Proceedings of Machine Learning Research vol 134:1–26, 2021 34th Annual Conference on Learning Theory

Online Learning with Simple Predictors and a Combinatorial Characterization of Minimax in 0/1 Games

Steve Hanneke STEVE.HANNEKE@GMAIL.COM
Toyota Technological Institute at Chicago

Roi Livni RLIVNI@TAUEX.TAU.AC.IL
Tel Aviv University

Shay Moran SMORAN@TECHNION.AC.IL
Technion and Google Research

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

Which classes can be learned properly in the online model? — that is, by an algorithm that on each round uses a predictor from the concept class. While there are simple and natural cases where improper learning is useful and even necessary, it is natural to ask how complex must the improper predictors be in such cases. Can one always achieve nearly optimal mistake/regret bounds using "simple" predictors?

In this work, we give a complete characterization of when this is possible, thus settling an open problem which has been studied since the pioneering works of Angluin (1987) and Littlestone (1988). More precisely, given any concept class $\mathbb{C}$ and any hypothesis class $\mathbb{H}$ we provide nearly tight bounds (up to a log factor) on the optimal mistake bounds for online learning $\mathbb{C}$ using predictors from $\mathbb{H}$. Our bound yields an exponential improvement over the previously best known bound by Chase and Freitag (2020).

As applications, we give constructive proofs showing that (i) in the realizable setting, a near-optimal mistake bound (up to a constant factor) can be attained by a sparse majority-vote of proper predictors, and (ii) in the agnostic setting, a near optimal regret bound (up to a log factor) can be attained by a randomized proper algorithm. The latter was proven non-constructively by Rakhlin, Sridharan, and Tewari (2015). It was also achieved by constructive but improper algorithms proposed by Ben-David, Pal, and Shalev-Shwartz (2009) and Rakhlin, Shamir, and Sridharan (2012).

A technical ingredient of our proof which may be of independent interest is a generalization of the celebrated Minimax Theorem (von Neumann, 1928) for binary zero-sum games with arbitrary action-sets: a simple game which fails to satisfy Minimax is "*Guess the Larger Number*". In this game, each player picks a natural number and the player who picked the larger number wins. Equivalently, the payoff matrix of this game is infinite triangular. We show that this is the only obstruction: if the payoff matrix does not contain triangular submatrices of unbounded sizes then the Minimax Theorem is satisfied. This generalizes von Neumann's Minimax Theorem by removing requirements of finiteness (or compactness) of the action-sets, and moreover it captures precisely the types of games of interest in online learning: namely, Littlestone games.

Keywords: Online Learning, Equivalence Queries, Littlestone Dimension, Minimax Theorem, Mistake Bound, VC Dimension

1. Introduction

An improper learning algorithm is an algorithm that learns a class $\mathbb{C}$ using hypotheses h that are not necessarily in $\mathbb{C}$. While at a first sight this may seem like a counter-intuitive thing to do, improper algorithms are extremely powerful and using them often circumvents computational issues and sample complexity barriers (Srebro, Rennie, and Jaakkola, 2005; Candès and Recht, 2009; Anava, Hazan, Mannor, and Shamir, 2013; Hazan, Livni, and Mansour, 2015; Hanneke, 2016; Hazan and Ma, 2016; Hazan, Kale, and Shalev-Shwartz, 2017; Agarwal, Bullins, Hazan, Kakade, and Singh, 2019). In fact, there are extreme examples of learning tasks that can only be performed by improper algorithms (Daniely and Shalev-Shwartz, 2014; Daniely, Sabato, Ben-David, and Shalev-Shwartz, 2015; Angluin, 1987; Montasser, Hanneke, and Srebro, 2019).

