Selectivity Functions of Range Queries are Learnable

Xiao Hu, Yuxi Liu, Haibo Xiu, Pankaj K. Agarwal, Debmalya Panigrahi, Sudeepa Roy, and Jun Yang
{xh102,yuxi.liu,haibo.xiu,pankaj,debmalya,sudeepa,junyang}@cs.duke.edu
Duke University, Durham, NC, USA

ABSTRACT

This paper explores the use of machine learning for estimating the selectivity of range queries in database systems. Using classic learning theory for real-valued functions based on fat-shattering dimension, we show that the selectivity function of a range space with bounded VC-dimension is learnable. As many popular classes of queries (e.g., orthogonal range search, inequalities involving linear combination of attributes, distance-based search, etc.) represent range spaces with finite VC-dimension, our result immediately implies that their selectivity functions are also learnable. To the best of our knowledge, this is the first attempt at formally explaining the role of machine learning techniques in selectivity estimation, and complements the growing literature in empirical studies in this direction. Supplementing these theoretical results, our experimental results demonstrate that, empirically, even a basic learning algorithm with generic models is able to produce accurate predictions across settings, matching state-of-art methods designed for specific queries, and using training sample sizes commensurate with our theory.

1 INTRODUCTION

In this paper, we formally model and study the problem of learning selectivity functions for selection queries in database (DB) systems. The selectivity of a selection query on a database is defined as the probability that a randomly chosen tuple from the database satisfies the query predicate. Estimating query selectivity is a core problem in the query optimization pipeline, and has a rich history of research over many decades (see, e.g., [10]). In recent years, the focus has shifted from traditional optimization methods to machine learning (ML) techniques (e.g., [4, 11]), with the latter outperforming the former in empirical studies. In this paper, we establish a learning-theoretic framework for the selectivity-estimation problem, show that the estimation problem is indeed *learnable* for popular classes of selection queries from a small set of training samples using this framework.

Our Contributions. First, we formalize the learnability of the selectivity-estimation problem. Recall that a database is a collection of tuples, and a selection query is a predicate that selects a subset of these tuples. The selectivity of a selection query is the probability that a randomly selected tuple satisfies the query. In order to learn the selectivity function, we employ the agnostic-learning framework [5], an extension of the classical PAC learning framework for real-valued functions, where we are given a set of sample queries and their respective selectivities from a fixed distribution (the *training set*), and our goal is to design an algorithm that can output the selectivity of a new query from the same distribution with high accuracy (see Figure 1 for an example).

Classical PAC learning theory asserts that a Boolean function is learnable if its VC-dimension is bounded. Generalizing this notion, it has been shown that a real-valued function is learnable using finitely many samples if its *fat shattering dimension* (defined in Section 2) is bounded [1, 2, 8]. This reduces the question of learnability of selectivity functions to bounding their respective fat shattering dimensions. We further note that selectivity functions correspond to selection queries on the underlying data. Each selection query, in turn, is a binary function on the data (i.e., which data items satisfy the query predicate), and the complexity of a class of binary functions is captured by its *VC-dimension* [12]. Our main result (see Theorem 2.1 below) shows that if a class of selection queries has bounded VC-dimension, then the fat shattering dimension of the corresponding selectivity function must also be bounded, and therefore, the selectivity function for such queries is learnable.

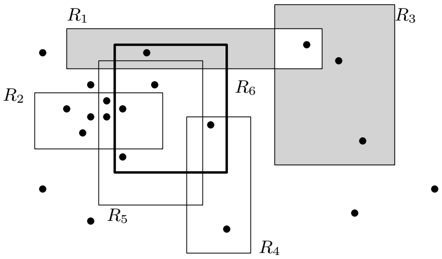

Figure 1: An illustration of the learned selectivity problem. There are 20 points in the dataset D and 5 training queries R_1, R_2, R_3, R_4, R_5 with their selectivities given by $s_D(R_1) = 0.1$, $s_D(R_2) = 0.3$, $s_D(R_3) = 0.15$, $s_D(R_4) = 0.1$ and $s_D(R_5) = 0.25$. The goal is to estimate the selectivity of an unknown query R_6 (in bold), and the correct answer in this example is 0.25. The shaded area will be explained in Section 2.3.

