Online Learning and Disambiguations of Partial Concept Classes

Tsun-Ming Cheung
McGill University, Montreal, Canada

Hamed Hatami
McGill University, Montreal, Canada

Pooya Hatami
Ohio State University, Columbus, OH, USA

Kaave Hosseini
University of Rochester, NY, USA

Abstract

In a recent article, Alon, Hanneke, Holzman, and Moran (FOCS '21) introduced a unifying framework to study the learnability of classes of *partial* concepts. One of the central questions studied in their work is whether the learnability of a partial concept class is always inherited from the learnability of some "extension" of it to a total concept class.

They showed this is not the case for PAC learning but left the problem open for the stronger notion of online learnability.

We resolve this problem by constructing a class of partial concepts that is online learnable, but no extension of it to a class of total concepts is online learnable (or even PAC learnable).

2012 ACM Subject Classification Theory of computation → Online learning theory

Keywords and phrases Online learning, Littlestone dimension, VC dimension, partial concept class, clique vs independent set, Alon-Saks-Seymour conjecture, Standard Optimal Algorithm, PAC learning

Digital Object Identifier 10.4230/LIPIcs.ICALP.2023.42

Category Track A: Algorithms, Complexity and Games

Related Version *arXiv*: `https://arxiv.org/abs/2303.17578`

Funding *Hamed Hatami*: Supported by an NSERC grant.
Pooya Hatami: Supported by NSF grant CCF-1947546.

Acknowledgements We wish to thank Mika Göös for clarifying the reductions in [3, 4, 5, 2].

1 Introduction

In many practical learning problems, the learning task is tractable because we are only required to predict the labels of the data points that satisfy specific properties. In the setting of binary classification problems, instead of learning a total concept $h : \mathcal{X} \to \{0, 1\}$, we are often content with learning a partial version of it $\widetilde{h} : \mathcal{X} \to \{0, 1, \star\}$, where $\widetilde{h}(x) = \star$ means that both 0 and 1 are acceptable predictions. This relaxation of allowing unspecified predictions renders a wider range of learning tasks tractable.

Consider, for example, predicting whether a person approves or disapproves of various political stances by observing their previous voting pattern. This person might not hold a strong opinion about particular political sentiments, and it might be impossible to predict their vote on those issues based on their previous history. However, the learning task might become possible if we allow both "approve" and "disapprove" as acceptable predictions in those cases where a firm conviction is lacking.

A well-studied example of this phenomenon is learning half-spaces with a large margin. In this problem, the domain is the set of points in a bounded region in an arbitrary Euclidean space, and the concepts are half-spaces that map each point to 1 or 0 depending on whether they belong to the half-space or not. It is well-known that when the dimension of the underlying Euclidean space is large, one needs many samples to learn a half-space. However, in the large margin setting, we are only required to correctly predict the label of a point if its distance from the defining hyperplane is bounded from below by some margin. Standard learning algorithms for this task, such as the classical Perceptron algorithm, due to Rosenblatt [9], show that this relaxation of the learning requirement makes the problem tractable even for high-dimensional Euclidean spaces. Motivated by such examples, Alon, Hanneke, Holzman, and Moran [1] initiated a systematic study of the learnability of partial concept classes $\mathbb{H} \subseteq \{0,1,\star\}^{\mathcal{X}}$. They focused on the two frameworks of *probably approximately correct (PAC) learning* and *online learning*. We refer to [1] for the definition of PAC learnability of partial concept classes. We define online learnability in Definition 4.

PAC learning is an elegant theoretical framework characterized by the combinatorial parameter of the Vapnik–Chervonenkis (VC) dimension. The fundamental theorem of PAC learning states that a total binary concept class is PAC learnable if and only if its VC dimension is finite. Similarly, online learnability of total concept classes is characterized by a combinatorial parameter called the Littlestone dimension (LD). We formally define the VC dimension and the Littlestone dimension in Definitions 14 and 15 respectively. Alon, Hanneke, Holzman, and Moran [1] proved that these characterizations of PAC and online learnability extend to the setting of partial concept classes.

▶ **Theorem 1** ([1, Theorems 1 and 15]). *Let $\mathbb{H} \subseteq \{0,1,\star\}^{\mathcal{X}}$ be a partial concept class.*

- *$\mathbb{H}$ is PAC learnable if and only if $\mathrm{VC}(\mathbb{H}) < \infty$.*
- *$\mathbb{H}$ is online learnable if and only if $\mathrm{LD}(\mathbb{H}) < \infty$.*

It follows from the definitions of VC and LD dimensions that for every partial concept class $\mathbb{H} \subseteq \{0,1,\star\}^{\mathcal{X}}$, we have $\mathrm{VC}(\mathbb{H}) \leq \mathrm{LD}(\mathbb{H})$. In particular, online learnability always implies PAC learnability.

One of the central questions studied in [1] is whether the learnability of a partial concept class is always inherited from the learnability of some total concept class. To make this question precise, we need to define the notion of disambiguation of a partial concept class. While we defer the formal definitions to Section 2.2, one may understand a *strong disambiguation* of a partial class as simply an assignment of each $\star$ to either 1 or 0 for each partial concept in the class. When $\mathcal{X}$ is infinite, it is more natural to consider the weaker notion of *disambiguation* that we shall define in Definition 17. When $\mathcal{X}$ is finite, the notions of disambiguation and strong disambiguation coincide.

Consider the problem of learning the partial concept class $\mathbb{H} \subseteq \{0,1,\star\}^{\mathcal{X}}$ in PAC learning or online learning. If the partial concept class $\mathbb{H}$ has a disambiguation $\overline{\mathbb{H}} \subseteq \{0,1\}^{\mathcal{X}}$ that is PAC learnable, then $\mathbb{H}$ is PAC learnable. This follows from $\mathrm{VC}(\mathbb{H}) \leq \mathrm{VC}(\overline{\mathbb{H}})$, or simply by running the PAC learning algorithm of $\overline{\mathbb{H}}$ on $\mathbb{H}$. Similarly, if a disambiguation $\overline{\mathbb{H}}$ of $\mathbb{H}$ is online learnable, then $\mathbb{H}$ is online learnable.

Is the learnability of every partial concept class inherited from the learnability of some disambiguation to a total concept class?

▶ **Question 2** (Informal [1]). *Does every learnable partial class have a learnable disambiguation?*

Equipped with the VC dimension characterization of Theorem 1, [1] proved that for PAC learning, the answer to Question 2 is *negative*.

