Proceedings of Machine Learning Research vol 132:1–11, 2021 32nd International Conference on Algorithmic Learning Theory

Near-tight closure bounds for the Littlestone and threshold dimensions

Badih Ghazi BADIHGHAZI@GMAIL.COM
Google Research, Mountain View
Noah Golowich* NZG@MIT.EDU
MIT EECS
Ravi Kumar RAVI.K53@GMAIL.COM
Google Research, Mountain View
Pasin Manurangsi PASIN@GOOGLE.COM
Google Research, Mountain View

Editors: Vitaly Feldman, Katrina Ligett and Sivan Sabato

Abstract

We study closure properties for the Littlestone and threshold dimensions of binary hypothesis classes. Given classes $\mathcal{H}_1, \ldots, \mathcal{H}_k$ of binary functions with bounded Littlestone (respectively, threshold) dimension, we establish an upper bound on the Littlestone (respectively, threshold) dimension of the class defined by applying an arbitrary binary aggregation rule to $\mathcal{H}_1, \ldots, \mathcal{H}_k$. We also show that our upper bounds are nearly tight. Our upper bounds give an exponential (in k) improvement upon analogous bounds shown by Alon et al. (COLT 2020), thus answering an open question posed by their work.

Keywords: Littlestone dimension, threshold dimension, closure property

1. Introduction

Let X be a set and $\mathcal{H}_1, \ldots, \mathcal{H}_k$ be hypothesis classes consisting of binary classifiers $h : X \to \{0, 1\}$; for instance, each of $\mathcal{H}_1, \ldots, \mathcal{H}_k$ may be a collection of experts. Given an arbitrary *aggregation rule* $G : \{0, 1\}^k \to \{0, 1\}$ (e.g., the majority vote among the k experts), we study the maximum possible complexity of the class $G(\mathcal{H}_1, \ldots, \mathcal{H}_k)$, defined as the set of all classifiers $x \mapsto G(h_1(x), \ldots, h_k(x))$ for some choices $h_1 \in \mathcal{H}_1, \ldots, h_k \in \mathcal{H}_k$, as a function of the complexities of $\mathcal{H}_1, \ldots, \mathcal{H}_k$.

Such a closure property has long been known when complexity is measured via the VC dimension: Dudley (1978) showed that if the VC dimension of each of $\mathcal{H}_1, \ldots, \mathcal{H}_k$ is at most d, then the VC dimension of $G(\mathcal{H}_1, \ldots, \mathcal{H}_k)$, is at most $O(dk \log k)$. (This was also shown independently when G is the k-wise OR function in Blumer et al. (1989, Lemma 3.2.3).) Recently Alon et al. (2020) proved similar, but quantitatively weaker, closure properties for the *Littlestone dimension* (Littlestone, 1988) (Definition 1), which characterizes online learnability of a class (Ben-David et al., 2009), and *threshold dimension* (Shelah, 1978; Hodges, 1997) (Definition 8), which is known to be exponentially related to the Littlestone dimension and was used by Alon et al. (2019) to show that privately PAC-learnable classes are online learnable (i.e., have finite Littlestone dimension). The upper bounds of Alon et al. (2020) exhibit an exponential dependence on k, and it was asked in Alon et al. (2020) whether this dependence could be improved. Our main contribution is to resolve this question in the affirmative, proving tighter upper bounds with a nearly linear dependence on k and to show that this is nearly the best possible. In particular:

* This work was done while interning at Google Research. Supported at MIT by a Fannie & John Hertz Foundation Fellowship and an NSF Graduate Fellowship.

1. When the Littlestone dimension of each of $\mathcal{H}_1, \ldots, \mathcal{H}_k$ is at most d, we show that the Littlestone dimension of $G(\mathcal{H}_1, \ldots, \mathcal{H}_k)$ is at most $O(dk \log k)$ (Proposition 3), improving upon the bound of $\tilde{O}(2^{2k}k^2d)$ of Alon et al. (2020). Moreover, our upper bound is tight up to a constant factor (Proposition 5).

2. When the threshold dimension of each of $\mathcal{H}_1, \ldots, \mathcal{H}_k$ is at most d, we show that the threshold dimension of $G(\mathcal{H}_1, \ldots, \mathcal{H}_k)$ is at most $2^{O(dk \log k)}$ (Proposition 9), and that it can be at least $2^{\Omega(dk)}$ (Proposition 11). These bounds improve upon the upper and lower bounds of $2^{2^{O(k)}d}$ and $2^{\Omega(d)}$, respectively, shown in Alon et al. (2020).

Related work. Several papers have developed a more refined description of the closure properties of the VC dimension. Eisenstat and Angluin (2007) showed that the $O(dk \log k)$ upper bound on the VC dimension of $G(\mathcal{H}_1, \ldots, \mathcal{H}_k)$ by Dudley (1978); Blumer et al. (1989) is tight up to a constant factor even when G is the k-wise OR function for any $d \geq 5$. This condition was improved to $d \geq 2$, which is the best possible for $G = \text{OR}$, in Eisenstat (2009). The question of closure properties has also been investigated for the specific case that $\mathcal{H}_1 = \cdots = \mathcal{H}_k$ is the class of half-spaces in $\mathbb{R}^d$. It is known that for $d \geq 4$, the VC dimension of the class of k-fold unions of half-spaces in $\mathbb{R}^d$ is $\Theta(kd \log k)$ (Csikós et al., 2018), while for $d \leq 3$, the VC dimension of the class of k-fold unions of half-spaces in $\mathbb{R}^d$ is $\Theta(k)$ (Dobkin and Gunopulos, 1995).

Closure properties similar to the ones considered in this paper are also known to hold for the Rademacher complexity (Bartlett and Mendelson, 2003), and its sequential variant (Rakhlin et al., 2015). Apart from Alon et al. (2020), we are not aware of any prior work studying closure properties for the Littlestone or threshold dimensions.

