

# Quadratic Upper Bound for Recursive Teaching Dimension of Finite VC Classes

**Lunjia Hu**  
**Ruihan Wu**  
**Tianhong Li**

HULJ14@MAILS.TSINGHUA.EDU.CN  
WRH14@MAILS.TSINGHUA.EDU.CN  
LTH14@MAILS.TSINGHUA.EDU.CN

*Institute for Interdisciplinary Information Sciences, Tsinghua University, China*

**Liwei Wang**

WANGLW@CIS.PKU.EDU.CN

*Key Laboratory of Machine Perception, School of Electronics Engineering and Computer Sciences, Peking University, China*

## Abstract

In this work we study the quantitative relation between the recursive teaching dimension (RTD) and the VC dimension (VCD) of concept classes of finite sizes. The RTD of a concept class  $\mathcal{C} \subseteq \{0, 1\}^n$ , introduced by [Zilles et al. \(2011\)](#), is a combinatorial complexity measure characterized by the worst-case number of examples necessary to identify a concept in  $\mathcal{C}$  according to the recursive teaching model.

For any finite concept class  $\mathcal{C} \subseteq \{0, 1\}^n$  with  $\text{VCD}(\mathcal{C}) = d$ , [Simon and Zilles \(2015\)](#) posed an open problem  $\text{RTD}(\mathcal{C}) = O(d)$ , i.e., is RTD linearly upper bounded by VCD? Previously, the best known result is an exponential upper bound  $\text{RTD}(\mathcal{C}) = O(d \cdot 2^d)$ , due to [Chen et al. \(2016\)](#). In this paper, we show a quadratic upper bound:  $\text{RTD}(\mathcal{C}) = O(d^2)$ , much closer to an answer to the open problem. We also discuss the challenges in fully solving the problem.

**Keywords:** Recursive teaching dimension; VC dimension; Recursive teaching model

## 1. Introduction

Sample complexity is one of the most important concepts in machine learning. Basically, it is the amount of data needed to achieve a desired learning accuracy. Sample complexity has been extensively studied in various learning models. In PAC-learning, sample complexity is characterized by the VC dimension (VCD) of the concept class ([Blumer et al., 1989](#); [Vapnik and Chervonenkis, 1971](#)). PAC-learning is a passive learning model. In this model, the role of the teacher is limited to providing labels to data *randomly* drawn from the underlying distribution.

Different from PAC-learning, there are important models in which a teacher is involved more actively in the learning process. For example, in the classical teaching model ([Goldman and Kearns, 1995](#); [Shinohara and Miyano, 1991](#)), the teacher chooses a set of labeled examples so that the learner, after receiving the examples, can distinguish the target concept from all other concepts in the concept class. In this model, the key complexity measure of a concept class is the teaching dimension, which is defined as the worst-case number of examples needed to be selected by the teacher ([Goldman and Kearns, 1995](#)). Teaching dimension finds applications in many learning problems ([Angluin, 2004](#); [Hanneke, 2007](#); [Dasgupta, 2005](#); [Hegedűs, 1995](#); [Goldman and Mathias, 1993](#); [Anthony et al., 1995](#)).

Another model of teaching, the recursive teaching model, is proposed by [Zilles et al. \(2011\)](#). The idea underlying the recursive teaching model is to let the teacher exploit a hierarchical structure in the concept class. Concretely, the hierarchy of a concept class is a nesting, starting from the concept that requires the smallest amount of data to teach, and then applying this process recursively to the rest of the concepts. The complexity measure of a concept class in the recursive teaching model is called recursive teaching dimension (RTD). RTD is defined as the worst-case number of examples needed to be selected by the teacher for any target concept during the recursive process ([Zilles et al., 2011](#)). See also Section 2 for a formal definition.

Although less intuitive, RTD exhibits surprising properties. The most interesting property is the quantitative relation between RTD and VCD of a finite concept class. An example is intersection-closed concept classes. For such a concept class  $\mathcal{C}$ ,  $\text{RTD}(\mathcal{C}) \leq \text{VCD}(\mathcal{C})$ . On the other hand, there exist cases where  $\text{RTD}(\mathcal{C}) > \text{VCD}(\mathcal{C})$ . However, the best known worst-case lower bound is  $\text{RTD}(\mathcal{C}) \geq \frac{5}{3}\text{VCD}(\mathcal{C})$ , which is proven by giving an explicit construction ([Chen et al., 2016](#)). For more special cases such that  $\text{RTD}(\mathcal{C}) = \text{VCD}(\mathcal{C})$  or  $\text{RTD}(\mathcal{C}) \leq \text{VCD}(\mathcal{C})$ , please refer to ([Doliwa et al., 2014](#)).