▶ **Theorem 3** ([1, Theorem 11]). *For every $n \in \mathbb{N}$, there exists a partial concept class $\mathbb{H}_n \subseteq \{0,1,\star\}^{[n]}$ with $\mathrm{VC}(\mathbb{H}_n) = 1$ such that any disambiguation $\overline{\mathbb{H}}$ of $\mathbb{H}_n$ has $\mathrm{VC}(\overline{\mathbb{H}}) \geq (\log n)^{1-o(1)}$. Moreover, for $\mathcal{X} = \mathbb{N}$, there exists $\mathbb{H}_\infty \subseteq \{0,1,\star\}^{\mathcal{X}}$ with $\mathrm{VC}(\mathbb{H}_\infty) = 1$ such that $\mathrm{VC}(\overline{\mathbb{H}}) = \infty$ for every disambiguation $\overline{\mathbb{H}}$ of $\mathbb{H}_\infty$.*

While Theorem 3 gives a strong negative answer to Question 2 in the case of PAC learning, the question was left open for online learning. Roughly speaking, this question strengthens the bounded-VC assumption on $\mathbb{H}$ to bounded *Littlestone dimension* (LD), which pertains to *online learnability* of $\mathbb{H}$.

The authors in [1] also proposed a second open problem that replaces the bounded-VC dimension assumption by the assumption of *polynomial growth*. This assumption is weaker than bounded LD dimension but stronger than bounded VC dimension.

As we discuss below, our main result resolves these two open problems.

Online learnability

Online learning is performed in a sequence of consecutive rounds, where at round t, the learner is presented with an instance $x_t \in \mathcal{X}$ and is required to predict its label. After predicting the label, the correct label $y_t \in \{0,1\}$ is revealed to the learner. Note that even for partial concept classes, we require that the correct label is 0 or 1. The learner's goal is to make as few prediction mistakes as possible during this process. We assume that the true labels are always *realizable*, i.e. there is a partial concept $h \in \mathbb{H}$ with $h(x_i) = y_i$ for all $i = 1, \ldots, t$.

▶ **Definition 4** (Online Learnability). *A partial concept class $\mathbb{H} \subseteq \{0,1,\star\}^{\mathcal{X}}$ is online learnable if there is a* mistake bound *$m := m(\mathbb{H}) \in \mathbb{N}$ such that for every $T \in \mathbb{N}$, there exists a learning algorithm that on every realizable sequence $(x_i, y_i)_{i=1,\ldots,T}$ makes at most m mistakes.*

Online learnability for total classes is equivalent to the bounded Littlestone dimension. In Theorem 1, Alon, Hanneke, Holzman, and Moran [1] showed that the same equivalence carries out in the setting of partial classes. They asked the following formulation of Question 2.

If a partial class is online learnable, is there a disambiguation of it that is online learnable?

More precisely, they pose the following question:

▶ **Problem 5** ([1]). *Let $\mathbb{H}$ be a partial class with $\mathrm{LD}(\mathbb{H}) < \infty$. Does there exist a disambiguation $\overline{\mathbb{H}}$ of $\mathbb{H}$ with $\mathrm{LD}(\overline{\mathbb{H}}) < \infty$? Is there one with $\mathrm{VC}(\overline{\mathbb{H}}) < \infty$?*

We give a negative answer to Problem 5:

▶ **Theorem 6** (Main Theorem). *For every $n \in \mathbb{N}$, there exists a partial concept class $\mathbb{H}_n \subseteq \{0,1,\star\}^{[n]}$ with $\mathrm{LD}(\mathbb{H}_n) \leq 2$ such that every disambiguation $\overline{\mathbb{H}}$ of $\mathbb{H}_n$ satisfies $\mathrm{LD}(\overline{\mathbb{H}}) \geq \mathrm{VC}(\overline{\mathbb{H}}) = \Omega(\log\log n)$. Consequently, for $\mathcal{X} = \mathbb{N}$, there exists $\mathbb{H}_\infty \subseteq \{0,1,\star\}^{\mathcal{X}}$ with $\mathrm{LD}(\mathbb{H}_\infty) \leq 2$ and $\mathrm{LD}(\overline{\mathbb{H}}) \geq \mathrm{VC}(\overline{\mathbb{H}}) = \infty$ for every disambiguation $\overline{\mathbb{H}}$ of $\mathbb{H}_\infty$.*

Polynomial growth

A general strategy to prove a super-constant lower bound on the VC dimension of a total concept class $\mathbb{H} \subseteq \{0,1\}^n$ is to show that the class is of super-polynomial size. This is the approach utilized in Theorem 3 and Theorem 6. For a total concept class $\mathbb{H} \subseteq \{0,1\}^n$ with VC dimension d, one has $2^d \leq |\mathbb{H}| \leq O(n^d)$: the lower bound is immediate from the definition of VC dimension, and the upper bound is the consequence of the celebrated Sauer-Shelah-Perles (SSP) lemma.

▶ **Theorem 7** (Sauer-Shelah-Perles lemma [10]). *Let* $\mathbb{H} \subseteq \{0,1\}^n$ *and* $\mathrm{VC}(\mathbb{H}) = d$. *Then*

$$|\mathbb{H}| \leq \binom{n}{\leq d} \coloneqq \sum_{i=0}^{d} \binom{n}{i} = O(n^d).$$

The direct analog of the SSP lemma is not true for partial concept classes: [1] proved that there exists $\mathbb{H} \subseteq \{0,1,\star\}^{[n]}$ with $\mathrm{VC}(\mathbb{H}) = 1$ such that every disambiguation $\overline{\mathbb{H}}$ has size $|\overline{\mathbb{H}}| \geq n^{\Omega(\log n)}$. This result, combined with the SSP lemma for total classes, immediately implies Theorem 3.

Interestingly, under the stronger assumption of the bounded Littlestone dimension, the polynomial growth behavior of the original SSP lemma remains valid.

▶ **Theorem 8** ([1]). *Every partial concept class* $\mathbb{H} \subseteq \{0,1,\star\}^{[n]}$ *with* $\mathrm{LD}(\mathbb{H}) \leq d$ *has a disambiguation* $\overline{\mathbb{H}}$ *with* $|\overline{\mathbb{H}}| \leq O(n^d)$.