However, while many of the improper algorithms proposed in the theoretical literature use sophisticated representations (e.g., Haussler, Littlestone, and Warmuth, 1994; Littlestone, 1988), other natural improper algorithms use "simple" hypotheses which often can be described as simple combinations of functions from $\mathbb{C}$. For instance, many algorithms (e.g., boosting) use sparse weighted majority-votes of concepts from the class. It is therefore natural to ask:

> Can any given (learnable) class $\mathbb{C}$ be learned by algorithms which use "simple" hypotheses?

Online Classification. While the above question has been extensively studied (and answered[1]) in the batch setting, in the online setting it remains largely open. Note that already in the realizable mistake-bound model there are learnable classes which cannot be learned properly: one simple example is the class of all singletons over $\mathbb{N}$. Indeed, in each round any proper learner must use a singleton $1_{\{n\}}$, and therefore the adversary can always force a mistake by presenting the example $(n, 0)$. Note however, that if the learner could use the all-zero function $1_\emptyset$ (which is not in the class), then she would only make one mistake before learning the target concept.

In general, the optimal mistake-bound for learning $\mathbb{C}$ is achieved by the *Standard Optimal Algorithm* (SOA) of Littlestone, which is improper (as the above example shows). It is interesting to note that it is not known whether the hypothesis space $\mathbb{H}$ used by the SOA is simple in any natural sense: e.g., it is not known whether its Littlestone dimension is bounded in terms of the Littlestone dimension of $\mathbb{C}$, or even whether its VC dimension is. (In the above example, the SOA only uses the class $\mathbb{H} = \mathbb{C} \cup \{1_\emptyset\}$ which has the same Littlestone and VC dimensions as $\mathbb{C}$.) One of the results in this work provides a nearly optimal algorithm (up to a numerical multiplicative factor) which uses hypotheses that are sparse majority votes of functions from $\mathbb{C}$.

1.1. Our Contribution

In this section we survey the main contributions of this work. Some of the statements involve standard technical terms which are defined in Section 2, where we also give complete formal statements of our results.

1. Indeed, by *Uniform Convergence*, any *Empirical Risk Minimizer* attains a near-optimal sample complexity in the PAC model. (Vapnik and Chervonenkis, 1971)

MAIN RESULT I: WHEN DO HYPOTHESES FROM $\mathbb{H}$ SUFFICE TO LEARN $\mathbb{C}$?

Our first main result provides a complete combinatorial characterization of a near-optimal mistake bound for online learning $\mathbb{C}$ using hypotheses from $\mathbb{H}$. This settles an open problem which was studied since the early days of Computational Learning Theory (Angluin, 1987; Littlestone, 1988; Angluin, 1990; Hellerstein et al., 1996; Balcázar et al., 2001, 2002a,b; Hayashi et al., 2003; Angluin and Dohrn, 2020; Chase and Freitag, 2020).

Two Basic Lower Bounds. The early works of Angluin (1987, 2004); Balcázar et al. (2002b) and Littlestone (1988) presented two basic lower bounds for online learning $\mathbb{C}$ using $\mathbb{H}$. In the seminal work which introduced the mistake-bound model Littlestone (1988) defined the Littlestone[2] dimension and noticed that it provides a lower bound on the number of mistakes, even if the algorithm is allowed to use any hypothesis $h : \mathcal{X} \to \{0, 1\}$ (i.e. $\mathbb{H} = \{0, 1\}^{\mathcal{X}}$).

