Proceedings of Machine Learning Research vol 291:1–16, 2025 38th Annual Conference on Learning Theory

Lower Bounds for Greedy Teaching Set Constructions

Spencer Compton COMPTONS@STANFORD.EDU
Stanford University

Chirag Pabbaraju CPABBARA@STANFORD.EDU
Stanford University

Nikita Zhivotovskiy ZHIVOTOVSKIY@BERKELEY.EDU
UC Berkeley

Editors: Nika Haghtalab and Ankur Moitra

Abstract

A fundamental open problem in learning theory is to characterize the best-case teaching dimension $\mathrm{TS}_{\min}$ of a concept class $\mathcal{C}$ with finite VC dimension d. Resolving this problem will, in particular, settle the conjectured upper bound on Recursive Teaching Dimension posed by Simon and Zilles (COLT 2015). Prior work used a natural greedy algorithm to construct teaching sets recursively, thereby proving upper bounds on $\mathrm{TS}_{\min}$, with the best known bound being $O(d^2)$ (Hu, Wu, Li, and Wang, COLT 2017). In each iteration, this greedy algorithm chooses to add to the teaching set the k labeled points that restrict the concept class the most. In this work, we prove lower bounds on the performance of this greedy approach for small k. Specifically, we show that for $k = 1$, the algorithm does not improve upon the halving-based bound of $O(\log(|\mathcal{C}|))$. Furthermore, for $k = 2$, we complement the upper bound of $O\left(\log(\log(|\mathcal{C}|))\right)$ from Moran, Shpilka, Wigderson, and Yehudayoff (FOCS 2015) with a matching lower bound. Most consequentially, our lower bound extends up to $k \leq \lceil cd \rceil$ for small constant $c > 0$: suggesting that studying higher-order interactions may be necessary to resolve the conjecture that $\mathrm{TS}_{\min} = O(d)$.

1. Introduction

One of the well-known open problems in learning theory is to characterize the size of the smallest teaching set, called the *best-case teaching dimension* and denoted by $\mathrm{TS}_{\min}$, for a concept class of finite VC dimension. More specifically, given a concept class $\mathcal{C}$ of VC dimension d defined on a finite domain $\mathcal{X}$, we ask whether there exists a concept $c \in \mathcal{C}$ whose teaching set (i.e., a set of domain elements on which c differs from every other concept in $\mathcal{C}$) has size $O(d)$. Addressing this question would immediately solve the COLT Open Problem posed by Simon and Zilles (2015), where the conjecture is that the *Recursive Teaching Dimension* is $O(d)$. We refer to (Doliwa et al., 2014) for exact definitions and a detailed account of the topic.

The early works on determining the teaching dimension trace back to Cover (1965), who showed that $\mathrm{TS}_{\min} = O(d)$ for the concept class induced by half-spaces. In fact, Cover used the geometric structure of half-spaces to bound the *average-case teaching dimension* by $O(d)$, which implies the desired $\mathrm{TS}_{\min} = O(d)$. A generalization of this approach was studied in (Doliwa et al., 2014), where the authors analyzed *shortest-path closed* concept classes and showed that their average teaching set has size $O(d)$, implying the same bound on $\mathrm{TS}_{\min}$. However, there is a known limitation to using average teaching set size to bound $\mathrm{TS}_{\min}$, since the average-case teaching dimension need not be bounded by the VC dimension (Kushilevitz et al., 1996; Doliwa et al., 2014). A related line of work also leverages additional structure on the concept class, such as intersection-closed classes (Kuhlmann, 1999; Doliwa et al., 2014), which likewise guarantee $\mathrm{TS}_{\min} = O(d)$. However, these structural assumptions are class-specific and do not provide a general solution to bounding $\mathrm{TS}_{\min}$ for arbitrary VC classes.

A line of work exploiting *no properties* of the concept class beyond its finite VC dimension was initiated by Kuhlmann (1999), who proved that $\mathrm{TS}_{\min} = 1$ for classes with $d = 1$. For general d, there is a simple halving-based bound of $\mathrm{TS}_{\min} = O\big(\log(|\mathcal{C}|)\big)$, which is sensitive to the size of $\mathcal{C}$. The first result to improve on this basic result was given by Moran, Shpilka, Wigderson, and Yehudayoff (2015), who showed that $\mathrm{TS}_{\min} = O\big(d2^d \log(\log(|\mathcal{C}|))\big)$. In fact, the dependence on $|\mathcal{C}|$ is unnecessary and the bound of Moran et al. (2015) has been subsequently improved to $O(d2^d)$ by Chen, Cheng, and Tang (2016), and later to the current state-of-the-art bound $O(d^2)$ by Hu, Wu, Li, and Wang (2017).

Remarkably, the proofs of all the above-mentioned general bounds on $\mathrm{TS}_{\min}$ rely on the same greedy strategy, formally given as Algorithm 1. The idea is as follows: pick an integer k (the "greediness" parameter), and at each step find a subset $T^\star \subset \mathcal{X}$ of size k together with a binary pattern $b^\star \in \{0,1\}^k$ such that the number of concepts in $\mathcal{C}$ agreeing with $b^\star$ on $T^\star$ is minimized (but still nonempty). Then add $T^\star$ to the current teaching set, and restrict $\mathcal{C}$ to those concepts matching $b^\star$ on $T^\star$. This restriction is denoted by $\mathcal{C}|_{T^\star,b^\star}$. Repeat this procedure until the remaining concept class contains exactly one concept, which is then characterized by the constructed teaching set.

As a proof of concept, the result of Kuhlmann (1999) for $d = 1$ follows from applying Algorithm 1 with $k = 1$. The definition of the VC dimension in his proof ensures that the algorithm finishes constructing the teaching set right after the first iteration, thereby establishing $\mathrm{TS}_{\min} = 1$ when $d = 1$. Similarly, the analysis in (Moran et al., 2015) applies Algorithm 1 with $k = 2$, the bound of Chen et al. (2016)[1] with $k = 2^d(d-1)+1$, while the bound of Hu et al. (2017) can be modified slightly so that it corresponds to the application of the algorithm with $k = O(d)$.

The main aim of this work is to study the limitations of Algorithm 1 as a general method for constructing teaching sets for different values of k. Prior to this work, even for $k = 1$, the size of the constructed teaching sets remained unclear for general concept classes. We show that, in the worst case, Algorithm 1 with $k = 1$ cannot achieve any better dependence on the size of the teaching set than the halving-based bound $O(\log(|\mathcal{C}|))$. Since our focus is on lower bounds, we establish the tightness of some previous results, including the somewhat exotic $O(\log(\log(|\mathcal{C}|)))$ dependence on the teaching set size in (Moran et al., 2015) when the algorithm is run with $k = 2$. We summarize our findings in the following table:

	Upper Bound	**Lower Bound**
$k = 1$	$O\big(\log\|\mathcal{C}\|\big)$; folklore	$\Omega\big(\log\|\mathcal{C}\|\big)$; Theorem 1
$2 \le k \le \lceil cd \rceil$ for some $c > 0$	$O_{d,k}\big(\log(\log\|\mathcal{C}\|)\big)$; (Moran et al., 2015)[2]	$\Omega_{d,k}\big(\log(\log\|\mathcal{C}\|)\big)$; Theorem 4
$k = \lceil c'd \rceil$ for a larger $c' > 0$	$O\big(d^2\big)$; (Hu et al., 2017)	$\Omega\big(d\big)$; trivial

$O_{d,k}(\cdot)$ denotes ignoring multiplicative factors in d, k. Prior to our work, only the trivial lower bound of $\Omega(d)$ (e.g., $\mathcal{C} = \{0,1\}^d$) was known for all settings.

1. Using a fixed k will incur a slightly worse guarantee of $O(d^2 2^d)$, compared to their $O(d2^d)$ guarantee with exponentially decreasing k.
2. The upper bound in (Moran et al., 2015) is shown for $k = 2$, but their proof implies the same bound for all $k > 2$.

