TRANSACTIONS OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 00, Number 0, Pages 000–000
S 0002-9947(XX)0000-0

MODEL THEORY AND AGNOSTIC ONLINE LEARNING VIA EXCELLENT SETS

M. MALLIARIS AND S. MORAN

Abstract. We use algorithmic methods from online learning to explore some important objects at the intersection of model theory and combinatorics, and find natural ways that algorithmic methods can detect and explain (and improve our understanding of) stable structure in the sense of model theory. The main theorem deals with existence of ϵ-excellent sets (which are key to the Stable Regularity Lemma, a theorem characterizing the appearance of irregular pairs in Szemerédi's celebrated Regularity Lemma). We prove that ϵ-excellent sets exist for any $\epsilon < \frac{1}{2}$ in k-edge stable graphs in the sense of model theory (equivalently, Littlestone classes); earlier proofs had given this only for $\epsilon < 1/2^{2^k}$ or so. We give two proofs: the first uses regret bounds from online learning, the second uses Boolean closure properties of Littlestone classes and sampling. We also give a version of the dynamic Sauer-Shelah-Perles lemma appropriate to this setting, related to definability of types. We conclude by characterizing stable/Littlestone classes as those supporting a certain abstract notion of majority: the proof shows that the two distinct, natural notions of majority, arising from measure and from dimension, densely often coincide.

In the recent papers [4], [7], ideas from model theory played a role in the conjecture, and then the proof, that Littlestone classes (which model theorists would call stable) are precisely those which can be PAC learned in a differentially private way. We direct the reader to those papers for precise statements and further literature review. The present work may be seen as complementary to that work in that it shows, perhaps even more surprisingly, that ideas and techniques can travel profitably in the other direction.

In the introduction below, we briefly present three points of view (combinatorics, online learning, model theory) which inform this work. The aim is to allow the paper to be readable by people in all three communities. Before this, we explain the results, deferring some definitions to the introduction below.

The technical contributions of the paper are as follows:

MM's research partially supported by NSF CAREER 1553653 and NSF-BSF 2051825.

SM is Robert J. Shillman Fellow and support by ISF grant 1225/20, by BSF grant 2018385, by an Azrieli Faculty Fellowship, by Israel PBC-VATAT, by the Technion Center for Machine Learning and Intelligent Systems (MLIS), and by the the European Union (ERC, GENERALIZATION, 101039692). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

(1) The main theorem: Theorem 3.1, page 9 below. Informally, stable classes $\equiv$ Littlestone classes have large ϵ-excellent sets for *any* $\epsilon < \frac{1}{2}$ (and this is a characterization).

The technical significance of this result is that it extends a useful proof from [24] which required $\epsilon < \frac{1}{2^d}$ to essentially any ϵ for which "ϵ-excellent" is well defined. It is nice to have this answer, but the longer-term effect of this result, in the opinion of the authors, may come instead from the methods of proofs. In fact, we give two distinct proofs. The first proof, in §3, uses *no-regret* and *multiplicative-weights algorithms* from machine learning. These methods have a long history in computer science. We found it surprising that these applied rather naturally to extract model theoretic information and we wonder what else may be done with this. The second proof, in §4, uses the VC theorem and closure properties of Littlestone (stable) classes under Boolean operations. There are some interesting quantitative questions here.

(2) Theorem 6.11, page 20 below. Informally, this gives a new characterization of stable/Littlestone classes as those admitting a certain axiomatic notion of majority.

This theorem (or perhaps its definition) resolves an apparent discrepancy between two notions of majority arising in this world: the dimensional majority used in rank or Littlestone dimension, and the counting majority used in goodness and excellence. It shows that the two can always be made to coincide in a precise way. Earlier analogues of this discrepancy could be seen, for instance, in proofs of stable regularity which tended to choose either one or the other (cf. [24], [25] on the one hand and [23] on the other) according to whether the focus was on bounds or definability.

(3) Theorem 5.1, page 15 below. Informally, we can detect whether or not a class is stable (Littlestone) by counting the number of *algorithms* needed to simulate it on an unknown tree of fixed height.

The statement is a mild extension of a known result in machine learning for Littlestone classes (removing the assumption of "oblivious," i.e. replacing sequences with trees). The known result is what explains the contribution of finite Littlestone dimension in §3, so some exposition of this was necessary in order to make that proof clear to the model-theoretic reader; we took the occasion to prove the extension and to make clear the connection with definability of types.

Although it is technically the simplest contribution, we feel that the discovery of a context in which the correct analogue of *types* is *algorithms* is significant.

We thank the anonymous referee for many thoughtful and helpful suggestions.

1. Introduction

In this section we give a high-level overview from three different, though interconnected, points of view, and informally describe the work. We defer some formal definitions to later sections.

1.1. **Context from combinatorics.** Szemerédi's celebrated Regularity Lemma for finite graphs says essentially that any huge finite graph G can be well approximated by a much smaller random graph. The lemma gives a partition of any such G into pieces of essentially equal size so that edges are distributed uniformly between most pairs of pieces (that is, most pairs are ϵ-regular for some $\epsilon > 0$ given in advance[1]). Szemerédi's original proof allowed for some pairs to be irregular, and he asked if this was necessary [33]. As described in [17, §1.8], it was observed by several researchers including Alon, Duke, Leffman, Rödl and Yuster [3] and Lovász, Seymour, Trotter that irregular pairs are unavoidable due to the counterexample of half-graphs. A k-half graph has distinct vertices $a_1, \ldots, a_k$, $b_1, \ldots, b_k$ such that there is an edge between (a_i, b_j) if and only if $i < j$.

Malliaris and Shelah showed that half-graphs characterize the existence of irregular pairs in Szemerédi's lemma, by proving a stronger regularity lemma for k-edge stable graphs called the Stable Regularity Lemma [24]. (A graph is called k-edge stable if it contains no k-half graph. This should remind a model theoretic reader of the negation of the order property. The Stable Regularity Lemma says that a finite k-edge stable graph can be equipartitioned into $\leq m$ pieces (where m is polynomial in $\frac{1}{\epsilon}$) such that *all* pairs of pieces are regular, with densities close to 0 or 1. Two of these conditions, the improved size of the partition and the densities of regular pairs being near 0 or 1, are already expected from finite VC dimension, see [3], [22], though here by a different proof. (See also section 1.3 for more on VC dimension, which has also a venerable history in model theory. After stable regularity, there began an active line of work looking via model theory at these questions; the introduction to [35] surveys this literature.) For an exposition of stable regularity, and the model theoretic ideas behind it, see [25].

A central idea in the stable regularity lemma was that k-edge stability for small k means it is possible to find large "indivisible" sets, so-called ϵ-*excellent* sets.

To informally recall the definition (formal definitions will be given in 2.2 below), let $0 < \epsilon < \frac{1}{2}$. Let $G = (V, E)$ be a finite graph. Following [24], say $B \subseteq V$ is ϵ-*good* if for any $a \in V$, one of $\{b \in B : (a, b) \in E\}$, $\{b \in B : (a, b) \notin E\}$ has size $< \epsilon|B|$. If the first [most $b \in B$ connect to a], write $\mathbf{t}(a, B) = 1$, and if the second [most $b \in B$ do not connect to a] write $\mathbf{t}(a, B) = 0$. Say that $A \subseteq V$ is ϵ-*excellent* if for any $B \subseteq V$ which is ϵ-good, one of $\{a \in A : \mathbf{t}(a, B) = 1\}$, $\{a \in A : \mathbf{t}(a, B) = 0\}$ has size $< \epsilon|A|$. Informally, any $a \in A$ has a majority opinion about any ϵ-good B by definition of good, and excellence says that additionally, a majority of elements of A have the same majority opinion. Observe that if A is ϵ-excellent it is ϵ-good, because any set of size one is ϵ-good.

A partition into excellent sets is quickly seen to have no irregular pairs (for a related ϵ).

Notice that while, e.g., $\frac{1}{4}$-good implies $\frac{1}{3}$-good, the same is a priori not true for ϵ-excellent, because the definition of ϵ-excellence quantifies over ϵ-good sets. See Example 2.6 below. For the stable regularity lemma, it was sufficient to show that large ϵ-excellent sets exist in k-edge stable graphs for $\epsilon < \frac{1}{2^{2^k}}$ or so. In this language, one contribution of the present paper is a new proof for existence of ϵ-excellent sets

[1]When A, B are finite sets of vertices, let $e(A, B)$ denote the number of edges between A and B, and let $d(A, B) = e(A, B)/|A||B|$ denote the density. Recall that (A, B) is called ϵ-regular if for all $A' \subseteq A$ with $|A'| \geq \epsilon|A|$, and all $B' \subseteq B$ with $|B'| \geq \epsilon|B|$, we have $|d(A, B) - d(A', B')| < \epsilon$.

in k-edge stable graphs, which works for any $\epsilon < \frac{1}{2}$, i.e. any ϵ for which excellence is well defined.

1.2. **Context from online learning.** Online learning is a well-studied model for algorithms making real-time predictions on sequentially arriving data. In the basic version of online learning, a game between a learner and an adversary is played sequentially for T rounds as follows. At each stage $t = 1, \ldots, T$ the adversary presents the learner with a point $x_t \in X$ of the adversary's choosing and asks the learner to predict the label $y_t \in \{0, 1\}$. Then, the learner makes a prediction $\hat{y}_t$ after which the correct label y_t is revealed to the learner.[2] The learner receives a penalty of 1 for a mistake. The adversary's goal is to play so as to maximize the number of mistakes, and the learner's goal is to minimize them. The performance of the learner is measured in terms of its *regret*:

$$\sum_{t=1}^{T} 1[\hat{y}_t \neq y_t] - \min_{h \in \mathcal{H}} \sum_{t=1}^{T} 1[h(x_t) \neq y_t]$$

that is, the excess number of mistakes the learner makes compared to the best function $h \in \mathcal{H}$. Say $\mathcal{H}$ is (agnostic) online learnable if the learner has an algorithm whose regret against any adversary is sublinear in T. An important special case is online learning in the realizable case, in which the adversary is restricted to produce sequences $\{(x_t, y_t)\}_{t=1}^{T}$ that are realizable by (or consistent with) the class $\mathcal{H}$: that is, for which there exists $h \in \mathcal{H}$ such that $h(x_t) = y_t$ for all $t = 1, \ldots, T$. Notice that in the realizable case, the regret of the learner specializes to the absolute number of mistakes the learner made on the input sequence.

The online learning setting shifts the basic context from graphs to *hypothesis classes*, i.e. pairs $(X, \mathcal{H})$ where X is a finite or infinite set and $\mathcal{H} \subseteq \mathcal{P}(X)$ is a set of subsets of X, called hypotheses or predictors. We will identify elements $h \in \mathcal{H}$ with their characteristic functions, and write "$h(x) = 1$" for "'$x \in h$" and $h(x) = 0$ otherwise. Any such hypothesis class can be naturally viewed as a bipartite graph on the disjoint sets of vertices X and $\mathcal{H}$ with an edge between $x \in X$ and $h \in \mathcal{H}$ if and only if $h(x) = 1$. However, something which may be lost in this translation is a powerful understanding in the computer science community of the role of dynamic/adaptive/predictive arguments. This perspective is an important contribution to the proofs below, and seems to highlight some understanding currently missing in the other contexts.