Based on these insights, [Simon and Zilles \(2015\)](#) posed an open problem on the quantitative relation between RTD and VCD of general concept classes: For any finite concept class  $\mathcal{C} \subseteq \{0, 1\}^n$  with  $\text{VCD}(\mathcal{C}) = d$ , is  $\text{RTD}(\mathcal{C})$  linearly upper bounded by  $d$ , i.e., does  $\text{RTD}(\mathcal{C}) \leq \kappa d$  hold for a universal constant  $\kappa$ ?

At the time when this open problem was posed, the only known result for general concept classes  $\mathcal{C} \subseteq \{0, 1\}^n$  is  $\text{RTD}(\mathcal{C}) = O(d \cdot 2^d \log \log |\mathcal{C}|)$  ([Moran et al., 2015](#)). This bound is exponential in VCD and depends on the size of the concept class.

Before our work, the best known upper bound is due to [Chen et al. \(2016\)](#), who proved that  $\text{RTD}(\mathcal{C}) = O(d \cdot 2^d)$ , which is the first upper bound for  $\text{RTD}(\mathcal{C})$  that depends only on  $\text{VCD}(\mathcal{C})$ , but not on the size of the concept class.

In this paper, we continue this line of research and extend the techniques developed in ([Kuhlmann, 1999](#); [Moran et al., 2015](#); [Chen et al., 2016](#)). Our main result is a quadratic upper bound  $\text{RTD}(\mathcal{C}) = O(d^2)$  for any finite concept class  $\mathcal{C} \subseteq \{0, 1\}^n$  with  $\text{VCD}(\mathcal{C}) = d$ . In particular, we prove  $\text{RTD}(\mathcal{C}) \leq 39.3752d^2 - 3.6330d$ . Comparing to previous results, our bound is much closer to the linear upper bound in the open problem.

As pointed out by [Simon and Zilles \(2015\)](#), a solution to their open problem will have important implications: It provides deeper understanding not only of the relationship between the complexity of teaching and the complexity of passive supervised learning, but also on the well-known sample compression conjecture ([Warmuth, 2003](#); [Littlestone and Warmuth, 1986](#)), which states that for every concept class of VCD  $d$ , there is a compression scheme that can compress the samples to a subset of size at most  $d$ . (See also ([David et al., 2016](#)) for recent progress.)

The rest of the paper is organized as follows. Section 2 presents the background and all the definitions. In Section 3 we propose our main results and proofs. Section 4 provides discussions on the challenges in fully solving the open problem.

## 2. Preliminaries

Let  $X$  be a finite instance space and  $\mathcal{C}$  a concept class over  $X$ , i.e.,  $\mathcal{C} \subseteq \{0, 1\}^X$ . For notational simplicity, we always assume  $X = [n]$  where  $[n] = \{1, 2, \dots, n\}$ , and consider concept class  $\mathcal{C} \subseteq \{0, 1\}^n$ .

The VC dimension of a concept class  $\mathcal{C} \subseteq \{0, 1\}^n$ , denoted by  $\text{VCD}(\mathcal{C})$ , is the maximum size of a shattered subset of  $[n]$ , where  $A \subseteq [n]$  is said to be shattered by  $\mathcal{C}$  if  $|\{c|_A : c \in \mathcal{C}\}| = 2^{|A|}$ . Here  $c|_A$  is the projection of  $c$  on  $A$ . In other words, for every  $b \in \{0, 1\}^{|A|}$ , there is  $c \in \mathcal{C}$  so that  $c|_A = b$ .

For a given concept class  $\mathcal{C} \subseteq \{0, 1\}^n$  and a concept  $c \in \mathcal{C}$ , we say  $A \subseteq [n]$  is a teaching set for  $c$  if  $A$  distinguishes  $c$  from all other concepts in  $\mathcal{C}$ . That is,  $c|_A \neq c'|_A$  for all  $c' \in \mathcal{C}$ ,  $c' \neq c$ .

The size of the *smallest* teaching set for  $c$  with respect to  $\mathcal{C}$  is denoted by  $\text{TD}(c; \mathcal{C})$ . In the classical teaching model (Goldman and Kearns, 1995; Shinohara and Miyano, 1991), the teaching dimension of a concept class  $\mathcal{C}$ , denoted by  $\text{TD}(\mathcal{C})$ , is defined as  $\text{TD}(\mathcal{C}) = \max_{c \in \mathcal{C}} \text{TD}(c; \mathcal{C})$ .  $\text{TD}(\mathcal{C})$  can be seen as the worst-case teaching complexity (Kuhlmann, 1999), as it considers the hardest concept to distinguish from other concepts. However, defining teaching complexity using the hardest concept is often restrictive; and  $\text{TD}(\mathcal{C})$  does not always capture the idea of cooperation in teaching and learning. In fact, a simple concept class may have the maximum possible complexity (Zilles et al., 2011). Instead, one can consider the best-case teaching dimension of  $\mathcal{C}$ .

