  is asymptotically determined by the highest of the  $(\rho, \epsilon)$ -summary bounds on  $\pi_1, \dots, \pi_z$ .

## 2 Disagreement coefficients

### 2.1 Existing Definitions on Distributions

Disagreement coefficient was introduced by Hanneke [8] to analyze active learning algorithms (although a similar concept had been coined earlier [1] in statistics).

Let  $\mathcal{D}$  be a distribution over  $\mathbb{R}^d$ . For any region  $A \subseteq \mathbb{R}^d$ , we denote by  $\Pr_{\mathcal{D}}[A]$  the probability of  $x \in A$  when  $x$  is drawn from  $\mathcal{D}$ . Let  $\mathcal{R}$  be a family of geometric ranges. Given a subset  $\mathcal{R}' \subseteq \mathcal{R}$ , define the *disagreement region*  $DIS(\mathcal{R}')$  of  $\mathcal{R}'$  as

$$DIS(\mathcal{R}') = \{x \in \mathbb{R}^d \mid \exists h_1, h_2 \in \mathcal{R}' \text{ s.t. } x \in h_1 \text{ and } x \notin h_2\}.$$

That is,  $DIS(\mathcal{R}')$  includes every such point  $x \in \mathbb{R}^d$  that does not fall in all the ranges in  $\mathcal{R}'$ , and in the meantime, does not fall outside all the ranges in  $\mathcal{R}'$ , either. Given a range  $h \in \mathcal{R}$  and a real value  $r > 0$ , define its *r-ball*  $B_{\mathcal{D}}(h, r)$  as the set of all ranges  $h' \in \mathcal{R}$  satisfying  $\Pr_{\mathcal{D}}[DIS(\{h, h'\})] \leq r$ . It is worth mentioning that  $DIS(\{h, h'\})$  is simply the symmetric difference between  $h$  and  $h'$ .

Now, fix a range  $h$ , and consider increasing  $r$  continuously; this can only expand the set  $B_{\mathcal{D}}(h, r)$ , and hence, also  $DIS(B_{\mathcal{D}}(h, r))$ . Interestingly, even though  $\Pr_{\mathcal{D}}[DIS(B_{\mathcal{D}}(h, r))]$  is monotonically increasing, the ratio  $\Pr_{\mathcal{D}}[DIS(B_{\mathcal{D}}(h, r))]/r$  may remain bounded by a certain quantity. Given a real value  $\sigma \geq 0$ , the *disagreement coefficient*  $\theta_{\mathcal{D}}^h(\sigma)$  of  $h$  measures this quantity with respect to all  $r > \sigma$ :

$$\theta_{\mathcal{D}}^h(\sigma) = \max \left\{ 1, \sup_{r > \sigma} \frac{\Pr_{\mathcal{D}}[DIS(B_{\mathcal{D}}(h, r))]}{r} \right\}. \tag{3}$$

The function  $\theta_{\mathcal{D}}^h(\sigma)$  has several useful properties:

1. By definition,  $\theta_{\mathcal{D}}^h(\sigma)$  is between 1 and  $1/\sigma$ , regardless of  $\mathcal{D}$  and  $h$ .
2. The supremum in (3) ensures that  $\theta_{\mathcal{D}}^h(\sigma)$  is monotonically decreasing.
3. For any  $c \geq 1$ , it holds that  $\theta_{\mathcal{D}}^h(\sigma) \leq c \cdot \theta_{\mathcal{D}}^h(c\sigma)$  (see Corollary 7.2 of [9]).

### 2.2 New Definitions on Range Spaces

The above definitions rely on  $\mathcal{D}$ , and are not suitable for our problem settings where the input  $X$  is a finite set. Next, we present a way to adapt the definitions to a range space  $(X, \mathcal{R}|_X)$  for analyzing geometric algorithms.

We impose a uniform distribution over  $X$ : let  $\mathcal{U}(X)$  be the distribution of a random point drawn uniformly from  $X$ . By replacing  $\mathcal{D}$  with  $\mathcal{U}(X)$  in (3), we rewrite (3) into the following for any  $\sigma \geq 0$ :

$$\theta_{\mathcal{U}(X)}^h(\sigma) = \max \left\{ 1, \sup_{r > \sigma} \frac{\Pr_{\mathcal{U}(X)}[DIS(B_{\mathcal{U}(X)}(h, r))]}{r} \right\}. \tag{4}$$

Set

$$\sigma_{min} = \frac{\min_{h \in \mathcal{R}} |X \cap h|}{n} \tag{5}$$

We define the *disagreement coefficient of the range space*  $(X, \mathcal{R}|_X)$  as a function  $\theta_X(\sigma) : [\sigma_{min}, \infty) \rightarrow \mathbb{R}$  where

$$\theta_X(\sigma) = \min_{h \in \mathcal{R} \text{ s.t. } \overline{X}(h) \leq \sigma} \left\{ \theta_{\mathcal{U}(X)}^h(\sigma) \right\}. \tag{6}$$

It is clear from the above discussion that  $1 \leq \theta_X(\sigma) \leq 1/\sigma$  and  $\theta_X(\sigma)$  is monotonically decreasing.

As a remark, the finiteness of  $X$  gives a simpler interpretation of the  $r$ -ball  $B_{\mathcal{U}(X)}(h, r)$ : it is the set of ranges  $h' \in \mathcal{R}$  such that  $DIS(\{h, h'\})$  covers no more than  $r|X|$  points in  $X$ . Also,  $\Pr_{\mathcal{U}(X)}[A]$  for any region  $A \subseteq \mathbb{R}^d$  is simply  $|X \cap A|/|X|$ .

### 3 Small $(\rho, \epsilon)$ -summaries based on disagreement coefficients

Given a range space  $(X, \mathcal{R})$  with VC-dimension  $\lambda$ , we will show how to find a  $(\rho, \epsilon)$ -summary whose size can be bounded using disagreement coefficients. Our algorithm is randomized, and succeeds with probability at least  $1 - \delta$  for a real-valued parameter  $0 < \delta < 1$ . Set  $n = |X|$ . We require that  $\rho \geq \sigma_{min}$ ; otherwise, manually increasing  $\rho$  to  $\sigma_{min}$  achieves the same approximation guarantee.