We say that a partial concept class $\mathbb{H} \subseteq \{0,1,\star\}^{\mathcal{X}}$ has *polynomial growth with parameter* $d \in \mathbb{N}$ if for every finite $\mathcal{X}' \subseteq \mathcal{X}$, there is a disambiguation $\overline{\mathbb{H}|_{\mathcal{X}'}}$ of $\mathbb{H}|_{\mathcal{X}'}$ of size at most $O(|\mathcal{X}'|^d)$. Note that by Theorem 8, every partial concept class with Littlestone dimension d has polynomial growth with parameter d.

Alon, Hanneke, Holzman, and Moran asked the following question:

▶ **Problem 9** ([1]). *Let* $\mathbb{H} \subseteq \{0,1,\star\}^{\mathcal{X}}$ *be a partial concept class with polynomial growth. Does there exist a disambiguation* $\overline{\mathbb{H}}$ *of* $\mathbb{H}$ *such that* $\mathrm{VC}(\overline{\mathbb{H}}) < \infty$*?*

Note that Problem 9 cannot be resolved (in the negative) by a naive application of the SSP lemma to disambiguations of $\mathbb{H}$ or its restrictions. However, Theorem 6 combined with Theorem 8 refutes Problem 9 as well.

▶ **Theorem 10.** *For every* $n \in \mathbb{N}$, *there is* $\mathbb{H} \subseteq \{0,1,\star\}^{[n]}$ *with polynomial growth with parameter* 2 *such that every disambiguation* $\overline{\mathbb{H}}$ *of* $\mathbb{H}$ *has* $\mathrm{VC}(\overline{\mathbb{H}}) = \Omega(\log\log n)$.

Consequently, for $\mathcal{X} = \mathbb{N}$, *there exists* $\mathbb{H}_\infty \subseteq \{0,1,\star\}^{\mathcal{X}}$ *with polynomial growth with parameter* 2 *such that every disambiguation* $\overline{\mathbb{H}_\infty}$ *of* $\mathbb{H}_\infty$ *has* $\mathrm{VC}(\overline{\mathbb{H}_\infty}) = \infty$.

The Alon-Saks-Seymour Problem

The proof of Theorem 3 in [1] hinges on the breakthrough result of Göös [4] and its subsequent improvements [2] that led to almost optimal super-polynomial bounds on the "biclique partition number versus chromatic number" problem of Alon, Saks, and Seymour. The *biclique partition number* of a graph G, denoted by $\mathrm{bp}(G)$, is the smallest number of complete bipartite graphs (bicliques) that partition the edge set of G. Alon, Saks, and Seymour conjectured that the chromatic number of a graph with biclique partition number k is at most $k+1$. Huang and Sudakov refuted the Alon-Saks-Seymour conjecture in [6] by establishing a superlinear gap between the two parameters. Later in a breakthrough, Göös [4] proved a superpolynomial separation.

Our main result, Theorem 6, also builds on the aforementioned graph constructions. However, unlike previous works, our theorem demands a reasonable upper bound on the number of vertices. Since the constructions result from a complex sequence of reductions involving query complexity, communication complexity, and graph theory [3, 4, 5, 2], it is necessary to scrutinize them to ensure that the required parameters are met. We present a reorganized and partly simplified sequence of constructions in Section 3.3 that establishes the following theorem.

▶ **Theorem 11** (Small-size refutation of the Alon-Saks-Seymour conjecture). *There exists a graph G on $2^{\Theta(k^4 \log^3 k)}$ vertices that admits a biclique partition of size $2^{O(k \log^4 k)}$ but its chromatic number is at least $2^{\Omega(k^2)}$.*

Theorem 11 is essentially due to [2]. Our contribution to this theorem is obtaining an explicit and optimized bound on the size of G.

Standard Optimal Algorithm

Theorem 6 provides an example partial class with Littlestone dimension ≤ 2, such that the VC dimension of every disambiguation is $\Omega(\log \log n)$. Whether one can improve the $\Omega(\log \log n)$ lower bound is unclear. In particular, it is an interesting question whether every disambiguation of a partial class of Littlestone dimension at most 2 has VC dimension $O(\log \log n)$. One natural candidate approach for obtaining such an upper bound would be to utilize the Standard Optimal Algorithm (SOA).

SOA is an online learning algorithm devised by Littlestone [7] that can learn classes with bounded Littlestone dimensions. Alon, Hanneke, Holzman, and Moran, in their proof of Theorem 8, showed that applying SOA to a partial concept class $\mathbb{H}$ with Littlestone dimension d yields a disambiguation of size $|\overline{\mathbb{H}}| \leq O(n^d)$ and consequently VC dimension $O(d \log n)$. This shows that the lower bound of Theorem 6 on VC dimension of disambiguations cannot be improved beyond $O(\log n)$. It is hence natural to ask whether it is possible to obtain an improved upper bound on the VC dimension of the SOA-based disambiguation.

We answer this question in the negative by constructing a family of partial concept classes $\mathbb{H}$ of Littlestone dimension d where the disambiguation obtained by the SOA algorithm has VC dimension $\Omega(d \log(n/d))$.

▶ **Theorem 12.** *For every natural numbers $d \leq n$, there exists a partial concept class $\mathbb{H}_{n,d} \subseteq \{0,1,\star\}^{[n]}$ with $d \leq \mathrm{LD}(\mathbb{H}_{n,d}) \leq d+1$ such that the SOA disambiguation of $\mathbb{H}_{n,d}$ has* VC *dimension $\Omega(d \log(n/d))$.*

2 Preliminaries and Background

For a positive integer k, we denote $[k] \coloneqq \{1, \ldots, k\}$. We adopt the convention that $\{0,1\}^0$ or $\{0,1,\star\}^0$ contains the empty string only, which we denote by $()$.

We adopt the standard computer science asymptotic notations, such as Big-O, and use the asymptotic tilde notations to hide poly-logarithmic factors.

2.1 VC Dimension and Littlestone Dimension

Let $\mathbb{H} \subseteq \{0,1,\star\}^{\mathcal{X}}$ be a partial concept class. When the domain $\mathcal{X}$ is finite, we sometimes view $\mathbb{H}$ as a partial matrix $\mathbf{M}_{\mathcal{X} \times \mathbb{H}}$, where each row corresponds to a point $x \in \mathcal{X}$ and each column corresponds to a concept $h \in \mathbb{H}$, and the entries are defined as $\mathbf{M}(x,h) = h(x)$.