**Definition 1 (Best-Case Teaching Dimension)** *The best-case teaching dimension of a concept class  $\mathcal{C}$ , denoted by  $\text{TD}_{\min}(\mathcal{C})$ , is defined as*

$$\text{TD}_{\min}(\mathcal{C}) = \min_{c \in \mathcal{C}} \text{TD}(c; \mathcal{C}).$$

In the recursive teaching model (Zilles et al., 2011), the teacher exploits a hierarchy of the concept class  $\mathcal{C}$ . It recursively removes from the given concept class all concepts whose teaching dimension with respect to the remaining concepts is smallest. The recursive teaching dimension  $\text{RTD}$  of  $\mathcal{C}$  is defined as the largest value of the smallest teaching dimensions encountered in the recursive process.

**Definition 2 (Recursive Teaching Dimension (Zilles et al., 2011))** *For a given concept class  $\mathcal{C}$ , define a sequence  $\mathcal{C}_0, \mathcal{C}_1, \dots, \mathcal{C}_T$  such that  $\mathcal{C}_0 = \mathcal{C}$ , and  $\mathcal{C}_{t+1} = \mathcal{C}_t \setminus \{c \in \mathcal{C}_t : \text{TD}(c; \mathcal{C}_t) = \text{TD}_{\min}(\mathcal{C}_t)\}$ . Here  $T$  is the smallest integer so that  $\mathcal{C}_{T+1} = \emptyset$ . The recursive teaching dimension of  $\mathcal{C}$ , denoted by  $\text{RTD}(\mathcal{C})$ , is defined as  $\text{RTD}(\mathcal{C}) = \max_{0 \leq t \leq T} \text{TD}_{\min}(\mathcal{C}_t)$ .*

Our goal is to bound  $\text{RTD}(\mathcal{C})$  in terms of  $\text{VCD}(\mathcal{C})$ . It turns out that rather than studying VC dimension and shattering directly, considering the number of projection patterns is more helpful (Kuhlmann, 1999; Moran et al., 2015).

**Definition 3 ( $(x, y)$ -class)** *We say a concept class  $\mathcal{C} \subseteq \{0, 1\}^n$  is an  $(x, y)$ -class for positive integers  $x, y$ , if for any  $A \subseteq [n]$  such that  $|A| \leq x$ ,  $|\{c|_A : c \in \mathcal{C}\}| \leq y$ .*

In the rest of this paper we will frequently use the following observations. A concept class  $\mathcal{C}$  with  $\text{VCD}(\mathcal{C}) = d$  is a  $(d + 1, 2^{d+1} - 1)$ -class. More importantly,  $\mathcal{C}$  is an  $(x, \lfloor (\frac{e}{d})^d \rfloor)$ -class for every  $x > d$ , due to Sauer's lemma stated below.

**Theorem 4 (Sauer-Shelah Lemma (Sauer, 1972; Shelah, 1972))** *Let  $\mathcal{C} \subseteq \{0, 1\}^n$  be a concept class with  $\text{VCD}(\mathcal{C}) = d$ . Then for any  $A \subseteq [n]$  such that  $|A| > d$ ,*

$$|\{c|_A : c \in \mathcal{C}\}| \leq \sum_{k=0}^d \binom{|A|}{k} \leq \left(\frac{e|A|}{d}\right)^d.$$

Our main result is based on an analysis of the largest possible best-case teaching dimension of all finite  $(x, y)$ -classes.

**Definition 5** Define  $f(x, y) = \sup_{\mathcal{C}} \text{TD}_{\min}(\mathcal{C})$ , where the supremum is taken over all finite  $(x, y)$ -classes  $\mathcal{C}$ .

[Kuhlmann \(1999\)](#) proved  $f(2, 3) = 1$ , and [Moran et al. \(2015\)](#) proved  $f(3, 6) \leq 3$ .

### 3. Main Results

In this section, we state and prove our main result. We show that for any finite concept class  $\mathcal{C}$ ,  $\text{RTD}(\mathcal{C})$  is quadratically upper bounded by  $\text{VCD}(\mathcal{C})$ .

**Theorem 6** For any concept class  $\mathcal{C} \subseteq \{0, 1\}^n$  with  $\text{VCD}(\mathcal{C}) = d$ ,

$$\text{RTD}(\mathcal{C}) = O(d^2).$$

We first give an informal description of the proof, in which we extend the techniques developed in ([Kuhlmann, 1999](#); [Moran et al., 2015](#); [Chen et al., 2016](#)). The key idea of our approach is to analyze  $f(x, y)$ , the largest possible best-case teaching dimension for  $(x, y)$ -classes. The first step is to show a recursive formula for  $f(x, y)$ . The observation is that for a monotone increasing function  $\phi(x)$  that grows substantially slower than  $2^x$ , we have

$$f(x + 1, \phi(x + 1)) \leq f(x, \phi(x)) + O(x).$$

The recursive formula immediately leads to a quadratic upper bound  $f(x, \phi(x)) \leq O(x^2)$ .