#### 3.1 Algorithms

##### 3.1.1 Computing a $(\rho, \epsilon)$ -Summary

We will shrink  $\mathcal{R}$  progressively by removing a range  $h$  from  $\mathcal{R}$  once we are sure we can provide an accurate estimate for  $\bar{X}(h)$ . Define  $\mathcal{R}_0 = \mathcal{R}$ . We perform at most  $\lceil \log(1/\rho) \rceil$  rounds. Given  $\mathcal{R}_{i-1}$ , Round  $i \geq 1$  is executed as follows:

1.  $m_i \leftarrow$  the number of points  $x \in X$  such that  $x$  falls in *all* the ranges in  $\mathcal{R}_{i-1}$
2.  $X_i \leftarrow X \cap DIS(\mathcal{R}_{i-1})$
3. draw a set  $S_i$  of points uniformly at random from  $X_i$  with

$$|S_i| = O\left(\frac{|X_i|}{n} \cdot \frac{2^i}{\epsilon^2} \left(\lambda \log \frac{1}{\rho} + \log \frac{\log(1/\rho)}{\delta}\right)\right) \quad (7)$$

4.  $\mathcal{R}_i = \{h \in \mathcal{R}_{i-1} \mid \bar{S}_i(h) \cdot |X_i| + m_i < n/2^i\}$

The algorithm terminates when either  $i = \lceil \log(1/\rho) \rceil$  or  $\mathcal{R}_i = \emptyset$ . Suppose that in total  $t$  rounds are performed. The final  $(\rho, \epsilon)$ -summary consists of sets  $S_1, S_2, \dots, S_t$ , and  $2t + 1$  integers  $n, m_1, m_2, \dots, m_t, |X_1|, |X_2|, \dots, |X_t|$ .

##### 3.1.2 Performing Estimation

Given a range  $h \in \mathcal{R}$ , we deploy the summary to estimate  $\bar{X}(h)$  in two steps:

1.  $j \leftarrow$  the largest  $i \in [1, t]$  such that  $h \in \mathcal{R}_i$
2. return  $\bar{S}_j(h) \cdot \frac{|X_j|}{n} + \frac{m_j}{n}$  as the estimate

Regarding Step 1, whether  $h \in \mathcal{R}_i$  can be detected as follows. First, if  $h \notin \mathcal{R}_{i'}$  for any  $i' < i$ , then immediately  $h \notin \mathcal{R}_i$ . Otherwise, compute  $\bar{S}_i(h)$ , and declare  $h \in \mathcal{R}_i$  if and only if  $\bar{S}_i(h) \cdot |X_i| + m_i < n/2^i$ .

#### 3.2 Analysis

We now proceed to prove the correctness of our algorithms, and bound the size of the produced summary. It suffices to consider  $\epsilon \leq 1/3$  (otherwise, lower  $\epsilon$  to  $1/3$  and then apply the argument below).

The subsequent discussion is carried out under the event that, for every  $i \in [1, t]$ ,  $S_i$  is a relative  $(\rho_i, \epsilon/4)$ -approximation of  $X_i$  with respect to the ranges in  $\mathcal{R}$  where

$$\rho_i = \frac{n(1 + \epsilon)}{2^i \cdot |X_i|}.$$

By the result of [17] (reviewed in Section 1.1), with  $|S_i|$  shown in (7), the event happens with a probability at least  $1 - \delta \cdot \frac{t}{\lceil \log(1/\rho) \rceil} \geq 1 - \delta$ .

### 3.2.1 Correctness

To show that our algorithm indeed outputs a  $(\rho, \epsilon)$ -summary, we prove in the full version:

► **Lemma 1.** *The following are true for all  $i \in [1, t]$ : (i) for every range  $h \in \mathcal{R}_i$ ,  $\bar{X}(h) < (1 + \epsilon)/2^i$ ; (ii) for every range  $h \notin \mathcal{R}_i$ ,  $\bar{X}(h) \geq (1 - \epsilon)/2^i$ .*

Now consider the estimation algorithm in Section 3.1.2. Given the value  $j$  obtained at Step 1 for the input range  $h \in \mathcal{R}$ , the above lemma suggests that

$$(1 - \epsilon)/2^{j+1} \leq \bar{X}(h) < (1 + \epsilon)/2^j.$$

This, together with  $S_j$  being a  $(\rho_j, \epsilon/4)$ -approximation of  $X_j$ , ensures that our estimate satisfies the  $(\rho, \epsilon)$ -guarantee for  $h$ . The details can be found in the full version.

### 3.2.2 Bounding the Size

To bound the size of our  $(\rho, \epsilon)$ -summary, we will focus on bounding  $\sum_{i=1}^t |S_i|$ , because the rest of the summary clearly needs  $O(t) = O(\log(1/\rho))$  extra integers. Let us start with a trivial bound that follows directly from  $|X_i| \leq n$ :

$$\begin{aligned} \sum_{i=1}^t |S_i| &= O\left(\sum_{i=1}^t \frac{2^i}{\epsilon^2} \left(\lambda \log \frac{1}{\rho} + \log \frac{\log(1/\rho)}{\delta}\right)\right) \\ &= O\left(\frac{1}{\rho \epsilon^2} \left(\lambda \log \frac{1}{\rho} + \log \frac{\log(1/\rho)}{\delta}\right)\right). \end{aligned} \tag{8}$$

Next, we use disagreement coefficients to prove a tighter bound. Fix  $h \in \mathcal{R}$  to be an arbitrary range such that  $\bar{X}(h) \leq \rho$  ( $h$  definitely exists because  $\rho \geq \sigma_{min}$ ).

► **Lemma 2.**  $\mathcal{R}_i \subseteq B(h, \rho + (1 + \epsilon)/2^i)$ .

**Proof.** It suffices to prove that, for any  $h' \in \mathcal{R}_i$ ,  $\Pr_{U(X)}[DIS(\{h, h'\})] \leq \rho + (1 + \epsilon)/2^i$ , or equivalently,  $|X \cap DIS(\{h, h'\})| \leq n(\rho + (1 + \epsilon)/2^i)$ .