The Littlestone dimension of a function class $\mathcal{F}$ is closely related to the sample complexity of online learning for the class $\mathcal{F}$. In particular the Littlestone dimension of $\mathcal{F}$ exactly characterizes the mistake bound for online learning the class $\mathcal{F}$ in the realizable setting (Littlestone, 1988; Shalev-Shwartz, 2012). In the agnostic setting, if the Littlestone dimension of $\mathcal{F}$ is denoted by d, then the optimal regret $\text{Reg}(T)$ for an online learning algorithm applied to $\mathcal{F}$ scales as $\Omega(\sqrt{dT}) \leq \text{Reg}(T) \leq O(\sqrt{dT \log T})$ (Ben-David et al., 2009; Shalev-Shwartz, 2012). A version of this upper bound which requires few calls to an (appropriately chosen) oracle for $\mathcal{F}$ has also been developed (Rakhlin et al., 2012).

2. Closure bounds for the Littlestone dimension

2.1. Preliminaries

In this section we mostly follow the notation of Rakhlin and Sridharan (2014); Rakhlin et al. (2015). For a positive integer n, we use $[n]$ to denote $\{1, \ldots, n\}$. For a positive integer t and a sequence $\epsilon_1, \epsilon_2, \ldots, \epsilon_t, \ldots$, we let $\epsilon_{1:t}$ denote the tuple $(\epsilon_1, \ldots, \epsilon_t)$. As a convention let $\epsilon_{1:0}$ denote the empty sequence. Let $\{0,1\}^X$ be the set of all classifiers $f : X \to \{0,1\}$.

For a set X, an X-*valued tree* $\mathbf{x}$ of depth n is a collection of functions $\mathbf{x}_t : \{0,1\}^{t-1} \to X$ for $1 \leq t \leq n$. Consider a binary hypothesis class $\mathcal{F} \subseteq \{0,1\}^X$. The class $\mathcal{F}$ is said to *shatter* a tree $\mathbf{x}$ of depth n if

$$\forall (\epsilon_1, \ldots, \epsilon_n) \in \{0,1\}^n, \ \exists f \in \mathcal{F} \ \text{s.t.} \ \ f(\mathbf{x}_t(\epsilon_{1:t-1})) = \epsilon_t \ \forall t \in [n].$$

Definition 1 (Littlestone dimension) *The* Littlestone dimension *of a class $\mathcal{F}$, denoted* $\mathrm{Ldim}(\mathcal{F})$*, is the depth of the largest X-valued binary tree $\mathbf{x}$ that is shattered by $\mathcal{F}$.*

A set $\mathcal{V}$ of $\{0,1\}$-valued trees of depth n is called a *0-cover* for $\mathcal{F}$ on a given X-valued tree $\mathbf{x}$ of depth n if:

$$\forall f \in \mathcal{F},\ \forall(\epsilon_1, \dots, \epsilon_{n-1}) \in \{0,1\}^{n-1},\ \exists \mathbf{v} \in \mathcal{V} \ \text{ s.t. } f(\mathbf{x}_t(\epsilon_{1:t-1})) = \mathbf{v}_t(\epsilon_{1:t-1})\ \forall t \in [n].$$

The *0-covering number* of $\mathcal{F}$ on the tree $\mathbf{x}$ is defined as:

$$\mathcal{N}_0(\mathcal{F}, \mathbf{x}) := \min\left\{|\mathcal{V}| : \mathcal{V} \text{ is a 0-cover for } \mathcal{F} \text{ on } \mathbf{x}\right\}.$$

Lemma 2 (Rakhlin et al. (2014), Theorem 7; "Sauer–Shelah lemma for 0-covering number in trees") *For any X-valued tree $\mathbf{x}$ of depth n, we have*

$$\mathcal{N}_0(\mathcal{F}, \mathbf{x}) \leq \sum_{i=0}^{d} \binom{n}{i} \leq \left(\frac{en}{d}\right)^d,$$

when $\mathrm{Ldim}(\mathcal{F}) = d \leq n$*.*

Finally, recall that the *VC dimension* of a class $\mathcal{F}$ is the length of the longest sequence $x_1, \dots, x_n \in X$ so that for any $b = (b_1, \dots, b_n) \in \{0,1\}^n$, there is an $f \in \mathcal{F}$ so that $f(x_i) = b_i$ for all $i \in [n]$.

2.2. Improved bounds

Let X be a set. For a function $G : \{0,1\}^k \to \{0,1\}$ and classifiers $h_1, \dots, h_k : X \to \{0,1\}$, let $G(h_1, \dots, h_k) : X \to \{0,1\}$ be the mapping $x \mapsto G(h_1(x), \dots, h_k(x))$. Then for binary hypothesis classes $\mathcal{H}_1, \dots, \mathcal{H}_k \subseteq \{0,1\}^X$, we define

$$G(\mathcal{H}_1, \dots, \mathcal{H}_k) := \{G(h_1, \dots, h_k) : h_i \in \mathcal{H}_i\}.$$

In Proposition 3, we prove an upper bound for $\mathrm{Ldim}(G(\mathcal{H}_1, \dots, \mathcal{H}_k))$ in terms of $\max_{1 \leq j \leq k} \mathrm{Ldim}(\mathcal{H}_j)$ that grows quasi-linearly with k. The proof follows as a consequence of the bound on the 0-covering number given by the Sauer–Shelah lemma for trees (Lemma 2), in a similar manner to Dudley's (Dudley, 1978, Proposition 7.12) proof of a closure property for VC classes using the classic Sauer–Shelah lemma (Sauer, 1972; Vapnik and Chervonenkis, 1968). In Alon et al. (2020, Section 2.1.1), the authors state that they are not aware of a proof of Proposition 3 using the related definition of *thicket shatter function*. We discuss the relation between 0-covering number and thicket shatter function further in Appendix A.

Proposition 3 *Let $G : \{0,1\}^k \to \{0,1\}$ be a Boolean function, let $\mathcal{H}_1, \dots, \mathcal{H}_k \subseteq \{0,1\}^X$ be binary hypothesis classes, and let $d \in \mathbb{N}$ be such that* $\mathrm{Ldim}(\mathcal{H}_i) \leq d$ *for all $i \in [k]$. Then*

$$\mathrm{Ldim}(G(\mathcal{H}_1, \dots, \mathcal{H}_k)) \leq O(kd \log k).$$

Before proving Proposition 3 we state the following lemma; its (straightforward) proof is deferred to Appendix A.