The second step is to select an appropriate function  $\phi(\cdot)$ . We choose  $\phi(x) = \alpha^x$  for certain  $\alpha \in (1, 2)$ . Next, we relate the VC dimension to  $f(x, y)$ . We show that for any finite concept class  $\mathcal{C}$  with  $\text{VCD}(\mathcal{C}) = d$ ,  $\mathcal{C}$  must be an  $(x, \alpha^x)$ -class for some  $x$  not much larger than  $d$ . In fact, it suffices when  $x$  is a constant times of  $d$ . Combining the above arguments, we have shown that the best-case teaching dimension of  $\mathcal{C}$  is upper bounded by  $O(d^2)$ . Finally, a standard argument yields  $\text{RTD}(\mathcal{C}) = O(d^2)$ .

Now we give the formal proof of [Theorem 6](#). The next lemma gives the recursive formula of  $f(x, y)$ .

**Lemma 7** For any positive integer  $x, y, z$  such that  $y \leq 2^x - 1$  and  $z \leq 2y + 1$ , the following inequality holds:

$$f(x + 1, z) \leq f(x, y) + \left\lceil \frac{(y + 1)(x - 1) + 1}{2y - z + 2} \right\rceil.$$

**Proof** For convenience, let  $k = \left\lceil \frac{(y+1)(x-1)+1}{2y-z+2} \right\rceil$ . For any concept class  $\mathcal{C} \subseteq \{0, 1\}^n$ , we only need to show that if  $\mathcal{C}$  is an  $(x + 1, z)$ -class, then

$$\text{TD}_{\min}(\mathcal{C}) \leq f(x, y) + k.$$

If  $n < k$ , the theorem is trivial, because

$$\text{TD}_{\min}(\mathcal{C}) \leq n < k \leq f(x, y) + k.$$

Assume  $n \geq k$  in the rest of the proof. For any  $Y \subseteq [n]$ ,  $|Y| = k$ , and any  $b \in \{0, 1\}^k$ , define

$$\mathcal{C}^{Y,b} := \{c \in \mathcal{C} : c|_Y = b\}.$$

Following the approach of (Kuhlmann, 1999; Moran et al., 2015; Chen et al., 2016), we choose  $Y^*, b^*$  among all possible  $Y, b$  such that  $\mathcal{C}^{Y^*, b^*}$  is nonempty and has the *smallest* size. Without loss of generality, we assume  $b^* = \mathbf{0}$ .

If  $\mathcal{C}^{Y^*, b^*}$  is an  $(x, y)$ -class, our proof is finished, because we can find a concept  $c \in \mathcal{C}^{Y^*, b^*}$  so that  $c$  has a teaching set  $T \subseteq [n] \setminus Y^*$  of size no more than  $f(x, y)$  which distinguishes  $c$  from all other concepts in  $\mathcal{C}^{Y^*, b^*}$ . Then  $T \cup Y^*$  is a teaching set that distinguishes  $c$  from all other concepts in  $\mathcal{C}$ . The fact that  $|T \cup Y^*| \leq f(x, y) + k$  completes the proof.

Finally we show  $\mathcal{C}^{Y^*, b^*}$  is an  $(x, y)$ -class. The proof idea is as follows.

Suppose  $\mathcal{C}^{Y^*, b^*}$  is not an  $(x, y)$ -class. Then we can find  $Y'$  and  $b'$  such that  $|Y'| = |Y^*|$  and  $\mathcal{C}^{Y', b'}$  is a strict nonempty subset of  $\mathcal{C}^{Y^*, b^*}$ . This contradicts the assumption that  $\mathcal{C}^{Y^*, b^*}$  is the smallest nonempty set over all choices of  $Y, b$  such that  $|Y| = k$ . We construct  $Y', b'$  as this: if  $\mathcal{C}^{Y^*, b^*}$  is not an  $(x, y)$ -class, then we can find two sets  $Z$  and  $W$  such that

1.  $|Z| = |W|$ ;
2. The intersection of  $Z$  and  $Y^*$  is the empty set;
3.  $W$  is a subset of  $Y^*$ .

Let  $Y' = (Y^* \setminus W) \cup Z$ , then  $|Y'| = |Y^*|$ . In addition, we can also find a string  $s$ , whose length equals to  $|Z|$ , such that  $\mathcal{C}^{(Y^* \setminus W) \cup Z, \mathbf{0} \circ s}$  is a strict (and nonempty) subset of  $\mathcal{C}^{Y^*, b^*}$ , where  $\mathbf{0}$  is the zero string of length  $|Y^* \setminus W|$ . (Note that we assumed w.l.o.g.  $b^*$  is the zero string.) Therefore, choosing  $b' = \mathbf{0} \circ s$  leads to the desired contradiction. The proof detail is as follows.