Next, we define the VC dimension and the Littlestone dimension of partial classes, which generalize the definitions of these notions for total classes. As shown in [1], the VC and Littlestone dimensions for partial classes capture PAC and online learnability, respectively.

▶ **Definition 13** (Shattered set). *A finite* set *of points $C = \{x_1, \ldots, x_n\} \subseteq \mathcal{X}$ is* shattered *by a partial concept class $\mathbb{H} \subseteq \{0,1,\star\}^{\mathcal{X}}$ if for every pattern $y \in \{0,1\}^n$, there exists $h \in \mathbb{H}$ with $h(x_i) = y_i$ for all $i \in [n]$.*

▶ **Definition 14** (VC dimension). *The* VC *dimension of a partial class* $\mathbb{H}$*, denoted by* $\mathrm{VC}(\mathbb{H})$*, is the maximum* d *such that there exists a size-*d *subset of* $\mathcal{X}$ *that is shattered by* $\mathbb{H}$*. If no such largest* d *exists, define* $\mathrm{VC}(\mathbb{H}) = \infty$*.*

Viewed as a matrix, the VC dimension of $\mathbb{H}$ is the maximum d such that the associated partial matrix $\mathbf{M}_{\mathcal{X}\times\mathbb{H}}$ contains a zero/one submatrix of dimensions $d \times 2^d$, where the columns enumerate all d-bit zero/one patterns.

The Littlestone dimension is defined through the shattering of decision trees instead of sets. Consider a full binary decision tree of height d where every non-leaf v is labelled with an element $x_v \in \mathcal{X}$. We identify every node of this tree by the string $v \in \bigcup_{k=0}^{d}\{0,1\}^k$ that corresponds to the path from the root to the node. That is, the root is the empty string, its children are the two elements in $\{0,1\}$, and more generally, the children of a node $\vec{v} \in \{0,1\}^k$ are the two strings $\vec{v}0$ and $\vec{v}1$ in $\{0,1\}^{k+1}$.

We say that such a tree is *shattered* by a partial concept class $\mathbb{H}$ if for every leaf $y \in \{0,1\}^d$, there exists $h \in \mathbb{H}$ such that $h(x_{y[<i]}) = y_i$ for each $i \in [d]$, where $y[< i]$ is the first $(i-1)$-th bits of y. In other words, applying the decision tree to h will result in the leaf y.

▶ **Definition 15** (Littlestone dimension). *The* Littlestone dimension *of a partial concept class* $\mathbb{H}$*, denoted by* $\mathrm{LD}(\mathbb{H})$*, is the maximum* d *such that there is an* $\mathcal{X}$*-labelled height-*d *full binary decision tree that is shattered by* $\mathbb{H}$*. If no such largest* d *exists, define* $\mathrm{LD}(\mathbb{H}) = \infty$*.*

The *dual* of a concept class $\mathbb{H}$ is the concept class with the roles of points and concepts exchanged. Concretely, the dual class of $\mathbb{H} \in \{0,1,\star\}^{\mathcal{X}}$, denoted by $\mathbb{H}^\top$, is the collection of functions $f_x : \mathbb{H} \to \{0,1,\star\}$ for every $x \in \mathcal{X}$, which is defined by $f_x(h) = h(x)$ for each $h \in \mathbb{H}$. When $\mathcal{X}$ is finite, taking the dual corresponds to transposing the matrix of the concept class. The VC-dimension of the dual-class is related to that of the primal class by the inequality

$$\mathrm{VC}(\mathbb{H}^\top) \le 2^{\mathrm{VC}(\mathbb{H})+1} - 1$$

(see [8]), which translates to a lower bound of the VC-dimension of the primal class.

2.2 Disambiguations

We start by formally defining *strong disambiguation* and *disambiguation*. As mentioned earlier, the two notions coincide when the domain $\mathcal{X}$ is finite.

▶ **Definition 16** (Strong Disambiguation). *A* strong disambiguation *of a partial concept class* $\mathbb{H} \subseteq \{0,1,\star\}^{\mathcal{X}}$ *is a total concept class* $\overline{\mathbb{H}} \subseteq \{0,1\}^{\mathcal{X}}$ *such that for every* $h \in \mathbb{H}$*, there exists a* $\bar{h} \in \overline{\mathbb{H}}$ *that is consistent with* h *on the points* $h^{-1}(\{0,1\})$*.*

▶ **Definition 17** (Disambiguation). *A* disambiguation *of a partial concept class* $\mathbb{H} \subseteq \{0,1,\star\}^{\mathcal{X}}$ *is a total concept class* $\overline{\mathbb{H}} \subseteq \{0,1\}^{\mathcal{X}}$ *such that for every* $h \in \mathbb{H}$ *and every finite* $S \subseteq h^{-1}(\{0,1\})$*, there exists* $\bar{h} \in \overline{\mathbb{H}}$ *that is consistent with* h *on* S*.*

A learning algorithm can often provide a disambiguation of a partial concept class by assigning the prediction of the algorithm to unspecified values. Relevant to our work is the disambiguation by the Standard Optimal Algorithm of Littlestone. It was observed in [1] that this algorithm can provide "efficient" disambiguations of partial classes with bounded Littlestone dimensions. We describe this disambiguation next.

Consider a partial concept class $\mathbb{H} \subseteq \{0,1,\star\}^{\mathcal{X}}$ with a countable domain $\mathcal{X}$ and an ordering $x_1, x_2, \ldots$ of $\mathcal{X}$. Given $\vec{b} \in \{0,1,\star\}^k$, let $\mathbb{H}|_{\vec{b}}$ be the set of concepts h where $h(x_i) = b_i$ for every $i \in [k]$. For convenience, we identify $\mathbb{H}|_{()} = \mathbb{H}$. For the purpose of the algorithm, we adopt the convention $\mathrm{LD}(\emptyset) = -1$.

The SOA obtains a disambiguation iteratively and assigns a 0/1 value to each $\star$ in $\mathbb{H}$: for each $k \in \mathbb{N}$, consider $\mathbb{H}|_{\vec{b}}$ for every $\vec{b} \in \{0,1\}^{k-1}$. Pick $c \in \{0,1\}$ which maximizes $\mathrm{LD}(\mathbb{H}|_{\vec{b}c})$, breaking ties by favoring $c = 0$, and assign c to $h(x_k) = \star$ for every $h \in \mathbb{H}|_{\vec{b}\star}$.