Taking a geometric view, each input range consists of a subset of points in $\mathbb{R}^d$ that lies inside the query region. The above result implies that the selectivity functions of many popular ranges such as rectangles, halfspaces, and balls are learnable (see Section 2.3 for details).

While our framework establishes the learnability of the selectivity of above query types from a small set of training examples, it does not by itself prescribe any specific model or learning algorithm. As part of establishing the learnability of our selectivity query, we also need a procedure that, given a set of training samples and a family of data distributions (e.g. histograms, discrete distributions), constructs a data distribution from the given family that "best fits" the training samples. Our framework then guarantees that the learned data distribution estimates the selectivity of any query chosen from the same distribution as the training samples with high accuracy. For specific query types (e.g., orthogonal range queries), there already exists a large body of work on the selectivity-estimation problem, and our framework now gives them a solid foundation. To demonstrate the power of our framework beyond justifying existing methods, we further propose a simple, generic approach that embodies our theoretical results, and empirically validates its efficiency using extensive experiments. It is important to note that we are not designing this generic approach to "beat" existing methods with novel or sophisticated features; in fact, we intentionally avoid sophisticated features so that experimental comparison can focus on illustrating the power of our unifying framework instead of the artifacts of extra features. Despite the simplicity of our approach, our experimental results show that it performs comparably to the state-of-the-art methods for orthogonal range queries. Furthermore, for query classes that have seen less previous research, such as linear inequality and distance-based queries, our generic approach also work effectively, demonstrating the generality of the our theoretical framework. Due to lack of space, these results are omitted from this abstract and can be found in [7].

2 LEARNABILITY OF QUERY SELECTIVITY

A *range space* Σ is a pair $(X, \mathcal{R})$, where X is a set of *objects* and $\mathcal{R}$ is a collection of subsets of X called *ranges*. For example, $X = \mathbb{R}^d$ and $\mathcal{R}$ can be the set of all d-dimensional rectangles, halfspaces, or balls. Let D be a probability distribution over X. For a given D, we define the *selectivity function* $s_D : \mathcal{R} \to [0, 1]$ as $s_D(R) = \Pr_{x \sim D}[x \in R]$.

Our goal is to *learn* the selectivities of the ranges in a range space Σ under an unknown data distribution from a finite sample of ranges and their respective selectivities. Formally, we define this learning task as follows.

2.1 The Learning Framework

Learnability. Following the agnostic learning model proposed by Haussler [5] (see also [1, 2]), which generalizes the PAC model, we define learnability in a more general setting. Let $\mathscr{H}$ be a family of functions from a domain Y to $[0, 1]$. Set $Z = Y \times [0, 1]$. For a function $H \in \mathscr{H}$, we define the *loss function* $\ell_H : Z \to [0, 1]$. For $z = (y, w) \in Z$, $\ell_H(z) = (H(y) - w)^2$. For a probability distribution Q over Z and for a function $H \in \mathscr{H}$, we define

$$\mathrm{er}_Q(H) = \int_Z \ell_H(z) dQ(z) \tag{1}$$

to be the mean square loss of H with respect to distribution Q.

A *learning procedure* $\mathcal{A}$ is mapping from finite sequences in Z to $\mathscr{H}$. Given a *training sample* $\mathbf{z}^n = (z_1, z_2, \cdots, z_n) \in Z^n$, $\mathcal{A}$ returns a function $\mathcal{A}(\mathbf{z}^n)$. Given $\epsilon, \delta \in (0, 1)$ and an integer $n > 0$, we say that $\mathcal{A}$ (ϵ, δ)-learns (agnostically) from n random training samples with respect to $\mathscr{H}$ if

$$\sup_Q \Pr[\mathrm{er}_Q(\mathcal{A}(\mathbf{z}^n)) \geq \inf_{H \in \mathscr{H}} \mathrm{er}_Q(H) + \epsilon] \leq \delta,$$

where Pr denotes the probability with respect to a random sample $\mathbf{z}^n \in Z^n$, each of $z_1, z_2, \cdots, z_n$ is drawn independently from Z at random according to Q, and supremum is taken over all distributions defined on Z. For $\epsilon > 0$, $\mathscr{H}$ is called *ϵ-learnable* if there exists a function $n_0 : [0, 1]^2 \to \mathbb{N}$ and a learning procedure $\mathcal{A}$ such that for all $\delta > 0$ and for all $n \geq n_0(\epsilon, \delta)$, $\mathcal{A}$ (ϵ, δ)-learns from n examples with respect to $\mathscr{H}$; $n_0(\epsilon, \delta)$ is referred to as the minimum training set size for $\mathscr{H}$. Finally, $\mathscr{H}$ is *learnable* if it is ϵ-learnable for all $\epsilon > 0$.