Lemma 4 *Suppose that $\mathcal{F} \subset \{0,1\}^X$ shatters a tree $\mathbf{x}$ of depth n. Then any 0-cover $\mathcal{V}$ for $\mathcal{F}$ on the tree $\mathbf{x}$ has size at least 2^n.*

Proof [of Proposition 3] It is without loss of generality to assume $d \geq 3$. Let us write $N = \mathrm{Ldim}(G(\mathcal{H}_1, \ldots, \mathcal{H}_k))$. We may assume $N \geq d$; if not, then the conclusion of the proposition is already established. Let $\mathbf{x}$ be an X-valued complete binary tree of depth N that is shattered by $\mathcal{F}$. By Lemma 2, for each $i \in [k]$, since $\mathrm{Ldim}(\mathcal{H}_i) \leq d$ for each i, we have

$$\mathcal{N}_0(\mathcal{H}_i, \mathbf{x}) \leq \sum_{i=0}^{d} \binom{N}{i} \leq \left(\frac{eN}{d}\right)^d.$$

Now, for each $i \in [k]$, let $\mathcal{V}_i$ be a 0-cover for $\mathcal{H}_i$ on $\mathbf{x}$ of size $|\mathcal{V}_i| \leq (eN/d)^d$.

We next construct a 0-cover for $G(\mathcal{H}_1, \ldots, \mathcal{H}_k)$ of size at most $\prod_{i=1}^k |\mathcal{V}_i|$ as follows: for each tuple $\tau = (\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(k)}) \in \mathcal{V}_1 \times \cdots \times \mathcal{V}_k$, construct a tree $\mathbf{w}^{(\tau)}$ defined by

$$\mathbf{w}_t^{(\tau)}(\epsilon_{1:t-1}) := G(\mathbf{v}_t^{(1)}(\epsilon_{1:t-1}), \ldots, \mathbf{v}_t^{(k)}(\epsilon_{1:t-1})) \qquad \forall (\epsilon_1, \ldots, \epsilon_{N-1}) \in \{0,1\}^N, t \in [N].$$

To see that the collection $\mathcal{W} := \{\mathbf{w}^{(\tau)}\}_{\tau \in \mathcal{V}_1 \times \cdots \times \mathcal{V}_k}$ indeed forms a 0-cover, consider any $g \in G(\mathcal{H}_1, \ldots, \mathcal{H}_k)$. Then there are $h_1 \in \mathcal{H}_1, \ldots, h_k \in \mathcal{H}_k$ so that $g(x) = G(h_1(x), \ldots, h_k(x))$. Also fix any sequence $(\epsilon_1, \ldots, \epsilon_{N-1}) \in \{-1, 1\}^{N-1}$. Since $\mathcal{V}_i$ is a 0-cover for $\mathcal{H}_i$ on $\mathbf{x}$, for each $i \in [k]$, there is some $\mathbf{v}^{(i)} \in \mathcal{V}_i$ so that for all $t \in [N]$, $h_i(\mathbf{x}_t(\epsilon_{1:t-1})) = \mathbf{v}_t^{(i)}(\epsilon_{1:t-1})$. Thus, for $\tau = (\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(k)})$, for each $t \in [N]$, we have

$$\mathbf{w}_t^{(\tau)}(\epsilon_{1:t-1}) = G(\mathbf{v}_t^{(1)}(\epsilon_{1:t-1}), \ldots, \mathbf{v}_t^{(k)}(\epsilon_{1:t-1})) = G(h_1(\mathbf{x}_t(\epsilon_{1:t-1})), \ldots, h_k(\mathbf{x}_t(\epsilon_{1:t-1}))) = g(\mathbf{x}_t(\epsilon_{1:t-1})).$$

Hence $\mathcal{W}$ is indeed a 0-cover of $G(\mathcal{H}_1, \ldots, \mathcal{H}_k)$ on $\mathbf{x}$.

Since $\mathbf{x}$ is shattered by $G(\mathcal{H}_1, \ldots, \mathcal{H}_k)$, by Lemma 4, we have that $|\mathcal{W}| \geq 2^N$. Summarizing,

$$2^N \leq |\mathcal{W}| \leq (eN/d)^{kd},$$

so $N \leq kd \log(eN/d)$, i.e., $N \leq O(kd \log k)$. ∎

We remark that an alternative proof of Proposition 3 follows by applying (Ben-David et al., 2009, Lemma 12) to cover each $\mathcal{H}_i$ by a finite class of at most d experts.

Yet another way to upper bound $\mathrm{Ldim}(G(\mathcal{H}_1, \ldots, \mathcal{H}_k))$ in the context of Proposition 3 is to use Rakhlin et al. (2015, Proposition 9 and Corollary 6). In particular, Rakhlin et al. (2015, Corollary 6) gives a similar closure property for the sequential Rademacher complexities, and Rakhlin et al. (2015, Proposition 9) implies that sequential Rademacher complexities are closely related to the Littlestone dimension. However, this technique would give an upper bound of $O(\log^4(kd) \cdot k^2 d)$, which is worse than that of Proposition 3.

Proposition 5 shows that the upper bound of Proposition 3 is tight up to a constant factor.

Proposition 5 *There is a positive constant C so that for any positive integers $k, d \geq C$, there is a domain X and a class $\mathcal{H} \subset \{0,1\}^X$ so that:*

1. $\mathrm{Ldim}(\mathcal{H}) \leq d$.

2. *Defining $G : \{0,1\}^{3k} \to \{0,1\}$ to be the $3k$-wise OR function,* $\mathrm{Ldim}(G(\mathcal{H}, \ldots, \mathcal{H})) \geq \Omega(kd \log k)$.

Before proving Proposition 5, we need a couple of lemmas. Lemma 6 was used in Eisenstat and Angluin (2007) to show an analogous result to Proposition 5 for the VC dimension.