Consider the idea of a mistake tree: we have a full binary tree whose internal nodes are labeled by elements of X, and which is realized by $\mathcal{H}$ in the following sense. We can think of the process of traversing a root-to-leaf path, that is, a branch, in a tree of height d as being described by a sequence of pairs $(x_i, y_i) \in X \times \{0, 1\}$, for $1 \leq i \leq d$, recording that at step i the node we are at is labeled by x_i and we then travel right (if $y_i = 1$) or left (if $y_i = 0$) to a node labeled by x_{i+1}, and so on. Say that $h \in \mathcal{H}$ realizes a given branch $(x_1, y_1), \ldots (x_d, y_d)$ if $h(x_i) = y_i$ for all $1 \leq i \leq d$, and say that a given tree is shattered by $\mathcal{H}$ if each branch is realized

[2]Note that $\hat{y}_t = \hat{y}_t(x_1, \ldots, x_t)$ may depend on everything that happened up to round t. When the learner uses randomness to make predictions, the adversary isn't able to see the model's internal randomness. Thus, the learner shares only the probability p_t that the predicted outcome $\hat{y}_t$ will be 1. The actual prediction $\hat{y}_t$ is made randomly, based on this probability, and only after the adversary has chosen y_t.

by some $h \in \mathcal{H}$. Call such a tree a *mistake tree of height d* for $\mathcal{H}$. (For a more extensive discussion see [32] §18.1-18.2.)

The Littlestone dimension d of $\mathcal{H}$, denoted $\mathsf{Ldim}(\mathcal{H})$, is the depth of the largest mistake tree, i.e., the largest complete [binary] tree that is shattered by $\mathcal{H}$, or ∞ if no such bound exists. Notice that existence of such a tree helps the adversary force at least d mistakes in the realizable case. $\mathcal{H}$ is called a Littlestone class if it has finite Littlestone dimension; for reasons explained in the next subsection, we may prefer to say that $(X, \mathcal{H})$ is a Littlestone pair. Littlestone [19] and Ben-David, Pál, and Shalev-Shwartz [5] proved that Ldim characterizes online learnability of the class. Quantitatively, the optimal regret is equal to $\Theta(\sqrt{\mathsf{Ldim} \cdot T})$ ([5, 1]) and in the realizable case the optimal mistake bound is Ldim [19].[3]

1.3. **Context from model theory.** Consider again the case of a finite bipartite graph G with vertex set $X \cup Y$ and edge relation R (abstracting the study of a formula $\varphi(\bar{x}, \bar{y})$). Following logical notation we write $R(a, b)$ or $\neg R(a, b)$ to denote an edge or a non-edge. Define a [full] *special tree* of height n to have internal nodes $\{a_\eta : \eta \in {}^{n>}2\}$ from X and indexed by binary sequences of length $< n$ and leaves $\{b_\rho : \rho \in {}^{n}2\}$ from Y and indexed by binary sequences of length exactly n, which satisfy the following.[4] For any a_η and b_ρ, if η is an initial segment of ρ (notation: $\eta \trianglelefteq \rho$), then $R(a_\eta, b_\rho)$ if $\eta^\frown\langle 1\rangle \trianglelefteq \rho$ and $\neg R(a_\eta, b_\rho)$ if $\eta^\frown\langle 0\rangle \trianglelefteq \rho$. A key ingredient in the proof of the Stable Regularity Lemma was the following special case of Shelah's Unstable Formula Theorem [31, II.2.2]. For a bipartite graph G as above, if G has a full special tree of height n, then it has a half-graph of size about $\log n$, with the a's chosen from X and the b's chosen from Y (i.e., it is not k-edge stable in the sense above). Moreover, if G has a half-graph of size k, it has a full special tree of height about $\log k$. (These bounds are due to Hodges, see [15] or [14] Lemma 6.7.9, and see §7 # 2 below.)

It was noticed by Chase and Freitag [8] that the condition of model-theoretic stability (Shelah's 2-rank; e.g., in this language, no full special tree of height n for some finite n) corresponds to finite Littlestone dimension, and they used this to give natural examples of Littlestone classes using stable theories. An analogous connection between so-called *NIP theories* and VC dimension had also been previously observed by Laskowski [18] and led to results in learning theory, particularly in the context of compression schemes, see for instance Livni and Simon [21], and on related definability questions, see for instance Eshel and Kaplan [10], but also some of the first polynomial bounds for VC dimension for sigmoidal neural networks, see Karpinski and Macintyre [16].

The following discussion reflects an understanding developed in the online learning papers Alon-Livni-Malliaris-Moran [4], Bun-Livni-Moran [7], where model theoretic ideas had played a role in the proof that the Littlestone classes are precisely those which can be PAC-learned in a differentially private way. One contribution of the model theoretic point of view for online learning in general, and for our present argument in particular, is that a condition in online learning which appears inherently asymmetric, namely the Littlestone dimension (it treats elements and hypotheses as different kinds of objects; they play fairly different roles in the

[3] More precisely, Ldim is the optimal bound achievable in the realizable setting by *deterministic* learners. The optimal bound achievable by randomized learner is characterized by the randomized Littlestone dimension [11].

[4] On this notation, see Convention 2.13.

partitioning) is equivalent to a condition which is extremely symmetric, namely existence of half-graphs (when switching the roles of X and Y in a half-graph, it suffices to rotate the picture). Thus if $\mathcal{H}$ is a Littlestone class, the "dual" class obtained by setting $X' = \mathcal{H}$ and $\mathcal{H}' = \{\{h \in \mathcal{H} : h(x) = 1\} : x \in X\}$ is also. In online learning, k-edge stability also has a natural meaning: Threshold dimension k, that is, there do not exist elements $a_1, \ldots, a_k$ from X and hypotheses $h_1, \ldots h_k$ from $\mathcal{H}$ such that $h_j(a_i) = 1$ if and only if $i < j$. In what follows, we sometimes refer to $(X, \mathcal{H})$ as a *Littlestone pair*, rather than simply saying that $\mathcal{H}$ is a Littlestone class, to emphasize this line of thought.

2. Prior results and a characterization

Convention 2.1. *Following convention, we say "$(X, \mathcal{H})$ is a hypothesis class" to mean that X is a set and $\mathcal{H}$ is a set of subsets of X (sometimes identified with their characteristic functions).*

In the language of online learning, $(X, \mathcal{H})$ is a Littlestone class if it has finite Littlestone dimension (Ldim) in the sense of [32] Chapter 21. A combinatorial or model theoretic reader may take the Littlestone dimension simply to be the maximal height of a special tree (§1.3 above), an instance of Shelah's 2-rank, see [8] or [26].

The original facts about good and excellent sets in [24] were proved for k-edge stable graphs (§1.1 above); the translation to Littlestone classes is immediate, but we record this here for completeness.

Definition 2.2. *Let $0 < \epsilon < \frac{1}{2}$ and let $(X, \mathcal{H})$ be a hypothesis class.*

1. *Say $B \subseteq X$ is* ϵ-good *if for any $h \in \mathcal{H}$, one of $\{b \in B : h(b) = 1\}$, $\{b \in B : h(x) = 0\}$ has size $< \epsilon|B|$. Write $\mathbf{t}(h, B) = 1$ in the first case, $\mathbf{t}(h, B) = 0$ in the second.*
2. *Say that $H \subseteq \mathcal{H}$ is* ϵ-excellent *if for any $B \subseteq X$ which is ϵ-good, one of $\{h \in H : \mathbf{t}(h, B) = 1\}$, $\{h \in H : \mathbf{t}(h, B) = 0\}$ has size $< \epsilon|H|$. Write $\mathbf{t}(H, B) = 1$ or $\mathbf{t}(H, B) = 0$ to record this.*
3. *Define "H is an ϵ-good subset of $\mathcal{H}$" and "A is an ϵ-excellent subset of X" in the parallel way switching the roles of X and $\mathcal{H}$.*

Remark 2.3. The definition of ϵ-good is monotonic in ϵ: it becomes weaker as ϵ increases (below $\frac{1}{2}$). This is a priori not the case for excellence: as ϵ increases, the ϵ-good sets B quantified over may increase, as the following illustrates.

Example 2.4. *We first give an example of a set which is ϵ-good but not ϵ-excellent. Let $A = \{a_1, a_2, a_3, a_4, a_5\}$ and $B = \{b_1, b_2, b_3, b_4, b_5\}$ and consider the bipartite graph with vertex set A, B. Let ϵ be slightly larger than $1/5$, say, $\epsilon = 13/60$. Suppose the edges are given as follows:*

- *$a_1 \sim b_1$, and $a_1 \not\sim b_2, b_3, b_4, b_5$, so $\mathbf{t}(a_1, B) = 0$.*
- *$a_2 \sim b_2$, and $a_2 \not\sim b_1, b_3, b_4, b_5$, so $\mathbf{t}(a_2, B) = 0$.*
- *$a_3 \not\sim b_3$, and $a_3 \sim b_1, b_2, b_4, b_5$, so $\mathbf{t}(a_3, B) = 1$.*
- *$a_4 \not\sim b_4$, and $a_4 \sim b_1, b_2, b_3, b_5$, so $\mathbf{t}(a_4, B) = 1$.*
- *$a_5 \not\sim b_5$, and $a_5 \sim b_1, b_2, b_3, b_4$, so $\mathbf{t}(a_5, B) = 1$.*

Then A and B are both ϵ-good but since $(1-\epsilon)|A| > |\{a \in A : \mathbf{t}(a, B) = 1\}| > \epsilon|A|$, i.e. $1 - \epsilon > 2/5 > \epsilon$, A is not ϵ-excellent (nor is B for the parallel reason).

Example 2.5. *Let ϵ_1 be slightly above $1/5$, say $\epsilon_1 = 13/60$, and let ϵ_2 be slightly above $1/3$, say $21/60$. Next we give an example of a set B which $(\star)$ is ϵ_2-good, but no $B' \subseteq B$ with $|B'| > 1$ is ϵ_1-good.*

Let B have vertices $b_0, b_1, b_2, b_3, b_4, b_5$. Suppose that:

- *every vertex in the graph either connects to at most two vertices of B, or connects to all but at most two vertices of B.*
- *for every two distinct vertices of B there is an element of the graph which connects only to them.*

To verify $(\star)$ note that if $B' \subseteq B$ has size 2 or 4 it can be split in half, if it has size 3 or 6 it can be split 1/3, 2/3, and if it has size 5 it can be split 2/5, 3/5. So B' is not ϵ_1-good unless it is a singleton. On the other hand, B is clearly ϵ_2-good.