Algorithm 1 Greedy algorithm for constructing teaching sets

Input: Concept class $\mathcal{C}$, greediness parameter k
Output: Teaching set S for some concept $c \in \mathcal{C}$
Procedure `Greedy` $(\mathcal{C}, k)$:
  $S \leftarrow \emptyset$
  while $|\mathcal{C}| > 1$ **do**
    $T^\star, b^\star \leftarrow \underset{\substack{T \subseteq \mathcal{X},\, 1 \le |T| \le k \\ b \in \{0,1\}^{|T|} \\ |\mathcal{C}|_{T,b}| > 0}}{\arg\min} \left|\mathcal{C}|_{T,b}\right|$ ▷ Greedily compute smallest restriction[3]
    $S \leftarrow S \cup T^\star$ ▷ Add $T^\star$ to teaching set
    $\mathcal{C} \leftarrow \mathcal{C}|_{T^\star, b^\star}$ ▷ Restrict $\mathcal{C}$ to teaching set constructed so far
  end
  return S, c (where $\mathcal{C} = \{c\}$)

The main consequence of our result is to refute the plausible-seeming agenda of resolving the $\mathrm{TS}_{\min} = O(d)$ conjecture by more sharply analyzing the natural greedy Algorithm 1 for smaller $k = o(d)$; we show that Algorithm 1 may fail to construct small teaching sets even when $k = \lceil c\, d \rceil$ for a sufficiently small absolute constant $c > 0$. This suggests that a better study of higher-order interactions, or some approach that exploits the overall structure of the concept class, might be necessary. Note how this does not imply the greedy approach is suboptimal for large k; by definition, the greedy algorithm optimally finishes in one round when $k = \mathrm{TS}_{\min}$. Our construction reveals an unexpected phase transition: if indeed $\mathrm{TS}_{\min} = O(d)$ holds, then by selecting *a sufficiently large* constant $c' > 0$, Algorithm 1 with $k = \lceil c'\, d \rceil$ should be capable of constructing the desired teaching set in a single iteration, as dictated by its definition. However, our findings show that for $k = \lceil c\, d \rceil$ with *a small* constant c, the algorithm fails to achieve the desired bound.

The remainder of the paper is organized as follows. In Section 2 we analyze the basic case of $k = 1$, and in Section 3 we consider other values of k. Both concept classes have small $\mathrm{TS}_{\min}$: they are barriers for the greedy algorithm with small k, not for the general $\mathrm{TS}_{\min} = O(d)$ conjecture (see Appendix C). To keep the paper concise, it is assumed that the reader is familiar with standard definitions and results, such as VC dimension; further details can be found in standard textbooks.

2. Lower bound for $k = 1$

Our first main result is a lower bound on the size of the teaching set returned by Algorithm 1 and the procedure GREEDY for $k = 1$. The geometric construction employed in this proof serves both as a foundation and as an illustration for our analysis when $k \geq 2$, which is presented in Section 3. In fact, the proof of Theorem 1 is driven by the illustration in Figure 1. Once the construction and the procedure of Algorithm 1 for $k = 1$ are understood, the remainder of the section is devoted to formalizing the intuitively clear argument.

Before we present our statement and construction, we recall that each rectangle classifies points in its interior and along its border as 1, and points in its exterior as 0.

3. Here, we may assume that ties are broken in favor of a T that has smaller size.

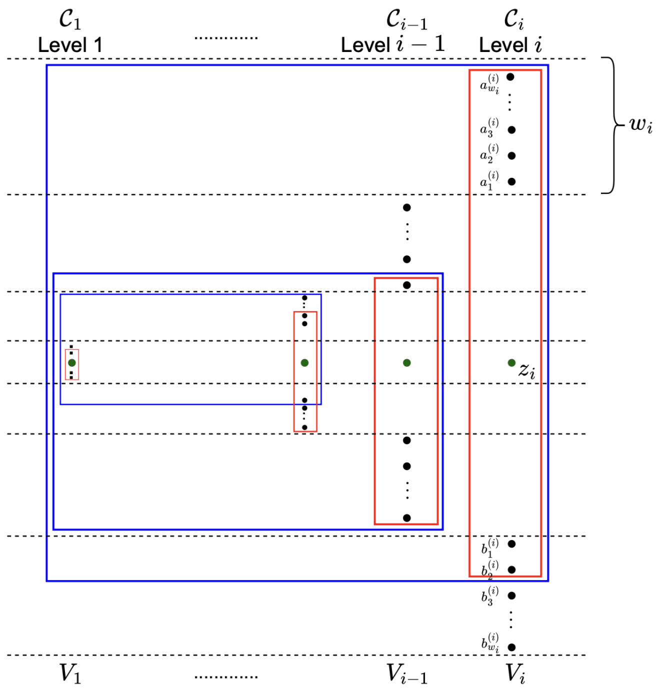

Figure 1: The arrangement of the point sets $V_1 \cup V_2 \cup \dots$ in the two-dimensional plane. The black horizontal dashed lines delineate the "vertical ranges" of these sets. The red rectangles denote concepts in $\mathcal{C}_i^{(\text{up, }1)}$ and $\mathcal{C}_i^{(\text{down, }1)}$, whereas the blue rectangles denote concepts in $\mathcal{C}_i^{(\text{up, }2)}$ and $\mathcal{C}_i^{(\text{down, }2)}$. For example, observe how the red rectangle in Level i, which belongs to $\mathcal{C}_i^{(\text{up, }1)}$, is enlarged to include all the points in $V_1 \cup \dots \cup V_{i-1}$, and yields the corresponding blue rectangle from $\mathcal{C}_i^{(\text{up, }2)}$.

Theorem 1 (Rectangles Lower Bound for $k = 1$) *There exists a family $\{\mathcal{F}_N\}$ of concept classes (here $N = 1, 2, \dots$) such that*

1. *$\mathcal{F}_N$ consists of indicators of axis-aligned rectangles in $\mathbb{R}^2$ (i.e., VC dimension at most 4),*
2. *$\mathcal{F}_N$ has size $2^{\Theta(N)}$ and is defined on a domain $\mathcal{X} \subseteq \mathbb{R}^2$ of size $2^{\Theta(N)}$,*
3. *$\textsc{Greedy}(\mathcal{F}_N, 1)$ returns a teaching set of size at least $N = \Omega(\log(|\mathcal{F}_N|))$.*

The remainder of the section is devoted to the proof of Theorem 1.

Construction of the class. We first describe the construction of the concept class $\mathcal{F}_N$ for every $N \geq 1$. *Domain.* Consider a collection of sets of points $V_1 \cup \dots \cup V_N$. Each V_i consists of a center point z_i, $w_i \triangleq 2^{10i}$ points extending vertically up, and w_i points extending vertically down; thus, $|V_i| = 2w_i + 1$. Each set V_i will be placed horizontally next to V_{i+1} (see Figure 1). The domain

$\mathcal{X} \subseteq \mathbb{R}^2$ will precisely be $\mathcal{X} = V_1 \cup V_2 \cup \cdots \cup V_N$, so that

$$|\mathcal{X}| = \sum_{i=1}^{N} |V_i| = \sum_{i=1}^{N} (2^{10i+1} + 1) \leq \sum_{i=1}^{N} 2^{10i+2} \leq 2^{10N+3},$$

$$\text{and also } |\mathcal{X}| = \sum_{i=1}^{N} (2^{10i+1} + 1) \geq 2^{10N+1}.$$

Thus, $|\mathcal{X}| = 2^{\Theta(N)}$ as required.

Concept class. We now describe the concept class $\mathcal{F}_N$. We construct $\mathcal{F}_N$ as the union of $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_N$, where we think of $\mathcal{C}_i$ to be defined at the "i^{th} level". It is helpful to refer to Figure 1. To define $\mathcal{C}_i$, denote the w_i points vertically above the center z_i as $a_1^{(i)}, \ldots, a_{w_i}^{(i)}$, and those vertically below z_i as $b_1^{(i)}, \ldots, b_{w_i}^{(i)}$. Let $\mathcal{C}_i^{(\text{up, }1)}$ consist of all rectangles that precisely contain $a_1^{(i)}, \ldots, a_{w_i}^{(i)}$, the center z_i, and then additionally a (potentially empty) prefix of $b_1^{(i)}, \ldots, b_{w_i}^{(i)}$. Similarly, let $\mathcal{C}_i^{(\text{down, }1)}$ consist of all rectangles that precisely contain $b_1^{(i)}, \ldots, b_{w_i}^{(i)}$, the center z_i, and then additionally a (potentially empty) prefix of $a_1^{(i)}, \ldots, a_{w_i}^{(i)}$. We have that $|\mathcal{C}_i^{(\text{up, }1)}| = |\mathcal{C}_i^{(\text{down, }1)}| = w_i + 1$, and both contain a common rectangle that contains all the points in V_i.