Assume for the sake of contradiction that  $\mathcal{C}^{Y^*, b^*}$  is not an  $(x, y)$ -class. Then there exists  $Z \subseteq [n]$  such that  $|Z| \leq x$  and  $|\{c|_Z : c \in \mathcal{C}^{Y^*, b^*}\}| \geq y + 1$ . Note that  $Z \setminus Y^*$  cannot be an empty set since  $y + 1 > 1$ . Without loss of generality, we assume  $Z \cap Y^* = \emptyset$ ; otherwise simply consider  $Z \setminus Y^*$  instead of  $Z$ .

Now define

$$\mathcal{C}_Z^{Y^*, b^*} := \{c|_Z : c \in \mathcal{C}^{Y^*, b^*}\},$$

and for every  $w \in Y^*$  define

$$\mathcal{C}_Z^{w, 1} := \{c|_Z : c \in \mathcal{C}, c|_{\{w\}} = 1\}.$$

Recall that  $\mathcal{C}$  is an  $(x + 1, z)$ -class,  $|Z| \leq x$ , and we assumed  $b^* = \mathbf{0}$ . Therefore, the projection of  $\mathcal{C}$  on the set  $Z \cup \{w\}$  has no more than  $z$  patterns. Thus

$$|\mathcal{C}_Z^{Y^*, b^*}| + |\mathcal{C}_Z^{w, 1}| \leq z.$$

Since  $|\mathcal{C}_Z^{Y^*, b^*}| \geq y + 1$ , we have  $|\mathcal{C}_Z^{w, 1}| \leq z - y - 1$ . Now, pick a subset  $\tilde{\mathcal{C}}_Z^{Y^*, b^*} \subseteq \mathcal{C}_Z^{Y^*, b^*}$  so that  $|\tilde{\mathcal{C}}_Z^{Y^*, b^*}| = y + 1$ . We have for every  $w \in Y^*$ ,  $|\tilde{\mathcal{C}}_Z^{Y^*, b^*} \setminus \mathcal{C}_Z^{w, 1}| \geq 2y - z + 2$ . Thus,

$$\begin{aligned} \sum_{w \in Y^*} |\tilde{\mathcal{C}}_Z^{Y^*, b^*} \setminus \mathcal{C}_Z^{w, 1}| &\geq k(2y - z + 2) \\ &> (y + 1)(x - 1) \\ &= |\tilde{\mathcal{C}}_Z^{Y^*, b^*}| \cdot (x - 1) \\ &\geq |\tilde{\mathcal{C}}_Z^{Y^*, b^*}| \cdot (|Z| - 1). \end{aligned}$$

It then follows from the Pigeonhole Principle that there exists  $W \subseteq Y^*$  such that  $|W| = |Z|$  and  $\bigcap_{w \in W} (\tilde{\mathcal{C}}_Z^{Y^*, b^*} \setminus \mathcal{C}_Z^{w, 1}) \neq \emptyset$ . Pick any string  $s \in \bigcap_{w \in W} (\tilde{\mathcal{C}}_Z^{Y^*, b^*} \setminus \mathcal{C}_Z^{w, 1})$ , and consider the set  $\mathcal{C}^{(Y^* \setminus W) \cup Z, \mathbf{0} \circ s}$  defined as

$$\mathcal{C}^{(Y^* \setminus W) \cup Z, \mathbf{0} \circ s} := \{c \in \mathcal{C} : c|_{(Y^* \setminus W)} = \mathbf{0}, c|_Z = s\}.$$

It is clear that  $\mathcal{C}^{(Y^* \setminus W) \cup Z, \mathbf{0} \circ s}$  is a nonempty and proper subset of  $\mathcal{C}^{Y^*, b^*}$ . This leads to a contradiction with the choice of  $Y^*, b^*$ .  $\blacksquare$

Using the recursive formula established in Lemma 7, we are able to give upper bound on the best-case teaching complexity for all  $(x, y)$ -classes.