Lemma 6 (Eisenstat and Angluin (2007)) *There is a constant k_0 so that for all $k \geq k_0$, for $X := \{1, 2, \ldots, \lfloor k \log k \rfloor\}$, there is a class $\mathcal{F} \subset \{0,1\}^X$ so that:*

1. *For each $f_1, f_2 \in \mathcal{F}$, there are at most 4 distinct points $x \in X$ so that $f_1(x) = f_2(x) = 1$.*
2. *Let $G : \{0,1\}^{3k} \to \{0,1\}$ be the $3k$-wise OR function. Then:*

$$\{(h(1), \ldots, h(\lfloor k \log k \rfloor)) : h \in G(\mathcal{F}, \ldots, \mathcal{F})\} = \{0,1\}^{\lfloor k \log k \rfloor}.$$

(It is said that $G(\mathcal{F}, \ldots, \mathcal{F})$ shatters the set X.)

To state the next lemma, we define the notion of a subtree of a binary tree $\mathbf{x}$. A *subtree of depth 1* is simply a node of $\mathbf{x}$. For any integer $h > 1$, a *subtree of depth h* is obtained by specifying a node v of $\mathbf{x}$, together with a subtree of depth $h-1$ of each of the trees rooted at the left and right children of v.

Lemma 7 (Shelah (1978); Hodges (1997)) *Suppose the nodes of a binary tree $\mathbf{x}$ of depth at least $kd - (k-1)$ are colored with k colors. Then there is a monochromatic subtree of $\mathbf{x}$ of height d.*

The specific statement of Lemma 7 may be found in (Jung et al., 2020, Lemma 16). Now we are ready to prove Proposition 5.

Proof [of Proposition 5] Let $C = \max\{k_0, 5\}$, where k_0 is the constant of Lemma 6. Fix any $k \geq C$, and as in Lemma 6 let G be the $3k$-wise OR function. By Lemma 6, there is a domain X and a class $\mathcal{F} \subset \{0,1\}^X$ so that items 1 and 2 hold. Item 2 tells us that $\mathcal{F}$ shatters the set $\{1, 2, \ldots, \lfloor k \log k \rfloor\}$, and hence the VC dimension of $G(\mathcal{F}, \ldots, \mathcal{F})$ is at least $\lfloor k \log k \rfloor$. Since VC dimension is a lower bound on the Littlestone dimension, it follows that $\mathrm{Ldim}(G(\mathcal{F}, \ldots, \mathcal{F})) \geq \lfloor k \log k \rfloor$.

On the other hand, we claim that $\mathrm{Ldim}(\mathcal{F}) \leq 5$. If there were a binary tree $\mathbf{x}$ of depth 6 shattered by $\mathcal{F}$, note that there must be at least 2 functions $f_1, f_2 \in \mathcal{F}$ so that

$$f_b(\mathbf{x}_1) = f_b(\mathbf{x}_2(1)) = f_b(\mathbf{x}_3(1,1)) = f_b(\mathbf{x}_4(1,1,1)) = f_b(\mathbf{x}_5(1,1,1,1)) = 1, \quad b \in \{1,2\}.$$

Since $\mathcal{F}$ shatters $\mathbf{x}$ it is straightforward that the 5 points $\mathbf{x}_1, \mathbf{x}_2(1), \ldots, \mathbf{x}_5(1,1,1,1)$ are distinct, which contradicts item 1.

Next fix any positive integer ℓ. Similar to the argument of Eisenstat and Angluin (2007, Lemma 4), we now define the set $Y := \{1, 2, \ldots, \ell\} \times X$, and the class $\mathcal{H} \subset \{0,1\}^Y$ by

$$\mathcal{H} = \left\{ h \in \{0,1\}^Y : \exists f_1, \ldots, f_\ell \in \mathcal{F} \text{ s.t. } \forall x \in X,\ \forall j \in [\ell],\ h((j,x)) = f_j(x) \right\}.$$

Notice that $\mathcal{H}$ is the product of ℓ distinct copies of $\mathcal{F}$. We claim that[1] $\mathrm{Ldim}(\mathcal{H}) \leq 5\ell$. To see this, suppose for the purpose of contradiction that there is some binary tree $\mathbf{x}$ of depth $5\ell + 1 = 6\ell - (\ell - 1)$ that is shattered by $\mathcal{H}$. Next we define the following ℓ-coloring of the nodes of $\mathbf{x}$: each node of $\mathbf{x}$

1. There is also an alternative proof that $\mathrm{Ldim}(\mathcal{H}) \leq \ell \cdot \mathrm{Ldim}(\mathcal{F})$ using the online mistake bound characterization of Ldim (Littlestone, 1988). Specifically, given an online learner for $\mathcal{F}$, we may devise an online learner for $\mathcal{H}$ as follows. We keep ℓ copies of the learner for $\mathcal{F}$; when we receive a sample $(j,x) \in Y$, we output the prediction of the j-th learner on sample x. It is simple to observe that the mistake bound for the new learner is at most ℓ times that of the original learner, implying that $\mathrm{Ldim}(\mathcal{H}) \leq \ell \cdot \mathrm{Ldim}(\mathcal{F})$.

is labeled by some $(j, x) \in [\ell] \times \mathcal{X}$; color that node with the color j. By Lemma 7, there is a monochromatic subtree, which we call $\mathbf{x}'$, of $\mathbf{x}$ of depth 6. Since $\mathbf{x}$ is shattered by $\mathcal{H}$, $\mathbf{x}'$ is shattered by $\mathcal{F}$. But this contradicts the fact that $\mathrm{Ldim}(\mathcal{F}) \leq 5$.