Consider enlarging every rectangle in $\mathcal{C}_i^{(\text{up, }1)}$ to additionally include all the points in $V_1 \cup V_2 \cup \cdots \cup V_{i-1}$. These enlarged rectangles form $\mathcal{C}_i^{(\text{up, }2)}$. Similarly, enlarge every rectangle in $\mathcal{C}_i^{(\text{down, }1)}$ to include all the points in $V_1 \cup V_2 \cup \cdots \cup V_{i-1}$. These enlarged rectangles form $\mathcal{C}_i^{(\text{down, }2)}$. Again, we have that $|\mathcal{C}_i^{(\text{up, }2)}| = |\mathcal{C}_i^{(\text{down, }2)}| = w_i + 1$, and both contain a common rectangle that contains all the points in $V_1 \cup \cdots \cup V_i$.

The concept class $\mathcal{C}_i$ is then simply $\mathcal{C}_i^{(\text{up, }1)} \cup \mathcal{C}_i^{(\text{up, }2)} \cup \mathcal{C}_i^{(\text{down, }1)} \cup \mathcal{C}_i^{(\text{down, }2)}$. Subtracting out the common concepts, we have that $|\mathcal{C}_i| = 4(w_i + 1) - 2 = 4w_i + 2$. This gives us that

$$|\mathcal{F}_N| = \sum_{i=1}^{N} |\mathcal{C}_i| = \sum_{i=1}^{N} (4w_i + 2) = \sum_{i=1}^{N} (2^{10i+2} + 2) \leq \sum_{i=1}^{N} 2^{10i+3} \leq 2^{10N+4},$$

$$\text{and also } |\mathcal{F}_N| = \sum_{i=1}^{N} (2^{10i+2} + 2) \geq 2^{10N+2}.$$

Thus, $|\mathcal{F}_N| = 2^{\Theta(N)}$ also as required. Note also that every rectangle in $\mathcal{C}_i$ contains the center z_i, and either all the points in $V_1 \cup V_2 \cup \cdots \cup V_{i-1}$ or none of them. Additionally, any point in $V_1 \cup V_2 \cup \cdots \cup V_{i-1}$ is contained in exactly half the rectangles in $\mathcal{C}_i$, and any point in V_i that is not the center is contained in at least $2w_i + 2$ rectangles in $\mathcal{C}_i$.

The following property of the construction will also be useful ahead. It says that the number of concepts in the i^{th} level is much larger than the *total* number of concepts in all lower levels.

Claim 2 (A level dominates all lower levels) *For any $i \in \{2, \ldots, N\}$,*

$$\sum_{j=1}^{i-1} |\mathcal{C}_j| < \frac{1}{2} |\mathcal{C}_i|.$$

Proof We have that

$$\sum_{j=1}^{i-1}|\mathcal{C}_j| = \sum_{j=1}^{i-1}(4w_j+2) = \sum_{j=1}^{i-1}(2^{10j+2}+2)$$
$$\leq 2^{10(i-1)+4} < \frac{1}{2}(4\cdot 2^{10i}+2) = \frac{1}{2}(4w_i+2) = \frac{1}{2}|\mathcal{C}_i|.$$

∎

We can show a lower bound for the teaching set constructed by the greedy algorithm with greediness parameter $k=1$, when it is instantiated for our constructed concept class $\mathcal{F}_N = \bigcup_{i=1}^N \mathcal{C}_i$.

Lemma 3 *The teaching set returned by* GREEDY$(\mathcal{F}_N, 1)$ *has size at least* N.

Proof For $i = 0, 1, 2, \dots, N-1$, we claim that at the beginning of the i^{th} iteration of the while loop of Algorithm 1 (where $i=0$ refers to the first iteration),

$$\mathcal{C} = \bigcup_{j=1}^{N-i} \mathcal{C}_i \quad \text{and } S = \{z_{N-i+1}, z_{N-i+2}, \dots, z_N\}. \tag{1}$$

We argue this inductively. When $i=0$, we are just entering the while loop for the very first time, and so $\mathcal{C} = \mathcal{F}_N = \bigcup_{j=1}^N \mathcal{C}_i$, and also $S=\emptyset$. Now, suppose that the claim holds for some $i \geq 0$: we will show that it continues to hold for $i+1$. In particular, we will argue that in the i^{th} iteration of the while loop, $T^\star$ is chosen to be $\{z_{N-i}\}$ and $b^\star$ to be 0 in Algorithm 1. This will prove the claim, since (i) $T^\star$ gets appended to S in Algorithm 1, (ii) all the concepts in $\mathcal{C}_{N-i}$ get removed from $\mathcal{C}$ upon restricting to $T^\star, b^\star$ in Algorithm 1, since every concept in $\mathcal{C}_{N-i}$ labels z_{N-i} (which is the center of V_{N-i}) as 1, and (iii) no concepts in $\mathcal{C}_1, \dots, \mathcal{C}_{N-i-1}$ are removed, since all such concepts label $\{z_{N-i}\}$ as 0.

For any x, let $\mathcal{C}(x,0)$ and $\mathcal{C}(x,1)$ denote the concepts in $\mathcal{C}$ that label x as 0 and 1 respectively; here, $\mathcal{C} = \bigcup_{j=1}^{N-i} \mathcal{C}_i$ is the effective concept class at the beginning of the i^{th} iteration of the while loop. Observe that for any $x \notin V_1 \cup \dots \cup V_{N-i}$, $\mathcal{C}(x,1) = \emptyset$ (such an x is strictly to the right of the remaining rectangles in $\mathcal{C}$). Thus, $T = \{x\}, b=1$ is not under consideration for the greedy choice. Furthermore, observe also that $\mathcal{C}(x,0) = \mathcal{C}$. Since $\mathcal{C}$ has at least two concepts (by virtue of entering the loop), these two concepts must then differ on some $y \in V_1 \cup \dots \cup V_{N-i}$, which means that the $\arg\min$ can also not be attained at $T=x, b=0$ (since, e.g., restricting to $T=\{y\}, b=1$ reduces the size of the class by at least 1).

We can thus restrict our attention to $x \in V_1 \cup \dots \cup V_{N-i}$. Note that $|\mathcal{C}(z_{N-i},0)| > 0$ (the rectangles to the left of z_{N-i} do not contain it). Note also that $|\mathcal{C}(z_{N-i},0)| < |\mathcal{C}(z_{N-i},1)|$ because all the concepts in $\mathcal{C}_{N-i}$ label z_{N-i} as 1. Even though all the remaining concepts label z_{N-i} as 0, we still know from Claim 2 that these are strictly less than half the number of concepts in $\mathcal{C}_{N-i}$.

Hence, our task is to show that for any $x \neq z_{N-i}$, $|\mathcal{C}(z_{N-i},0)| < |\mathcal{C}(x,0)|$ and $|\mathcal{C}(z_{N-i},0)| < |\mathcal{C}(x,1)|$. If this is the case, then $T = \{z_{N-i}\}, b=0$ will indeed realize the greedy choice in Algorithm 1. Note that this is equivalent to showing that $|\mathcal{C}(z_{N-i},1)| > |\mathcal{C}(x,1)|$ and $|\mathcal{C}(z_{N-i},1)| > |\mathcal{C}(x,0)|$.