Returning to the selectivity function of range space $\Sigma = (X, \mathcal{R})$, let $\mathscr{D}$ be a set of distributions defined on X. Set $\mathcal{S}_{\Sigma,\mathscr{D}} = \{s_D \mid D \in \mathscr{D}\}$, a family of functions from $\mathcal{R}$ to $[0, 1]$. Set $Z = \mathcal{R} \times [0, 1]$. Our main result is a characterization of learnability of $\mathcal{S}_{\Sigma,\mathscr{D}}$ in terms of the VC-dimension of Σ, defined below.

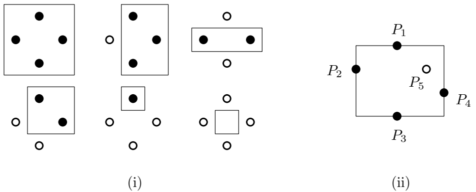

Figure 2: VC-dimension of $\Sigma = (\mathbb{R}^2, \mathcal{R}_\square)$, where $\mathcal{R}_\square$ is the set of all two-dimensional rectangles, is 4. (i) is an illustration of a set of 4 points shattered by $\mathcal{R}_\square$. On the other hand, no set $Y = \{p_1, p_2, p_3, p_4, p_5\}$ in $\mathbb{R}^2$ can be shattered by $\mathcal{R}_\square$ in (ii): let $\{P_1, P_2, P_3, P_4\} \subseteq Y$ be the subset of (at most 4) points of Y with extreme x- and y-coordinates. Then any rectangle containing P_1, P_2, P_3, P_4 also contains P_5.

VC dimension. A subset $P \subseteq X$ is *shattered* by $\mathcal{R}$ if $\{P \cap R \mid R \in \mathcal{R}\} = 2^P$. The *VC-dimension* of $\mathcal{R}$, denoted by $\mathsf{VC\text{-}dim}(\Sigma)$, is the size of the largest subset of X that can be shattered by Σ. An example is given in Figure 2. If the VC-dimension of Σ is not bounded by a constant, then $\mathsf{VC\text{-}dim}(\Sigma) = \infty$. Our main result, stated in the theorem below, is that $\mathcal{S}_{\Sigma,\mathscr{D}}$ is learnable if and only if $\mathsf{VC\text{-}dim}(\Sigma)$ is finite.

Theorem 2.1. *Let $\Sigma = (X, \mathcal{R})$ be a range space, let $\mathscr{D}$ be a set of distributions defined on X, and let $\epsilon \in (0, 1)$ be a parameter. If $\mathsf{VC\text{-}dim}(\Sigma) = \lambda$, for some constant $\lambda > 0$, then the family $\mathcal{S}_{\Sigma,\mathscr{D}}$ of selectivity functions is ϵ-learnable with a training set of size $\tilde{O}\left(\frac{1}{\epsilon^{\lambda+3}}\right)$.[1] Conversely, if $\mathsf{VC\text{-}dim}(\Sigma) = \infty$, $\mathcal{S}_{\Sigma,\mathscr{D}}$ is not (agnostically) learnable.*

Remark. Note that we do not assume training sample $z_i = (R_i, s_i) \in Z$ to be of the form $s_i = s_D(R_i)$ for some data distribution $D \in \mathscr{D}$. They are drawn from some distribution Q defined on $\mathcal{R} \times [0, 1]$, and the goal is to learn the selectivity function in $\mathcal{S}_{\Sigma,\mathscr{D}}$ that minimizes the mean square loss. This is important, which allows us to decouple training samples from the family of functions, and the problem just becomes to find a function from the given family that minimizes the expected loss. This model is more general than the one assuming training sample in a form of $z_i = (R_i, s_D(R_i))$ for some data distribution $D \in \mathscr{D}$, for example, capturing the noisy input for learning the selectivity functions.