Example 2.6. *To find an example where excellence changes with ϵ, we combine the two previous examples.*

Let $A = \{a_1, a_2, a_3, a_4, a_5\}$ and let $B = \{b_0, b_1, b_2, b_3, b_4, b_5\}$. Consider the bipartite graph with vertex set $A \cup C$ on one side and B on the other, with edges given according to the following pattern:

- *$a_1 \sim b_0, b_1$, and $a_1 \not\sim b_2, b_3, b_4, b_5$.*
- *$a_2 \sim b_2$, and $a_2 \not\sim b_0, b_1, b_3, b_4, b_5$.*
- *$a_3 \not\sim b_3$, and $a_3 \sim b_0, b_1, b_2, b_4, b_5$.*
- *$a_4 \not\sim b_4$, and $a_4 \sim b_0, b_1, b_2, b_3, b_5$.*
- *$a_5 \not\sim b_5$, and $a_5 \sim b_0, b_1, b_2, b_3, b_4$.*
- *For every pair of distinct elements of B, there is an element of C which connects to them and to no other elements of B.*
- *There are no other edges.*

Let $\epsilon_1 = 13/60$ and $\epsilon_2 = 21/60$. A is ϵ_1-good and thus ϵ_2-good. Also, A is ϵ_1-excellent because the only ϵ_1-good subsets of B are the singletons. However B is ϵ_2-good, and so witnesses that A is not ϵ_2-excellent.

Below we use the possible "gaps" in excellence arising from nonmonotonicity as a motivation to give new proofs and new theorems.[5]

The key point about excellent sets that they exist characteristically in stable (Littlestone) classes:

Fact 2.7 ([24] Claim 5.4 or [25] Claim 1.8, in our language)**.** *Let $(X, \mathcal{H})$ be a Littlestone class, $\mathsf{Ldim}(\mathcal{H}) = d$ and $0 < \epsilon < \frac{1}{2^d}$. For any finite $H \subseteq \mathcal{H}$ there is $A \subseteq H$, $|A| \geq \epsilon^d |H|$ such that A is ϵ-excellent.*

The proof of Fact 2.7 proceeds by noting that if $H = H_\emptyset$ is not ϵ-excellent, then there is some ϵ-good $A = A_\emptyset$ which witnesses this failure, splitting $H_\emptyset$ naturally into $H_{\langle 0 \rangle}$ and $H_{\langle 1 \rangle}$ according to $\mathbf{t}$. If either of these is excellent, we stop; if not, continue inductively to label the internal nodes and leaves of a full binary tree with A's and H's respectively. Suppose we arrive to height d. To extract a Littlestone tree, or equivalently a full special tree (p. 5 above), choose a h_ρ from each H_ρ and then show it is possible to choose a suitable a_η from each A_η by using $\epsilon < 2^{-d}$, the

[5]It also may be interesting to investigate the limits of these gaps. Added in revision: note that in the course of significant recent work on hypergraphs, [35, Corollary 5.10] Terry and Wolf consider the relation of goodness and *regularity*, showing that a pair of ϵ-good sets is essentially $\sqrt{\epsilon}$-regular, which Malliaris and Shelah [24, Claim 5.17] show for ϵ-excellent sets.

definition of good (really, of $\mathbf{t}$) and the union bound: that is, for each given η, each h_ρ with ρ extending η rules out at most an ϵ-fraction of elements of A_η. So by the choice of ϵ, there is a remaining element in A_η acceptable to all such h_ρ.[6]

Since finding such a Littlestone tree contradicts $\mathsf{Ldim}(\mathcal{H}) = d$, it must be that for some ρ of length $< d$, H_ρ is excellent.

Note moreover that the same proof works to show existence of ϵ-good sets, simply by taking the sets A to be singletons (note that any singleton is trivially ϵ-good). In this case, the union bound is not needed and the proof works for any $\epsilon < \frac{1}{2}$.

Fact 2.8 ([24], see above)**.** *Let $(X, \mathcal{H})$ be a Littlestone class, $\mathsf{Ldim}(\mathcal{H}) = d$, $0 < \epsilon < \frac{1}{2}$. For every finite $A \subseteq \mathcal{H}$ there is $B \subseteq A$, $|B| \geq \epsilon^d |A|$ such that B is ϵ-good.*

Definition 2.9. *Let $(X, \mathcal{H})$ be a hypothesis class. Define the* dual *hypothesis class to be $(X', \mathcal{H}')$ where $X' = \mathcal{H}$ and $\mathcal{H}' = \{\{h \in \mathcal{H} : h(x) = 1\} : x \in X\}$.*

Definition 2.9 is perhaps most natural in the language of graphs:

Definition 2.10. *A hypothesis class $(X, \mathcal{H})$ gives rise to a bipartite graph $\mathrm{Bip}(X, \mathcal{H})$ with an edge between x and h if and only if $h(x) = 1$.*

Likewise any bipartite graph gives rise to a hypothesis class after we specify which side is the domain and which side is the hypotheses. The dual class $(X', \mathcal{H}')$ arises from the same bipartite graph $\mathrm{Bip}(X, \mathcal{H})$ but with the opposite specification.

We conclude this section by observing in Claim 2.11 that existence of large good sets (both in X and in $\mathcal{H}$) is characteristic of Littlestone classes. At this point, a similar result for excellence could also be stated, for all sufficiently small ϵ.

Claim 2.11. *The following are equivalent for any hypothesis class $(X, \mathcal{H})$.*

1. *For every $\epsilon < \frac{1}{2}$ there is a constant $c = c(\epsilon) > 0$ such that for every finite $A \subseteq \mathcal{H}$ there exists $B \subseteq A$, $|B| \geq c|A|$ such that B is ϵ-good.*
2. *For some $\epsilon < \frac{1}{2}$ and some constant $c > 0$, for every finite $A \subseteq \mathcal{H}$ there exists $B \subseteq A$, $|B| \geq c|A|$ such that B is ϵ-good.*
3. *$\mathcal{H}$ is a Littlestone class.*
4. *For some finite k, $\mathrm{Bip}(X, \mathcal{H})$ does not contain any k-half graph as an induced subgraph.*[7]
5. *For some finite ℓ, $\mathrm{Bip}(X', \mathcal{H}')$ does not contain any ℓ-half graph as an induced subgraph.*
6. *The dual class $(X', \mathcal{H}')$ (in the sense of* 2.9*) is a Littlestone class.*
7. *For every $\epsilon < \frac{1}{2}$ there is a constant $c = c(\epsilon) > 0$ such that for every finite $A \subseteq X$ there exists $B \subseteq A$, $|B| \geq c|A|$ such that B is ϵ-good.*
8. *For some $\epsilon < \frac{1}{2}$ and some constant $c > 0$, for every finite $A \subseteq X$ there exists $B \subseteq A$, $|B| \geq c|A|$ such that B is ϵ-good.*

Proof. (1) implies (2) is immediate, and (3) implies (1) is Fact 2.8.

To show (2) implies (3), suppose we are given ϵ and c from (2). Since any finite hypothesis class is necessarily Littlestone, we may assume X is infinite. Choose n large enough so that $\lfloor \epsilon \lfloor cn \rfloor \rfloor \geq 1$ and so that for any $k \geq \lceil \epsilon n \rceil$, we have $\min\{\lfloor \frac{k}{2} \rfloor - 1, \lfloor \frac{k-1}{2} \rfloor\} \geq \epsilon k$. If $\mathcal{H}$ is not a Littlestone class, then we know it has infinite (i.e. not finite) Threshold dimension and so for our chosen n, there are elements $\{x_i : i < n\}$ from X and $H := \{h_j : j < n\}$ from $\mathcal{H}$ such that $x_i \in h_j$ if and only if $i < j$.

[6] Observe that this appeal to the union bound won't work for Example 2.6 above.

[7] Such graphs are called k-stable graphs.

But for any $H' \subseteq H$ of size $k \geq cn$, we can pick out all the cuts of H' using the x_i's. In particular, if $H' = \{h_{i_\ell} : \ell < k\}$, let $m = \lceil \epsilon k \rceil$ (by choice of n, this is not larger than whichever of $k/2$ or $(k-1)/2$ is an integer). Then x_m partitions H' into $\{h_{i_\ell} : 0 \leq \ell < m\}$ and $\{h_{i_\ell} : m \leq \ell < k\}$, both of which have size $\geq \epsilon k$, contradicting its being ϵ-good.

The equivalence of (3) and (4) is by Shelah's unstable formula theorem, as explained in §1.3 above.

The equivalence of (4) and (5) is simply because a bipartite graph (A, B) contains an induced k-half graph if and only if the bipartite graph (B, A) does.[8]

The remainder of the proof then goes by a parallel argument. □

Remark 2.12. *Attentive readers may wonder about in the bounds in* 2.11 *relating mistake trees and half-graphs. The known finite bounds, tracing back to Hodges* [15]*, can be stated in combinatorial language as: from a tree of height d one obtains a half-graph of size about* $\log d$*, and from a half-graph of length k one obtains a tree of height about* $\log k$*. This is the subject of open problem* 2 *in* §7 *below.*

Convention 2.13. *We clarify some notational points which hopefully will not cause confusion if explicitly pointed out. The word "label" in online learning, and in this paper, usually refers to a value such as* 0 *or* 1 *attached to an element of X (say, the value of some partial characteristic function). Nonetheless, we write "X-labeled tree" to mean a tree in which we associate to each node an element of X. In set theoretic notation, for each integer n, $n = \{0, \ldots, n-1\}$. Also, ${}^x y$ denotes the set of functions from x to y, as distinguished from y^x which is the* size *of the set of functions from x to y. For logicians, a tree of height T has levels* 0 *to $T-1$, whereas in online learning the same tree would have levels* 1 *to T.*

3. Existence via regret bounds

The aim of this section is to prove the following theorem, using regret bounds in online learning. (As noted, this improves [24], Claim 5.4 by allowing for $\epsilon < \frac{1}{2}$ rather than $\epsilon < \frac{1}{2^d}$, however, when $\epsilon < \frac{1}{2^d}$ that proof obtains ϵ^d as the constant c, which is not obtained by the methods here.)

Theorem 3.1. *For every $\epsilon < \frac{1}{2}$ and positive integer d there is a constant $c = c(d, \epsilon)$ such that if $\mathsf{Ldim}(\mathcal{H}) = d$ and $H \subseteq \mathcal{H}$ is finite, then there is $A \subseteq H$, $|A| \geq \epsilon^c |H|$ such that A is ϵ-excellent. Moreover c is upper bounded by d_ϵ from* 3.8 *below.*

The key ingredient in the proof is Theorem 3.7. To begin we re-present the definitions of good and excellent in the language of probability.

Discussion 3.2. *Although to our knowledge new,* 3.3 *is a natural opening move in our context. For instance, it allows for an extension of the existing model theoretic definitions into the randomized online learning setting where learner and adversary are possibly playing distributions, as explained* §1.2 *above.*

In order to reason about distributions we need to define an appropriate probability space. For the sake of simplicity in the present paper we focus on the case when both

[8]Thus to repeat the point made in §1.3, in connecting the property of Littlestone, which is not obviously self-dual, with half-graphs which are, the Unstable Formula Theorem allows us to immediately conclude $\mathcal{H}$ is Littlestone if and only if its dual is, though of course with possibly different Littlestone dimensions.

$\mathcal{X}$ and $\mathcal{H}$ are finite or countable. This allows us to use the trivial sigma algebra which consists of the entire power set, and hence avoid stating (standard) measure theoretic assumptions. When applying these results in the uncountable setting the reader is cautioned to also verify the (standard) measure theoretic assumptions inherited from the quoted results on bounds.