Case 1: $x \in V_1 \cup V_2 \cup \cdots \cup V_{N-i-1}$, that is x is in a lower level than z_{N-i}.
Recall that exactly half of the concepts in $\mathcal{C}_{N-i}$ label such an x as 1. Then, even if all the remaining concepts were to label x as 1, we still have that

$$|\mathcal{C}(x,1)| \leq \frac{1}{2}|\mathcal{C}_{N-i}| + \sum_{j=1}^{N-i-1} |\mathcal{C}_j| < \frac{1}{2}|\mathcal{C}_{N-i}| + \frac{1}{2}|\mathcal{C}_{N-i}| = |\mathcal{C}_{N-i}| = |\mathcal{C}(z_{N-i},1)|,$$

where in the second inequality, we used Claim 2, and in the last equality, we used that all concepts in $\mathcal{C}_{N-i}$ label z_{N-i} as 1. By exactly the same calculation, we also get that $|\mathcal{C}(x,0)| < |\mathcal{C}(z_{n-i},1)|$.

Case 2: $x \in V_{N-i}$, that is, x is in the same level as z_{N-i} but is not the center.
We immediately have that $|\mathcal{C}(z_{N-i},1)| > |\mathcal{C}(x,1)|$, because only the concepts in $\mathcal{C}_{N-i}$ label these points as 1, and the concepts in $\mathcal{C}_{N-i}$ that label x as 1 are a strict subset of the concepts that label z_{N-i} as 1 (which is all of them). Furthermore, it is also the case that $|\mathcal{C}(z_{N-i},1)| > |\mathcal{C}(x,0)|$. To see this, observe how x is 0 in strictly less than half of C_{N-1}, so that

$$|\mathcal{C}(x,0)| < \frac{1}{2}|\mathcal{C}_{N-i}| + \sum_{j=1}^{N-i-1} |\mathcal{C}_j| < \frac{1}{2}|\mathcal{C}_{N-i}| + \frac{1}{2}|\mathcal{C}_{N-i}| = |\mathcal{C}_{N-i}| = |\mathcal{C}(z_{N-i},1)|.$$

This completes our inductive proof for (1). In particular, for $i = N-1$, we have shown that $S = \{z_2, z_3, \ldots, z_N\}$, and $\mathcal{C} = \mathcal{C}_1$. Again, $\mathcal{C}_1$ has at least two concepts that differ on some point in V_1, and so, the $(N-1)^{\text{th}}$ iteration of the while loop will find such a point in V_1 to add to S. Thus, the final teaching set that is returned has size at least N.

■

We now reflect on the structural properties of the construction that the proof effectively relied on. The number of concepts in level i heavily dominates (Claim 2) the total number of concepts in lower levels. Every concept in level i labels the center z_i as 1, which biases it. Since the concepts in lower levels form such a minority, they don't sway the bias of the center by much. On the other hand, for any point in a lower level, the concepts in level i maintain exactly a 50-50 balance. The rest of the concepts might sway this bias by a bit, but again, these concepts are too few when compared to the concepts in level i, so such points stay nearly unbiased.

3. Lower bound for $k \geq 2$

We will now show that for any $k \geq 2$, there is a concept class with VC dimension $d \leq 4k+1$ for which the $\log(\log(|\mathcal{C}|))$ dependence is unavoidable. This dependence on the concept class size appeared in (Moran et al., 2015), where the upper bound

$$\mathrm{TS}_{\min} = O\Big(d\, 2^d \log\big(\log(|\mathcal{C}|)\big)\Big)$$

was proven for $k = 2$. In particular, our next result shows that there exists a family of concept classes $\{\mathcal{F}_N\}$, each having VC dimension at most 9, such that GREEDY$(\mathcal{C}, 2)$ outputs a teaching set of size $\Omega\big(\log\big(\log(|\mathcal{F}_N|)\big)\big)$.

Our concept class will take inspiration from our $k=1$ construction, although we require a more complicated construction for this setting.

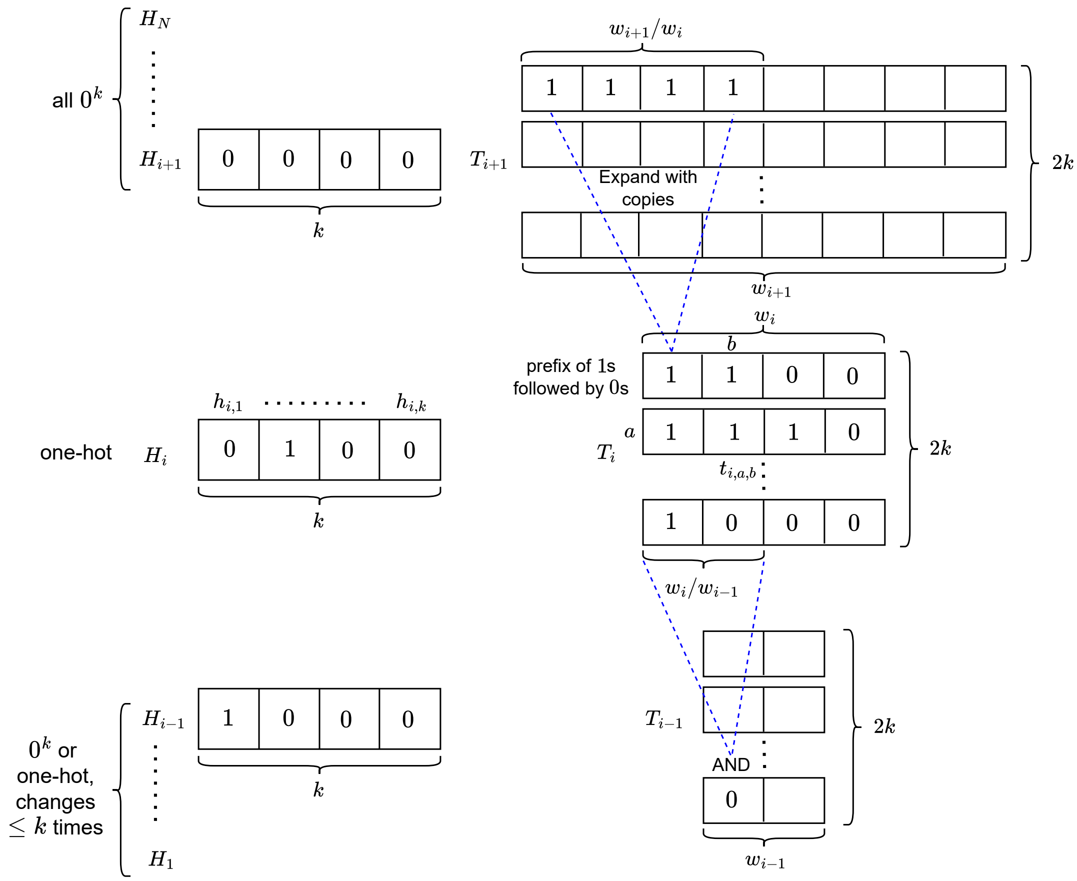

Figure 2: Example of a concept in $\mathcal{C}_i$ for Theorem 4. Given the prefixes for T_i, the values are deterministically expanded/contracted for other T_j. At most k values of $j \in \{1, \ldots, i-1\}$ may have H_j be different from H_{j+1}; in this concept a change occurred from H_{i-1} to H_i.

Theorem 4 (Lower Bound for $k \geq 2$) *For every positive integer $k \geq 2$, there exists a family $\{\mathcal{F}_N\}$ of concept classes (here $N = 1, 2, \ldots$) such that*

1. *$\mathcal{F}_N$ has VC dimension at most $4k + 1$,*
2. *$\mathcal{F}_N$ has size at most $2^{4k\log(8k)\cdot 2^{2N}}$ and is defined on a domain $\mathcal{X}$ of size at most $6k \cdot 2^{\log(8k)\cdot 2^{2N}}$,*
3. *GREEDY$(\mathcal{F}_N, k)$ returns a teaching set of size at least $kN = \Omega(\log(\log(|\mathcal{F}_N|)))$.*

Construction of the class. We begin by designing the concept class $\mathcal{F}_N$ for every $N \geq 1$.