This holds because  $|X \cap DIS(\{h, h'\})| \leq |(X \cap h) \cup (X \cap h')|$ . By definition of  $h$ , we know  $|X \cap h| \leq n\rho$ , while By Lemma 1, we know  $|X \cap h'| \leq n(1 + \epsilon)/2^i$ . Therefore,  $|X \cap DIS(\{h, h'\})| \leq n(\rho + (1 + \epsilon)/2^i)$ . ◀

► **Lemma 3.**  $|X_i|/n \leq \theta_{U(X)}^h(2\rho) \cdot (\rho + \frac{1+\epsilon}{2^{i-1}})$ .

**Proof.** Lemma 2 tells us that  $DIS(\mathcal{R}_{i-1}) \subseteq DIS(B_{\mathcal{U}(X)}(h, \rho + \frac{1+\epsilon}{2^{i-1}}))$ . Thus:

$$\begin{aligned} |X_i|/n &= \Pr_{\mathcal{U}(X)}(DIS(\mathcal{R}_{i-1})) \\ &\leq \Pr_{\mathcal{U}(X)}\left(DIS\left(B_{\mathcal{U}(X)}\left(h, \rho + \frac{1+\epsilon}{2^{i-1}}\right)\right)\right) \\ \text{(by (4))} &\leq \theta_{\mathcal{U}(X)}^h\left(\rho + \frac{1+\epsilon}{2^{i-1}}\right) \cdot \left(\rho + \frac{1+\epsilon}{2^{i-1}}\right) \end{aligned}$$

By  $1/2^{i-1} > \rho$ , and the fact that  $\theta_{\mathcal{U}(X)}^h$  is monotonically decreasing, the above leads to

$$\begin{aligned} \theta_{\mathcal{U}(X)}^h\left(\rho + \frac{1+\epsilon}{2^{i-1}}\right) \cdot \left(\rho + \frac{1+\epsilon}{2^{i-1}}\right) &\leq \theta_{\mathcal{U}(X)}^h(\rho + \rho) \cdot \left(\rho + \frac{1+\epsilon}{2^{i-1}}\right) \\ &= \theta_{\mathcal{U}(X)}^h(2\rho) \cdot \left(\rho + \frac{1+\epsilon}{2^{i-1}}\right). \end{aligned} \quad \blacktriangleleft$$

Therefore:

$$\begin{aligned} \sum_{i=1}^t \frac{|X_i| \cdot 2^i}{n} &\leq \theta_{\mathcal{U}(X)}^h(2\rho) \cdot \sum_{i=1}^t 2^i \cdot \left(\rho + \frac{1+\epsilon}{2^{i-1}}\right) \\ &= \theta_{\mathcal{U}(X)}^h(2\rho) \cdot \sum_{i=1}^t (2^i \cdot \rho + O(1)) \\ \text{(by } 1/2^i = \Omega(\rho)) &= \theta_{\mathcal{U}(X)}^h(2\rho) \cdot O(t) \\ &= \theta_{\mathcal{U}(X)}^h(2\rho) \cdot O(\log(1/\rho)) \\ &= \theta_{\mathcal{U}(X)}^h(\rho) \cdot O(\log(1/\rho)) \end{aligned} \quad (9)$$

where the last equality used the fact that  $\theta_{\mathcal{U}(X)}^h(2\rho) \leq 2 \cdot \theta_{\mathcal{U}(X)}^h(\rho)$ .

Remember that the above holds for *all*  $h \in \mathcal{R}$  satisfying  $\overline{X}(h) \leq \rho$ . By the definition in (6), we can improve the bound of (9) to

$$\sum_{i=1}^t \frac{|X_i| \cdot 2^i}{n} = \theta_X(\rho) \cdot O(\log(1/\rho)). \quad (10)$$

Combining the above with (7) gives  $\sum_{i=1}^t |S_i| = O(\frac{1}{\epsilon^2} \cdot \theta_X(\rho) \log(1/\rho) \cdot (\lambda \log(1/\rho) + \log \frac{\log(1/\rho)}{\delta}))$ . Putting this together with (8) and setting  $\delta$  to a constant gives:

► **Theorem 4.** *For any  $\rho \geq \sigma_{min}$  and any  $0 < \epsilon < 1$ , a range space  $(X, \mathcal{R}|_X)$  of VC-dimension  $\lambda$  has a  $(\rho, \epsilon)$ -summary which keeps  $O(\log(1/\rho))$  integers and  $O(\min\{\frac{1}{\rho}, \theta_X(\rho) \cdot \log \frac{1}{\rho}\} \cdot \frac{\lambda}{\epsilon^2} \log \frac{1}{\rho})$  points of  $X$ . Here,  $\sigma_{min}$  is defined in (5), and  $\theta_X$  is the disagreement coefficient function defined in (6).*

### 3.2.3 A Remark

Our  $(\rho, \epsilon)$ -summary is currently not sample-based, but this can be fixed by keeping – at Step 1 of the computation algorithm in Section 3.1 – an arbitrary point counted by  $m_i$ .

The  $(\rho, \epsilon)$ -summary after the fix also serves as a  $\rho$ -net. Thus, by setting  $\epsilon$  to a constant in Theorem 4, we know that for any  $\rho \geq \sigma_{min}$ , the range space  $(X, \mathcal{R}|_X)$  in Theorem 4 has an  $\rho$ -net of size  $O(\min\{\frac{1}{\rho}, \theta_X(\rho) \cdot \log \frac{1}{\rho}\} \cdot \lambda \log \frac{1}{\rho})$ . However, it should be pointed out that this bound on  $\rho$ -nets can be slightly improved, as is implied by Theorem 5.1 of [9] and made explicit in [16].



**4 Bridging distribution and finite-set disagreement coefficients**

This section will establish another theorem which will be used together with Theorem 4 to explain why we are able to obtain  $(\rho, \epsilon)$ -summarizes of  $o(1/\rho)$  size on practical datasets. Suppose that the input  $X$  has been generated by taking  $n$  points independently following the same distribution  $\mathcal{D}$  over  $\mathbb{R}^d$ . The learning literature (see, e.g., [9]) has developed a solid understanding on when the quantity  $\theta_{\mathcal{D}}^h(\sigma)$  is small. Unfortunately, those findings can rarely be applied to  $\theta_{\mathcal{U}(X)}^h(\sigma)$  because they are conditioned on requirements that must be met by  $\mathcal{D}$ , e.g., one common requirement is continuity.  $\mathcal{U}(X)$ , due to its discrete nature, seldom meets the requirements.