**Lemma 8** For every  $\alpha \in (1, 2)$ , and every positive integer  $x$ ,

$$f(x, \lfloor \alpha^x \rfloor) \leq \frac{(x-1)^2}{4-2\alpha} + \frac{3-2\alpha}{4-2\alpha} \cdot (x-1).$$

**Proof** Applying lemma 7 by setting  $y = \lfloor \alpha^x \rfloor$  and  $z = \lfloor \alpha^{x+1} \rfloor$ , we have

$$f(x+1, \lfloor \alpha^{x+1} \rfloor) \leq f(x, \lfloor \alpha^x \rfloor) + \left\lceil \frac{(\lfloor \alpha^x \rfloor + 1)(x-1) + 1}{2 \lfloor \alpha^x \rfloor - \lfloor \alpha^{x+1} \rfloor + 2} \right\rceil. \quad (1)$$

Since

$$\left\lceil \frac{(\lfloor \alpha^x \rfloor + 1)(x-1) + 1}{2 \lfloor \alpha^x \rfloor - \lfloor \alpha^{x+1} \rfloor + 2} \right\rceil \leq \frac{x-1}{2 - \frac{\lfloor \alpha^{x+1} \rfloor}{\lfloor \alpha^x \rfloor + 1}} + 1 \leq \frac{x-1}{2-\alpha} + 1 = \frac{x+1-\alpha}{2-\alpha},$$

Inequality (1) can be simplified to

$$f(x+1, \lfloor \alpha^{x+1} \rfloor) \leq f(x, \lfloor \alpha^x \rfloor) + \frac{x+1-\alpha}{2-\alpha}.$$

Observe that  $f(1, 1) = 0$  and apply the above inequality recursively, we obtain

$$f(x, \lfloor \alpha^x \rfloor) \leq \frac{(x-1)^2}{4-2\alpha} + \frac{3-2\alpha}{4-2\alpha} \cdot (x-1).$$

$\blacksquare$

Next we show that for a concept class  $\mathcal{C}$  with  $\text{VCD}(\mathcal{C}) = d$ ,  $\mathcal{C}$  must be an  $(x, \lfloor \alpha^x \rfloor)$ -class for  $x$  not much larger than  $d$ .

**Lemma 9** Given  $\alpha \in (1, 2)$ , define

$$\lambda^* := \inf\{\lambda \geq 1 : \lambda \ln \alpha - \ln \lambda - 1 \geq 0\}.$$

Then for any concept class  $\mathcal{C} \subseteq \{0, 1\}^n$  with  $\text{VCD}(\mathcal{C}) = d$ ,  $\mathcal{C}$  is an  $(x, \lfloor \alpha^x \rfloor)$ -class for every integer  $x \geq \lambda^* d$ .

**Proof** By Sauer's lemma, we only need to verify

$$\left(\frac{ex}{d}\right)^d \leq \alpha^x,$$

holds for all  $x \geq \lambda^*d$ . This follows from elementary calculus. We omit the details. ■

Now we give the main conclusion.

**Theorem 10** For any concept class  $\mathcal{C} \subseteq \{0, 1\}^n$  with  $\text{VCD}(\mathcal{C}) = d$ ,

$$\text{RTD}(\mathcal{C}) \leq 39.3752d^2 - 3.6330d.$$

**Proof** By Lemma 8 and Lemma 9, we have for any  $\alpha \in (1, 2)$  and any  $x \geq \lambda^*d$ , where  $\lambda^*$  is defined in Lemma 9, the following holds

$$\text{TD}_{\min}(\mathcal{C}) \leq \frac{(x-1)^2}{4-2\alpha} + \frac{3-2\alpha}{4-2\alpha} \cdot (x-1).$$

Observe that the VC dimension of a concept class does not increase after a concept is removed, we have

$$\text{RTD}(\mathcal{C}) \leq \frac{(x-1)^2}{4-2\alpha} + \frac{3-2\alpha}{4-2\alpha} \cdot (x-1). \quad (2)$$

To optimize the coefficients in the quadratic bound, we choose  $\lambda^* = 4.71607$ ,  $\alpha = (e\lambda^*)^{1/\lambda^*} \approx 1.71757$ ,  $x = \lceil \lambda^*d \rceil$ . Finally, observe that the RHS of (2) is an increasing function of  $x$  on the interval  $[\lambda^*, +\infty)$  given our choice of the parameters, we conclude that

$$\text{RTD}(\mathcal{C}) \leq \frac{(\lambda^*d)^2}{4-2\alpha} + \frac{3-2\alpha}{4-2\alpha} \cdot \lambda^*d \leq 39.3752d^2 - 3.6330d. \quad \blacksquare$$

#### 4. Discussion and Conclusion

In the previous section we show that for finite concept class  $\mathcal{C}$ ,  $\text{RTD}(\mathcal{C}) = O(\text{VCD}(\mathcal{C})^2)$ . In this section, we discuss our thoughts on the challenges in fully solving the open problem  $\text{RTD}(\mathcal{C}) = O(\text{VCD}(\mathcal{C}))$ .