Domain. Our domain $\mathcal{X}$ is an abstract finite set that consists of the union of N sets $V_1 \cup \cdots \cup V_N$. Each V_i is the union of two sets of domain points H_i and T_i. The set H_i has k *head points* $h_{i,1}, \ldots, h_{i,k}$. The set T_i has $2k$ rows $T_{i,1}, \ldots, T_{i,2k}$ of w_i *tail points* each, where $t_{i,a,b}$ denotes the tail point in the a^{th} row and b^{th} column of T_i (see Fig. 2). We set $w_i \triangleq 2^{\log(8k)\cdot 2^{2i}}$. Note that $w_{i+1} = w_i^4$; in particular, each w_i divides w_{i+1}, a property we will utilize. Roughly, the head points

H_i will play a similar role to the center point z_i in Theorem 1, and the tail points T_i will play a similar role to $V_i \setminus z_i$ in Theorem 1.

Concept class. The concept class $\mathcal{F}_N$ will be a union of $\mathcal{C}_1, \dots, \mathcal{C}_N$. Let us focus on the concepts in $\mathcal{C}_i$. We will specify these as a product of a concept class $\mathcal{A}_i$ on the combined set of head points $H_1, \dots, H_N$, and concept class $\mathcal{B}_i$ on the combined set of tail points $T_1, \dots, T_N$. That is, $\mathcal{C}_i \triangleq \mathcal{A}_i \otimes \mathcal{B}_i$, yielding a concept for every pair of concepts in $\mathcal{A}_i$ and $\mathcal{B}_i$.

The concept class $\mathcal{A}_i$ consists of all concepts c_h, where: (i) c_h labels every point in $H_{i+1}, \dots, H_N$ as 0, (ii) c_h labels H_i as a one-hot vector (that is, $c_h(h_{i,j}) = 1$ for exactly one $j \in [k]$), (iii) for any $j < i$, c_h labels H_j either as the zero vector 0^k, or as a one-hot vector, and (iv) there are at most k "changes" in the labeling of c_h on $H_1, \dots, H_{i-1}$; concretely, there are at most k indices in $\{1, 2, \dots, i-1\}$ where the labeling of c_h on H_j (as a k-dimensional vector) differs from the labeling on H_{j+1}. This is illustrated in the left half of Fig. 2. Property (iv) is enforced to control the VC dimension of $\mathcal{A}_i$ to be $O(k)$.

We now describe the concept class $\mathcal{B}_i$. We refer to the right half of Fig. 2 for better understanding. A concept c_t in $\mathcal{B}_i$ labels each row of T_i with a (possibly empty) prefix of 1s, followed by a (possibly empty) suffix of 0s—in total, there are $(w_i + 1)^{2k}$ possible choices of these prefix sizes over the $2k$ rows of T_i. We will have a concept c_t in $\mathcal{B}_i$ for each such choice; moreover, the labels of c_t on T_i will determine its labels on $T_1, \dots, T_{i-1}$ as well as $T_{i+1}, \dots, T_N$. Thus, the total size of $\mathcal{B}_i$ will be $(w_i + 1)^{2k}$. To describe the labels that a concept c realizes on these other tail sets, suppose that it labels the $2k$ rows of T_i as $(c_t(t_{i,1,1}), \dots, c_t(t_{i,1,w_i})), \dots, (c_t(t_{i,2k,1}), \dots, c_t(t_{i,2k,w_i}))$.

For $j > i$, starting with $j = i + 1$, the labeling of c_t on $T_{j,a}$ will be such that each $c_t(t_{j,a,b})$ is one of $\frac{w_j}{w_{j-1}}$ copies of the label that c_t assigns to a corresponding point in $T_{j-1,a}$. More concretely, for $j > i$,

$$c_t(t_{j,a,b}) = c_t\left(t_{j-1,a,\left\lceil \frac{b}{w_j/w_{j-1}} \right\rceil}\right). \tag{2}$$

Similarly, for $j < i$, starting with $j = i - 1$, each label that c_t assigns to a point in $T_{j,a}$ will be the logic AND of the labels it assigns to a batch of $\frac{w_{j+1}}{w_j}$ points in $T_{j+1,a}$. Concretely, for $j < i$,

$$c_t(t_{j,a,b}) = \mathrm{AND}\left(\left\{c_t\left(t_{j+1,a,\frac{w_{j+1}}{w_j}(b-1)+1}\right), \dots, c_t\left(t_{j+1,a,\frac{w_{j+1}}{w_j}(b-1)+\frac{w_{j+1}}{w_j}}\right)\right\}\right). \tag{3}$$

This construction is better understood visually via Fig. 2. We can see how a label of c_t in a particular row in T_i is copied over ("expanded") to w_{i+1}/w_i many slots in the corresponding row in T_{i+1}. Similarly, we can also see how the label of c_t in a row in T_{i-1} corresponds to the AND ("contraction") of w_i/w_{i-1} labels in the corresponding row in T_i. We note how these expansion/contraction operations maintain the property that every row of tail points is labeled with a prefix of 1s followed by a suffix of 0s.

We recall that $\mathcal{C}_i = \mathcal{A}_i \otimes \mathcal{B}_i$, and $\mathcal{F}_N = \bigcup_{i=1}^N \mathcal{C}_i$. We note also that the concept classes $\mathcal{C}_i$ are disjoint, as is witnessed by how for $i < j$, any concept in $\mathcal{C}_i$ labels H_j as 0^k, while every concept in $\mathcal{C}_j$ labels H_j as some one-hot vector.

Size of teaching set. We now analyze the size of the teaching set that GREEDY$(\mathcal{F}_N, k)$ returns.

Intuition. Our strategy will be similar to the proof of Theorem 1. We will aim to maintain that at the beginning of the i^{th} iteration of the while loop in GREEDY$(\mathcal{F}_N, k)$, the remaining concepts are exactly $\mathcal{C}_1 \cup \cdots \cup \mathcal{C}_{N-i}$. In particular, we will inductively prove that at iteration i, the algorithm picks $T^\star$ to be the k head points in H_{N-i}, and sets $b^\star = 0^k$. This removes all concepts in $\mathcal{C}_{N-i}$, yet none of the concepts in $\mathcal{C}_1, \dots, \mathcal{C}_{N-i-1}$. Proceeding thus for N iterations would then force the returned teaching set to be of size at least kN as desired. Our main intuition is as follows: for any choice of restriction other than $H_{N-i}, 0^k$ in Algorithm 1, there are at least k rows of T_{N-i} that are completely unrestricted, and hence these contribute at least $(w_{N-i}+1)^k$ concepts consistent with the restriction. However, our choice of w_{N-i} ensures that $(w_{N-i}+1)^k$ is strictly greater than the number of concepts in $\mathcal{C}_1 \cup \cdots \cup \mathcal{C}_{N-i-1}$; but these are precisely the concepts that remain if our restriction forces all of H_i to 0. We now make this formal.

For $i = 0, 1, 2, \dots, N-1$, we claim that at the beginning of the i^{th} iteration of the while loop in Algorithm 1 (where $i = 0$ refers to the first iteration),

$$\mathcal{C} = \bigcup_{j=1}^{N-i} \mathcal{C}_i \quad \text{and } S = H_{N-i+1} \cup H_{N-i+2} \cup \cdots \cup H_N. \tag{4}$$

When $i = 0$, we are just entering the while loop for the very first time, and so $\mathcal{C} = \mathcal{F}_N = \bigcup_{j=1}^{N} \mathcal{C}_i$, and also $S = \emptyset$. Now, suppose that the claim holds for some $i \geq 0$: we will show that it continues to hold for $i+1$. In particular, we will argue that in the i^{th} iteration of the while loop, $T^\star$ is chosen to be H_{N-i} and $b^\star$ to be 0^k in Algorithm 1. This will prove the claim, since (i) $T^\star$ gets appended to S in Algorithm 1, (ii) all the concepts in $\mathcal{C}_{N-i}$ get removed from $\mathcal{C}$ upon restricting to $T^\star, b^\star$ in Algorithm 1, since every concept in $\mathcal{C}_{N-i}$ labels the head points H_{N-i} as a one-hot vector, and (iii) no concepts in $\mathcal{C}_1, \dots, \mathcal{C}_{N-i-1}$ are removed, since all these concepts label H_{N-i} as 0^k.