We use the notation $\overline{\mathbb{H}}^{\mathrm{SOA}}$ for the SOA disambiguation of a partial concept class $\mathbb{H}$. As mentioned earlier, for a partial class with Littlestone dimension d, Theorem 8 gives an upper bound of $\binom{n}{\leq d} = O(n^d)$ on $\left|\overline{\mathbb{H}}^{\mathrm{SOA}}\right|$. The theorem follows from the mistake bound of SOA for online learning, which relies on the crucial property that at least one choice of $c \in \{0,1\}$ satisfies $\mathrm{LD}(\mathbb{H}|_{\vec{b}c}) \leq \mathrm{LD}(\mathbb{H}|_{\vec{b}}) - 1$ whenever $\mathbb{H}|_{\vec{b}} \neq \emptyset$.

3 Proofs

In this section, we present the proofs of Theorems 6, 10, 11, and 12.

3.1 Proofs of Theorems 6 and 10

As mentioned earlier, Theorem 10 is an immediate corollary of Theorem 6 and Theorem 8. We focus on proving Theorem 6.

Suppose $G = (V, E)$ is the graph supplied by Theorem 11 on $|V| = n = 2^{\Theta(k^4 \log^3 k)}$ vertices with a biclique partition of size $m = 2^{O(k \log^4 k)}$. We will use G to build a partial concept class $\mathbb{G} \subseteq \{0,1,\star\}^V$. This construction is simply the dual of the partial concept class of [1] in their proof of Theorem 6.

Let $\{B_1, \ldots, B_m\}$ be the size-m biclique partition of the edges of G. We fix an orientation $B_i = L_i \times R_i$ for each biclique. Define $\mathbb{G} \subseteq \{0,1,\star\}^V$ as follows. For each $i \in [m]$, associate a concept $h_i : V \to \{0,1,\star\}$ to the biclique B_i, defined by

$$h_i(v) = \begin{cases} 0 & \text{if } v \in L_i \\ 1 & \text{if } v \in R_i \\ \star & \text{otherwise} \end{cases}.$$

We first observe that the Littlestone dimension of this concept class is at most 2.

▷ Claim 18. $\mathrm{LD}(\mathbb{G}) \leq 2$.

Proof. We show that $\mathbb{G}$, viewed as a matrix, does not contain $\begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}$ as a submatrix and then show that the existence of this submatrix is necessary for having a Littlestone dimension greater than 2.

If $\begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}$ appears in $\mathbb{G}$ as a submatrix, then there exist $i \neq j$ and $u \neq v \in V(G)$ such that $h_i(v) = h_j(v) = 1$ and $h_i(u) = h_j(u) = 0$. However, this means that $v \in R_i \cap R_j$ and $u \in L_i \cap L_j$, which in turn implies that the edge $\{u, v\}$ is covered by both B_i and B_j, contradicting the assumption that each edge is covered exactly once.

On the other hand, for a class $\mathbb{H} \subseteq \{0,1,\star\}^{\mathcal{X}}$ with Littlestone dimension greater than 2, there exists a shattered $\mathcal{X}$-labelled height-3 full binary tree. In particular, there exists $h, h' \in \mathbb{H}$ and points $x_{()}, x_1, x_{10}$ such that

$$\begin{array}{lll} h(x_{()}) = 1, & h(x_1) = 0, & h(x_{10}) = 0, \\ h'(x_{()}) = 1, & h'(x_1) = 0, & h'(x_{10}) = 1. \end{array}$$

This means that the submatrix restricted to the columns $\{x_{()}, x_1\}$ and the rows $\{h, h'\}$ is $\begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}$. We conclude that $\mathrm{LD}(\mathbb{G}) \leq 2$. ◁

Proof of Theorem 6. Consider the partial concept class $\mathbb{G} \subseteq \{0,1,\star\}^V$ above. By Claim 18, we have $\mathrm{LD}(\mathbb{G}) \leq 2$. We show that for every disambiguation $\overline{\mathbb{G}}$ of $\mathbb{G}$, we have $\mathrm{VC}(\overline{\mathbb{G}}) \geq \Omega(\log\log n)$. The argument here is similar to the proof of Theorem 3.

Consider a disambiguation $\overline{\mathbb{G}}$ of $\mathbb{G}$. Note that if two columns u and v are identical in $\overline{\mathbb{G}}$, then there is no edge between u and v, as otherwise, some h_i would have assigned 0 to one of u and v and 1 to the other. Therefore, if two columns u and v are identical, we can color the corresponding vertices with the same color. Consequently, the number of distinct columns in $\overline{\mathbb{G}}$ is at least the chromatic number $\chi(G) \geq 2^{\Omega(k^2)}$. By the SSP lemma (Theorem 7), if $\mathrm{VC}(\overline{\mathbb{G}}^\top) \leq d$, then $\overline{\mathbb{G}}$ must have at most $O(m^d)$ distinct columns. Therefore,

$$2^{\Omega(k^2)} \leq O(m^d).$$

Substituting $m = 2^{\tilde{O}(k)}$ shows that $d = \tilde{\Omega}(k)$. Finally,

$$\mathrm{VC}(\overline{\mathbb{G}}) \geq \Omega(\log \mathrm{VC}(\overline{\mathbb{G}}^\top)) \geq \Omega(\log k) \geq \Omega(\log\log n).$$

This completes the proof of the first part of Theorem 6.

For the second part, we adopt the same construction in the proof of [1, Theorem 11]. Let $\mathbb{H}_\infty$ be a union of disjoint copies of $\mathbb{H}_n$ over $n \in \mathbb{N}$, each supported on a domain $\mathcal{X}_n$ mutually disjoint from others and the partial concepts of $\mathbb{H}_n$ extend outside of its domain by $\star$. Since any disambiguation $\mathbb{H}$ of $\mathbb{H}_\infty$ simultaneously disambiguates all $\mathbb{H}_n$, the Sauer-Shelah-Perles lemma implies that $\mathrm{VC}(\mathbb{H})$ must be infinite. ◀

3.2 Disambiguations via the SOA algorithm (Theorem 12)

This section is dedicated to the proof of Theorem 12.

Proof of Theorem 12. We prove the statement by showing that for every $r, d \in \mathbb{N}$, there exists a partial concept class $\mathbb{H}_{r,d}$ on $[n]$, where $n = d(2^r + r)$, such that $d \leq \mathrm{LD}(\mathbb{H}_{r,d}) \leq d+1$ and the SOA disambiguation has VC dimension $\geq dr$ and at least 2^{dr} distinct rows. The other cases of n follow by trivially extending the domain.