The second (and less known) lower bound was rooted in the seminal work of Angluin (1987) which introduced the equivalence-query model. This bound was later generalized by Hellerstein, Pillaipakkamnatt, Raghavan, and Wilkins (1996); Balcázar, Castro, Guijarro, and Simon (2002b) and today it is known as the *strong consistency dimension* (Balcázar, Castro, Guijarro, and Simon, 2002b) or the *dual Helly number* (Bousquet, Hanneke, Moran, and Zhivotovskiy, 2020). The idea behind this lower bound can be seen as a generalization of the argument showing that singletons over $\mathbb{N}$ are not properly learnable, which was discussed earlier. Assume there is a set $S \subseteq \mathcal{X} \times \{0, 1\}$ of labeled examples such that no function $h \in \mathbb{H}$ satisfies $(\forall(x, y) \in S) : h(x) = y$, but for some $k \in \mathbb{N}$ *every* $(x_1, y_1), \ldots, (x_k, y_k) \in S$ has some $c \in \mathbb{C}$ with $(\forall i \leq k) : c(x_i) = y_i$. In words, S is not realizable by the hypothesis class $\mathbb{H}$, but every subset of S of size k is realizable by the concept class $\mathbb{C}$. Given such a set S, the adversary can force k mistakes for any learner with hypothesis class $\mathbb{H}$, as follows. On each round, the learner proposes a hypothesis $h \in \mathbb{H}$, and since S is not realizable by $\mathbb{H}$ there must exist $(x, y) \in S$ with $h(x) \neq y$, so that this counts as a mistake. Since all subsets of S of size k are realizable by $\mathbb{C}$, this adversary also guarantees that all k examples it gives will still be consistent with some $c \in \mathbb{C}$. However, if instead there exists a subset of S of size k that is not realizable by $\mathbb{C}$, then this strategy for the adversary might fail. The *dual Helly number* K of $\mathbb{C}$ relative to $\mathbb{H}$ is the smallest k such that, for every set S not realizable by $\mathbb{H}$, there exists a subset of size at most k not realizable by $\mathbb{C}$; K is defined to be infinite if no such k exists. Thus, the above adversary can always force at least $\mathrm{K} - 1$ mistakes.

We show in Theorem 1 that the Littlestone dimension and dual Helly number are the only obstacles for learning $\mathbb{C}$ using $\mathbb{H}$:

Main Result I (Theorem 1)
There exists a deterministic algorithm which online learns $\mathbb{C}$ with at most $O(\mathrm{L} \cdot \mathrm{K} \cdot \log(\mathrm{K}))$ mistakes while only using hypotheses from $\mathbb{H}$, where L is the Littlestone dimension of $\mathbb{C}$ and K is the dual Helly number of $\mathbb{C}$ relative to $\mathbb{H}$.

This improves over the best previous result K^{L} by (Chase and Freitag, 2020) who were the first to show that $\mathbb{C}$ can be learned using hypotheses from $\mathbb{H}$ if and only if $\mathrm{K}, \mathrm{L} < \infty$. Our result further improves the upper bound to a polynomial dependence which is nearly tight in the sense that a lower bound of $\max\{\mathrm{L}, \mathrm{K} - 1\}$ on the optimal mistake bound always holds, and there exist classes $\mathbb{C}, \mathbb{H}$

2. The name, *Littlestone dimension*, was later coined by Ben-David, Pál, and Shalev-Shwartz (2009).

of any L and K for which the optimal number of mistakes is $\Omega(\mathrm{LK})$ (Angluin, 1987; Littlestone, 1988; Balcázar, Castro, Guijarro, and Simon, 2002b).

MAIN RESULT II: OPTIMAL MISTAKE-BOUNDS USING SPARSE-MAJORITIES

We have seen simple examples demonstrating that sometimes one has to use improper learners in order to achieve non-trivial mistake bounds in online learning $\mathbb{H}$. A natural question is "how improper" must an optimal algorithm be? Is there an algorithm which is close to being proper? Our second main result shows that it is possible to achieve a nearly optimal mistake-bound (up to a universal constant factor), using predictors that are sparse majority votes of functions in the class.

> **Main Result II (Theorem 3)**
> There exists a deterministic algorithm which online learns $\mathbb{C}$ with at most $O(\mathrm{L})$ mistakes while only using hypotheses of the form $\mathrm{Maj}(h_1, \ldots, h_p)$ for $h_i \in \mathbb{C}$, where L is the Littlestone dimension of $\mathbb{C}$, and p is a constant depending only on $\mathbb{C}$ (proportional to dual VC dimension).