On the other hand, clearly  $\mathcal{U}(X)$  approximates  $\mathcal{D}$  increasingly better as  $n$  grows. Thus, we ask the question:

How large  $n$  needs to be for  $\theta_{\mathcal{U}(X)}^h(\sigma)$  to be asymptotically the same as  $\theta_{\mathcal{D}}^h(\sigma)$ ?

We partially answer the question in the next theorem:

► **Theorem 5 (The Bridging Theorem).** *Let  $\mathcal{D}$  be a distribution over  $\mathbb{R}^d$ , and  $\mathcal{R}$  be a family of ranges. Denote by  $\lambda$  the VC-dimension of the range space  $(\mathbb{R}^d, \mathcal{R})$ .*

*Fix an arbitrary range  $h \in \mathcal{R}$ , an arbitrary integer  $n$ , a real value  $0 < \delta < 1$ , a real value  $\sigma$  satisfying  $n \geq \frac{c}{\sigma} (\log \frac{n}{\delta} + \lambda \log \frac{1}{\sigma})$  for some universal constant  $c$ . If we draw a set  $X$  of  $n$  points independently from  $\mathcal{D}$ , then with probability at least  $1 - \delta$ , it holds that  $\theta_{\mathcal{U}(X)}^h(\sigma) \leq 8 \cdot \theta_{\mathcal{D}}^h(2\sigma)$ .*

The rest of the section serves as a proof of the theorem. Let us first get rid of two easy cases:

- If  $\sigma \geq 1$ ,  $\theta_{\mathcal{U}(X)}^h(\sigma) = \theta_{\mathcal{D}}^h(2\sigma) = 1$  by definition of (4); and the theorem obviously holds.
- If  $\sigma < 1/n$ , observe that every range  $h' \in B_{\mathcal{U}(X)}(h, \sigma)$  covers exactly the same set of points in  $X$  as  $h$ . Hence,  $\Pr_{\mathcal{U}(X)}[DIS(B_{\mathcal{U}(X)}(h, r))] = 0$ . It follows from (4) that  $\theta_{\mathcal{U}(X)}^h(\sigma) = 1$ . The theorem again obviously holds because  $\theta_{\mathcal{D}}^h(2\sigma) \geq 1$ , by definition.

Hence, it suffices to consider  $1/n \leq \sigma < 1$ . Define  $S = \{i/n \mid i \text{ is an integer in } [\sigma n, n]\}$ . For  $\sigma \geq 1/n$ , (4) implies

$$\theta_{\mathcal{U}(X)}^h(\sigma) \leq \max \left\{ 1, 2 \cdot \max_{r \in S} \frac{\Pr_{\mathcal{U}(X)}[DIS(B_{\mathcal{U}(X)}(h, r))]}{r} \right\}. \tag{11}$$

Consider an arbitrary  $r \in S$ . We will show that, when  $n$  satisfies the condition in the theorem, with probability at least  $1 - \delta/n$ , it holds that

$$\frac{\Pr_{\mathcal{U}(X)}[DIS(B_{\mathcal{U}(X)}(h, r))]}{r} \leq 4 \cdot \theta_{\mathcal{D}}^h(2\sigma). \tag{12}$$

Once this is done, applying the union bound on all the  $r \in S$  will prove that (11) is at most  $8 \cdot \theta_{\mathcal{D}}^h(2\sigma)$  with probability at least  $1 - \delta$ , as claimed in the theorem.

We aim to establish the following equivalent form of (12):

$$\frac{|X \cap DIS(B_{\mathcal{U}(X)}(h, r))|}{nr} \leq 4 \cdot \theta_{\mathcal{D}}^h(2\sigma). \tag{13}$$

For the above purpose, the most crucial step is to prove:

► **Lemma 6.** *When  $n \geq \frac{c_1}{r} (\lambda \log \frac{1}{r} + \log \frac{n}{\delta})$  for some universal constant  $c_1$ , it holds with probability at least  $1 - \delta/(2n)$  that  $B_{\mathcal{U}(X)}(h, r) \subseteq B_{\mathcal{D}}(h, 2r)$ .*

**Proof.** The rationale of our proof is that any  $h' \notin B_{\mathcal{D}}(h, 2r)$  is unlikely to appear in  $B_{\mathcal{U}(X)}(h, r)$  when  $n$  is large. Indeed,  $h' \notin B_{\mathcal{D}}(h, 2r)$  indicates that a point  $x$  drawn from  $D$  has probability over  $2r$  to fall in  $DIS(\{h, h'\})$ . Hence,  $|X \cap DIS(\{h, h'\})|$  should be sharply concentrated around  $2r \cdot n$ , rendering  $h' \notin B_{\mathcal{U}(X)}(h, r)$ . The challenge, however, is that there can be an infinite number of ranges  $h'$  to consider. To tackle the challenge, we need to bring down the number of ranges somehow to  $n^{O(\lambda)}$ . We achieve the purpose by observing that we can define another range space with VC-dimension  $O(\lambda)$  to capture the disagreement regions of range pairs from  $\mathcal{R}$ , as shown below.

Define  $\mathcal{R}^{dis} = \{DIS(\{h, h'\}) \mid h, h' \in \mathcal{R}\}$ . We observe that the range space  $(\mathbb{R}^d, \mathcal{R}^{dis})$  has VC-dimension  $O(\lambda)$ . To explain why, for any  $h \in \mathcal{R}$ , define  $\bar{h} = \mathbb{R}^d \setminus h$ . Accordingly, define  $\bar{\mathcal{R}} = \{\bar{h} \mid h \in \mathcal{R}\}$ . The two range spaces  $(\mathbb{R}^d, \mathcal{R})$  and  $(\mathbb{R}^d, \bar{\mathcal{R}})$  have the same VC-dimension  $\lambda$ . Therefore, the range space  $(\mathbb{R}^d, \mathcal{R} \cup \bar{\mathcal{R}})$  has VC-dimension at most  $2\lambda + 1$ . Now apply a 2-fold intersection on  $(\mathbb{R}^d, \mathcal{R} \cup \bar{\mathcal{R}})$  to create  $(\mathbb{R}^d, \mathcal{R}_1)$  where  $\mathcal{R}_1 = \{h \cap h' \mid h, h' \in \mathcal{R} \cup \bar{\mathcal{R}}\}$ . By a result of [2], the VC dimension of  $(\mathbb{R}^d, \mathcal{R}_1)$  is bounded by  $O(\lambda)$ . Finally, apply a 2-fold union on  $(\mathbb{R}^d, \mathcal{R}_1)$  to create  $(\mathbb{R}^d, \mathcal{R}_2)$  where  $\mathcal{R}_2 = \{h \cup h' \mid h, h' \in \mathcal{R}_1\}$ . By another result of [2], the VC dimension of  $(\mathbb{R}^d, \mathcal{R}_2)$  is bounded by  $O(\lambda)$ . Notice that  $\mathcal{R}^{dis}$  is a subset of  $\mathcal{R}_2$ . It thus follows that the VC-dimension of  $(\mathbb{R}^d, \mathcal{R}^{dis})$  must be  $O(\lambda)$ .