For any $r, d \in \mathbb{N}$, define

$$\mathcal{F}_{r,d} = \{F \subseteq [d2^r] : |F| = d\}.$$

Note that $|\mathcal{F}_{r,d}| = \binom{d2^r}{d} \geq 2^{dr}$. We enumerate the sets in $\mathcal{F}_{r,d}$ as $F_1, \ldots, F_{\binom{d2^r}{d}}$ in the natural order.

Next, we define the partial concept class $\mathbb{H}_{r,d}$ on domain $[d(2^r + r)]$. The class consists of the partial concepts $h_{i,j}$ for $i \in [\binom{d2^r}{d}]$ and $j \in [dr]$ defined as follows:

$$h_{i,j}(x) = \begin{cases} 1 & \text{if } x \in F_i \\ 0 & \text{if } x \in [d2^r] \setminus F_i \\ \beta(i,j) & \text{if } x = d2^r + j \\ \star & \text{otherwise} \end{cases},$$

where $\beta(i,j)$ denotes j-th bit of the dr-bit binary representation of i if $i \in [2^{dr}]$, and $\beta(i,j) = \star$ otherwise.

We first prove that $d \leq \mathrm{LD}(\mathbb{H}_{r,d}) \leq d+1$. Note that there is a set of 2^d indices $I \subseteq [d2^r]$ which

$$\{F_i \cap [d] : i \in I\} = \mathcal{P}([d]),$$

therefore $[d]$ can be shattered by $\{h_{i,1} : i \in I\}$ and hence $\mathrm{LD}(\mathbb{H}_{r,d}) \geq \mathrm{VC}(\mathbb{H}_{r,d}) \geq d$. On the other hand, note that $|f^{-1}(1)| \leq d+1$ for any $f \in \mathbb{H}_{r,d}$, which implies that $\mathrm{LD}(\mathbb{H}_{r,d}) \leq d+1$.

Next, we consider the SOA disambiguation. We claim that $\{d2^r+1, \ldots, d(2^r+r)\}$ is shattered by $\{h_{i,1} : i \in [2^{dr}]\}$. There are no disambiguations for $x \in [d2^r]$. For $x > d2^r$, note that for any $\vec{b} \in \{0,1\}^{x-1}$, either $\mathbb{H}_{r,d}|_{\vec{b}} = \emptyset$ or

$$\mathbb{H}_{r,d}|_{\vec{b}} = \{h_{i,j} : j \in [dr]\},$$

where $i \in [d2^r]$ such that $F_i = \{k \in [d2^r] : b_k = 1\}$. We focus on the latter case and restrict to $i \in [2^{dr}]$. There is exactly one $c \in \{0,1\}$ such that $\mathbb{H}_{r,d}|_{\vec{b}c} \neq \emptyset$, namely $c = \beta(i, x - d2^r)$ and in this case $\mathbb{H}_{r,d}|_{\vec{b}c} = \{h_{i,c}\}$. This forces the algorithm to disambiguate every function f with $\vec{b} \in \{0,1\}^{x-1}$ by setting $f(x) = h_{i,c}(x) = \beta(i, x - d2^r)$. In this manner, every $h_{i,j}$ is eventually disambiguated into the same total function:

$$\overline{h_{i,j}}(x) = \begin{cases} 1 & \text{if } x \in F_i \\ 0 & \text{if } x \in [d2^r] \setminus F_i \\ \beta(i, x - d2^r) & \text{if } x > d2^r \end{cases}.$$

In particular, for every $i \in [2^{dr}]$, the bit string $(\overline{h_{i,1}}(d2^r+1), \ldots, \overline{h_{i,1}}(d2^r+dr))$ is the dr-bit binary representation of i. This provides a witness for which $\mathrm{VC}(\overline{\mathbb{H}_{r,d}}^{\mathrm{SOA}}) \geq dr$. ◀

As an illustration, we provide the matrix representation of $\mathbb{H}_{1,2}$ and some essential steps of the SOA disambiguation below in Figure 1.

$$\begin{pmatrix}
1 & 1 & 0 & 0 & 0 & \\
1 & 1 & 0 & 0 & & 0 \\
1 & 0 & 1 & 0 & 0 & \\
1 & 0 & 1 & 0 & & 1 \\
1 & 0 & 0 & 1 & 1 & \\
1 & 0 & 0 & 1 & & 0 \\
0 & 1 & 1 & 0 & 1 & \\
0 & 1 & 1 & 0 & & 1 \\
0 & 1 & 0 & 1 & & \\
0 & 1 & 0 & 1 & & \\
0 & 0 & 1 & 1 & & \\
0 & 0 & 1 & 1 & &
\end{pmatrix}$$

(a) Matrix representation of $\mathbb{H}_{1,2}$: all empty spaces are filled with stars.

$$\begin{pmatrix}
1 & 1 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 1 \\
1 & 0 & 1 & 0 & 0 & 1 \\
1 & 0 & 0 & 1 & 1 & 0 \\
1 & 0 & 0 & 1 & 1 & 0 \\
0 & 1 & 1 & 0 & 1 & 1 \\
0 & 1 & 1 & 0 & 1 & 1 \\
0 & 1 & 0 & 1 & & \\
0 & 1 & 0 & 1 & & \\
0 & 0 & 1 & 1 & & \\
0 & 0 & 1 & 1 & &
\end{pmatrix}$$

(b) The SOA disambiguation of $\mathbb{H}_{1,2}$: the shaded entries indicate where the shattering occurs.

Figure 1 $\mathbb{H}_{1,2}$ and its SOA disambiguation.

3.3 Small-size refutation of the Alon-Saks-Seymour conjecture (Theorem 11)

In this section, we present the construction of Theorem 11 in detail. The starting point is constructing a Boolean function due to [2] in query complexity. This Boolean function then goes through several reductions to be converted into a graph, as described below.

We first introduce some basic definitions related to the notion of *certificate complexity*. Let $f : \{0,1\}^n \to \{0,1\}$ be a Boolean function. For $b \in \{0,1\}$ and an input $x \in f^{-1}(b)$, a partial input $\rho \in \{0,1,\star\}^n$ is called a b-certificate if x is consistent with ρ and for every

$x' \in \{0,1\}^n$ consistent with ρ, we have $f(x') = b$. The size of ρ is the number of non-$\star$ entries of ρ. Define $\mathrm{C}_b(f,x)$ as the smallest size of a b-certificate for x. The b-certificate complexity of f, denoted $\mathrm{C}_b(f)$, is the maximum of $\mathrm{C}_b(f,x)$ over all $x \in f^{-1}(b)$.