2.2 Implications of Theorem 2.1

Before proving Theorem 2.1, we give some of its implications. We begin with the query classes mentioned in the introduction.

Orthogonal Range Queries: The range space $\Sigma_\square = (\mathbb{R}^d, \mathcal{R}_\square)$ for orthogonal range queries is defined as

$$\mathcal{R}_\square = \{\times_{i=1}^d [a_i, b_i] : a_i, b_i \in \mathbb{R}, a_i \leq b_i, \forall i \in [d]\}.$$

It is well known that $\mathsf{VC\text{-}dim}(\Sigma_\square) = 2d$ [9] (see Figure 2 for $d = 3$), therefore Theorem 2.1 implies that for any family $\mathscr{D}$ of distributions defined on $\mathbb{R}^d$ and for any $\epsilon > 0$, the selectivity functions are ϵ-learnable with training set of size $\tilde{O}\left(\frac{1}{\epsilon^{2d+3}}\right)$.

[1] $\tilde{O}(.)$ to hide lower order terms that are in polylog $\left(\frac{1}{\epsilon}, \frac{1}{\delta}\right)$ for constant λ.

Linear Inequality Queries: The range space $\Sigma_\setminus = (\mathbb{R}^d, \mathcal{R}_\setminus)$ for linear inequality queries is defined as

$$\mathcal{R}_\setminus = \{R_{\setminus(a,b)} : a \in \mathbb{R}^d, b \in \mathbb{R}\},$$

where $R_{\setminus(a,b)} = \{x \in \mathbb{R}^d : a \cdot x \geq b\}$. It is known that VC-dim$(\Sigma_\setminus) = d + 1$ [9], therefore Theorem 2.1 implies that for any family $\mathscr{D}$ of distributions defined on $\mathbb{R}^d$ and for any $\epsilon > 0$, the selectivity functions are ϵ-learnable with training set of size $\tilde{O}\left(\frac{1}{\epsilon^{d+4}}\right)$.

Distance-Based Queries: The range space $\Sigma_\circ = (\mathbb{R}^d, \mathcal{R}_\circ)$ for distance-based queries is defined as

$$\mathcal{R}_\circ = \{R_{\circ(a,b)} : a \in \mathbb{R}^d, b \in \mathbb{R}\},$$

where $R_{\circ(a,b)} = \{x \in \mathbb{R}^d : \|x - a\|_2 \leq b\}$ and $\|\cdot\|$ is the Euclidean norm. It is known that VC-dim$(\Sigma_\circ) \leq d + 2$ [9], therefore Theorem 2.1 implies that for any family $\mathscr{D}$ of distributions defined on $\mathbb{R}^d$ and for any $\epsilon > 0$, the selectivity functions are ϵ-learnable with training set of size $\tilde{O}\left(\frac{1}{\epsilon^{d+5}}\right)$.

Semi-algebraic Range Queries. A very general class of range queries is the so-called *semi-algebraic range query*. A d-dimensional semi-algebraic set is subset of $\mathbb{R}^d$ defined by a Boolean formula over polynomial inequality. For example, $R = \{(x, y) \in \mathbb{R}^2 \mid (x^2 + y^2 \leq 4) \wedge (x^2 + y^2 \geq 1) \wedge (y - 2x^2 \leq 0)\}$ is a semi-algebraic sets. All the three above examples are special cases of semi-algebraic range queries. Let $\mathbb{T}_{d,b,\Delta}$ be the set of all semi-algebraic sets defined by at most b d-variate polynomial inequalities, each of degree at most Δ. It is known that the VC-dimension of range space $(\mathbb{R}^d, \mathbb{T}_{d,b,\Delta})$ is a constant $\lambda := \lambda(d, b, \Delta)$. Hence the selectivity functions on $(\mathbb{R}^d, \mathbb{T}_{d,b,\Delta})$ are also learnable for any constants d, b, Δ.