Essentially, now the task is to draw a sufficiently large set  $X$  of points from  $\mathcal{D}$  to guarantee with probability at least  $1 - \delta/(2n)$ : for every range  $h \in \mathcal{R}^{dis}$  with  $\Pr_{\mathcal{D}}(h) > 2r$ , we ensure  $|X \cap h|/|X| > r$ . By applying a result of [17] on general range spaces, we know that  $|X|$  only needs to be  $\frac{c_1}{r}(\lambda \log \frac{1}{r} + \log \frac{n}{\delta})$  for some constant  $c_1$  which does not depend on  $r, \delta$ , and  $n$ .  $\blacktriangleleft$

Set  $r' = \Pr_{\mathcal{D}}(DIS(B_{\mathcal{D}}(h, 2r)))$ ; notice that, by definition of  $\theta_{\mathcal{D}}^h(2r)$ ,  $r' \leq 2r \cdot \theta_{\mathcal{D}}^h(2r)$ . We want to draw a sufficiently large set  $X$  of points from  $\mathcal{D}$  to guarantee, with probability at least  $1 - \delta/(2n)$ ,  $|X \cap DIS(B_{\mathcal{D}}(h, 2r))| \leq 2n \cdot \max\{r, r'\}$ . By Chernoff bounds,  $n$  only needs to be at least  $\frac{c_2}{r} \log \frac{n}{\delta}$  for some universal constant  $c_2$ .

Now, set  $c = \max\{c_1, c_2\}$  and  $n = \frac{c}{r}(\lambda \log \frac{1}{r} + \log \frac{n}{\delta})$ . With probability at least  $1 - \delta/n$ , we can derive (13) from the above discussion as follows:

$$\begin{aligned} \frac{|X \cap DIS(B_{\mathcal{U}(X)}(h, r))|}{nr} &\leq \frac{|X \cap DIS(B_{\mathcal{D}}(h, 2r))|}{nr} && \text{(by Lemma 6)} \\ &\leq \frac{2n \cdot \max\{r, r'\}}{nr} = 2 \cdot \max\{1, r'/r\} \\ &\leq 2 \cdot \max\{1, 2 \cdot \theta_{\mathcal{D}}^h(2r)\} = 4 \cdot \theta_{\mathcal{D}}^h(2r) \leq 4 \cdot \theta_{\mathcal{D}}^h(2\sigma) \end{aligned}$$

where the last inequality used  $r \geq \sigma$  and the fact that  $\theta_{\mathcal{D}}^h$  is monotonically decreasing. This establishes (13) and hence completes the proof of Theorem 5.

## 5 $o(1/\rho)$ -size summaries for halfspace ranges

We are ready to explain why a set of points generated from a stochastic distribution often admits  $(\rho, \epsilon)$ -summaries of  $o(1/\rho)$  size for fixed  $\epsilon$ . This requires specializing  $\mathcal{R}$  into a concrete range family. We will do so by constraining  $\mathcal{R}$  to be the set of halfspaces in  $\mathbb{R}^d$ , because this family has received considerable attention (as reviewed in Section 1.1).

We prove in the full version the next two technical lemmas regarding the disagreement coefficients on box-uniform and ball-uniform distributions:

**► Lemma 7.** *Let  $\mathcal{U}$  be the distribution where a point is drawn uniformly at random from the unit box  $[0, 1]^d$  with  $d = O(1)$ . For any halfspace  $h$  disjoint with the box, it holds that  $\theta_{\mathcal{U}}^h(\sigma) = O(\log^{d-1} \frac{1}{\sigma})$  for all  $\sigma > 0$ .*

► **Lemma 8.** Let  $\mathcal{U}$  be the distribution where a point is drawn uniformly at random from the unit ball  $\{x \in \mathbb{R}^d \mid \sum_{i=1}^d x[i]^2 \leq 1\}$  with  $d = O(1)$ . For any halfspace  $h$  disjoint with the ball, it holds that  $\theta_{\mathcal{U}}^h(\sigma) = O((\frac{1}{\sigma})^{\frac{d-1}{d+1}})$  for all  $\sigma > 0$ .

Next, we establish our main result for *non-uniform* distributions:

► **Theorem 9.** Let  $\mathcal{R}$  be the family of halfspaces in  $\mathbb{R}^d$  with a constant dimensionality  $d$ . Let  $\mathcal{D}$  be a distribution over  $\mathbb{R}^d$  such that the pdf  $\pi$  of  $\mathcal{D}$  satisfies Conditions C1 and C2 as prescribed in Section 1.2. Suppose that we draw a set  $X$  of  $n$  points independently from  $\mathcal{D}$ . Both of the following hold with probability at least  $1 - 1/n^2$ :

- When  $\text{supp}(\pi)$  is the unit box, for any  $0 < \epsilon < 1$  and any  $\rho \geq \frac{c \log n}{n}$  where  $c > 0$  is a constant,  $X$  has a  $(\rho, \epsilon)$ -summary that keeps  $O(\log(1/\rho))$  integers and  $O(\frac{1}{\epsilon^2} \log^{d+1} \frac{1}{\rho})$  points of  $X$ .
- When  $\text{supp}(\pi)$  is the unit ball, for any  $0 < \epsilon < 1$  and any  $\rho \geq \frac{c \log n}{n}$  where  $c > 0$  is a constant,  $X$  has a  $(\rho, \epsilon)$ -summary that keeps  $O(\log(1/\rho))$  integers and  $O(\frac{1}{\epsilon^2} \cdot (\frac{1}{\rho})^{\frac{d-1}{d+2}} \cdot \log^2 \frac{1}{\rho})$  points of  $X$ .