This provides another demonstration to the usefulness of majority-votes/ensemble-methods in classification and might be seen as a kind of online learning analogue of Hanneke's result for optimal PAC learning (Hanneke, 2016), which is also achievable by an algorithm based on majority votes of concepts in $\mathbb{C}$. A corollary of this result is that there exist randomized proper algorithms whose expected mistake-bound is $\tilde{O}(\varepsilon \cdot T + \frac{\mathrm{L}}{\varepsilon})$, where L is the Littlestone dimension of the class.

MAIN RESULT III: NEAR-OPTIMAL REGRET-BOUNDS USING RANDOMIZED PROPER ALGORITHMS

In a fascinating result, Rakhlin, Sridharan, and Tewari (2015) have established the existence of optimal randomized proper algorithms in agnostic online learning (under some additional topological restrictions). Interestingly, their result is non-constructive: they prove the existence of such an algorithm by viewing online learning as a repeated game between the learner and the adversary, and by analyzing the value of that game via a dual perspective, which involves an application of the Minimax Theorem w.r.t. the *repeated* game. On the other hand, Rakhlin, Shamir, and Sridharan (2012) and Ben-David, Pál, and Shalev-Shwartz (2009) gave constructive proofs of this fact, but the implied algorithms are randomized *improper* algorithms. The following result achieves the best of both worlds:

> **Main Result III (Theorem 5)**
> We give a constructive proof demonstrating that any class $\mathbb{C}$ can be online learned in the agnostic setting by a randomized proper algorithm whose expected regret is $\tilde{O}(\sqrt{\mathrm{L} \cdot T})$, where L is the Littlestone dimension, and T is the horizon.

Note that in the agnostic setting randomization is essential (see e.g. Ben-David, Pál, and Shalev-Shwartz, 2009; Cesa-Bianchi and Lugosi, 2006), and unlike the realizable setting, in the agnostic setting nearly optimal randomized proper learners can exist. Another advantage of our proof compared to the argument by Rakhlin, Sridharan, and Tewari (2015) is that their application of the

Minimax Theorem required additional assumptions such as separability and compactness, which are not needed in our proof.

One may view the above result as yet another demonstration of the benefits of randomized algorithms. Moreover, if one views online learning as a repeated zero-sum game where the learner's pure strategies are the hypotheses in $\mathbb{H}$, and the adversary's pure strategies are labelled examples, then a randomized proper online learner *simply corresponds to a mixed strategy in the repeated game*, and so this can be interpreted as another manifestation of the benefits of mixed (randomized) strategies in online learning.

In fact, the Minimax Theorem for zero-sum games is a key component in our derivation, and as another technical contribution of this work, we prove a generalization of it to infinite games which we discuss next.

MAIN RESULT IV: A GENERALIZATION OF THE MINIMAX THEOREM

Consider the following "guess the larger number" game between two players whom we call Alice and Bob. Each of the players privately picks a natural number; then, Alice and Bob reveal the numbers to each other, and the winner is the player who picked the larger number. Note that the payoff matrix of this game is an $\mathbb{N} \times \mathbb{N}$ triangular matrix.

This game does not satisfy the Minimax theorem. Indeed, given any mixed strategy P of Alice, namely a distribution over the natural numbers, Bob can pick a sufficiently large natural number n such that the probability that a random number $m \sim P$ chosen by Alice satisfies $m \geq n$ is arbitrarily small. Thus, for every mixed strategy played by Alice, Bob can find a response which wins with probability arbitrarily close to 1. By symmetry, also the opposite holds: for every mixed strategy played by Bob, Alice can find a response that wins with probability arbitrarily close to 1. We show that this game is the *only* obstruction to the Minimax Theorem in the following sense:

Main Result IV (Theorem 6 and Theorem 7)
The Minimax Theorem applies to every (possibly infinite) binary-valued zero-sum game, provided that its payoff matrix does not contain triangular[3] sub-matrices of unbounded sizes.