The *unambiguous* b-certificate complexity of f, denoted $\mathrm{UC}_b(f)$, is the smallest k such that

1. Every input $x \in f^{-1}(b)$ has a b-certificate ρ_x of size at most k;
2. For every $x \neq y$ in $f^{-1}(b)$, we have $\rho_x \neq \rho_y$.

The main result of [2] is the following separation between UC_1 and C_0.

▶ **Theorem 19** ([2, Theorem 1]). *There is a function $f : \{0,1\}^{12n^4\log^2 n} \to \{0,1\}$ such that $\mathrm{UC}_1(f) = O(n\log^3 n)$ and $\mathrm{C}_0(f) = \Omega(n^2)$.*

The next step of the construction is to transform the function separating the certificate complexities UC_1 and C_0 into a communication problem. This is achieved by the "lifting" trick: given a function $f : \{0,1\}^n \to \{0,1\}$ and a "gadget" function $g : \{0,1\}^k \times \{0,1\}^k \to \{0,1\}$, we define $f \circ g^n : \{0,1\}^{nk} \times \{0,1\}^{nk} \to \{0,1\}$ as

$$f \circ g^n([x_1,\dots,x_n],[y_1,\dots,y_n]) = f(g(x_1,y_1),\dots,g(x_n,y_n)).$$

For a communication problem $f : \{0,1\}^m \times \{0,1\}^m \to \{0,1\}$ and $b \in \{0,1\}$, let $\mathrm{Cov}_b(f)$ denote the minimum number of b-monochromatic rectangles required to cover all the b-entries of f. We denote by $\mathrm{UCov}_b(f)$ the minimum number of b-monochromatic rectangles required to *partition* all the b-entries of f. The following theorem provides a connection between the communication complexity parameters and the certificate complexity parameters.

▶ **Theorem 20** ([5, Theorem 33]). *There exists a gadget $g : \{0,1\}^k \times \{0,1\}^k \to \{0,1\}$ with $k = \Omega(\log n)$ such that for every $f : \{0,1\}^n \to \{0,1\}$, we have*

$$\log \mathrm{Cov}_b(f \circ g^n) = \Omega(k\,\mathrm{C}_b(f)).$$

Note that for every $b \in \{0,1\}$, we have $\log \mathrm{UCov}_b(f \circ g^n) \leq 2k\,\mathrm{UC}_b(f)$. This combined with Theorem 20 allows one to "lift" the UC_1 vs C_0 separation of Theorem 19 into a UCov_1 vs Cov_0 separation.

▶ **Corollary 21.** *There exists a function $f : \{0,1\}^{O(n^4\log^3 n)} \times \{0,1\}^{O(n^4\log^3 n)} \to \{0,1\}$ such that*

$$\log \mathrm{Cov}_0(f) = \Omega(n^2) \qquad \text{and} \qquad \log \mathrm{UCov}_1(f) = n\log^4 n.$$

Next, we show how to convert these communication parameters to graph parameters of the biclique partition number and chromatic number.

▶ **Lemma 22.** *Let $h : \{0,1\}^t \times \{0,1\}^t \to \{0,1\}$ be a Boolean function with $\mathrm{Cov}_0(h) = c$ and $\mathrm{UCov}_1(h) = m$. There exists a graph $G = (V,E)$ on at most 2^{2t} vertices with $\mathrm{bp}(G) \leq m^2$ and $\chi(G) \geq \sqrt{c}$.*

Proof. Define the graph G with $V \coloneqq h^{-1}(0)$ as follows. Two vertices $(x,y),(x',y') \in V$ are adjacent in G iff $h(x,y') = 1$ or $h(x',y) = 1$. By construction, if $\{(x_1,y_1),\dots,(x_\ell,y_\ell)\} \subseteq V$ is an independent set, then $\{x_1,\dots,x_\ell\} \times \{y_1,\dots,y_\ell\}$ is a 0-monochromatic rectangle for h. Thus every proper vertex coloring of G with $\chi(G)$ colors corresponds to a 0-cover of h with $\chi(G)$ many 0-monochromatic rectangles. Therefore, $\chi(G) \geq c$.

We next show that there exists a small set of bicliques such that every edge of E is covered at least once and at most twice by these bicliques. Let $h^{-1}(1) = \bigcup_{i=1}^{m}(A_i \times B_i)$ be a partition of $h^{-1}(1)$ into m many 1-monochromatic rectangles. Note that every 1-monochromatic rectangle $A_i \times B_i$ corresponds to a biclique $Q_i := S_i^- \times S_i^+$ in G, where

$$S_i^- := \{(x,y) \in V(G) : \ x \in A_i\} \text{ and } S_i^+ = \{(x,y) \in V(G) : \ y \in B_i\}.$$

Notice that each edge $\{(x,y),(x',y')\}$ of G is covered at least once by $Q_1,\dots,Q_m$, and it is covered at most twice, the latter happening when $h(x,y') = h(x',y) = 1$.

We have thus constructed a graph G on at most 2^{2t} vertices such that $\chi(G) \geq c$, and there are at most m bicliques where every edge in G appears in at least one and at most two bicliques.

Define H_2 as the subgraph of G that consists of all the edges covered by exactly two bicliques among $Q_1,\dots,Q_m$. For every $i,j \in [m]$, define $Q_{ij} = (S_i^- \cap S_j^+) \times (S_i^+ \cap S_j^-)$. Note that each Q_{ij} is a biclique of H_2, and moreover, each edge of H_2 appears in exactly one Q_{ij}. Hence, the biclique partition number of H_2 is at most m^2. Now, if $\chi(H_2) \geq \sqrt{c}$, we obtain H_2 as the desired graph. Suppose otherwise that $\chi(H_2) < \sqrt{c}$, and consider a proper vertex coloring of H_2 with $\sqrt{c}$ colors with color classes $V_1,\dots,V_{\sqrt{c}}$. Since $\chi(G) \geq c$, there must exist i such that the induced subgraph of G on V_i, denoted by $G[V_i]$, satisfies $\chi(G[V_i]) \geq \sqrt{c}$. Since V_i is an independent set of H_2, thus the restrictions of bicliques $Q_1,\dots,Q_m$ to V_i form a biclique partition of $G[V_i]$. ◀

Lemma 22 and Corollary 21 together imply Theorem 11.