Definition 3.3 (ϵ-Good and ϵ-Excellent Distributions)**.** *Let $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$.*

(1) *We say a distribution P over $\mathcal{X}$ is ϵ-good w.r.t $\mathcal{H}$ if*

$$(\forall h \in \mathcal{H}) : \Pr_{x \sim P}[h(x) = 1] \in [0, \epsilon] \cup [1 - \epsilon, 1].$$

(2) *Similarly, a distribution Q over $\mathcal{H}$ is ϵ-good if*

$$(\forall x \in \mathcal{X}) : \Pr_{h \sim Q}[h(x) = 1] \in [0, \epsilon] \cup [1 - \epsilon, 1].$$

(3) *Next, a distribution P over $\mathcal{X}$ is ϵ-excellent if*

$$(\forall\ \epsilon\text{-good } Q) : \Pr_{h \sim Q, x \sim P}[h(x) = 1] \in [0, \epsilon] \cup [1 - \epsilon, 1].$$

(4) *Finally, a distribution Q over $\mathcal{H}$ is ϵ-excellent if*

$$(\forall\ \epsilon\text{-good } P) : \Pr_{h \sim P, x \sim Q}[h(x) = 1] \in [0, \epsilon] \cup [1 - \epsilon, 1].$$

Distributions which trivially satisfy 3.3(1),(2) by concentrating on a single point exist in any hypothesis class. In Littlestone classes, one source of nontrivial examples comes from choosing some finite ϵ-good set A and taking a distribution which assigns measure 0 to the complement of A and is uniform on A. So indeed 3.3 naturally extends the usual notion of excellent and good:

Convention 3.4. *We say that a subset of $\mathcal{H}$ or of $\mathcal{X}$ is ϵ-good (ϵ-excellent) if the uniform distribution on it is ϵ-good (ϵ-excellent).*

Convention 3.5. *In the process of extracting large ϵ-excellent subsets of $\mathcal{H}$ we use full binary trees whose nodes are labelled by ϵ-good distributions; let us refer here to such trees as* ϵ-good trees.

Definition 3.6. *Observe that each hypothesis $h \in \mathcal{H}$ naturally realizes a branch in an ϵ-good tree $\mathcal{T}$. An ϵ-good tree $\mathcal{T}$ is said to be <u>shattered</u> by $\mathcal{H}$ if every branch is realized by some $h \in \mathcal{H}$.*

We now state the key technical result of the section.

Theorem 3.7. *Let $\mathcal{H}$ be a hypothesis class, let $\mathcal{T}$ be an ϵ-good complete binary tree that is shattered by $\mathcal{H}$, and let T denote the depth of $\mathcal{T}$. Then, for every online learning algorithm $\mathcal{A}$, the tree $\mathcal{T}$ witnesses a lower bound on the regret of $\mathcal{A}$ (relative to the set of experts $\mathcal{H}$) in the following sense. There exist distributions $\mathcal{D}_1, \ldots, \mathcal{D}_T$ over $X \times \{0,1\}$ such that an independent sequence of random examples $(x_t, y_t) \sim \mathcal{D}_t, t = 1, \ldots, T$ satisfies the following:*

- *The expected number of mistakes $\mathcal{A}$ makes on the random sequence is at least $\frac{T}{2}$.*
- *$\exists h \in \mathcal{H}$ whose expected number of mistakes on the random sequence is at most $\epsilon \cdot T$.*

Thus, the expected regret of $\mathcal{A}$ w.r.t $\mathcal{H}$ on the random sequence is at least $(\frac{1}{2} - \epsilon) \cdot T$.

Before we prove this theorem, let us demonstrate how one can use it to bound the maximum depth of an ϵ-good tree which is shattered by a Littlestone class $\mathcal{H}$. A central line of work in the subject has established that that every class $\mathcal{H}$ and for every $T \in \mathbb{N}$ there exists an algorithm $\mathcal{A}$ whose expected[9] regret w.r.t any sequence of examples $(x_1, y_1), \ldots, (x_T, y_T)$ is

$$O\big(\sqrt{d \cdot T}\big), \tag{3.1}$$

where d is the Littlestone dimension of $\mathcal{H}$, and the big oh notation conceals a fixed numerical constant.[10] Thus, by Theorem 3.7, it follows that if there exists a (complete) ϵ-good tree that is shattered by $\mathcal{H}$ of depth T then T must satisfy the following inequality:

$$\Big(\frac{1}{2} - \epsilon\Big) \cdot T \leq O\Big(\sqrt{d \cdot T}\Big).$$

Indeed, the LHS in the above inequality is a lower bound on the expected regret of $\mathcal{A}$, where as the RHS is an upper bound on it. A simple arithmetic manipulation yields that $T = O(d/(1/2 - \epsilon)^2)$. Thus, we get the following corollary:

Corollary 3.8. *Let $\mathcal{H}$ be a class with Littlestone dimension $d < \infty$ and let $\epsilon \in \big[0, \frac{1}{2}\big]$. Denote by d_ϵ the maximum possible depth of a complete ϵ-good tree which is shattered by $\mathcal{H}$. (Note that $d_0 = d$.) Then,*

$$d_\epsilon = O\Big(\frac{d}{\big(\frac{1}{2} - \epsilon\big)^2}\Big).$$

Proof of Theorem 3.7. Let $\mathcal{T}$ be a tree and $\mathcal{A}$ be an online algorithm as in the premise of the theorem. We begin with defining the distributions $\mathcal{D}_t$. We first note that the label in each distribution $\mathcal{D}_t$ is deterministic; that is, there exist a distribution D_t over X and a label $y_t \in \{0, 1\}$ such that a random example $(x, y) \sim \mathcal{D}_t$ satisfies that $y = y_t$ always (with probability $= 1$) and $x_t \sim D_t$. The distributions D_i and labels y_i correspond to a branch of $\mathcal{T}$ as follows:

- Initialize $t = 1$, set the "current" node v_t to be the root of the tree.
- For $t = 1, \ldots, T$
  (1) Let D_t denote the ϵ-good distribution D_{v_t} which is associated with v_t.
  (2) Define the label y_t to be 1 if and only if

$$\Pr_{(x_i)_{i=1}^t \sim \prod_{i=1}^t D_i, \mathcal{A}}\Big[\mathcal{A}\big(x_t; (x_{t-1}, y_{t-1}), \ldots, (x_1, y_1)\big) = 1\Big] \leq 1/2,$$

  where $\mathcal{A}$ is the given online algorithm. (Note that the above probability is taken w.r.t the sampling of the x_i's, as well as the randomness of $\mathcal{A}$ in case it is a randomized algorithm.) I.e. the adversary forces that the algorithm errs with probability at least $1/2$ on x_t when given an input sequence $(x_1, y_1), \ldots, (x_{t-1}, y_t), x_t$, where the x_i's are sampled from the D_i's.
  (3) Set v_{t+1} to be the root of the subtree corresponding to the label y_t.
- Output the sequence $(D_1, y_1), \ldots, (D_T, y_T)$.

[9]The algorithm $\mathcal{A}$ is randomized.

[10]The derivation of the (optimal) bound of $O(\sqrt{d \cdot T})$ is somewhat involved [1], however a slightly weaker bound of $O(\sqrt{d \cdot T \log T})$ can be proven using elementary arguments [5].

Let $(x_t)_{t=1}^T \sim \prod_{t=1}^T D_t$ and fix $t \leq T$. Let $\hat{y}_t = \mathcal{A}(x_t; (x_{t-1}, y_{t-1}), \ldots, (x_1, y_1))$ be the prediction of $\mathcal{A}$ on x_t. Thus, by construction $\hat{y}_t \neq y_t$ with probability at least $1/2$, and therefore, by linearity of expectation:

$$\mathop{\mathbb{E}}_{(x_t)_{t=1}^T \sim \prod_{t=1}^T D_t, \mathcal{A}} \Big[\sum_{t=1}^{T} 1[y_t \neq \hat{y}_t] \Big] \geq \frac{T}{2}.$$

It remains to show that there exists $h \in \mathcal{H}$ whose expected number of mistakes is at most $\epsilon \cdot T$. This follows by considering an hypothesis $h \in \mathcal{H}$ which realizes the branch corresponding to $(D_t, y_t)_{t=1}^T$. Indeed, for each fixed distribution D_t on the branch $\mathcal{B}$, the probability that h errs on $x_t \sim D_t$ is at most ϵ. Thus, by linearity of expectation, the expected number of mistakes is at most $\epsilon \cdot T$, and so there exists h as stated. □

First proof of Theorem 3.1. This is immediate from Theorem 3.7, that is, just as in the earlier proof of excellence, such a bound means that the set of elements in least one leaf cannot be split in a balanced way by any ϵ-good set, so must be ϵ-excellent.[11] □

Discussion 3.9. *To summarize and explain the use of Littlestone dimension hidden in this argument we emphasize that the above proof relies on deep ideas from online learning which are worth highlighting. We also emphasize that this discussion surveys much prior work and not only what we do here.*

There are two background ideas: the first one has to do with no-regret algorithms (such as the multiplicative-weights algorithm, see e.g. [20]*). Consider a set of $m < \infty$ experts (say weather forecasters), and every evening each of them tells us whether they think it will rain tomorrow or not. Then, we use this list of predictions to make a prediction of our own. Can we come up with a strategy that will guarantee that over T days our prediction will not be much worse than that of the best expert in hindsight? No-regret algorithms address this problem and provide regret bounds of roughly $\sqrt{T \cdot \log m}$, i.e. such algorithms make at most roughly $\sqrt{T \cdot \log m}$ mistakes more than the best of the m experts in hindsight. We stress that the m experts can be arbitrary algorithms; in particular their prediction at day t may be based on all information up to that point (i.e. prefix-dependence).*

The second background idea, which is how the Littlestone dimension arises in the derivation of equation (3.1)*, is that any (possibly infinite) Littlestone class $\mathcal{H}$ can be covered by a finite set of experts: that is, for every $T < \infty$ there is a set of T choose $\leq \mathsf{Ldim}(\mathcal{H})$ dynamic sets which simulate all hypotheses on $\mathcal{H}$ with respect to* sequences[12] *of length T. Thus, by applying no-regret algorithms on the (finite!) set of experts we may ensure our regret is small relative to the best $h \in \mathcal{H}$.*

Together these explain existence of no-regret algorithms for any Littlestone class. Our bound on d_ϵ, the height of an ϵ-good complete shattered tree, exploits these connections in the new setting of ϵ-good trees.

[11] In fact, since our quantification is over ϵ-good distributions and not just ϵ-good sets, it is excellent in an a priori stronger sense than in the earlier proof. We have not yet investigated how much stronger; as noted in 3.2, the move to distributions has other conceptual advantages.

[12] Below, we extend this to *trees* of height T.

4. Existence via closure properties

In this section we give a second, quite different proof of Theorem 3.1. The resulting bound is significantly weaker than the one stated in Corollary 3.8, but the reasoning may perhaps be more intuitive. In particular, it does not rely on the notion of regret from online learning. The key ideas are the VC theorem and that classes remain Littlestone even after augmenting by fixed Boolean functions.