So, let $T \subseteq \mathcal{X}, 1 \leq |T| \leq k$ and $b \in \{0,1\}^{|T|}$ be any candidate choice for the $\arg\min$ in Algorithm 1. Note that this also requires that at least one concept in $\mathcal{C}_1 \cup \cdots \cup \mathcal{C}_{N-i}$ labels T as b. Let us decompose T into head and tail points as follows:

$$T = \{\underbrace{x_1, \dots, x_{n_h}}_{\text{head points}}, \underbrace{y_1, \dots, y_{n_t}}_{\text{tail points}}\}, \tag{5}$$

where $0 \leq n_h, n_t \leq |T|$ and $n_h + n_t = |T| \leq k$. Similarly, let $b_h \in \{0,1\}^{n_h}$, $b_t \in \{0,1\}^{n_t}$ be the labeling in b for the head and tail points respectively. Then, it must be the case that there is some $j \in \{1, 2, \dots, N-i\}$, some concept $c_h \in \mathcal{A}_j$, and some concept $c_t \in \mathcal{B}_j$, such that c_h labels $\{x_1, \dots, x_{n_h}\}$ as b_h and c_t labels $\{y_1, \dots, y_{n_t}\}$ as b_t.

Claim 5 *Consider any arbitrary labeling b_t of $\{y_1, \dots, y_{n_t}\}$ (where $n_t \leq k$) that is consistent with some $c_t \in \mathcal{B}_1 \cup \cdots \cup \mathcal{B}_{N-i}$. Then, there are at least $(w_{N-i}+1)^k$ different concepts in $\mathcal{B}_{N-i}$ that are consistent with this labeling.*

Proof Each y_i belongs to some set T_j of tail points—let the row in T_j that contains y_i be denoted as r_i. Then consider the rows $r_1, r_2, \dots, r_{n_t}$. These are at most $n_t \leq k$ distinct rows. In particular, there are at least k rows from $\{1, 2, \dots, 2k\}$ that do not feature in $r_1, r_2, \dots, r_{n_t}$.

Now, observe how $\mathcal{B}_{N-i}$ actually has additional structure: the labels on different rows of $T_1, \dots, T_N$ do not interact with each other. Namely, if we denote by $\mathcal{B}_{N-i,r}$ the restrictions of the concepts in $\mathcal{B}_{N-i}$ to row r in $T_1, \dots, T_N$, then $\mathcal{B}_{N-i} = \mathcal{B}_{N-i,1} \otimes \mathcal{B}_{N-i,2} \otimes \cdots \otimes \mathcal{B}_{N-i,2k}$. The labeling b_t on $\{y_1, \dots, y_{n_t}\}$ possibly pins down $n_t \leq k$ sets in this product, in the worst case, to a single concept (there will always be at least one concept, since, e.g., for all the y_is that are in some common row r_i, there is at least one concept in $\mathcal{B}_{N-i,r_i}$ that is consistent with the labels on these y_is, because these labels must have—due to realizability—the necessary prefix structure enforced on concepts in $\bigcup_{j=1}^{N-i} \mathcal{B}_j$); but even so, there are at least k sets that are untouched. Each of these sets has size $w_{N-i}+1$ (we choose a prefix in the corresponding row in T_{N-i}, and expand/contract upward/downward). Thus, the total cardinality of the product (and hence the number of concepts in $\mathcal{B}_{N-i}$), even after restricting to b_t on $\{y_1,\ldots,y_{n_t}\}$, is at least $(w_{N-i}+1)^k$ as claimed. ■

Claim 6 *Consider any arbitrary labeling b_h of $\{x_1,\ldots,x_{n_h}\}$ (where $n_h \le k$, and excluding the case where $\{x_1,\ldots,x_{n_h}\} = H_{N-i}$ and $b_h = 0^k$) that is consistent with some $c_h \in \mathcal{A}_1 \cup \cdots \cup \mathcal{A}_{N-i}$. Then, there is at least one concept in $\mathcal{A}_{N-i}$ that is consistent with this labeling.*

Proof Let $S = \{x_1,\ldots,x_{n_h}\}$; note that any $x_i \in S$ that belongs to $H_{N-i+1} \cup \ldots H_N$ must be labeled as 0 in b_h. This is because S is realizable by $\mathcal{A}_1 \cup \cdots \cup \mathcal{A}_{N-i}$, and every concept in this union labels all head points above H_{N-i} as 0. So, from the point of view of proving the existence of a concept in $\mathcal{A}_{N-i}$ consistent with b_h of S, we can assume without loss of generality that $S \subseteq H_1 \cup \ldots H_{N-i}$. We will then construct a labeling on $H_1,\ldots,H_N$ consistent with b_h on S, such that this labeling corresponds to a concept in $\mathcal{A}_{N-i}$. The labeling on each of $H_{N-i+1},\ldots,H_N$ is simply the zero vector. We specify the labeling on $H_{N-i},\ldots,H_1$ according to the following cases:

Case 1: S consists of k points from some H_j, for $1 \le j \le N-i$. Note then that b_h can either be the zero vector, or a one-hot vector—this is because b_h is a labeling of S that is realizable by $\mathcal{A}_1 \cup \cdots \cup \mathcal{A}_{N-i}$, and every concept in this union labels a head set either as the zero vector, or a one-hot vector. If b_h is a one-hot vector, then simply assign each of $H_{N-i},\ldots,H_1$ to this one-hot vector. Otherwise, if b_h is the zero vector, then by the assumption in the claim, it must be the case that $j < N-i$. We then label H_{N-i} by some arbitrary one-hot vector, and each of $H_{N-i-1},\ldots,H_1$ as the zero vector. Either way, we ensure that the assignment to $H_{N-i},\ldots,H_1$ incurs at most 1 change, and hence the overall labeling to $H_1,\ldots,H_N$ constitutes a valid concept in $\mathcal{A}_{N-i}$.

Case 2: S has less than k points from each of $H_1,\ldots,H_{N-i}$. Consider

$$I = \{j \le N-i : S \text{ contains at least one point of } H_j\}.$$

Note that $|I| \le n_h \le k$. Furthermore, for any such $j \in I$, at most one element in S that is in H_j can be labeled 1 in b_h. For every $j \in I$, if among the elements in S that are in H_j, there is an x that is labeled as 1 in b_h, we assign H_j the one-hot vector that labels x as 1. Otherwise, we assign H_j to be the one-hot vector where the 1 is at an arbitrary head point in H_j not in S (such a point exists because S has strictly less than k points from any head). Now, consider the indices in $\{N-i, N-i-1,\ldots,1\}$ that remain to be assigned a label (these are precisely the indices not in I). The indices in I induce a partition of $\{N-i, N-i-1,\ldots,1\}$ into at most $k+1$ groups. Namely, if $I = \{j_1,\ldots,j_{|I|}\}$ in decreasing order, these groups are $\{N-i,\ldots,j_1\}$, $\{j_1-1,\ldots,j_2\},\ldots,\{j_{|I|}-1,\ldots,1\}$. For any every $j_l \in I$, consider the partition ending in j_l: we label the head set at every index in this partition identically as the one-hot vector we assigned to H_{j_l}. Finally, we label every head set in the last partition $\{j_{|I|}-1,\ldots,1\}$ identically also to the one-hot vector assigned to $H_{j_{|I|}}$. We can verify that this leads to assigning a one-hot vector to each of $H_{N-i}, \dots, H_1$, in a way that incurs at most k changes. Thus, our overall labeling to $H_1, \dots, H_N$ corresponds to a valid concept in $\mathcal{A}_{N-i}$. ■

Now, consider any scenario other than when $T = H_{N-i}, b = 0^k$. For the decomposition of T as in (5), we already argued that there must exist some $j \in \{1, 2, \dots, N-i\}$, $c_h \in \mathcal{A}_j$ and $c_t \in \mathcal{B}_j$ such that c_h labels $\{x_1, \dots, x_{n_h}\}$ as b_h and c_t labels $\{y_1, \dots, y_{n_t}\}$ as b_t. Then, $\{y_1, \dots, y_{n_t}\}$ together with the labeling b_t satisfy the condition of Claim 5. Thus, there are at least $(w_{N-i}+1)^k$ different concepts in $\mathcal{B}_{N-i}$ that are consistent with this labeling. Similarly, $\{x_1, \dots, x_{n_h}\}$ together with the labeling b_h satisfy the condition of Claim 6. Thus, there is at least one concept in $\mathcal{A}_{N-i}$ that is consistent with this labeling. Since $\mathcal{C}_{N-i} = \mathcal{A}_{N-i} \otimes \mathcal{B}_{N-i}$, we can conclude that there are at least $(w_{N-i}+1)^k$ concepts in $\mathcal{C}_{N-i}$ consistent with the labeling b on T.