Semi-algebraic sets enable us to handle range spaces in which X is not a set of points in $\mathbb{R}^d$. For example, let $\mathbb{B}$ be the set of all discs in $\mathbb{R}^2$. For a query disc B, let $R_B \subseteq \mathbb{B}$ be the set of discs that intersect B. Define $\mathcal{R}_\bullet = \{R_B \mid B \in \mathbb{B}\}$, and consider the range space $\Sigma_\bullet = (\mathbb{B}, \mathcal{R}_\bullet)$. We can map each disc in $\mathbb{B}$ to a point (x, y, z) in $\mathbb{R}^3$ where (x, y) is the center of the disc and z is its radius. Then for a query disc B centered at (c_x, c_y) and radius r, the range R_B maps to the set

$$\gamma_B = \{(x, y, z) \in \mathbb{R}^3 \mid (x - c_x)^2 + (y - c_y)^2 \leq (r + z)^2, z \geq 0\}.$$

Set $\mathbb{R}^3_{z\geq 0} = \mathbb{R}^2 \times \mathbb{R}_{z\geq 0}$ and $\hat{\mathcal{R}}_\bullet = \{\gamma_B \mid B \in \mathbb{B}\}$. Then $\Sigma_\bullet$ is mapped to $(\mathbb{R}^3_{z\geq 0}, \hat{\mathcal{R}}_\bullet)$. Since ranges in $\hat{\mathcal{R}}$ are semi-algebraic sets with $b = 1$ and $\Delta \leq 2$, VC-dim$(\mathbb{R}^3_{\geq 0}, \hat{\mathcal{R}}_\bullet)$ is finite and hence selectivity functions on $(\mathbb{B}, \mathcal{R})$ are learnable.

We conclude this discussion by giving an example of range space for which selectivity functions are not learnable.

Polygon range queries with arbitrary number of vertices. Let $\mathbb{C}$ be the set of all convex polygons in $\mathbb{R}^2$ with arbitrary number of vertices. Consider the range space $\Sigma = (\mathbb{R}^2, \mathbb{C})$. It is known that VC-dim$(\Sigma) = \infty$ [6], therefore Theorem 2.1 implies that selectivity functions on Σ are not learnable.

2.3 Proof of Theorem 2.1

We prove Theorem 2.1 using the notion of *fat-shattering dimension* introduced by Kearns and Schapire [8], which is a generalization of VC-dimension, and the results by Alon et al. [1] and Bartlett-Long [2]. As in Section 2.1, let $\mathscr{H}$ be a class of functions from a domain X into $[0, 1]$. Let $\gamma \in (0, 1/2)$ be a parameter. We say that $\mathscr{H}$ γ*-shatters* a subset $V \subseteq X$ if there is a witness function $\sigma : V \longrightarrow [0, 1]$ such that for every subset $E \subseteq V$, there is a function $H_E \in \mathscr{H}$ with

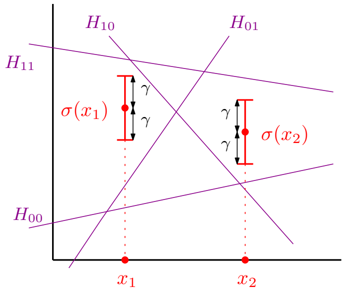

Figure 3: x_1, x_2 are γ-shattered by linear functions. we choose H to be the linear function whose bit sequence b_2b_1 corresponds to E (i.e., $b_i = 1$ if $x_i \in E$).

$$\begin{aligned} H_E(x) &\geq \sigma(x) + \gamma, \quad \forall x \in E,\\ H_E(x) &\leq \sigma(x) - \gamma, \quad \forall x \in V \setminus E. \end{aligned} \tag{2}$$

An example is shown in Figure 3.

The *γ-fat shattering dimension* of $\mathscr{H}$, denoted by fat$_{\mathscr{H}}(\gamma)$, is the size of the largest subset of X that can be γ-shattered by $\mathscr{H}$. If subsets of unbounded finite size can be γ-shattered by $\mathscr{H}$, then we set fat$_{\mathscr{H}}(\gamma) = \infty$. Note that if $\mathscr{H}$ is a class of functions from X into $\{0, 1\}$, then γ-fat shattering dimension is the same as VC-dimension. An advantage of γ-fat shattering dimension is that it is sensitive to the scale at which difference in the function values are considered important. Alon et al. [1] proved that if fat$_{\mathscr{H}}(c\epsilon)$ is finite, where $c \in (0, 1)$ is a suitable constant, then $\mathscr{H}$ is ϵ-learnable. The bound on the size of the training set was improved by Bartlett and Long [2]. In particular, their result implies that $\mathscr{H}$ is ϵ-learnable with training-set size

$$n_0(\epsilon, \delta) = O\left(\frac{1}{\epsilon^2}\left(\text{fat}_{\mathscr{H}}(\frac{\epsilon}{9}) \log^2 \frac{1}{\epsilon} + \log \frac{1}{\delta}\right)\right).$$