We next claim that $\mathrm{Ldim}(G(\mathcal{H}, \ldots, \mathcal{H})) \geq \ell\lfloor k \log k \rfloor$. This follows from the fact that the VC dimension of $G(\mathcal{H}, \ldots, \mathcal{H})$ is at least $\ell\lfloor k \log k \rfloor$, which in turn holds since $G(\mathcal{F}, \ldots, \mathcal{F})$ shatters the set X, meaning that $G(\mathcal{H}, \ldots, \mathcal{H})$ shatters Y, which is of size $\ell\lfloor k \log k \rfloor$. ■

3. Closure bounds for the threshold dimension

For positive integers i, j, write $\mathbb{1}[i \geq j]$ to be 1 if $i \geq j$ and 0 otherwise. Similarly write $\mathbb{1}[i = j]$ to be 1 if $i = j$ and 0 otherwise. The threshold dimension of a hypothesis class is defined as follows:

Definition 8 (Threshold dimension) *For a binary hypothesis class $\mathcal{F} \subset \{0,1\}^X$, the* threshold dimension *of $\mathcal{F}$, denoted $\mathrm{Tdim}(\mathcal{F})$, is the largest positive integer d so that there are $x_1, \ldots, x_d \in X$ and $f_1, \ldots, f_d \in \mathcal{F}$ such that $f_i(x_j) = \mathbb{1}[i \geq j]$ for all $i, j \in [d]$. In such a case, we say that $x_1, \ldots, x_d$ are* threshold shattered by $\mathcal{F}$ via $f_1, \ldots, f_d$.

Proposition 9 establishes an upper bound for $\mathrm{Tdim}(G(\mathcal{H}_1, \ldots, \mathcal{H}_k))$ in terms of $\max_{1 \leq j \leq k} \mathrm{Tdim}(\mathcal{H}_j)$. It improves upon an upper bound of Alon et al. (2020) that grows *doubly* exponentially in k. The proof technique is similar to that of Alon et al. (2020), except that in the application of Ramsey's theorem a coloring with only $2k$, as opposed to 2^{2k}, colors is used.

Proposition 9 *Let $G : \{0,1\}^k \to \{0,1\}$ be a Boolean function. Let $\mathcal{H}_1, \ldots, \mathcal{H}_k \subseteq \{0,1\}^X$ be binary hypothesis classes, and let $d \in \mathbb{N}$ be such that $\mathrm{Tdim}(\mathcal{H}_i) \leq d$ for all $i \in [k]$. Then*

$$\mathrm{Tdim}(G(\mathcal{H}_1, \ldots, \mathcal{H}_k)) \leq 2^{O(kd \log k)}.$$

Proof Let N be the smallest positive integer such that, for every coloring of the edges of the complete graph K_N in $c = 2k$ colors, there exists a monochromatic clique of size $r = 2d + 1$. It is well known in Ramsey theory (e.g., Greenwood and Gleason (1955)) that $N \leq c^{rc} = 2^{O(kd \log k)}$. We will show that $\mathrm{Tdim}(G(\mathcal{H}_1, \ldots, \mathcal{H}_k)) < N$.

Suppose contrapositively that $\mathrm{Tdim}(G(\mathcal{H}_1, \ldots, \mathcal{H}_k)) \geq N$. By definition of threshold dimension, there exists $x_1, \ldots, x_N \in X$ and $h_{i\ell} \in \mathcal{H}_\ell$ for $i \in [N], \ell \in [k]$ such that

$$G(h_{i1}(x_j), \ldots, h_{ik}(x_j)) = \mathbb{1}[i \geq j] \qquad \forall i, j \in [N].$$

Consider the complete graph K_N and a coloring with $2k$ colors defined as follows: for each $1 \leq p < q \leq N$, let $\ell \in [k]$ be the smallest index such that $h_{p\ell}(x_q) \neq h_{q\ell}(x_p)$; such ℓ must exist because $G(h_{p1}(x_q), \ldots, h_{pk}(x_q)) = 0 \neq 1 = G(h_{q1}(x_p), \ldots, h_{qk}(x_p))$. Then, let the color of the edge $\{p, q\}$ be $(\ell, h_{p\ell}(x_q))$.

By our choice of N, the graph must contain a monochromatic clique with vertices $i_1 < \cdots < i_{2d+1}$; let the color of its edges be (t, y) where $t \in [k]$ and $y \in \{0,1\}$. From how each edge's color is defined, the following holds for all distinct $u, v \in [2d+1]$:

$$h_{i_u t}(x_{i_v}) = \begin{cases} y & \text{if } u < v \\ 1 - y & \text{if } u > v. \end{cases}$$

Thus, if $y = 0$, then $x_{i_2}, x_{i_4}, \ldots, x_{i_{2d}}$ is threshold shattered by $\mathcal{H}_t$ (via the hypotheses $h_{i_3 t}, h_{i_5 t}, \ldots, h_{i_{2d+1} t}$). Otherwise, if $y = 1$, then $x_{i_{2d}}, x_{i_{2d-2}}, \ldots, x_{i_2}$ is threshold shattered by $\mathcal{H}_t$ (via $h_{i_{2d-1} t}, h_{i_{2d-1} t}, \ldots, h_{i_1 t}$). In both cases, we have $\mathrm{Tdim}(\mathcal{H}_t) \geq d$, which concludes our proof. ■

Next we establish a lower bound showing that Proposition 9 is nearly tight. We need the following lemma from Alon et al. (2020), which shows exponential dependence in d (but not necessarily in k) is necessary.

Lemma 10 (Alon et al. (2020), Theorem 2.2) *For every $d \geq 6$ there is a class $\mathcal{C}$ consisting of classifiers $f : \{0, 1, \ldots, 2^{\lfloor d/5 \rfloor} - 1\} \to \{0, 1\}$ so that $\mathrm{Tdim}(\mathcal{C}) \leq d$ yet*

$$\mathrm{Tdim}(\{f_1 \vee f_2 : f_1, f_2 \in \mathcal{C}\}) = 2^{\lfloor d/5 \rfloor}.$$

In fact, the class $\{f_1 \vee f_2 : f_1, f_2 \in \mathcal{C}\}$ realizes the thresholds $x \mapsto \mathbb{1}[b \geq x]$, for each $0 \leq b \leq 2^{\lfloor d/5 \rfloor} - 1$.