The constant  $c$  in the above does not depend on  $\mathcal{D}$ ,  $n$ ,  $\rho$ , and  $\epsilon$ .

**Proof.** We will prove only the case where  $\text{supp}(\pi)$  is the unit box because the unit-ball case is similar. Set  $\sigma^* = \frac{c \log n}{n}$  where  $c$  is some constant to be determined later. Thanks to Theorem 4, it suffices to prove that with probability at least  $1 - 1/n^2$ ,  $\theta_X(\rho) = O(\log^{d-1} \frac{1}{\rho})$  at every  $\rho \geq \sigma^*$ . We will argue that, with probability at least  $1 - 1/n^2$ , there exists a halfspace  $h \in \mathcal{R}$  such that  $\bar{X}(h) \leq \rho$  and  $\theta_{\mathcal{U}(X)}^h(\rho) = O(\log^{d-1} \frac{1}{\rho})$ . Once this is done, we know  $\theta_X(\rho) = O(\log^{d-1} \frac{1}{\rho})$  from (6).

Condition C<sub>2</sub> says that the pdf  $\pi$  satisfies  $\pi(x) \geq \gamma$  for any point  $x$  in  $\text{supp}(\pi)$  (i.e., the unit box), where  $\gamma$  is a positive constant. Remember that, by definition of  $\text{supp}(\pi)$ ,  $\pi(x) = 0$  for any  $x$  outside  $\text{supp}(\pi)$ .

Simply set  $h$  to a halfspace as stated in Lemma 7, i.e.,  $\theta_{\mathcal{U}}^h(\sigma) = O(\log^{d-1} \frac{1}{\sigma})$ . Let  $\pi_{\mathcal{U}}$  be the pdf of  $\mathcal{U}$ :  $\pi_{\mathcal{U}}(x)$  equals 1 if  $x \in [0, 1]^d$ , or 0 otherwise. Define  $\alpha$  as any constant such that  $\alpha \leq \gamma$ . We have  $\alpha \cdot \pi_{\mathcal{U}}(x) \leq \pi(x) \leq 1 \leq \frac{1}{\alpha} \cdot \pi_{\mathcal{U}}(x)$  for all  $x \in \mathbb{R}^d$ . Given this, Theorem 7.6 of [9] tells us that  $\theta_{\mathcal{D}}^h(\sigma) = O(\theta_{\mathcal{U}}^h(\sigma/\alpha))$ . It thus follows that  $\theta_{\mathcal{D}}^h(\sigma) = O(\log^{d-1} \frac{1}{\sigma})$  for all  $\sigma > 0$ .

Now, apply Theorem 5 on  $h$  by setting  $\delta = 1/n^2$  and  $\lambda = O(1)$ . The theorem shows that, when  $n \geq \frac{\beta \log n}{\rho}$  for some constant  $\beta$ ,  $\theta_{\mathcal{U}(X)}^h(\rho) \leq 8 \cdot \theta_{\mathcal{D}}^h(\rho) = O(\log^{d-1} \frac{1}{\rho})$  with probability at least  $1 - 1/n^2$ . We set  $c \geq \beta$  to ensure  $n \geq \frac{\alpha \log n}{\rho}$ . Note also that the choice of  $h$  guarantees  $\bar{X}(h) = 0 < \rho$ . This makes  $h$  a halfspace we are looking for, and concludes the proof. ◀

Some remarks are in order:

- (Composite Distributions) Let  $\mathcal{D}_1$  and  $\mathcal{D}_2$  be two distributions over  $\mathbb{R}^d$  with pdfs  $\pi_1$  and  $\pi_2$ , respectively (the support regions of  $\pi_1$  and  $\pi_2$  may overlap). Define a distribution  $\mathcal{D}$  with pdf  $\pi(x) = \gamma \cdot \pi_1(x) + (1 - \gamma) \cdot \pi_2(x)$ , for some constant  $0 < \gamma < 1$ . Theorem 7.7 of [9] tells us that, for any halfspace  $h \in \mathcal{R}$  and any  $\sigma > 0$ ,  $\theta_{\mathcal{D}}^h(\sigma) \leq \theta_{\mathcal{D}_1}^h(\frac{\sigma}{\gamma}) + \theta_{\mathcal{D}_2}^h(\frac{\sigma}{1-\gamma})$ . It thus follows from Lemma 7 that, when  $\mathcal{D}_1$  and  $\mathcal{D}_2$  are atomic distributions with support regions obtainable from the unit box through affine transformations,  $\theta_{\mathcal{D}}^h(\sigma) = O(\log^{d-2} \frac{1}{\sigma})$  for any  $h$  disjoint with  $\text{supp}(\mathcal{D}_1) \cup \text{supp}(\mathcal{D}_2)$ . The unit-box result of Theorem 9 can be easily shown to hold on this  $\mathcal{D}$  as well, by adapting the proof in a straightforward manner. The same is true for the unit-ball result of Theorem 9. All these results can now be extended to a composite distribution synthesized from a constant number of atomic distributions (see Section 1.2).

- (More Distributions with Near-Constant  $\theta$ ) What is given in Lemma 7 is only one scenario where  $\theta_{\mathcal{D}}^h(\sigma)$  is nearly a constant. There are other combinations of  $\mathcal{D}$  and  $\mathcal{R}$  where  $\theta_{\mathcal{D}}^h(\sigma) = \tilde{O}(1)$  for all  $h \in \mathcal{R}$ ; see [5, 7, 8, 9, 22] (in some of those combinations,  $\mathcal{R}$  may not contain all the halfspaces in  $\mathbb{R}^d$ ; e.g., a result of [8] concerns only the halfspaces whose boundary planes pass the origin). The proof of Theorem 9 can be adapted to show that  $X$  has a  $(\rho, \epsilon)$ -summary of size  $\tilde{O}(1/\epsilon^2)$  with high probability when  $\rho = \Omega(\max\{\frac{\log n}{n}, \min_{h \in \mathcal{R}} \Pr_{\mathcal{D}}(h)\})$ .
- (Time Complexity) In general, for any  $X$ , a  $(\rho, \epsilon)$ -summary (for the halfspace family  $\mathcal{R}$  in constant-dimensional space) can be found in polynomial time even by implementing the algorithm of Section 3.1 naively. The time can be improved to  $O(n \text{ polylog } n) + n^{1-\Omega(1)} \cdot s^{O(1)}$ , where  $s$  is the size of the returned summary, by utilizing specialized data structures [4, 18].