Returning to the selectivity functions, let $\Sigma = (X, \mathcal{R})$ be a range space, let $\mathscr{D}$ be a family of probability distributions on X and $\gamma \in (0, 1)$. Set $\mathcal{S} := \mathcal{S}_{\Sigma,\mathscr{D}}$ to be the selectivity functions defined by $\mathscr{D}$. Our main technical result is that if VC-dim$(\Sigma) = \lambda$, for some constant λ, then fat$_{\mathcal{S}}(\gamma) = \tilde{O}\left(\frac{1}{\gamma^{\lambda+1}}\right)$. By plugging this result into the results of [1, 2], we prove the first part of Theorem 2.1.

Let $\mathcal{T} \subseteq \mathcal{R}$ be a subset γ-shattered by $\mathcal{S}$. To bound fat$_{\mathcal{S}}(\gamma)$, it suffices to prove that $|\mathcal{T}| = \tilde{O}\left(\frac{1}{\gamma^{\lambda+1}}\right)$. First, we partition the ranges in $\mathcal{T}$ based on the values of their respective witnesses $\sigma(R)$[2]:

$$\mathcal{T}_j = \{R \in \mathcal{T} : \sigma(R) \in [(j-1)\cdot\gamma, j\cdot\gamma], \text{for} j \in [1/\gamma]\}.$$

Lemma 2.2. *Suppose Equation (2) is realized for some subset $E \in \mathcal{T}_j$ by s_D for some distribution $D \in \mathscr{D}$. Then, for any pair $R \in E, R' \in \mathcal{T}_j \setminus E$, we have*

$$s_D(R) - s_D(R') > \gamma. \tag{3}$$

[2] Note that although $\sigma(R) = 1$ is excluded by this definition if $1/\gamma$ is an integer, it is a well-defined partition since $\sigma(R)$ cannot be equal to 1 for any range $R \in \mathcal{T}$. This follows from the observation that if $\sigma(R) = 1$, then Equation (2) cannot be satisfied for $R \in E$ since $H_E(R) \leq 1$ and $\gamma > 0$.

Now, consider any fixed ordering $\pi = \langle R_1, R_2, \cdots, R_k \rangle$ of the ranges in $\mathcal{T}_j$, where $k = |\mathcal{T}_j|$. Let us also fix the subset:

$$E = \{R_{2i} \mid 1 \le i \le \lfloor k/2 \rfloor\} \tag{4}$$

to be the set of ranges with even index in π. We say that an object $x \in X$ crosses a pair of ranges R, R' if $x \in R \oplus R'$, where $\oplus$ is the symmetric difference (see Figure 1 for $R_1 \oplus R_3$). For $1 \le i < k$ and for every $x \in X$, we define an indicator random variable as follows:

$$I_{i,x} = \begin{cases} 1 & \text{if } x \in R_i \oplus R_{i+1}, \\ 0 & \text{otherwise,} \end{cases}$$

and let $I_x = \sum_{i=1}^{k-1} I_{i,x}$.

Since $\mathcal{T}$ is γ-shattered by $\mathcal{S}$, there is a distribution $D_\pi \in \mathscr{D}$ that satisfies (2) for E. The next lemma is a direct consequence of Lemma 2.2, by summing up over the pairs of ranges R_i, R_{i-1} for even i in $\mathcal{T}_j$:

Lemma 2.3. $\mathbb{E}_{x \sim D_\pi}[I_x] > \gamma(k-1)$.

The lower bound on $\mathbb{E}_{x \sim D_E}[I_x]$ in Lemma 2.3 holds for any ordering π of the ranges in $\mathcal{T}_j$; the distribution D_π obviously depends on π. We now complement this lower bound with an upper bound on $\mathbb{E}_{x \sim D_\pi}[I_x]$ for a *specific* ordering π of $\mathcal{T}_j$.