Proposition 11 shows that Proposition 9 is tight, up to possibly the factor of $\log k$ in the exponent:

Proposition 11 *For any positive integers $d \geq 6$ and k, there is a domain $\mathcal{X}$ and classes $\mathcal{H}_1, \ldots, \mathcal{H}_{3k} : \mathcal{X} \to \{0, 1\}$ and a function $G : \{0, 1\}^{3k} \to \{0, 1\}$ so that:*

1. *$\max\{\mathrm{Tdim}(\mathcal{H}_1), \ldots, \mathrm{Tdim}(\mathcal{H}_{3k})\} \leq d$.*
2. *$\mathrm{Tdim}(G(\mathcal{H}_1, \ldots, \mathcal{H}_{3k})) = 2^{k\lfloor d/5 \rfloor}$.*

Proof Fix $d \geq 6, k$ and write $D := 2^{\lfloor d/5 \rfloor}$. Consider the domain $\mathcal{X} := \{0, 1, \ldots, D^k - 1\}$. For $x \in \mathcal{X}$, we will write its base-D representation as $x = x_1 x_2 \cdots x_k$, so that $x_1, \ldots, x_k \in \{0, 1, \ldots D - 1\}$. Let $\mathcal{C}$, consisting of functions $f : \{0, 1, \ldots, D - 1\} \to \{0, 1\}$, be the class from Lemma 10. We now define k classes $\mathcal{H}_1, \ldots, \mathcal{H}_k$, as follows: for $1 \leq j \leq k$, let $\mathcal{H}_j := \{h_{j,f} : f \in \mathcal{C}\}$, where for $f \in \mathcal{C}$,

$$h_{j,f}(x_1 \cdots x_k) = f(x_j). \tag{1}$$

Also define classes $\mathcal{G}_1, \ldots, \mathcal{G}_k$ as follows: for $j \in [k]$, let $\mathcal{G}_j := \{g_{j,0}, \ldots, g_{j,D-1}\}$, where for $b \in \{0, 1, \ldots, D - 1\}$,

$$g_{j,b}(x_1 \cdots x_k) = \mathbb{1}[x_j = b].$$

Now define $\tilde{G} : \{0, 1\}^{2k} \to \{0, 1\}$ as follows: $\tilde{G}(y_1, \ldots, y_k, z_1, \ldots, z_k) = 1$ if and only if either (a) $y_1 = \cdots = y_k = 1$ or (b) in the case that there is a smallest index j with $z_j = 0$, it holds that $y_1 = \cdots = y_j = 1$. Finally define $G : \{0, 1\}^{3k} \to \{0, 1\}$, as follows:

$$G(y_1, \ldots, y_k, y'_1, \ldots, y'_k, z_1, \ldots, z_k) = \tilde{G}(y_1 \vee y'_1, \ldots, y_k \vee y'_k, z_1, \ldots, z_k).$$

On one hand, it is straightforward to see that for each $j \in [k]$, $\mathrm{Tdim}(\mathcal{H}_j) \leq d$, since $\mathrm{Tdim}(\mathcal{C}) \leq d$ from Lemma 10. It is also straightforward that $\mathrm{Tdim}(\mathcal{G}_j) \leq 1$ for each j: if the threshold dimension were at least 2, then there would be $x^{(1)}, x^{(2)} \in \mathcal{X}$ that are threshold shattered via g_{j,b_1}, g_{j,b_2} for some $b_1, b_2 \in \{0, 1, \ldots, D - 1\}$. However, $g_{j,b_1}(x^{(1)}) = g_{j,b_2}(x^{(1)}) = 1$ implies that $b_1 = x_j^{(1)} = b_2$ which contradicts $g_{j,b_1}(x^{(2)}) = 0 \neq 1 = g_{j,b_2}(x^{(2)})$.

On the other hand, we claim that $\mathrm{Tdim}(G(\mathcal{H}_1, \ldots, \mathcal{H}_k, \mathcal{H}_1, \ldots, \mathcal{H}_k, \mathcal{G}_1, \ldots, \mathcal{G}_k)) \geq D^k$. Now let $\mathcal{T}_j := \{\tau_{j,0}, \ldots, \tau_{j,D-1}\}$, where $\tau_{j,b}(x) = \mathbb{1}[b \geq x_j]$. Notice that

$$G(\mathcal{H}_1, \ldots, \mathcal{H}_k, \mathcal{H}_1, \ldots, \mathcal{H}_k, \mathcal{G}_1, \ldots, \mathcal{G}_k) = \tilde{G}(\mathcal{H}_1 \vee \mathcal{H}_1, \ldots, \mathcal{H}_k \vee \mathcal{H}_k, \mathcal{G}_1, \ldots, \mathcal{G}_k) \supseteq \tilde{G}(\mathcal{T}_1, \ldots, \mathcal{T}_j, \mathcal{G}_1, \ldots, \mathcal{G}_k),$$

where the inclusion above follows from Lemma 10: indeed, the lemma implies that for each $b \in \{0, 1, \ldots, D-1\}$, there are some $f_1, f_2 \in \mathcal{C}$ so that $(f_1 \vee f_2)(\cdot) = \mathbb{1}[b \geq \cdot]$. In particular, it follows that $(h_{j,f_1} \vee h_{j,f_2})(x) = \mathbb{1}[b \geq x_j] = \tau_{j,b}(x)$.

It therefore suffices to show that $\tilde{G}(\mathcal{T}_1, \ldots, \mathcal{T}_k, \mathcal{G}_1, \ldots, \mathcal{G}_k)$ can realize all threshold functions on $\mathcal{X}$. Indeed, for any $a = a_1 \cdots a_k \in \mathcal{X}$, for $a_1, \ldots, a_k \in \{0, 1, \ldots, D-1\}$, we have, for each $x \in \mathcal{X}$,

$$\mathbb{1}[a \geq x] = \tilde{G}(\tau_{1,a_1}(x), \ldots, \tau_{k,a_k}(x), g_{1,a_1}(x), \ldots, g_{k,a_k}(x)).$$

To see that the above holds, simply note that $a \geq x$ if and only if either (a) $a_j \geq x_j$ for each $1 \leq j \leq k$, or (b) for the smallest j such that $a_j \neq x_j$, we have $a_{j'} \geq x_{j'}$ for $1 \leq j' \leq j$. ■

4. Future work

There is a gap of $\log k$ between the exponent in the upper bound of Proposition 9 and in the exponent of the lower bound of Proposition 11. In the direction of closing this gap, we make the following observation: for positive integers r, c, let $R_c(r)$ denote the minimum positive integer N so that for every coloring of the edges of the complete graph K_N with c colors, there is a monochromatic clique of size r. Note that the proof of Proposition 9 shows the following: suppose that, for some odd integer $r = 2d+1$ there is some $h_r < \infty$ so that $R_c(r) \leq h_r^c$ for all $c \in \mathbb{N}$. Then for any $k \in \mathbb{N}$ and aggregation function $G : \{0,1\}^k \to \{0,1\}$, for any binary hypothesis classes $\mathcal{H}_1, \ldots, \mathcal{H}_k$ with threshold dimensions $\mathrm{Tdim}(\mathcal{H}_i) \leq d$ for $i \in [k]$, we have that $\mathrm{Tdim}(G(\mathcal{H}_1, \ldots, \mathcal{H}_k)) \leq h_{2d+1}^{2k} = 2^{O(k)}$, viewing d as a constant.