The key technical result in our proof is the quadratic upper bound in Lemma 8 which, loosely speaking, is that for  $\phi(x) < 2^x$

$$f(x, \phi(x)) = O(x^2), \quad (3)$$

which is based on the recursive formula

$$f(x+1, \phi(x+1)) \leq f(x, \phi(x)) + O(x), \quad (4)$$

In order to prove  $\text{RTD}(\mathcal{C}) = O(\text{VCD}(\mathcal{C}))$  (if it is true), one needs to strengthen (3) to

$$f(x, \phi(x)) = O(x).$$

If we still follow the recursive approach, we have to improve the recursive formula (4) to

$$f(x + 1, \phi(x + 1)) \leq f(x, \phi(x)) + O(1). \quad (5)$$

In our view, (5) is qualitatively different from (4); and this is the bottleneck of the current approach.

Another way to state the quadratic upper bound for  $f(x, y)$  in Lemma 8 is  $f(x, y) = O(\log^2 y)$  for  $y < 2^x$ . (To see this, observe  $f$  is non-increasing in  $x$  and non-decreasing in  $y$ .) The conjecture  $\text{RTD}(\mathcal{C}) = O(\text{VCD}(\mathcal{C}))$  is exactly equivalent to  $f(x, y) = O(\log y)$  for all  $y < 2^x$ . However, the only result we can show is that  $f(x, y) = O(\log y)$  for  $y$  not much larger than  $x$ .<sup>1</sup>

We also consider the relation between RTD and VCD via the probabilistic method. If we fix  $n$  and the size of the concept class as  $N$  ( $n, N$  sufficiently large), and draw  $N$  concepts from  $\{0, 1\}^n$  uniformly at random to form  $\mathcal{C}$ , it can be shown that with overwhelming probability  $\text{RTD}(\mathcal{C})$  is smaller than  $\text{VCD}(\mathcal{C})$ . Indeed, with probability at least  $1 - (\frac{1}{2})^{\Omega(\frac{n}{\log N})}$ ,  $\text{VCD}(\mathcal{C}) > \log_2(\frac{N}{\ln N})$  and  $\text{RTD}(\mathcal{C}) < \log_2(\frac{N}{\ln N})$  hold true simultaneously for sufficiently large  $n, N$ . Although this does not prove any bound, it tells us that the cases  $\text{RTD}(\mathcal{C}) \gg \text{VCD}(\mathcal{C})$  are rare.

So far we focus on the upper bounds for RTD in terms of VCD, and discuss the challenges in proving  $\text{RTD}(\mathcal{C}) = O(\text{VCD}(\mathcal{C}))$ . What if RTD is not linearly upper bounded by VCD? How to prove it? There are attempts along this line. [Kuhlmann \(1999\)](#) first showed there exist finite concept classes  $\mathcal{C}$  with  $\text{RTD}(\mathcal{C}) = \frac{3}{2}\text{VCD}(\mathcal{C})$ . Warmuth discovered the smallest such class ([Doliwa et al., 2014](#)). [Chen et al. \(2016\)](#), based on their insights and with the aid of SAT solvers, found finite concept classes  $\mathcal{C}$  with  $\text{RTD}(\mathcal{C}) = \frac{5}{3}\text{VCD}(\mathcal{C})$ . However, to prove RTD is not linearly bounded by VCD, we need a sequence  $\mathcal{C}_1, \mathcal{C}_2, \dots$  ( $\mathcal{C}_i \subseteq \{0, 1\}^{n_i}$ ) such that  $\frac{\text{RTD}(\mathcal{C}_i)}{\text{VCD}(\mathcal{C}_i)}$  grows beyond any constant.

In order that  $\frac{\text{RTD}(\mathcal{C}_i)}{\text{VCD}(\mathcal{C}_i)}$  grows unboundedly,  $n_i$  and  $|\mathcal{C}_i|$  have to grow unboundedly as well. This means that as the instance space is getting larger, there exist concept classes for which the ratio of RTD and VCD grows. However, currently there is no clue that larger  $n_i$  and  $|\mathcal{C}_i|$  would result in larger ratio between RTD and VCD in a *structural* way. The only known structural result is that for Cartesian product of two concept classes, the ratio does NOT grow. More concretely ([Doliwa et al., 2014](#)),

$$\text{RTD}(\mathcal{C}_1 \times \mathcal{C}_2) \leq \text{RTD}(\mathcal{C}_1) + \text{RTD}(\mathcal{C}_2),$$

and

$$\text{VCD}(\mathcal{C}_1 \times \mathcal{C}_2) = \text{VCD}(\mathcal{C}_1) + \text{VCD}(\mathcal{C}_2).$$

We believe any improvement along this line requires constructions more delicate in structure.