Remark 4.1. *Continuing Discussion 3.2, note that these results can continue to make sense in the case where $\mathcal{X}, \mathcal{H}$ are uncountable but in that case inheriting the assumptions required to apply the VC theorem.*

Remark 4.2. *The reader may choose to read this section either before or after §5. In the present order, Fact 4.5 may be taken as a black box on a first reading, since the new proof of that fact we give below uses the results of §5. Still, the existence of a combinatorial companion proof of Theorem 3.1 may be best appreciated in parallel to the proof just given, and the use of §5 in Fact 4.5 may be a good motivation for those results for readers not familiar with definability of types.*

Convention 4.3. *In the rest of this section, let $\epsilon < \frac{1}{2}$ be arbitrary but fixed.*

Definition 4.4. *Suppose we are given $k \in \mathbb{N}$ and some Boolean function $B : \{0,1\}^k \to \{0,1\}$. Given $(X, \mathcal{H})$, let $(X, \mathcal{H}^{(B)})$ denote the class*

$$\mathcal{H}^{(B)} = \Big\{ B(h_1, \ldots, h_k) : h_i \in \mathcal{H} \Big\},$$

where $B(h_1, \ldots, h_k)$ denotes the function which takes $x \in X$ to $B(h_1(x), \ldots h_k(x)) \in \{0,1\}$.

Informally, we enrich $\mathcal{H}$ by adding some additional hypotheses which come from applying our fixed B to k-tuples of elements of $\mathcal{H}$. (For example, we could start with $\mathcal{H}$ which is a set of subsets of X, and move to consider the hypothesis class whose elements are intersections of pairs of elements of $\mathcal{H}$; since we can recover any element of $\mathcal{H}$ as its intersection with itself, this is a kind of enrichment of $\mathcal{H}$.) Observe that if the Boolean function sends the constant-1 sequence to 1 and the constant-0 sequence to 0, then $\mathcal{H}^{(B)} \supseteq \mathcal{H}$.

We stress that although B can be arbitrary, it is fixed for any instance of this construction.

We will need the following fact. It was proven by [12] (improving upon a previous bound by [2]), but we sketch below a new proof using our techniques.

Fact 4.5 ([12], Proposition 3)**.** *If $\mathcal{H}$ is a Littlestone class (i.e. $\mathsf{Ldim}(X, \mathcal{H}) < \infty$) then also $\mathcal{H}^{(B)}$ is a Littlestone class, and*

$$\mathsf{Ldim}\big(X, \mathcal{H}^{(B)}\big) = O\Big(\mathsf{Ldim}(X, \mathcal{H}) \cdot k \cdot \log k\Big),$$

where the big oh notation conceals a universal numerical constant.

Model theorists may check their intuition against the assertion that if φ is stable and ψ is a fixed finite Boolean combination of instances of φ, then ψ is also stable.

Proof Sketch. Let us sketch a proof of Fact 4.5 using the language of §5 below (a derivation using dynamic sets has not appeared in the literature, and is intuitive and analogous to the corresponding fact for VC classes). If $\mathsf{Ldim}(X, \mathcal{H}) = d$ then for any integer T we have a set $\mathcal{E}_T$ of $\binom{T}{\leq d}$ dynamic sets which simulate $\mathcal{H}$ on any

X-labeled binary tree of height T. To see that $\mathcal{H}^{(B)}$ is also a Littlestone class it would suffice to show that the same is true for some d' replacing d. For each T, and for each k-tuple of dynamic sets $E_1, \ldots, E_k$ from $\mathcal{E}_T$, let $B(E_1, \ldots, E_k)$ denote the dynamic set which operates by applying B to the outputs of $E_1, \ldots, E_k$. Let $\mathcal{E}_T(B) = \{B(E_1, \ldots, E_k) : E_1, \ldots, E_k \in \mathcal{E}_T\}$. Observe that this collection of dynamic sets simulates $\mathcal{H}^{(B)}$ on any X-labeled binary tree of height T and its size will remain polynomial in T (at most roughly T^{dk}). On the other hand, had $\mathcal{H}^{(B)}$ not been Littlestone then one would need $2^T >> T^{dk}$ dynamic sets to cover it. □

With Fact 4.5 in hand, there is one more step: note we may also apply B dually to X rather than $\mathcal{H}$. To make sense of this, consider $(X, \mathcal{H})$ as a bipartite graph with an edge between $x \in X$ and $h \in \mathcal{H}$ if $h(x) = 1$. In this picture, $\mathcal{H}^{(B)}$ added some new points to the side of $\mathcal{H}$ and defined a rule for putting an edge between any such new point and any given element of X. To apply B dually, we carry out the parallel operation for X instead. That is, let $(X^{(B)}, \mathcal{H})$ be the class where X is enriched by new elements as follows: for any $x_1, \ldots, x_k \in X$ define an element $B(x_1, \ldots, x_k)$ and for any $h \in \mathcal{H}$, define $h(B(x_1, \ldots, x_k)) = 1$ if and only if $B(h(x_1), \ldots, h(x_k)) = 1$.

Recalling 2.11, the dual of a Littlestone class is a Littlestone class, so $(X^{(B)}, \mathcal{H})$ is Littlestone, though the Ldim may be quite a bit larger.[13]

Conclusion 4.6. *For any $k \in \mathbb{N}$ and any function $B : \{0,1\}^k \to \{0,1\}$, if $(X, \mathcal{H})$ is a Littlestone class, then $(X^{(B)}, \mathcal{H})$ is a Littlestone class too.*

Second proof of Theorem 3.1. Suppose we are given an ϵ-good tree $\mathcal{T}$ which is shattered by $\mathcal{H}$. Choose k large enough: $k = O(\mathsf{VCdim}(X, \mathcal{H})/(\frac{1}{2} - \epsilon)^2)$ will suffice. Let $B : \{0,1\}^k \to \{0,1\}$ be the majority vote operation given by $(x_1, \ldots, x_k) \mapsto 0$ if $\{1 \leq i \leq k : a_i = 0\} \geq \frac{1}{2}k$ and $(x_1, \ldots, x_k) \mapsto 1$ otherwise. Suppose we independently sample k elements $x_1, \ldots, x_k$ from one of the ϵ-good distributions labeling our given tree. Then by our choice of k, the VC theorem tells us that, with positive probability, the trace of each $h \in \mathcal{H}$ on this sample is close enough to its true proportion. Here "close enough" means that the error is less than $\frac{1}{2} - \epsilon$. In particular, with positive probability, for every $h \in \mathcal{H}$ the majority vote $B(h(x_1), \ldots, h(x_k))$ on this sample agrees with the opinion of the ϵ-good distribution on h. We can therefore sample k elements from each of the distributions labeling the nodes of the tree and with positive probability, all samples will be correct in this way.

The crucial point is now that any full binary ϵ-good tree $\mathcal{T}$ which is shattered by $\mathcal{H}$ can be transformed to a (standard) full binary tree $\mathcal{T}'$ of the same height which is shattered by the class $(X^{(B)}, \mathcal{H})$. That is, there exists a choice of k elements for each node in $\mathcal{T}$ such that the corresponding tree $\mathcal{T}'$ whose nodes are labelled by the k-wise majority votes of these elements (i.e. by the appropriate $B(x_1, \ldots, x_k)$) is shattered by $(X^{(B)}, \mathcal{H})$. (To emphasize, in our Boolean-augmented class, these majority votes are represented by actual elements, and that is how the tree becomes a standard tree.) This shows that the length of $\mathcal{T}'$ (and also of $\mathcal{T}$) is bounded by $\mathsf{Ldim}(X^{(B)}, \mathcal{H})$ which is finite by Conclusion 4.6. (Note that this argument implicitly gives an inequality between the approximate and virtual Littlestone dimension, which are defined elsewhere.) □

[13] It is known that $\mathsf{Ldim}(X, \mathcal{H}) \leq 2^{2^{\mathsf{Ldim}(\mathcal{H}, X)}}$ by applying Hodges' bound twice, see §7 # 2.

This argument is closer in spirit to similar arguments in VC theory concerning the variability of the VC dimension under natural operations. The obtained bounds however are much weaker than those of the previous section (at least double-exponentially weaker than the bound in Corollary 3.8).

5. Dynamic Sauer-Shelah-Perles lemmas for Littlestone classes

This section states and proves a mild variant of the celebrated Sauer-Shelah-Perles (SSP) lemma [30] replacing "sequences of length T" by "trees of height T" (informally, the adversary can change the elements we are given in response to our past choices). This should also allow the model-theoretic reader to understand the key use of Littlestone dimension in §3, where the case of sequences was already sufficient.

Let $\mathcal{H}$ be a class with Littlestone dimension $d < \infty$. Two results which could be considered variants of the SSP lemma are known for Littlestone classes: the first one, observed by Bhaskar [6] provides an upper bound of T choose $\leq d$ on the number of leaves in a binary-tree of height T with X-labeled nodes that are reachable by $\mathcal{H}$. The second, *dynamic* version is due to Ben-David, Pal, and Shalev-Shwartz [5]. This lemma is a key ingredient in the characterization of (agnostic) online-learnability by Littlestone dimension; it asserts the existence of T choose $\leq d$ online algorithms (or experts or dynamic-sets) such that for every sequence $x_1, \ldots, x_T$ and for every $h \in \mathcal{H}$ there exists an algorithm among the $\binom{T}{\leq d}$ algorithms which produces the labels $h(x_1), \ldots, h(x_T)$ when given the sequence $x_1, \ldots, x_T$ as input. Again, the version we shall prove is a mild extension of the Ben-David, Pal, Shalev-Shwartz lemma to the case of trees rather than sequences. We shall give the statement, then present the key terms, then give the proof.

Theorem 5.1. *Let $(X, \mathcal{H})$ be a Littlestone class of dimension d. For every $T \in \mathbb{N}$ there exists a collection $\mathbf{A}$ of $\binom{T}{\leq d}$ algorithms (dynamic sets) such that for every binary tree $\mathcal{T}$ of height T with X-labeled internal nodes, every branch in $\mathcal{T}$ which is realized by some $h \in \mathcal{H}$ is also realized by some algorithm from $\mathbf{A}$.*

Remark 5.2. *For simplicity, we define dynamic sets to be deterministic, but it is also reasonable for them to be random. In general, a randomized algorithm is simply a distribution over deterministic algorithms. When a randomized algorithm is a distribution over prefix-dependent deterministic algorithms, see below, then we may say it is prefix-dependent.*

Definition 5.3 (In the language of online learning)**.** *Fix $T \in \mathbb{N}$ and a set X. A dynamic set (or adaptive expert) $\mathcal{A}$ is a function which assigns to each internal node in each X-labeled binary tree of height $\leq T$ a value in $\{0, 1\}$ in a prefix-dependent way.*

To explain, notice that $\mathcal{A}$ naturally defines a walk in any such tree: it starts at the root which is labeled by some $a_\emptyset =: a_0$, it outputs $\mathcal{A}(\langle a_0 \rangle) =: t_0$, then travels left (if its output was 0) or right (if its output was 1) to a node labeled by $a_{\langle t_0 \rangle} =: a_1$, where it outputs $\mathcal{A}(\langle a_0, a_1 \rangle) =: t_1$ and so on. Prefix-dependence means that for any $\ell \leq T$, if in two different trees the sequences of values $a_0, \ldots, a_\ell$ produced in this way are the same, then also the output t_ℓ of $\mathcal{A}$ in both cases is the same.

It should now be clear what it means for an algorithm $\mathcal{A}$ to realize a branch in a tree (the directions it gives instruct us to walk along this root-to-leaf path). Note

that we can think of each $h \in \mathcal{H}$ as a very simple dynamic set in its guise as a characteristic function.