Lemma 2.4. *There is an ordering $R_1, R_2, \cdots, R_k$ of the ranges in $\mathcal{T}_j$ such that for any distribution D defined on X, we have:*

$$\mathbb{E}_{x \in D}[I_x] = O(k^{1-1/\lambda} \log k),$$

where $\lambda = \text{VC-dim}(X, \mathcal{R})$.

Proof. Let $\widetilde{\Sigma} = (X, \mathcal{T}_j)$ be the range space defined by the ranges in $\mathcal{T}_j$. Note that $\text{VC-dim}(\widetilde{\Sigma}) \le \text{VC-dim}(\Sigma) = \lambda$. Consider the dual range space $\widetilde{\Sigma}^*$ of $\widetilde{\Sigma}$, where $\widetilde{\Sigma}^* = (\mathcal{T}_j, \{\mathcal{R}_x = \{R \in \mathcal{T}_j : x \in R\} \mid x \in X\})$, i.e., the objects of $\widetilde{\Sigma}^*$ are the ranges of $\mathcal{T}_j$ and for each object $x \in X$, we have a dual range in $\widetilde{\Sigma}^*$ consisting of ranges of $\widetilde{\Sigma}$ that contain x. Note that $\widetilde{\Sigma}^{**} = \widetilde{\Sigma}$.

We compute the desired ordering of $\mathcal{T}_j$, using the following results by Chazelle and Welzl [3]: Let $\Xi = (V, \Gamma)$ be a finite range space with $|V| = m$. We say that a range $\gamma \in \Gamma$ crosses a pair $v_i, v_j \in V$ if $|\gamma \cap \{v_i, v_j\}| = 1$. The result in [3] (Theorem 4.3) proves that there is an ordering $v_1, v_2, \cdots, v_m$ of objects in V such that any range in $\mathcal{T}$ crosses $O\left(m^{1-1/\lambda^*} \log m\right)$ pairs (v_i, v_{i+1}) for $1 \le i < m$, where λ^* is the VC-dimension of the dual range space of Ξ. Applying this result to $\widetilde{\Sigma}^*$ and using the fact that $\widetilde{\Sigma}^{**} = \widetilde{\Sigma}$, we obtain an ordering $R_1, R_2, \cdots, R_k$ of $\mathcal{T}_j$ such that any range of $\widetilde{\Sigma}^*$ crosses $O\left(k^{1-1/\lambda} \log k\right)$ pairs (R_i, R_{i+1}). By the definition, a range $\mathcal{R}_x$ crosses R_i, R_{i+1} if $|\mathcal{R}_x \cap \{R_i, R_{i+1}\}| = 1$, which is equivalent to saying that $x \in R_i \oplus R_{i+1}$. Hence, for any $x \in X$, there are $O\left(k^{1-1/\lambda} \log k\right)$ pairs (R_i, R_{i+1}) crossed by x. Since this bound holds for every $x \in X$, we conclude that

$$\mathbb{E}_{x \sim D}[I_x] = O\left(k^{1-1/\lambda} \log k\right). \qquad \square$$

We are now ready to bound the size of $\mathcal{T}_j$.

Lemma 2.5. *For any $j \in \lceil 1/\gamma \rceil$, $|\mathcal{T}_j| = O\left((\frac{1}{\gamma} \log \frac{1}{\gamma})^\lambda\right)$.*

Proof. Plugging Lemmas 2.4 and 2.3 together, we conclude there exists a constant c such that

$$\gamma \cdot (k-1) \le c \cdot k^{1-1/\lambda} \log k,$$

which implies that $\frac{k^{1/\lambda}}{\log k} \le 2c/\gamma$, or $k = O\left((\frac{1}{\gamma} \log \frac{1}{\gamma})^\lambda\right)$. $\square$

Summing this bound over all $j \in \lceil 1/\gamma \rceil$, we conclude that $|\mathcal{T}| = \tilde{O}\left(\frac{1}{\gamma^{\lambda+1}}\right)$. Hence, the size of any set of query ranges in $\mathcal{R}$ that can be γ-shattered by $\mathcal{S}$ is $\tilde{O}\left(\frac{1}{\gamma^{\lambda+1}}\right)$, which implies the main technical result of this section.