Contrapositively, we have the following: suppose that for some integer d, for infinitely many values of $k \in \mathbb{N}$, there are binary hypothesis classes $\mathcal{H}_1, \ldots, \mathcal{H}_k$ and $G : \{0,1\}^k \to \{0,1\}$ satisfying $\mathrm{Tdim}(\mathcal{H}_i) \leq d$ for $i \in [k]$ and $\mathrm{Tdim}(G(\mathcal{H}_1, \ldots, \mathcal{H}_k)) = 2^{k\alpha(k)}$ for some function $\alpha(k)$ which goes to ∞ as $k \to \infty$. Then for infinitely many values of k, we have that $R_{2k}(2d+1) > 2^{2k\alpha(2k)}$, i.e., $\limsup_{c \to \infty} R_c(2d+1)^{1/c} = \infty$. This would resolve a long-standing open question in Ramsey theory (see, e.g., Xu and Radziszowski (2016, Conjecture 3.1.3), as well as Abbott and Hanson (1972); Chung (1973); Chung and Grinstead (1983)). This observation suggests that proving a lower bound stronger than Proposition 11 is likely to be quite difficult (if possible). On the other hand, we know of no Ramsey-theoretic implications of an *upper bound* that matches the lower bound of Proposition 11.

Alon et al. (2020) additionally established a similar closure property to the ones considered in this note for the sample complexity of private PAC learning. Their upper bound has a polynomial dependence on k; it would be interesting to determine if a stronger upper bound (say, nearly linear in k) could be established.

Acknowledgments

We are grateful to Noga Alon and Shay Moran who pointed us to several helpful references, including Eisenstat and Angluin (2007). We also thank anonymous reviewers for insightful comments.

References

H. Abbott and D. Hanson. A problem of Schur and its generalizations. *Acta Arithmetica*, 20(2): 175–187, 1972.

Noga Alon, Roi Livni, Maryanthe Malliaris, and Shay Moran. Private PAC learning implies finite Littlestone dimension. In *STOC*, page 852860, 2019. ISBN 9781450367059.

Noga Alon, Amos Beimel, Shay Moran, and Uri Stemmer. Closure properties for private classification and online prediction. In *COLT*, 2020.

Peter L. Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.*, 3(null):463482, March 2003. ISSN 1532-4435.

Shai Ben-David, Dávid Pál, and Shai Shalev-Shwartz. Agnostic online learning. In *COLT*, 2009.

Siddharth Bhaskar. Thicket density. *arXiv:1702.03956*, 2017.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *J. ACM*, 36(4):929–965, 1989.

F. R. K Chung. On the Ramsey numbers R(3,3,...3;2). *Discrete Mathematics*, (5), 1973.

F. R. K. Chung and C. M. Grinstead. A survey of bounds for classical Ramsey numbers. *Journal of Graph Theory*, 7(1):25–37, 1983.

Mónika Csikós, Andrey Kupavskii, and Nabil H. Mustafa. Optimal bounds on the VC-dimension. *CoRR*, abs/1807.07924, 2018.

David P. Dobkin and Dimitrios Gunopulos. Concept learning with geometric hypotheses. In *COLT*, pages 329–336, 1995.

Richard M. Dudley. Central limit theorems for empirical measures. *Ann. Probab.*, 6(6):899–929, 12 1978. doi: 10.1214/aop/1176995384.

David Eisenstat. k-fold unions of low-dimensional concept classes. *Inf. Process. Lett.*, 109(23-24): 1232–1234, 2009.

David Eisenstat and Dana Angluin. The VC dimension of k-fold union. *Inf. Process. Lett.*, 101: 181–184, 03 2007. doi: 10.1016/j.ipl.2006.10.004.

Robert E. Greenwood and Andrew M. Gleason. Combinatorial relations and chromatic graphs. *Canadian Journal of Mathematics*, 7:17, 1955. doi: 10.4153/CJM-1955-001-4.

Wilfrid Hodges. *A Shorter Model Theory*. Cambridge University Press, 1997.

Young Jung, Baekjin Kim, and Ambuj Tewari. On the equivalence between online and private learnability beyond binary classification. *arXiv:2006.01980*, 2020.

Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Mach. Learn.*, 2(4):285318, 1988. ISSN 0885-6125.

Alexander Rakhlin and Karthik Sridharan. Statistical Learning and Sequential Prediction, 2014. http://www.mit.edu/~rakhlin/courses/stat928/stat928_notes.pdf.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Sequential complexities and uniform laws of large numbers. *Probability Theory and Related Fields*, 2014.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning via sequential complexities. *Journal of Machine Learning Research*, 16:155–186, 2015.

Sasha Rakhlin, Ohad Shamir, and Karthik Sridharan. Relax and randomize : From value to algorithms. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 2141–2149. Curran Associates, Inc., 2012.

Norbert Sauer. On the density of families of sets. *J. Comb. Theory, Ser. A*, 13:145–147, 1972.

S. Shalev-Shwartz. *Online Learning and Online Convex Optimization*, volume 4. Foundations and Trends in Machine Learning, 2012.

Saharon Shelah. *Classification Theory and the Number of Non-Isomorphic Models*. North-Holland Pub. Co., 1978.