We will now argue that the choice of $T = H_{N-i}, b = 0^k$ retains strictly less than $(w_{N-i}+1)^k$ concepts from $\mathcal{C} = \cup_{j=1}^{N-i}\mathcal{C}_i$. In particular, recall that every concept in $\mathcal{C}_{N-i}$ labels H_{N-i} as a one-hot vector. Thus, all these concepts are removed in Algorithm 1. The number of remaining concepts can then be at most $\sum_{j=1}^{N-i-1}|\mathcal{C}_j|$. As with the rectangles construction in the section above, it is also the case here that the number of concepts in $\mathcal{C}_i$ dominates the total number of concepts in $\mathcal{C}_1, \dots, \mathcal{C}_{i-1}$. In Appendix A, we show an even stronger domination result implying $\sum_{j=1}^{N-i-1}|\mathcal{C}_j| < (w_{N-i}+1)^k$:

Claim 7 ($\mathcal{C}_i$ dominates $\mathcal{C}_1, \dots, \mathcal{C}_{i-1}$) *For any $i \in \{1, \dots, N\}$,*

$$\sum_{j=1}^{i}|\mathcal{C}_j| \le w_i^{4k}.$$

In particular, for $i \in \{2, \dots, N\}$, $\sum_{j=1}^{i-1}|\mathcal{C}_j| \le w_{i-1}^{4k} < (w_i+1)^k < (w_i+1)^{2k} = |\mathcal{B}_i| < |\mathcal{C}_i|$.

We remark that our choice of $w_i = 2^{\log(8k)\cdot 2^{2i}}$ was made so as to enable Claim 7. This completes the inductive proof of (4). In particular, for $i = N-1$, we have shown that $S = H_2 \cup H_3 \cup \dots H_N$, and $\mathcal{C} = \mathcal{C}_1$. In the $(N-1)^{\text{th}}$ iteration, the algorithm will choose some k points from $\mathcal{X} \setminus S$ to add to the teaching set S (repeating a point that is already in S is strictly suboptimal, as is choosing less than k points). Thus, the final teaching set that is returned has size at least kN.

VC dimension of the concept class. In Appendix B, we show that the VC dimension of $\mathcal{F}_N$ is at most $4k+1$. Roughly, this follows from how any shattered set may not contain too many head points, or we could choose a labeling that forces more than k changes, nor may the set contain too many tail points, or we could choose an impossible labeling due to the prefix structure. This proves Part 1 of Theorem 4.

Size of the concept class and domain. Using the domination result of Claim 7, we get

$$|\mathcal{F}_N| = \sum_{j=1}^{N}|\mathcal{C}_j| \le w_N^{4k} = 2^{4k\log(8k)\cdot 2^{2N}}. \tag{6}$$

Similarly, the size of the domain $\mathcal{X}$ is at most

$$\sum_{i=1}^{N}(k+2kw_i) = k\sum_{i=1}^{N}(2w_i+1) \le 3k\sum_{i=1}^{N}w_i \le 6kw_N = 6k\cdot 2^{\log(8k)\cdot 2^{2N}}.$$

This proves Part 2 of Theorem 4.

Concluding $kN = \Omega(\log(\log(|\mathcal{F}_N|)))$. Finally, we may conclude how $|\mathcal{F}_N|$ relates to kN as,

$$\begin{aligned}\frac{1}{14} \cdot \log(\log(|\mathcal{F}_N|)) &\le \frac{1}{14} \cdot (\log(4k\log(8k)) + 2N) \qquad \text{(using (6))}\\ &\le \frac{1}{14} \cdot (\log(4kN \cdot \log(8kN)) + 2kN) \le \frac{1}{14} \cdot (12kN + 2kN) = kN.\end{aligned}$$

This completes the proof of Part 3 of Theorem 4, completing its proof.

Acknowledgments

This work was supported by the National Defense Science & Engineering Graduate (NDSEG) Fellowship Program, Tselil Schramm's NSF CAREER Grant no. 2143246, Gregory Valiant's and Moses Charikar's Simons Foundation Investigator Awards, and NSF award AF-2341890. Nikita Zhivotovskiy would like to thank Benny Sudakov and István Tomon for a number of insightful discussions regarding upper bounds for teaching sets.

References

Xi Chen, Yu Cheng, and Bo Tang. On the recursive teaching dimension of VC classes. In *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, pages 151–159, 2016.

Thomas M Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, (3):326–334, 1965.

Thorsten Doliwa, Gaojian Fan, Hans Ulrich Simon, and Sandra Zilles. Recursive teaching dimension, VC-dimension and sample compression. *The Journal of Machine Learning Research*, 15 (1):3107–3131, 2014.

Lunjia Hu, Ruihan Wu, Tianhong Li, and Liwei Wang. Quadratic upper bound for recursive teaching dimension of finite VC classes. In *Proceedings of the 30th Conference on Learning Theory (COLT)*, pages 1147–1156. PMLR, 2017.

Christian Kuhlmann. On teaching and learning intersection-closed concept classes. In *Computational Learning Theory: 4th European Conference, EuroCOLT'99 Nordkirchen, Germany, March 29–31, 1999 Proceedings 4*, pages 168–182. Springer, 1999.

Eyal Kushilevitz, Nathan Linial, Yuri Rabinovich, and Michael Saks. Witness sets for families of binary vectors. *Journal of Combinatorial Theory, Series A*, 73(2):376–380, 1996.

Shay Moran, Amir Shpilka, Avi Wigderson, and Amir Yehudayoff. Teaching and compressing for low VC-dimension. In *Proceedings of the 56th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 40–51. IEEE, 2015.

Hans U Simon and Sandra Zilles. Open problem: Recursive teaching dimension versus VC dimension. In *Conference on Learning Theory*, pages 1770–1772. PMLR, 2015.

Appendix A. Proof of Claim 7

Since $\mathcal{C}_j = \mathcal{A}_j \otimes \mathcal{B}_j$, we have that $|\mathcal{C}_j| = |\mathcal{A}_j||\mathcal{B}_j|$. Recall that $|\mathcal{B}_j| = (w_j+1)^{2k}$. To bound $|\mathcal{A}_j|$, we recall the conditions (i),(ii), (iii) and (iv) from above that a concept in $\mathcal{A}_j$ must satisfy. In particular, a concept can make labeling changes in at most k locations in $\{1, 2, \dots, j-1\}$—there are at most $(j-1)^k$ possible choices for these locations. Each choice of change locations partitions $H_1, \dots, H_j$ into at most $k+1$ buckets, where the labeling on each bucket should be the same. Furthermore, we may label each bucket with one of $k+1$ choices, corresponding to 0^k and one of k one-hot vectors (with the exception of the last bucket that includes H_j, which may only be assigned a one-hot vector). In total, we have that

$$|\mathcal{C}_j| = |\mathcal{A}_j||\mathcal{B}_j| = (w_j+1)^{2k}|\mathcal{A}_j| \le (w_j+1)^{2k}(j-1)^k(k+1)^{k+1} \le (4jkw_j)^{2k},$$

so that for any $i \in \{1, \dots, N\}$,

$$\sum_{j=1}^{i} |\mathcal{C}_j| \le (4ki)^{2k} \sum_{j=1}^{i} w_j^{2k} \le (8kiw_i)^{2k}.$$

In the last inequality, we used that $w_{j+1} \ge 2w_j$. Finally, using that for any $j \ge 1$, $8kj \le 2^{\log(8k)\cdot 2^{2j}} = w_j$, we get that

$$\sum_{j=1}^{i} |\mathcal{C}_j| \le w_i^{4k}. \tag{7}$$

In particular, for $i \in \{2, \dots, N\}$,

$$\sum_{j=1}^{i-1} |\mathcal{C}_j| \le w_{i-1}^{4k} = \left(w_{i-1}^{4}\right)^k = w_i^k < (w_i+1)^k < |\mathcal{C}_i|.$$

■

We remark that our choice of $w_i = 2^{\log(8k)\cdot 2^{2i}}$ was made so as to satisfy $(8k(i-1)w_{i-1})^{2k} \le w_i^k$ in the calculation above.