Remark 5.4. *In online learning one distinguishes between adaptive and oblivious experts (or between experts with and without memory): an oblivious expert is simply an $X \to \{0,1\}$ function, whereas an adaptive expert has memory and can change its prediction based on previous observations. The above definition captures adaptive experts. The above definition slightly deviates from the standard definition of adaptive experts. In the standard definition, one usually only considers sequences (or oblivious trees), rather than general trees. Notice that the distinction between oblivious and general trees can be expressed analogously with respect to the adversary: the adversary, who presents the examples to the online learner, can be oblivious – in which case it decided on the sequence of examples in advance, or it can be adaptive – in which case it decides which example to present at time t based on the predictions the online learning algorithm made up to time t. In this language, our version of the dynamic SSP applies also to adaptive adversaries, whereas the previous version was restricted to oblivious adversaries.*

Definition 5.5 (In more set-theoretic language)**.** *Let $T \in \mathbb{N}$, and $\kappa = |X|^T$. Consider the set $\mathcal{E} = \langle e_i : i < \kappa \rangle$ of all T-element sequences of elements of X. A* dynamic set *assigns to each enumeration e_i a function $f_i : T \to \{0,1\}$, and the assignment must be* coherent *in the sense that if $e_i \restriction \beta = e_j \restriction \beta$ then $f_i \restriction \beta = f_j \restriction \beta$.*

Example 5.6. *Let $X = \mathbb{N}$. Let $\mathcal{A}$ be the algorithm which receives a_t at time t and outputs 1 if a_t is the largest prime it has seen so far and 0 otherwise.*

Definition 5.7. *Given a possibly partial characteristic function g with $\mathrm{dom}(g) \subseteq X$ and $\mathrm{range}(g) \subseteq \{0,1\}$, define the "version space" $H_g = \{h \in \mathcal{H} : g \subseteq h\}$.*

Remark 5.8. *Observe that if $\mathcal{H}$ is a Littlestone class, $H \subseteq \mathcal{H}$ and f is a possibly partial, possibly empty characteristic function with $\mathrm{dom}(f) \subseteq X$ and $a \in X$, then*

$$(\star)\ \ \min\{\mathsf{Ldim}(H_{f \cup \{(a,0)\}}), \mathsf{Ldim}(H_{f \cup \{(a,1)\}})\} < \mathsf{Ldim}(H_f)$$

i.e., on one side of any partition by a half-space the dimension must drop, since $\mathsf{Ldim}(H_f) \leq \mathsf{Ldim}(\mathcal{H}) = d$ is defined and finite. (This property is what enables the notion of Littlestone majority vote.)

Proof of Theorem 5.1. To define our algorithms some notation will be useful. By "tree" in this proof we always mean an X-labeled binary tree of height T. Given an algorithm $\mathcal{A}$ and a tree $\mathcal{T}$, let $\sigma = \sigma(\mathcal{A}, \mathcal{T}) = \langle a_i : i < T \rangle$ and let $\tau = \tau(\mathcal{A}, \mathcal{T}) = \langle t_i : i < T \rangle$ denote the sequence of elements of X associated to the nodes traversed, and the corresponding outputs of $\mathcal{A}$, respectively. Again, given $\mathcal{A}$ and $\mathcal{T}$, let $\gamma = \gamma(\mathcal{A}, \mathcal{T}) = \langle g_i : i < T \rangle$ be the sequence of partial characteristic functions given by $g_i = \{(a_j, t_j) : j < i\}$.

We define the $\binom{T}{\leq d}$ algorithms as follows. Each algorithm $\mathcal{A}$ is parametrized by a set $A \subseteq \{0, \ldots, T-1\}$ of size $\leq d$ (and there is an algorithm for each such set). Given any tree $\mathcal{T}$, the algorithm proceeds as follows. Upon reaching a node at level i labeled by a_i, it computes the values $\mathsf{Ldim}(H_{g_i \cup \{(a_i,0)\}})$ and $\mathsf{Ldim}(H_{g_i \cup \{(a_i,1)\}})$. Informally, it asks how the Littlestone dimension of the set H_{g_i} will change according to the decision on a_i. It then makes its decision by cases. If $i \in A$, then the algorithm chooses the value of t_i which will make $\mathsf{Ldim}(H_{g_i \cup \{(a_i,t_i)\}})$ smaller, and in case of ties chooses 0. If $i \notin A$, then the algorithm chooses the value of t_i which

will make $\mathsf{Ldim}(H_{g_i \cup \{(a_i,t_i)\}})$ larger, and in case of ties chooses 1. This finishes the definition of our class $\mathbf{A}$. Clearly the algorithms involved are all prefix dependent.

Let us verify that for any tree $\mathcal{T}$ and any $h \in \mathcal{H}$ there is an algorithm in $\mathbf{A}$ realizing the same branch as h. Let $(b_0, s_0), \ldots, (b_{t-1}, s_{t-1})$ denote the root-to-leaf path traversed by h. For each $i < T$, let $f_i = \{(b_j, s_j) : j < i\}$ denote the partial characteristic function in play as we arrive to b_i. (Notice that necessarily each $f_i \subseteq h$.) Let us consider how we may use A to signal what to do. Let $d^i = \mathsf{Ldim}(H_{f_i})$, let $d^i_0 = \mathsf{Ldim}(H_{f_i \cup \{(a_i,0)\}}$ and let $d^i_1 = \mathsf{Ldim}(H_{f_i \cup \{(a_i,1)\}}$. There are several cases. If we know that at stage i the Ldim does not drop then by 5.8 the choice is determined. If we know that the Ldim drops and $d^i_0 \neq d^i_1$ then the choice is determined by knowing whether we chose the larger or smaller. If we know that Ldim drops and $d^i_0 = d^i_1$ then the choice is determined by knowing whether or not we went left. With this in mind, define $B \subseteq \{0, \ldots, T-1\}$ to be $B = \{i < T : (\ \mathsf{Ldim}(H_{f_i}) \geq \mathsf{Ldim}(H_{f_i \cup \{(a_i,1-s_i)\}}) > \mathsf{Ldim}(H_{f_i \cup \{(a_i,s_i)\}})\ )$ or $(\mathsf{Ldim}(H_{f_i \cup \{(a_i,s_i)\}}) = \mathsf{Ldim}((H_{f_i \cup \{(a_i,1-s_i)\}}) = \mathsf{Ldim}(H_{f_i}) - 1$ and $s_i = 1)\}$. In English, B is the set of all $i < T$ at which either there was only one way to make the dimension drop as much as possible, or both ways the dimension dropped by the same amount and we went left. Since at every $i \in B$ the Littlestone dimension drops, necessarily $|B| \leq d$.

Consider the algorithm $\mathcal{A} \in \mathbf{A}$ parameterized by B. We argue by induction on $i < T$ that $g_i = f_i$, that is, $a_i = b_i$ and $t_i = s_i$. To start, $a_0 = b_0$ is the label of the root. If $i \notin B$, then at this stage along the path traversed by h, either the Littlestone dimension did not drop as much as possible or $s_i = 1$. In the first case, there is only one value of t_i which will keep the dimension larger, and that is $t_i = s_i$. If the dimension went down equally for both successors, $\mathcal{A}$ will choose $t_i = 1 = s_i$. If $i \in B$, then here the Littlestone dimension must drop as much as possible, so either there is only one way to achieve this and so $t_i = s_i$, or both successors drop equally and $t_i = 0 = s_i$. This completes the proof. □

Remark 5.9. *A model theoretic reader will see definability of φ-types for stable φ.*

We now verify that this is a characterization.

Lemma 5.10. *Suppose $\mathsf{Ldim}(X, \mathcal{H})$ is not finite. Then for every $d \in \mathbb{N}$, for all sufficiently large $T \in \mathbb{N}$ and every collection $\mathbf{A}$ of $\binom{T}{\leq d}$ dynamic sets, there is some binary tree $\mathcal{T}$ of height T with X-labeled internal nodes and some $h \in \mathcal{H}$ which realizes a branch in $\mathcal{T}$ not realized by any algorithm from $\mathbf{A}$.*

Proof. Choose T so that $2^T > T^d$. Since Ldim is not finite, we may construct a full binary tree of height T whose nodes are labeled by X and such that every branch is realized by some $h \in \mathcal{H}$. Every algorithm $\mathcal{A} \in \mathbf{A}$ realizes one and only one branch in $\mathcal{T}$, so there are not enough of them to cover all branches. □

To conclude, observe $\mathbf{A}$ simulates $\mathcal{H}$ in an even stronger way: its algorithms can continue to simulate the realization of branches by $\mathcal{H}$ even when we weaken the notion of realization to allow a certain number of mistakes. (This also gives a simple derivation of the "oblivious" SSP lemma from our "tree" version here.)

Corollary 5.11. *Suppose $\mathsf{Ldim}(\mathcal{H}) = d \in \mathbb{N}$, let $T \in \mathbb{N}$, and let $\mathbf{A}$ be the family of $\binom{T}{\leq d}$ algorithms constructed for $\mathcal{H}$ in Theorem 5.1. Let $\mathcal{T}$ be any binary tree of height T with X-labeled internal nodes. Given any branch $(a_0, t_0), \ldots, (a_{T-1}, t_{T-1})$*

and any $h \in \mathcal{H}$, let $S = \{i < T : h(a_i) \neq t_i\}$ be the set of "mistakes" made by h for this branch. Then there is $\mathcal{A} \in \mathbf{A}$ which makes the same set of mistakes for this branch.

Proof. Consider a new tree $\mathcal{T}_*$ where all the nodes at level i have the same label a_i (the tree is "oblivious"). So branches through $\mathcal{T}_*$ amount to choosing subsets of $\{a_i : i < T\}$. This particular tree also falls under the jurisdiction of Theorem 5.1, and so gives our corollary. □

6. Majorities in Littlestone classes

So far we have been guided by the thesis that Littlestone classes are characterized by frequent, large sets with well-defined notions of majority. However, there are at least two candidate notions of majority which are quite distinct: the majority arising from the counting measure, which we have been exploring via ϵ-excellent and ϵ-good, and the notion of majority arising from Littlestone rank.

In this section we prove that these two notions of majority "densely often agree" in Littlestone classes and indeed this is true of any simple axiomatic notion of majority, as defined below.

Definition 6.1. *Say that $H \subseteq \mathcal{H}$ is* Littlestone-opinionated *if for any $a \in X$, one and only one of*

$$\mathsf{Ldim}(\{h \in H : h(a) = 0\}), \mathsf{Ldim}(\{h \in H : h(a) = 1\})$$

is strictly smaller than $\mathsf{Ldim}(H)$.

As a warm-up, we prove several claims. As above, a *partition of H by a half-space* means that for some element $a \in X$ we separate H into $\{h \in H : h(a) = 0\}$ and $\{h \in H : h(a) = 1\}$.) So in this language, H is Littlestone-opinionated if in any partition by a half-space, exactly one of the two pieces retains the Ldim.