Vladimir N. Vapnik and Alexey Ja. Chervonenkis. Uniform convergence of frequencies of occurence of events to their probabilities. *Soviet Math Dokl.*, 9(4), 1968.

Xiaodong Xu and Stanisław P. Radziszowski. On some open questions for Ramsey and Folkman numbers. In Ralucca Gera, Stephen Hedetniemi, and Craig Larson, editors, *Graph Theory: Favorite Conjectures and Open Problems - 1*, pages 43–62. Springer International Publishing, Cham, 2016.

Appendix A. 0-covering numbers vs. thicket shatter function

In this section we discuss an alternative to the 0-covering number for which a Sauer–Shelah lemma holds as well. This alternative to the 0-covering number is known as the thicket shatter function (Bhaskar, 2017):

Definition 12 (Thicket shatter function) *For an X-valued tree $\mathbf{x}$ and function class $\mathcal{F}$, let $\rho(\mathcal{F}, \mathbf{x})$ denote the number of sequences $\epsilon = (\epsilon_1, \ldots, \epsilon_n) \in \{0,1\}^n$ so that there is some $f \in \mathcal{F}$ with*

$$f(\mathbf{x}_t(\epsilon_{1:t-1})) = \epsilon_t \;\; \forall t \in \{1, 2, \ldots, n\}. \tag{2}$$

In the event that (2) holds, we will say that the sequence ϵ admits a solution in $\mathcal{F}$ for the tree $\mathbf{x}$.

Analogously to Lemma 2, Bhaskar (2017, Theorem 4.1) showed that if the Littlestone dimension of $\mathcal{F}$ is at most d, then for any tree $\mathbf{x}$ of depth n, we have $\rho(\mathcal{F}, \mathbf{x}) \leq \sum_{i=0}^{d} \binom{n}{i}$. Lemma 13 shows that this statement follows directly from Lemma 2.

Lemma 13 *For an X-valued tree $\mathbf{x}$ and a function class $\mathcal{F} \subset \{0,1\}^X$, it holds that $\rho(\mathcal{F}, \mathbf{x}) \leq \mathcal{N}_0(\mathcal{F}, \mathbf{x})$.*

Proof Let us give $\{0,1\}^n$ the lexicographic ordering with $(0,\dots,0)$ first, $(0,\dots,0,1)$ second, $(0,\dots,1,0)$ third, and so on. Let $\mathcal{V}$ be a 0-cover for $\mathcal{F}$ on the tree $\mathbf{x}$.

Fix any sequence $\epsilon = (\epsilon_1,\dots,\epsilon_n) \in \{0,1\}^n$ that admits a solution in $\mathcal{F}$ (i.e., (2) holds). There must be some $\mathbf{v}^{(\epsilon)} \in \mathcal{V}$ so that for $t \in [n]$, we have $\mathbf{v}_t^{(\epsilon)}(\epsilon_{1:t-1}) = \epsilon_t$. Fix any $\epsilon' < \epsilon$ (using the lexicographic ordering) which also admits a solution in $\mathcal{F}$, and choose t_0 as small as possible so that $\epsilon'_{t_0} < \epsilon_{t_0}$. For all $t < t_0$, it follows that $\epsilon'_t = \epsilon_t$. Then we have

$$\mathbf{v}_{t_0}^{(\epsilon')}(\epsilon_{1:t_0-1}) = 0 \neq 1 = \mathbf{v}_{t_0}^{(\epsilon)}(\epsilon_{1:t_0-1}).$$

Hence $\mathbf{v}^{(\epsilon)} \neq \mathbf{v}^{(\epsilon')}$, and hence, for all $\epsilon \in \{0,1\}^n$ admitting a solution in $\mathcal{F}$, the $\mathbf{v}^{(\epsilon)}$ are distinct. ■

As an immediate corollary of Lemma 13, we obtain Lemma 4, since a tree $\mathbf{x}$ that is shattered by $\mathcal{F}$ satisfies $\rho(\mathcal{F},\mathbf{x}) = 2^n$.

Finally, we show in Proposition 14 that there are trees $\mathbf{x}$ so that $\rho(\mathcal{F},\mathbf{x})$ and $\mathcal{N}_0(\mathcal{F},\mathbf{x})$ may be very far apart. (This fact is not used to prove any other results in this article.)

Proposition 14 *For any $n \in \mathbb{N}$, there is a function class $\mathcal{F}$ and a tree $\mathbf{x}$ of depth n so that $\rho(\mathcal{F},\mathbf{x}) = 1$ yet $\mathcal{N}_0(\mathcal{F},\mathbf{x}) \geq 2^{n-1}$.*

Proof Let us label all $2^n - 1$ nodes of the tree $\mathbf{x}$ with different elements of X; in particular, for each $1 \leq t \leq n$, denote the 2^{t-1} nodes of layer t by $x_{t,1},\dots,x_{t,2^{t-1}}$, with all $x_{t,j}$ distinct. For simplicity we may assume that $X = \{x_{t,j} : t \in [n], 1 \leq j \leq 2^{t-1}\}$. Now, choose $\mathcal{F}$ to be the set of all functions $f : X \to \{0,1\}$ so that $f(x_{1,1}) = f(x_{2,1}) = \cdots = f(x_{n,1}) = 0$. Then $\rho(\mathcal{F},\mathbf{x}) = 1$ since the only ϵ admitting a solution in $\mathcal{F}$ for the tree $\mathbf{x}$ (i.e., satisfying (2)) is $\epsilon = (0,\dots,0)$.

On the other hand, letting $\epsilon_1 := 1$, then for any $\epsilon_2,\dots,\epsilon_n \in \{0,1\}$, there is some $f \in \mathcal{F}$ so that

$$f(x_{1,1}) = f(\mathbf{x}_1) = 0, \quad f(\mathbf{x}_t(\epsilon_{1:t-1})) = \epsilon_t \quad \forall t \geq 2.$$

Now the argument of Lemma 13 establishes that there must be a unique element of a 0-cover for each sequence of the form $(1,\epsilon_2,\dots,\epsilon_n)$. Thus $\mathcal{N}_0(\mathcal{F},\mathbf{x}) \geq 2^{n-1}$. ■