## 6 A lower bound with disagreement coefficients

In this section, we will prove a lower bound on the sizes of  $(\rho, 1/2)$ -summaries in relation to disagreement coefficients. Our core result is:

► **Theorem 10.** *Let  $\mathcal{R}$  be the family of all halfplanes in  $\mathbb{R}^2$ . Fix any integer  $w$  as the number of bits in a word. Choose arbitrary integers  $\eta, q$ , and  $k$  such that  $\eta \geq 4$ ,  $q$  is a multiple of  $\eta$ , and  $1 \leq k \leq q/(4\eta)$ . There must exist a set  $\mathcal{C}$  of range spaces  $(X, \mathcal{R}|_X)$ , each satisfying the following conditions:*

- $X$  is a set of  $q + k$  points in  $\mathbb{R}^2$ .
- The disagreement coefficient of  $(X, \mathcal{R}|_X)$  satisfies  $\theta_X(\frac{k}{q+k}) = \frac{k+q}{k+q/\eta}$ .
- Any encoding, which encodes a  $(\frac{k}{q+k}, 1/2)$ -summary for each range space in  $\mathcal{C}$ , must use at least  $\eta \cdot w$  bits on at least one range space in  $\mathcal{C}$ .

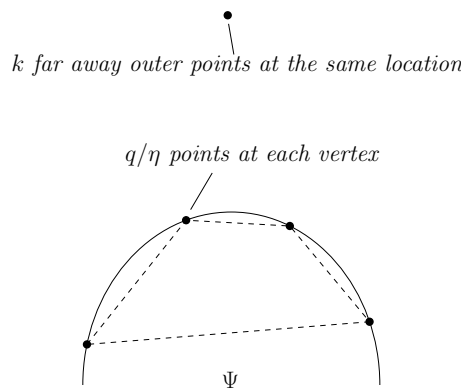
Therefore, for  $\rho = \frac{k}{q+k}$ , if one wishes to store a  $(\rho, 1/2)$ -summary for each range space in  $\mathcal{C}$ , at least  $\eta \cdot w$  bits (namely,  $\eta$  words) are needed on at least one range space. Since  $\theta_X(\rho) = \frac{k+q}{k+q/\eta} \leq \eta$ , this establishes  $\theta_X(\rho)$  as a space lower bound for  $(\rho, 1/2)$ -summaries. In the theorem, any  $X$  has dimensionality  $d = 2$  and any  $(X, \mathcal{R})$  has VC-dimension at most 3; hence, Theorem 4 is tight up to polylog factors on constant  $\lambda$  and  $\epsilon$ .

The flexibility of  $\eta, q$ , and  $k$  allows the lower bound to hold in a variety of more concrete settings. For example, by adjusting  $k$  and  $q$ , one sees that  $\theta_X(\rho)$  is a lower bound for the whole range of  $\rho \in (0, O(1)]$ . On the other hand, by focusing on any specific  $\rho \in (0, O(1)]$  but adjusting  $\eta$ , one sees that  $\theta_X(\rho)$  remains as a lower bound when  $\theta_X(\rho)$  goes from  $O(1)$  to  $\Omega(1/\rho)$ .<sup>1</sup>

**Proving Theorem 10.** Fix integers  $\eta, q$ , and  $k$  as stated in Theorem 10. Define  $n = q + k$ . Next, we construct a class  $\mathcal{X}$  of point sets, each consisting of  $n$  points in  $\mathbb{R}^2$ . First, place  $k$  points at coordinates  $(0, \infty)$ . Call them the *outer* points; they belong to all the sets in  $\mathcal{X}$ .

For each set  $X \in \mathcal{X}$ , we generate  $q$  extra *inner* points. For this purpose, place an arbitrary polygon  $\Psi$  with  $\eta$  vertices, making sure that all the vertices fall on the upper arc of the unit circle (i.e., the arc is  $\{(x[1], x[2]) \mid x[1]^2 + x[2]^2 = 1 \text{ and } x[2] \geq 0\}$ ). Then, given each vertex

<sup>1</sup> This rules out, for example, a claim of the form: “when  $\theta_X(\rho) \geq \sqrt{1/\rho}$ , there is a  $(\rho, 1/2)$ -summary of size  $O(\sqrt{\theta_X(\rho)})$ ”.



■ **Figure 2** A set of points in  $\mathcal{X}$  ( $\eta = 4$ ).

of  $\Psi$ , we add  $q/\eta$  inner points to  $X$ , all of which are located at that vertex. See Figure 2 for an example with  $\eta = 4$ . This finishes the construction of  $X$ . It is important to note that a different  $\Psi$  is used for each  $X$ . Thus,  $\mathcal{X}$  includes an infinite number of inputs, each corresponding to a possible  $\Psi$ . Our construction ensures a nice property:

► **Lemma 11.** *Fix any  $X \in \mathcal{X}$ . Given any  $(k/n, 1/2)$ -summary of  $X$ , we are able to infer all the vertices of  $\Psi$  used to construct  $X$ .*

**Proof.** We say that a halfplane in  $\mathcal{R}$  is *upward* if it covers the point  $(0, \infty)$ . Our aim is to prove that, the summary allows us to determine whether an arbitrary upward halfplane covers any inner point of  $X$ . This implies that we can reconstruct all the vertices of  $\Psi$  using the summary.<sup>2</sup>