Thus, this result identifies the size of a triangular submatrix as a combinatorial dimension which replaces the assumption that the action-sets are finite in the classical Minimax Theorem of von Neumann.

In the context of online learning, this implies that the Minimax Theorem applies to any game which corresponds to agnostic online learning for a given Littlestone class $\mathbb{C}$: i.e., the learner's strategies are hypotheses in $\mathbb{C}$ and the adversary's strategies are labelled examples. This follows due to the connection between the Littlestone dimension and the *Threshold dimension*, which implies that the payoff matrix of this game does not contain triangular submatrices of unbounded sizes (Shelah (1978), see Alon, Livni, Malliaris, and Moran (2019) for an elementary proof using learning theoretic terminology).

3. We define a triangular matrix to be 1 above the diagonal and 0 below it. Also, the sub-matrix may re-order the rows and columns to witness this triangular form.

2. Formal Definitions and Main Results

2.1. Definitions and Notation

Fix any non-empty set $\mathcal{X}$, known as the *instance space*. Also define $\mathcal{Y} = \{0, 1\}$, the *label space*. Our results will be expressed in terms of an abstract *concept class* $\mathbb{C}$ and *hypothesis class* $\mathbb{H}$, which can be any non-empty sets of *concepts*, that is, functions $h : \mathcal{X} \to \mathcal{Y}$.

For any set $\mathbb{H}$ of concepts, and any sequence $S \in (\mathcal{X} \times \mathcal{Y})^*$ or set $S \subseteq (\mathcal{X} \times \mathcal{Y})$, define $\mathbb{H}_S = \{h \in \mathbb{H} : \forall (x, y) \in S, h(x) = y\}$. We say S is $\mathbb{H}$-realizable if $\mathbb{H}_S \neq \emptyset$, and otherwise we say S is $\mathbb{H}$-unrealizable. To simplify notation, also abbreviate $\mathbb{H}_{(x,y)} = \mathbb{H}_{\{(x,y)\}}$ for $(x, y) \in \mathcal{X} \times \mathcal{Y}$.

For any sequence $x_1, x_2, \ldots$, we use the notation $x_{1:t} = (x_1, \ldots, x_t)$, or for $(x_1, y_1), (x_2, y_2), \ldots$, we write $(x_{1:t}, y_{1:t}) = ((x_1, y_1), \ldots, (x_t, y_t))$. Also, generally define $\log(z) = \max\{\ln(z), 1\}$ for $z \geq 1$.

Online Learning. An *online learning algorithm* is formally defined as a function $\mathbb{A} : (\mathcal{X} \times \mathcal{Y})^* \times \mathcal{X} \to \mathcal{Y}$, with the interpretation that for a sequence of examples $(x_1, y_1), (x_2, y_2), \ldots$, the algorithm's prediction at time t is

$$\hat{y}_t = \mathbb{A}(x_{1:(t-1)}, y_{1:(t-1)}, x_t),$$

and we say the algorithm makes a *mistake* at time t if $\hat{y}_t \neq y_t$. We will also write

$$\hat{h}_t(\cdot) = \mathbb{A}(x_{1:(t-1)}, y_{1:(t-1)}, \cdot)$$

as the function $\mathcal{X} \to \mathcal{Y}$ such that $\hat{h}_t(x) = \mathbb{A}(x_{1:(t-1)}, y_{1:(t-1)}, x)$. Generally, for any concept class $\mathbb{C}$, learning algorithm $\mathbb{A}$, and $T \in \mathbb{N}$, define the algorithm's *hypothesis class*

$$\mathbb{H}(\mathbb{C}, \mathbb{A}, T) = \{\mathbb{A}(S) : S \in (\mathcal{X} \times \mathcal{Y})^t, 0 \leq t \leq T, S \text{ is } \mathbb{C}\text{-realizable}\},$$

and also $\mathbb{H}(\mathbb{C}, \mathbb{A}) = \bigcup_T \mathbb{H}(\mathbb{C}, \mathbb{A}, T)$.