Lemma 2.6. *Let $\Sigma = (X, \mathcal{R})$ be a range space with $\text{VC-dim}(\Sigma) = \lambda$, let $\mathscr{D}$ be a family of probability distribution over X, and let $\mathcal{S} := \mathcal{S}_{\Sigma,\mathscr{D}}$ be the family of selectivity functions on Σ by $\mathscr{D}$. For any $\gamma \in (0,1)$, the γ-fat shattering dimension of $\mathcal{S}$ is $\tilde{O}\left(\frac{1}{\gamma^{\lambda+1}}\right)$.*

Finally, plugging Lemma 2.6 into the results of Alon et al. [1] and Bartlett-Long [2], we obtain the first part of Theorem 2.1.

We next turn to the second part of Theorem 2.1. As in Section 2.1, let $\mathscr{H}$ be a class of functions from a domain X into $[0,1]$. Let $\gamma \in [0,1]$ be a parameter. Alon et al. [1] proved that if $\text{fat}_{\mathscr{H}}(\epsilon) = \infty$, then $\mathscr{H}$ is not $(\epsilon^2/8 - \tau)$-learnable for any $\tau > 0$. Returning to the selectivity functions $\mathcal{S} := \mathcal{S}_{\Sigma,\mathscr{D}}$ defined on the range space $\Sigma = (X, \mathcal{R})$ and a family of probability distribution on X as $\mathscr{D}$. Our second technical result is that if $\text{VC-dim}(\Sigma) = \infty$, then $\text{fat}_{\mathcal{S}}(\gamma) = \infty$ for any $\gamma \in (0, 1/2)$.

Lemma 2.7. *Let $\Sigma = (X, \mathcal{R})$ be a range space, let $\mathscr{D}$ be a family of probability distribution over X, and let $\mathcal{S} := \mathcal{S}_{\Sigma,\mathscr{D}}$ be the family of selectivity functions on Σ by $\mathscr{D}$. If $\text{VC-dim}(\Sigma) = \infty$, the γ-fat shattering dimension of $\mathcal{S}$ is also ∞, for any $\gamma \in (0, 1/2)$.*

The above lemma proves second part of Theorem 2.1, thereby completing the proof of Theorem 2.1.

REFERENCES

[1] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. 1997. Scale-sensitive dimensions, uniform convergence, and learnability. *JACM* 44, 4 (1997), 615–631.

[2] Peter L Bartlett and Philip M Long. 1995. More theorems about scale-sensitive dimensions and learning. In *Proceedings of the eighth annual conference on Computational learning theory*. 392–401.

[3] B. Chazelle and E. Welzl. 1989. Quasi-optimal range searching in spaces of finite VC-dimension. *Discrete & Computational Geometry* 4, 5 (1989), 467–489.

[4] S. Hasan, S. Thirumuruganathan, J. Augustine, N. Koudas, and G. Das. 2020. Deep Learning Models for Selectivity Estimation of Multi-Attribute Queries. In *Proc. 39th ACM SIGMOD Int. Conf. Management Data*. 1035–1050.

[5] D. Haussler. 1992. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Inf. Comput.* 100, 1 (1992), 78–150.

[6] D. Haussler and E. Welzl. 1987. ϵ-nets and simplex range queries. *Discret. Comput. Geom.* 2, 2 (1987), 127–151.

[7] Xiao Hu, Yuxi Liu, Haibo Xiu, Pankaj K Agarwal, Debmalya Panigrahi, Sudeepa Roy, and Jun Yang. 2022. Selectivity Functions of Range Queries are Learnable. (2022).

[8] M. J. Kearns and R. E. Schapire. 1994. Efficient distribution-free learning of probabilistic concepts. *J. Comput. System Sci.* 48, 3 (1994), 464–497.

[9] M. J. Kearns and U. Vazirani. 1994. *An introduction to computational learning theory*. MIT press.

[10] Y. Matias, J. S. Vitter, and M. Wang. 1998. Wavelet-based histograms for selectivity estimation. In *Proc. 17th ACM SIGMOD Int. Conf. Management Data*. 448–459.

[11] Y. Park, S. Zhong, and B. Mozafari. 2020. Quicksel: Quick selectivity learning with mixture models. In *Proc. 39th ACM SIGMOD Int. Conf. Management Data,*. 1017–1033.

[12] V. N. Vapnik and A. Y. Chervonenkis. 2015. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity*. Springer, 11–30.