▶ Remark 23. In addition to providing effective bounds on the size of the graph, Lemma 22 also simplifies the original chain of reductions utilized in prior work [2, 4, 3, 11] toward achieving a super-polynomial separation between the biclique partition and chromatic numbers. We will briefly describe the original proof below and highlight the differences.

(i) Similar to our proof of Theorem 11, the chain of reduction begins with the function f provided by Corollary 21, such that

$$\log \mathrm{Cov}_0(f) = \Omega(n^2) \qquad \text{and} \qquad \log \mathrm{UCov}_1(f) = n \log^4 n.$$

(ii) Yannakasis [11] (see also [4, Figure 1]) showed how to use f to construct a graph F on $\mathrm{UCov}_1(f) = 2^{O(n \log^4 n)}$ vertices such that every Clique-Stable set separator of F is of size at least $\mathrm{Cov}_0(f) = 2^{\Omega(n^2)}$. Here, a Clique-Stable set separator is a collection of cuts in F such that for every disjoint pair (C,I) of a clique C and a stable set I in F, there is a cut (A,B) in the collection with $C \subseteq A$ and $I \subseteq B$.

(iii) Bousquet et. al., [3, Lemma 23] show how to use F to construct a new graph G with the so-called oriented biclique packing number at most $2^{n \log^4 n}$ and chromatic number $\chi(G) \geq 2^{\Omega(n^2)}$.

(iv) The graph G is then turned into a separation between the biclique partition number and chromatic number in a different graph H via a final reduction in [3].

The above chain of reductions is not sufficient for our application because the graph G of Step (iii) has a vertex for each pair (C,I) of a clique C and a stable set I of F, and as a result, there are no effective upper-bounds on the number of vertices of G. Our proof of Theorem 11 bypasses Step (ii) and employs a more direct approach to construct a small-size graph G that has similar properties to the graph G of Step (iii).

4 Concluding remarks

A few natural questions remain unanswered. The first question is whether a similar example $\mathbb{H}$ for Theorem 6 with the stronger assumption $\mathrm{LD}(\mathbb{H}) = 1$ exists.

▶ **Problem 24.** *Let $\mathbb{H}$ be a partial class with $\mathrm{LD}(\mathbb{H}) = 1$. Does there exist a disambiguation of $\mathbb{H}$ by a total class $\overline{\mathbb{H}}$ such that $\mathrm{LD}(\overline{\mathbb{H}}) < \infty$? Is there one with $\mathrm{VC}(\overline{\mathbb{H}}) < \infty$?*

Theorem 10 shows that for partial classes, having polynomial growth is not a sufficient condition for PAC learnability. A natural candidate reinstatement of the theorem is to work with the more restrictive assumption of linear growth.

▶ **Problem 25.** *Let $\mathbb{H} \subseteq \{0, 1, \star\}^{\mathcal{X}}$ have polynomial growth with parameter 1. Does there exist a disambiguation $\overline{\mathbb{H}}$ of $\mathbb{H}$ with $\mathrm{VC}(\overline{\mathbb{H}}) < \infty$?*

Another question is whether one can improve the lower bound of $\Omega(\log \log n)$ in Theorem 6 to $\Omega(\log n)$.

▶ **Problem 26.** *Can the lower bound in Theorem 6 be improved to $\mathrm{VC}(\overline{\mathbb{H}}) \geq \Omega(\log n)$?*

Forbidding combinatorial patterns

A natural method to prove upper bounds on the VC dimension of a concept class is establishing that it does not contain a specific combinatorial pattern. For example, the construction for Theorem 3 in [1] utilized the fact that the concept class (viewed as a matrix) does not contain the combinatorial patterns $\begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}$ and $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, which are patterns that are in any concept class $\mathbb{H}$ with $\mathrm{VC}(\mathbb{H}) \geq 2$. Similarly, the dual construction in Theorem 6 forbids the pattern $\begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}$, a compulsory pattern for any concept class $\mathbb{H}$ with $\mathrm{LD}(\mathbb{H}) \geq 3$.

▶ **Problem 27.** *Suppose $\mathbb{H} \subseteq \{0, 1, \star\}^{[n]}$ does not contain the pattern $\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$. Does every disambiguation $\overline{\mathbb{H}}$ of $\mathbb{H}$ satisfy $\mathrm{VC}(\overline{\mathbb{H}}) = O(1)$?*

References

1. Noga Alon, Steve Hanneke, Ron Holzman, and Shay Moran. A theory of PAC learnability of partial concept classes. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 658–671. IEEE, 2022.
2. Kaspars Balodis, Shalev Ben-David, Mika Göös, Siddhartha Jain, and Robin Kothari. Unambiguous DNFs and Alon-Saks-Seymour. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science – FOCS 2021*, pages 116–124. IEEE Computer Soc., Los Alamitos, CA, [2022] ©2022. doi:10.1109/FOCS52979.2021.00020.
3. N. Bousquet, A. Lagoutte, and S. Thomassé. Clique versus independent set. *European J. Combin.*, 40:73–92, 2014. doi:10.1016/j.ejc.2014.02.003.
4. Mika Göös. Lower bounds for clique vs. independent set. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science—FOCS 2015*, pages 1066–1076. IEEE Computer Soc., Los Alamitos, CA, 2015. doi:10.1109/FOCS.2015.69.
5. Mika Göös, Shachar Lovett, Raghu Meka, Thomas Watson, and David Zuckerman. Rectangles are nonnegative juntas. *SIAM J. Comput.*, 45(5):1835–1869, 2016.
6. Hao Huang and Benny Sudakov. A counterexample to the Alon-Saks-Seymour conjecture and related problems. *Combinatorica*, 32(2):205–219, 2012.

7 Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2(4):285–318, 1988. doi:10.1023/a:1022869011914.

8 Jiří Matoušek, editor. *Lectures on Discrete Geometry*. Springer New York, 2002. doi:10.1007/978-1-4613-0039-7.

9 Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65 6:386–408, 1958.

10 Norbert Sauer. On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13(1):145–147, 1972.

11 Mihalis Yannakakis. Expressing combinatorial optimization problems by linear programs. *J. Comput. System Sci.*, 43(3):441–466, 1991. doi:10.1016/0022-0000(91)90024-Y.