Our understanding of the quantitative relation between RTD and VCD is still preliminary. Even for the simple special case  $\text{VCD} = 2$ , we do not have a complete characterization: The best known upper bound for  $\mathcal{C} \subseteq \{0, 1\}^n$  with  $\text{VCD}(\mathcal{C}) = 2$  is  $\text{RTD}(\mathcal{C}) \leq 6$  ([Chen et al., 2016](#)); and the worst-case lower bound is  $\text{RTD}(\mathcal{C}) \geq 3$  ([Kuhlmann, 1999](#); [Doliwa et al., 2014](#)). The current knowledge of the four cases of  $\text{VCD}(\mathcal{C}) = 2$  (i.e., (3, 7), (3, 6), (3, 5), (3, 4)-classes) is not complete either: For (3, 7)-classes, [Chen et al. \(2016\)](#) proved

$$3 \leq \max_{\mathcal{C} \in (3,7)} \text{RTD}(\mathcal{C}) \leq 6;$$

For (3, 6)-classes, [Moran et al. \(2015\)](#) proved

$$2 \leq \max_{\mathcal{C} \in (3,6)} \text{RTD}(\mathcal{C}) \leq 3;$$

---

1. By induction on  $c$ , it's easy to show that  $\forall x, y, c \in \mathbb{N}^*, f(x + c, y + c) \leq \max\{f(x, y), \lfloor \log_2(y + c) \rfloor\}$ .

Using a similar argument as in the proof of Lemma 7 and optimizing the parameters with respect to the specific case of  $(3, 5)$ -classes (choosing  $x = 2, y = 3, z = 5, k = 1$ ), we can show

$$\max_{\mathcal{C} \in (3,5)} \text{RTD}(\mathcal{C}) = 2;$$

And hence for  $(3, 4)$ -classes  $\max_{\mathcal{C} \in (3,4)} \text{RTD}(\mathcal{C}) = 2$ .

The relationship between RTD and VCD is intriguing. Analyzing special cases and sub-classes with specific structure may provide insights for finally solving the open problem.

## Acknowledgments

Liwei Wang was partially supported by National Basic Research Program of China (973 Program) (grant no. 2015CB352502) and NSFC (61573026).

## References

- Dana Angluin. Queries revisited. *Theoretical Computer Science*, 313(2):175–194, 2004.
- Martin Anthony, Graham Brightwell, and John Shawe-Taylor. On specifying Boolean functions by labelled examples. *Discrete Applied Mathematics*, 61(1):1–25, 1995.
- 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.
- Xi Chen, Yu Cheng, and Bo Tang. On the recursive teaching dimension of VC classes. In *NIPS*, pages 2164–2171, 2016.
- Sanjoy Dasgupta. Coarse sample complexity bounds for active learning. In *NIPS*, volume 18, pages 235–242, 2005.
- Ofir David, Shay Moran, and Amir Yehudayoff. On statistical learning via the lens of compression. In *NIPS*, 2016.
- Thorsten Doliwa, Gaojian Fan, Hans Ulrich Simon, and Sandra Zilles. Recursive teaching dimension, VC-dimension and sample compression. *Journal of Machine Learning Research*, 15(1):3107–3131, 2014.
- Sally A Goldman and Michael J Kearns. On the complexity of teaching. *Journal of Computer and System Sciences*, 50(1):20–31, 1995.
- Sally A Goldman and H David Mathias. Teaching a smart learner. In *Proceedings of the sixth annual conference on Computational learning theory*, pages 67–76. ACM, 1993.
- Steve Hanneke. Teaching dimension and the complexity of active learning. In *International Conference on Computational Learning Theory*, pages 66–81. Springer, 2007.
- Tibor Hegedűs. Generalized teaching dimensions and the query complexity of learning. In *Proceedings of the eighth annual conference on Computational learning theory*, pages 108–117. ACM, 1995.

- Christian Kuhlmann. On teaching and learning intersection-closed concept classes. In *European Conference on Computational Learning Theory*, pages 168–182. Springer, 1999.
- Nick Littlestone and Manfred Warmuth. Relating data compression and learnability. Technical report, Technical report, University of California, Santa Cruz, 1986.
- Shay Moran, Amir Shpilka, Avi Wigderson, and Amir Yehudayoff. Compressing and teaching for low VC-dimension. In *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*, pages 40–51. IEEE, 2015.
- Norbert Sauer. On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13(1):145–147, 1972.
- Saharon Shelah. A combinatorial problem; stability and order for models and theories in infinitary languages. *Pacific Journal of Mathematics*, 41(1):247–261, 1972.
- Ayumi Shinohara and Satoru Miyano. Teachability in computational learning. *New Generation Computing*, 8(4):337–347, 1991.
- Hans Ulrich Simon and Sandra Zilles. Open problem: Recursive teaching dimension versus VC dimension. In *COLT*, pages 1770–1772, 2015.
- VN Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, 1971.
- Manfred K Warmuth. Compressing to VC dimension many points. In *COLT*, volume 3, pages 743–744. Springer, 2003.
- Sandra Zilles, Steffen Lange, Robert Holte, and Martin Zinkevich. Models of cooperative teaching and learning. *Journal of Machine Learning Research*, 12(Feb):349–384, 2011.