Appendix B. Bounding the VC Dimension of $\mathcal{F}_N$ in Theorem 4

Lemma 8 *The VC dimension of $\mathcal{F}_N$ is at most $4k+1$*

Proof

We divide our analysis into how large of a shattered set may exist among head points and tail points:

Claim 9 (Few Head Points) *Any set shattered by $\mathcal{F}_N$ may contain at most $2k+1$ head points.*

Proof If the shattered set contains at least two points from a single head H_i, then we know that this set cannot be shattered, as no concept in the class labels both these points simultaneously as 1. Thus, the set can only contain at most a single point from each head. Let these head points be $\{x_1, \ldots, x_m\}$—we will now construct a label pattern on these points that cannot be realized by $\mathcal{F}_N$ if m is larger than $2k+1$. Without loss of generality, suppose that the points are sorted in decreasing order of the index of the corresponding head that they appear in (i.e., $x_1 \in H_{i_1}, \ldots, x_m \in H_{i_m}$, where $i_1 > i_2 > \cdots > i_m$). Let us pair up these points into $\lfloor \frac{m}{2} \rfloor$ pairs as $(x_1, x_2), (x_3, x_4), \ldots, (x_{2\lfloor \frac{m}{2} \rfloor - 1}, x_{2\lfloor \frac{m}{2} \rfloor})$. We will determine labels for points in each pair based on the columns that these points lie in (see Fig. 2). If x_i and x_{i+1} are in the same column, then we will ask for x_i to be labeled as 1, and x_{i+1} to be labeled as 0. Otherwise, x_i and x_{i+1} are in different columns, and we will ask for both of them to be labeled as 1. Notice that the suggested label pattern necessitates a label change at the corresponding heads at each pair; since there are $\lfloor \frac{m}{2} \rfloor$ pairs, and no concept is allowed more than k label changes, $\lfloor \frac{m}{2} \rfloor \leq k \implies m \leq 2k+1$. ■

We move on to arguing that no shattered set can contain too many tail points. For this, we will require a structural property about the labels that can be realized at the same row $a \in \{1, 2, \ldots, 2k\}$ for two different tail sets T_i and T_j, where $i < j$. For $y \in \{1, 2, \ldots, w_j\}$, let $f(i,j,y)$ denote the column in $T_{i,a}$ that the point $t_{j,a,y}$ contracts down to; namely $f(i,j,y) \triangleq \left\lceil \frac{y}{w_j/w_i} \right\rceil$.

Observation 1 *For any concept c in $\mathcal{F}_N$, integers $1 \leq i < j \leq N$, row $1 \leq a \leq 2k$, and column $1 \leq b \leq w_i$, it holds that*

$$c(t_{i,a,b}) = \mathrm{AND}(\{c(t_{j,a,y}) : f(i,j,y) = b\}). \tag{8}$$

Proof First, suppose $c \in \mathcal{C}_1 \cup \cdots \cup \mathcal{C}_i$. Then, by the way that c is constructed, the label that it assigns to $t_{i,a,b}$ is copied over to all the points $\{t_{j,a,y} : f(i,j,y) = b\}$, and hence (8) holds. Now, suppose that $c \in \mathcal{C}_j \cup \cdots \cup \mathcal{C}_N$. Again, by construction, the labels that c assigns to $\{t_{j,a,y} : f(i,j,y) = b\}$ are contracted all the way down via ANDs to $t_{i,a,b}$. Finally, suppose that $c \in \mathcal{C}_l$ where $l \in \{i+1, \ldots, j-1\}$. This case essentially follows by combining the reasoning for the preceding two cases. In more detail, consider the "contraction path" of the set $Y = \{t_{j,a,y} : f(i,j,y) = b)\}$ down to the point $t_{i,a,b}$—this path intersects $T_{l,a}$ at a subset of columns $S \subset \{1, 2, \ldots, w_l\}$, such that every $d \in S$ maps to a distinct batch E_d of w_j/w_l points in Y, where these batches are disjoint, and together comprise all of Y. Furthermore, the label that c assigns to $t_{l,a,d}$ gets copied out back up to E_d. Because of this, $c(t_{l,a,d})$ is indeed the AND of the labels that c assigns to E_d. Moreover, the labels that c assigns to $T_{l,a}$ at the columns in S are also contracted down to $t_{i,a,b}$ via ANDs. Together, we get that $c(t_{i,a,b}) = \mathrm{AND}(\{c(t_{j,a,y}) : f(i,j,y) = b\})$, as desired.

■

We can now conclude that no set that is shattered by $\mathcal{F}_N$ may have more than $2k$ tail points.

Claim 10 (Few Tail Points) *Any set shattered by $\mathcal{F}_N$ may contain at most $2k$ tail points.*

Proof For the sake of contradiction, if a shattered set contains at least $2k+1$ tail points, then there must be at least two points that correspond to the same row a; these points are either within the same T_i, or across some T_i and T_j. Consider a pair of two such points, t_{i,a,b_1} and t_{j,a,b_2}, for $i \leq j$. If $i = j$, then without loss of generality, supposing $b_1 < b_2$, it is impossible for t_{i,a,b_1} to be labeled 0 while t_{j,a,b_2} is labeled 1—this is because of the prefix nature of how concepts label rows of tail points. Otherwise, suppose $i < j$. We will use Observation 1 to show how it is not possible to realize at least one pattern of labels on t_{i,a,b_1}, t_{j,a,b_2}

Case 1: $f(i, j, b_2) \le b_1$**.** We claim that no concept simultaneously labels t_{j,a,b_2} as 0 and t_{i,a,b_1} as 1. To see this, consider a concept c that labels t_{j,a,b_2} as 0, and let $f(i, j, b_2) = b_3 \le b_1$. From Observation 1, we then know that $c(t_{i,a,b_3}) = 0$. We then conclude that c cannot label t_{i,a,b_1} as 1, by the prefix nature required of c.

Case 2: $f(i, j, b_2) > b_1$**.** We claim that no concept simultaneously labels t_{j,a,b_2} as 1 and t_{i,a,b_1} as 0. To see this, consider a concept c that labels t_{j,a,b_2} as 1. Because $f(i, j, b_2) > b_1$, by Observation 1, we know that $c(t_{i,a,b_1})$ is an AND of labels that c assigns to a batch of points strictly to the left of t_{j,a,b_2}. All these labels must be 1 by the prefix nature of c, and hence we conclude that $c(t_{i,a,b_1}) = 1$. ■

Combining Claims 9 and 10, we conclude that any set shattered by $\mathcal{F}_N$ must be of size at most $4k + 1$, which bounds the VC dimension of $\mathcal{F}_N$ at $4k + 1$. ■

Appendix C. Bounding $\mathrm{TS}_{\min}$ for our constructions

Our constructions are not counterexamples to the general $\mathrm{TS}_{\min} = O(d)$ conjecture as they have small $\mathrm{TS}_{\min}$.

Concept class from Theorem 1 ($k = 1$). The $\mathrm{TS}_{\min}$ for this class is at most 2: only one concept simultaneously labels z_1 as 1, and the point immediately above as 0.

Concept class from Theorem 4 ($k \ge 2$). The $\mathrm{TS}_{\min}$ for this class is at most $2k + 1$: set the rightmost point in the k rows of T_N as 1, and set one point to 1 in each of $H_{N-k}, \ldots, H_N$ in a way that forces all the allowed changes and thus determines all other values.