Claim 6.2. *Suppose $\mathsf{Ldim}(\mathcal{H}) = d$ and $0 < \epsilon < \frac{1}{2}$. Then for any finite $H \subseteq \mathcal{H}$ there is $A \subseteq H$ of size $\geq \epsilon^d |H|$ such that A is both ϵ-good and Littlestone-opinionated and these two notions of majority agree, i.e. for any $a \in X$*

$$\mathsf{Ldim}(\{h \in A : h(a) = t\}) = \mathsf{Ldim}(A) \text{ iff } |\{h \in A : h(a) = t\}| \geq (1 - \epsilon)|A|.$$

Proof. It suffices to observe that given any finite $H \subseteq \mathcal{H}$ which (a) is not ϵ-good, (b) is ϵ-good but is not Littlestone-opinionated, or (c) is both ϵ-good and Littlestone-opinionated but the two notions of majority do not always agree, we can find $G \subseteq H$ (arising from by a partition of H by a half-space) with $|G| \geq \epsilon |H|$ and $\mathsf{Ldim}(G) < \mathsf{Ldim}(H)$, because this initiates a recursion which cannot continue more than $\mathsf{Ldim}(H) \leq \mathsf{Ldim}(\mathcal{H}) = d$ steps.

Why? In case (a), there is a partition into two pieces of size $\geq \epsilon |H|$; choose the one of smaller Ldim. In case (b), there is a partition into two pieces each of Ldim strictly smaller than $\mathsf{Ldim}(H)$; choose the one of larger counting measure. In case (c), there is a partition where the majorities disagree, and we can choose the piece of larger counting measure and thus smaller Ldim. This completes the proof. □

Definition 6.3. *Call P a* good property *(or: ϵ-good property) for $\mathcal{H}$ if it is a property of finite subsets of $\mathcal{H}$ which implies ϵ-good and which satisfies: for some constant $c = c(P) > 0$, for any finite $H \subseteq \mathcal{H}$ there is $B \subseteq H$ of size $\geq \epsilon^c |H|$ with property P.*

Corollary 6.4. *By* 3.1 *above, if $\mathcal{H}$ is Littlestone and $\epsilon < \frac{1}{2}$ then "ϵ-excellent" is an ϵ-good property for $\mathcal{H}$.*

Lemma 6.5. *Let $\mathcal{H}$ be a Littlestone class of dimension d and $0 < \epsilon < \frac{1}{2}$. Let P be a good property for $\mathcal{H}$ and $c = c(P)$. Then for any finite $H \subseteq \mathcal{H}$ there is $A \subseteq H$ of size $\geq \epsilon^{(c+1)d}|H|$ such that A has property P (so is also ϵ-good) and is Littlestone-opinionated, and for any $a \in X$,*

$$\mathsf{Ldim}(\{h \in A : h(a) = t\}) = \mathsf{Ldim}(H) \text{ iff } |\{h \in A : h(a) = t\}| \geq (1-\epsilon)|A|$$

i.e. the ϵ-good majority and the Littlestone majority agree.

Proof. Modify the recursion in the previous proof as follows. At a given step, if H does not have property P, replace it by a subset C of size $\geq \epsilon^c|H|$ which does. Since P implies ϵ-good, if we are not finished, then we are necessarily in case (b) or (c) and at the cost of an additional factor of ϵ we can find $B \subseteq C$ where the Ldim drops. In each such round, we replace H by $B \subseteq H$ with $|B| \geq \epsilon^{c+1}|H|$ and $\mathsf{Ldim}(B) < \mathsf{Ldim}(H)$. □

Discussion 6.6. Note that this majority agreement deals with half-spaces, which is arguably the interesting case for "Littlestone-opinionated" as it relates to the SOA. In 6.5, it is a priori not asserted that every subset of A which is large in the sense of counting measure (but does not arise from a half-space) has large Ldim.

Definition 6.7 (Axiomatic largeness)**.** *Define $\mathcal{M}$ to be an axiomatic notion of relative largeness for the class $\mathcal{H}$ if it satisfies the following properties.*[14]

1. *$\mathcal{M}$ is a subset of $\mathcal{P} = \{(B, A) : B \subseteq A \subseteq \mathcal{H}\}$.*
2. *Define $\mathcal{P}_{half} := \{(B, A) \in \mathcal{P} : B$ arises as the intersection of A with a half-space $\}$.*
3. *If $(B, A) \in \mathcal{M}$, say "B is a large subset of A." We may write $B \subseteq_{\mathcal{M}} A$.*
4. *The rules are:*
   (a) (monotonicity in the set) *if $C \subseteq B \subseteq A$ and $C \subseteq_{\mathcal{M}} A$ then $B \subseteq_{\mathcal{M}} A$.*
   (b) (monotonicity in the superset) *if $C \subseteq B \subseteq A$ and $C \subseteq_{\mathcal{M}} A$ then $C \subseteq_{\mathcal{M}} B$.*
   (c) (identity) $(A, A) \in \mathcal{M}$.
   (d) (non-contradiction) *If $(B, A) \in \mathcal{P}_{half}$ and $C = A \setminus B$ [so also $(C, A) \in \mathcal{P}_{half}$] then at most one of (B, A) and (C, A) belongs to $\mathcal{M}$.*
   (e) (chain condition) *There is $n = n(\mathcal{M}) < \omega$ such that if $\langle A_i : i < m\rangle$ is a set of subsets of $\mathcal{H}$ and $(A_{i+1}, A_i) \notin \mathcal{M}$ for all $i < m-2$ then $m \leq n$. In other words, the length of a descending chain*

$$A_{m-1} \subseteq \cdots\cdots \subseteq A_0$$

   of non-large subsets is upper bounded by n.

Example 6.8. *Suppose $\mathcal{H}$ is a Littlestone class. Then*

$$\mathcal{M} = \{(B, A) \in S : \mathsf{Ldim}(A) = \mathsf{Ldim}(B)\}$$

satisfies Definition 6.7.

Proof. Conditions (3)(a),(b),(c) are immediate; (d) follows by the definition of Ldim. Condition (e) is clear because if $(A_i, A_{i+1}) \notin \mathcal{M}$ then $\mathsf{Ldim}(A_{i+1}) < \mathsf{Ldim}(A_i)$ so $n(\mathcal{M}) \leq d$. □

[14]This captures *relative* majority or largeness since "being large in" is a two-place relation.

Example 6.9. *For model theoretic readers, note that for suitable hypotheses classes the Shelah $R(x = x, \Delta, \lambda)$ ranks, when defined and restricted to multiplicity one, can illustrate* 6.7 *for other values of Δ, λ.*

In 6.10 we don't need to assume a priori that $\mathcal{H}$ is Littlestone, though the proof will show that it is.

Lemma 6.10. *Suppose $\mathcal{H}$ admits a notion of relative largeness $\mathcal{M}$. Let $0 < \epsilon < \frac{1}{2}$ and let P be an ϵ-good property for $\mathcal{H}$. Let $c = c(P)$ and $n = n(\mathcal{M})$. Then for any nonempty finite $H \subseteq \mathcal{H}$ there is $A \subseteq H$ of size $\geq \epsilon^{(c+1)n}|H|$ such that:*

1. *A has property P, and thus is ϵ-good, so for any partition of A by a half-space into $B \cup C$, at least one (so exactly one) of B, C has size $< \epsilon|A|$.*
2. *For any partition of A by a half-space into $B \cup C$, at least one (so exactly one) of (B, A), (C, A) belongs to $\mathcal{M}$.*
3. *The two notions agree, i.e. $(B, A) \in \mathcal{M}$ if and only if $|B| \geq (1 - \epsilon)|A|$.*

Proof. Let $n = n(\mathcal{M})$ and set $A_0 = H$. By induction on $t \geq 0$ we shall prove that if A_t does not satisfy contitions 1, 2, and 3 then either it contains a subset of size $\geq \epsilon^c|A_t|$ which does, or there is $A_{t+1} \subseteq A_t$ such that $|A_{t+1}| \geq \epsilon^{c+1}|A_t|$ and $(A_{t+1}, A_t) \notin \mathcal{M}$. Our chain condition 6.7(4)(e) will then ensure $t \leq n$.

For each $t \geq 0$ proceed as follows. If A_t has property P, define $A'_t = A_t$. If not, replace A_t by a subset of size $\geq \epsilon^c|A_t|$ which does, and set this to be A'_t. A priori, we have no information on whether $(A'_t, A_t) \in \mathcal{M}$. Since A'_t has property P, if A'_t does not already satisfy 1, 2, and 3, then condition 2 or 3 must fail; in either case, there must be some half-space which partitions A'_t into two non-trivial sets at least one of which, call it B, has size at least $\epsilon|A'_t|$ and satisfies $(B, A'_t) \notin \mathcal{M}$. Set $A_{t+1} = B$. Then $|B| \geq \epsilon^{c+1}|A_t|$ and by condition 6.7(4)(b), $(A_{t+1}, A_t) \notin \mathcal{M}$. This completes the inductive step and the proof. □

Theorem 6.11. *The following are equivalent for $(X, \mathcal{H})$.*

1. *$\mathcal{H}$ admits a notion of relative largeness $\mathcal{M}$.*
2. *$\mathcal{H}$ is a Littlestone class.*
3. *For every $\mathcal{M}$ and $0 < \epsilon < \frac{1}{2}$ there is $n = n(\epsilon, \mathcal{M})$ such that every finite nonempty $H \subseteq \mathcal{H}$ has a subset A which satisfies:*
   (a) *$|A| \geq \epsilon^n|H|$, and*
   (b) *A is ϵ-good, and*
   (c) *for every partition of A by a half-space into $B \cup C$, $(B, A) \in \mathcal{M}$ if and only if $|B| \geq \epsilon|A|$ if and only if $|B| \geq (1 - \epsilon)|A|$.*

   i.e. the counting majority and the $\mathcal{M}$-majority are well defined and agree.
4. *In item (3) we may replace (b) by "A has property P" when P is an ϵ-good property for $\mathcal{H}$, at the cost of changing the exponent n in (a) to $(c + 1)n$ for $c = c(P)$.)*

Proof. (2) implies (1) is Example 6.8. (1) implies (3) [or (4)] is Lemma 6.10. Clearly (4) implies (3). For (3) implies (2), note that (3) tells us a fortiori that we can always find large ϵ-good subsets, so $\mathcal{H}$ must be a Littlestone class by 2.11. □

Remark 6.12. *Although we have formulated these largeness properties for subsets of $\mathcal{H}$, the symmetric results whould hold for subsets of X.*

Discussion 6.13. A key point in the proof of the stable regularity lemma (in our language) is that because finite Littlestone dimension implies finite VC dimension, if

we randomly partition ϵ-excellent sets then the pieces are likely to remain excellent (for a related ϵ). This is what allowed for an equitable partition into excellent sets. There is a priori no reason the analogous fact should be true for Littlestone-opinionated sets. However, this section finds (densely often) sets where the counting majority and the Littlestone majority agree. Randomly partitioning these, we retain goodness, so necessarily also retain the ability to correctly predict Littlestone majority in accordance with the original set, despite perhaps not being Littlestone-opinionated.