Given an upward halfplane  $h$ , we use the summary to obtain an estimate – denoted as  $\tau$  – of  $\overline{X}(h)$ . If  $\tau \geq 2k/n$ , we return “yes” (i.e.,  $h$  covers at least one inner point); otherwise, we return “no”. To see that this is correct, first note that  $\overline{X}(h)$  must be at least  $k/n$ , and hence  $0.5\overline{X}(h) \leq \tau \leq 1.5\overline{X}(h)$ . Therefore, if  $h$  covers no inner points,  $\overline{X}(h) = k/n$ , indicating  $\tau < 1.5k/n$ . Otherwise,  $\overline{X}(h) \geq \frac{k+q/\eta}{n} \geq 5k/n$ , indicating  $\tau \geq 2.5k/n$ . ◀

We prove in the full version:

► **Lemma 12.** *For each  $X \in \mathcal{X}$ , the disagreement coefficient of  $(X, \mathcal{R}|_X)$  satisfies  $\theta_X(k/n) = \frac{n}{k+q/\eta}$ .*

The set  $\mathcal{C}$  is simply the set  $\{(X, \mathcal{R}) \mid X \in \mathcal{X}\}$ . Recall that each  $X \in \mathcal{X}$  corresponds to a distinct  $\eta$ -vertex polygon  $\Psi$ . Hence, by Lemma 11, the  $(k/n, 1/2)$ -summaries associated with the range spaces in  $\mathcal{C}$  serve as an encoding of all such  $\Psi$ 's.

So far the number of  $\Psi$ 's is infinite, which does not fit the purpose of arguing for a space lower bound in RAM with a finite word length  $w$ . This can be easily fixed by creating  $2^w$  choices for each vertex of  $\Psi$ , such that each of the  $\eta$  vertices can independently take a choice of its own. This generates  $2^{\eta w}$  polygons for  $\Psi$ , and hence, the same number of inputs in  $\mathcal{C}$ . Lemmas 11 and 12 are still valid. Therefore, any encoding, which encodes a  $(k/n, 1/2)$ -summary for each range space in  $\mathcal{C}$ , can be used to distinguish all those  $2^{\eta w}$  choices of  $\Psi$ . The encoding, therefore, must use  $\eta \cdot w$  bits for at least one range space. This completes the proof of Theorem 10. ◀

<sup>2</sup> To see this, consider any vertex  $v$  of  $\Psi$ , and use the summary to distinguish the line  $\ell$  tangent to the arc at  $v$  and a line that is parallel to  $\ell$ , but moves slightly away from the arc.

## References

- 1 Kenneth S. Alexander. Rates of Growth and Sample Moduli for Weighted Empirical Processes Indexed by Sets. *Probability Theory and Related Fields*, 75:379–423, 1987.
- 2 Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM (JACM)*, 36(4):929–965, 1989.
- 3 Hervé Brönnimann, Bernard Chazelle, and Jiří Matoušek. Product Range Spaces, Sensitive Sampling, and Derandomization. *SIAM Journal on Computing*, 28(5):1552–1575, 1999.
- 4 Timothy M. Chan. Optimal Partition Trees. *Discrete & Computational Geometry*, 47(4):661–690, 2012.
- 5 Ran El-Yaniv and Yair Wiener. Active Learning via Perfect Selective Classification. *Journal of Machine Learning Research*, 13:255–279, 2012.
- 6 Esther Ezra. Small-size relative  $(p, \epsilon)$ -approximations for well-behaved range spaces. In *Proceedings of Symposium on Computational Geometry (SoCG)*, pages 233–242, 2013.
- 7 Wayne A. Fuller. *Sampling Statistics*. Wiley, 2009.
- 8 Steve Hanneke. A bound on the label complexity of agnostic active learning. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 353–360, 2007.
- 9 Steve Hanneke. Theory of Active Learning, 2014. Manuscript downloadable at <http://www.stevehanneke.com>.
- 10 Sariel Har-Peled, Haim Kaplan, Micha Sharir, and Shakhar Smorodinsky. Epsilon-Nets for Halfspaces Revisited. *CoRR*, abs/1410.3154, 2014. [arXiv:1410.3154](https://arxiv.org/abs/1410.3154).
- 11 Sariel Har-Peled and Micha Sharir. Relative  $(p, \epsilon)$ -Approximations in Geometry. *Discrete & Computational Geometry*, 45(3):462–496, 2011.
- 12 David Haussler. Decision Theoretic Generalizations of the PAC Model for Neural Net and Other Learning Applications. *Inf. Comput.*, 100(1):78–150, 1992.
- 13 David Haussler and Emo Welzl. Epsilon-Nets and Simplex Range Queries. *Discrete & Computational Geometry*, 2:127–151, 1987.
- 14 János Komlós, János Pach, and Gerhard J. Woeginger. Almost Tight Bounds for epsilon-Nets. *Discrete & Computational Geometry*, 7:163–173, 1992.
- 15 Andrey Kupavskii, Nabil H. Mustafa, and János Pach. New Lower Bounds for epsilon-Nets. In *Proceedings of Symposium on Computational Geometry (SoCG)*, pages 54:1–54:16, 2016.
- 16 Andrey Kupavskii and Nikita Zhivotovskiy. When are epsilon-nets small? *CoRR*, abs/1711.10414, 2017. [arXiv:1711.10414](https://arxiv.org/abs/1711.10414).
- 17 Yi Li, Philip M. Long, and Aravind Srinivasan. Improved Bounds on the Sample Complexity of Learning. *Journal of Computer and System Sciences (JCSS)*, 62(3):516–527, 2001.
- 18 Jiří Matoušek. Efficient Partition Trees. *Discrete & Computational Geometry*, 8:315–334, 1992.
- 19 Jiří Matoušek, Raimund Seidel, and Emo Welzl. How to Net a Lot with Little: Small epsilon-Nets for Disks and Halfspaces. In *Proceedings of Symposium on Computational Geometry (SoCG)*, pages 16–22, 1990.
- 20 János Pach and Gábor Tardos. Tight lower bounds for the size of epsilon-nets. In *Proceedings of Symposium on Computational Geometry (SoCG)*, pages 458–463, 2011.
- 21 D. Pollard. Rates of uniform almost-sure convergence for empirical processes indexed by unbounded classes of functions. *Manuscript*, 1986.
- 22 Liwei Wang. Smoothness, Disagreement Coefficient, and the Label Complexity of Agnostic Active Learning. *Journal of Machine Learning Research*, 12:2269–2292, 2011.