Discussion 6.14. It is interesting to inspect relative largeness from the perspective of online learning. Indeed, note that any such notion gives rise to an online learning strategy with a bounded mistake bound: the online learner maintains a version space $\mathcal{H}_i \subseteq \mathcal{H}$, starting with $\mathcal{H}_0 = \mathcal{H}$. For each input example x_i received, the learner predicts $\hat{y}_i$ such that

$$(\{h \in \mathcal{H}_i : h(x_i) = \hat{y}_i\}, \mathcal{H}_i) \in \mathcal{M}$$

and note that there can be at most such $\hat{y}_i$, if no such $\hat{y}_i$ exists then the learner predicts $\hat{y}_i = 0$. Then, upon receiving the true label y_i, if $y_i = \hat{y}_i$ then the learner sets $\mathcal{H}_{i+1} = \mathcal{H}_i$ and else, when $y_i \neq \hat{y}_i$, the learner sets $\mathcal{H}_{i+1} = \{h \in \mathcal{H}_i : h(x_i) = y_i\}$. Observe that given any sequence $(x_1, y_1), \ldots, (x_T, y_T)$, this learner makes at most $n(\mathcal{M})$ mistakes: indeed, if the learner makes a mistake on x_i then $(\mathcal{H}_{i+1}, \mathcal{H}_i) \notin \mathcal{M}$, and $\mathcal{H}_{i+1}$ is obtained by intersecting $\mathcal{H}$ with a halfspace.

This point of view offers an alternative explanation for the fact that only Littlestone classes admit notions of relative largeness. Moreover, it implies that for every notion of relative largeness $\mathcal{M}$, we have that $n(\mathcal{M})$ is at least the Littlestone dimension. This follows because the Littestone dimension is equal to the optimal mistake-bound. Thus, the notion of relative largeness which arises from the Littlestone dimension is optimal in the sense that it minimizes $n(\mathcal{M})$.

7. Some open problems

To conclude the paper we mention several natural open problems and directions for further work; some appear challenging, some more accessible.

(1) For VC classes, recall that we have the usual Sauer-Shelah-Perles lemma, and Haussler's covering lemma which says that every VC class of VC-dimension d can be ϵ-covered by roughly $\frac{1}{\epsilon^d}$ hypotheses [13].[15] (The SSP lemma can be thought of as the special case of Haussler's covering lemma where the domain has size n and when $d = \frac{1}{n}$.) This is clearly useful for learning. It is natural to ask whether there is a dynamic version of this covering lemma for Littlestone classes. That is, is there a function $f = f(\epsilon, d)$ such that for any Littlestone class $\mathcal{H}$ of Ldim d we can always find $\leq f(\epsilon, d)$ dynamic sets which approximately cover the whole class $\mathcal{H}$, meaning that for every sequence $x_1, \ldots, x_n$ from X and every $h \in \mathcal{H}$ there is a dynamic set in our list which is ϵ-close to it. This is also related to [1].

(2) In the classical case, there is a fundamental relationship between the Littlestone dimension and the half-graph/Threshold dimension, as explained by

[15] Informally, there is a list of approximately $\frac{1}{\epsilon^d}$ hypotheses such that every other hypothesis in our class is ϵ-close to one of the hypotheses in our list.

Shelah's unstable formula theorem: both are finite together, and bounds are known [31], [14]. However, the question of determining tight quantitative bounds in the finite remains open. The known bounds given by Hodges [15] say roughly that from a half-graph of length k there is a tree of height about $\log k$, and that from a tree of height n there is a half-graph of length about $\log n$. Determining whether these bounds are tight may be challenging, and seems worth emphasizing here. To reiterate what we said in section one, a useful aspect of this relationship is connecting a symmetric or self-dual quantity with Littlestone dimension (see 1.3 above).

(3) As mentioned in the text, applying the Hodges bounds twice tells us that if $\mathsf{Ldim}(\mathcal{H}) = d$, the dual class has Littlestone dimension bounded by about 2^{2^d}. Can this be improved?

(4) It seems worth while to explore further the significance of dynamic Sauer-Shelah-Perles lemmas for model theory. As a soft question, are there useful model theoretic explanations for the existing regret bounds for multiplicative-weights algorithms mentioned in §3?

(5) We have not sorted out the extent of the nonmonotonicity of excellence as ϵ varies; this could give another approach to existence of excellent sets. Nor have we tried to optimize the constants in either of the two existence proofs in the present paper.

(6) Is there a helpful "outside" characterization of the good properties in the sense of 6.3?

References

1. Noga Alon and Omri Ben-Eliezer and Yuval Dagan and Shay Moran and Moni Naor and Eylon Yogev, Samir Khuller and Virginia Vassilevska Williams, Adversarial laws of large numbers and optimal regret in online classification, STOC '21: 53rd Annual ACM SIGACT Symposium on Theory of Computing, Virtual Event, Italy, June 21-25, 2021, 447–455, ACM, 2021
2. Noga Alon and Amos Beimel and Shay Moran and Uri Stemmer, Jacob D. Abernethy and Shivani Agarwal, Closure Properties for Private Classification and Online Prediction, Conference on Learning Theory, COLT 2020, 9-12 July 2020, Virtual Event [Graz, Austria], Proceedings of Machine Learning Research, 125, 119–152, PMLR, 2020
3. N. Alon, R. Duke, H. Leffman, V. Rödl, R. Yuster, "The algorithmic aspects of the regularity lemma." FOCS 33(1992), 479-481; Journal of Algorithms 16 (1994), 80-109.
4. Noga Alon, Roi Livni, Maryanthe Malliaris, and Shay Moran. Private PAC learning implies finite Littlestone dimension. In Proceedings of the 51st Annual ACM Symposium on the Theory of Computing, STOC '19, New York, NY, USA, 2019. ACM.
5. Shai Ben-David and Dávid Pál and Shai Shalev-Shwartz, Agnostic Online Learning COLT 2009 - The 22nd Conference on Learning Theory, Montreal, Quebec, Canada, June 18-21, 2009
6. Siddharth Bhaskar, Thicket Density, arXiv, Number = arXiv:1702.03956, Year = 2017
7. Mark Bun, Roi Livni, and Shay Moran. An equivalence between private classification and online prediction. In Sandy Irani, editor, 61th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2020, November 16-19, 2020. IEEE Computer Society, 2020.
8. Hunter Chase and James Freitag. Model theory and machine learning. The Bulletin of Symbolic Logic, 25(03):319–332, Feb 2019. ArXiv preprint arXiv:1801.06566, 2018.
9. H. Chase and J. Freitag, "Bounds in query learning," in Conference on Learning Theory. PMLR, 2020, pp. 1142–1160.
10. Shlomo Eshel and Itay Kaplan. "On uniform definability of types over finite sets for NIP formulas." Journal of Mathematical Logic. Vol. 21, No. 3, 2150015. ArXiv:1904.10336.

11. Yuval Filmus, Steve Hanneke, Idan Mehalel, Shay Moran. "Optimal Prediction Using Expert Advice and Randomized Littlestone Dimension." The Thirty Sixth Annual Conference on Learning Theory, COLT 2023. Vol. 195, Pages 773–836.
12. Ghazi, Badih and Golowich, Noah and Kumar, Ravi and Manurangsi, Pasin, Proceedings of the 32nd International Conference on Algorithmic Learning Theory, Vitaly Feldman and Katrina Ligett and Sivan Sabato, 16–19 Mar, 686–696, PMLR, Proceedings of Machine Learning Research, Near-tight closure bounds for the Littlestone and threshold dimensions, 132, 2021
13. David Haussler, Sphere Packing Numbers for Subsets of the Boolean n-Cube with Bounded Vapnik-Chervonenkis Dimension, J. Comb. Theory, Ser. A, 69, 2, 217–232, 1995
14. W. Hodges, *Model Theory*, Encyclopedia of Mathematics, Cambridge University Press, 1993.
15. W. Hodges, "Encoding orders and trees in binary relations." Mathematika 28 (1981) 67–81.
16. M. Karpinski and A. Macintyre. "Polynomial bounds for VC dimension of sigmoidal and general Pfaffian neural networks." J. Computer and System Sciences (1997) 54(1):169–176.
17. J. Komlós and M. Simonovits, "Szemerédi's Regularity Lemma and its applications in graph theory." (1996) In *Combinatorics: Paul Erdös is Eighty*, Vol. 2 (D. Miklós, V. T. Sós and T. Szönyi, eds), Bolyai Society Math. Studies, Keszthely, Hungary, pp. 295-352.
18. M. C. Laskowski. "Vapnik-Chervonenkis classes of definable sets." J. London Math Soc (1992) 2(2):377–284.
19. Nick Littlestone, Learning Quickly When Irrelevant Attributes Abound: A New Linear-threshold Algorithm, Machine Learning, 2, 4, 285–318, 1988
20. Nick Littlestone and Manfred K. Warmuth, The Weighted Majority Algorithm, Inf. Comput., 108, 2, 212–261, 1994
21. Roi Livni and Pierre Simon. "Honest compressions and their application to compression schemes." In *Conference on Learning Theory* (2013) pps. 77–92.
22. L. Lovász and B. Szegedy, "Regularity partitions and the topology of graphons." *An Irregular Mind (Szemerédi is 70)*, Bolyai Society Mathematical Studies, 21 (2010) pp. 415-446.
23. M. Malliaris and A. Pillay. "The stable regularity lemma revisited." Proc. Amer. Math. Soc. 144 (2016) 1761–1765.
24. M. Malliaris and S. Shelah. "Regularity lemmas for stable graphs." Trans. Amer. Math Soc, 366 (2014), 1551–1585. First public version: arXiv:1102.3904 (Feb 2011).
25. M. Malliaris and S. Shelah. "Notes on the stable regularity lemma." To appear, Bulletin of Symbolic Logic. Arxiv:2012.09794.
26. A. Onshuus, Mathscinet review of [8].
27. A. Rakhlin and K. Sridharan and A. Tewari, Online learning via sequential complexities. J. Mach. Learn. Res. 16, 1, 155–186, 2015
28. A. Rakhlin and K. Sridharan and A. Tewari, Sequential complexities and uniform martingale laws of large numbers, Probability Theory and Related Fields 161, 1-2, 111–153, 2015, Springer,
29. A. Rakhlin and K. Sridharan and A. Tewari, Online learning: Random averages, combinatorial parameters, and learnability, Advances in Neural Information Processing Systems, 1984–1992, 2010
30. N. Sauer, On the Density of Families of Sets, J. Comb. Theory, Ser. A, 1, 145–147, 13, 1972
31. S. Shelah, *Classification Theory*, North-Holland, 1978. Revised edition, 1990.
32. Shai Shalev-Schwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014. Available on the first author's webpage.
33. E. Szemerédi, "Regular partitions of graphs." In *Colloques Internationaux C.N.R.S.* No. 260, Problèmes Combinatoires et Théorie des Graphes, Orsay (1976), 399–401.
34. C. Terry and J. Wolf. "Stable arithmetic regularity in the finite-field model." Bull. London Math. Soc. 51 (2019), 1, 70–88.
35. C. Terry and J. Wolf. "Irregular triads in 3-uniform hypergraphs." ArXiv:2111.01737. [reference added in revision]
36. R. Walker. "Tree Dimension and the Sauer-Shelah Dichotomy." ArXiv:2203.12211. [reference added in revision]
https://arxiv.org/abs/2203.12211

Department of Mathematics, University of Chicago
Email address: mem@math.uchicago.edu

Departments of Mathematics, Computer Science, and Data and Decision Sciences, Technion.
Email address: smoran@technion.ac.il