We will also discuss online learning algorithms that produce *randomized* predictors. In this case, the formal definition is a function $\mathbb{A} : (\mathcal{X} \times \mathcal{Y})^* \times \mathcal{X} \to [0, 1]$, with the interpretation that $\mathbb{A}(x_{1:(t-1)}, y_{1:(t-1)}, x_t)$ is the *probability* of predicting 1 at time t. Thus,

$$\left|\mathbb{A}(x_{1:(t-1)}, y_{1:(t-1)}, x_t) - y_t\right|$$

represents the *probability* of a mistake at time t. Clearly, a deterministic predictor is just the special case where $\mathbb{A}(x_{1:(t-1)}, y_{1:(t-1)}, x_t) \in \{0, 1\}$ always. As above, $\bar{h}_t = \mathbb{A}(x_{1:(t-1)}, y_{1:(t-1)})$ denotes the function $\mathcal{X} \to [0, 1]$ such that $\bar{h}_t(x) = \mathbb{A}(x_{1:(t-1)}, y_{1:(t-1)}, x)$, the hypothesis class of $\mathcal{A}$ is

$$\mathbb{H}(\mathbb{C}, \mathbb{A}, T) = \{\mathbb{A}(S) : S \in (\mathcal{X} \times \mathcal{Y})^t, 0 \leq t \leq T, S \text{ is } \mathbb{C}\text{-realizable}\},$$

and $\mathbb{H}(\mathbb{C}, \mathbb{A}) = \bigcup_T \mathbb{H}(\mathbb{C}, \mathbb{A}, T)$. Also define $\mathbb{H}(\mathbb{A}, T) = \mathbb{H}(\mathcal{Y}^{\mathcal{X}}, \mathbb{A}, T)$.

For any concept class $\mathbb{C}$, learning algorithm $\mathbb{A}$, and $T \in \mathbb{N}$, define the algorithm's *mistake bound*

$$\mathrm{MB}(\mathbb{C}, \mathbb{A}, T) = \max\left\{\sum_{t=1}^{T} \mathbb{1}[\mathbb{A}(x_{1:(t-1)}, y_{1:(t-1)}, x_t) \neq y_t] : (x_1, y_1), \ldots, (x_T, y_T) \text{ is } \mathbb{C}\text{-realizable}\right\}.$$

Also define $\mathrm{MB}(\mathbb{C}, \mathbb{A}) = \sup_T \mathrm{MB}(\mathbb{C}, \mathbb{A}, T)$. For any hypothesis class $\mathbb{H}$, define the *optimal mistake bound for learning* $\mathbb{C}$ *with* $\mathbb{H}$: $\mathrm{MB}(\mathbb{C}, \mathbb{H}, T) = \min\{\mathrm{MB}(\mathbb{C}, \mathbb{A}, T) : \mathbb{H}(\mathbb{C}, \mathbb{A}, T) \subseteq \mathbb{H}\}$ and $\mathrm{MB}(\mathbb{C}, \mathbb{H}) = \sup_T \mathrm{MB}(\mathbb{C}, \mathbb{H}, T)$.

Complexity Measures. We express our results in terms of well-known combinatorial complexity measures from the learning theory literature. The main quantity appearing in most of our results is the *Littlestone dimension* (Littlestone, 1988), denoted $\mathrm{L}(\mathbb{C})$, defined as the largest $n \in \mathbb{N} \cup \{0\}$ for which $\exists \{x_{\mathbf{y}} : \mathbf{y} \in \mathcal{Y}^t, t \in \{0, \ldots, n-1\}\} \subseteq \mathcal{X}$ (interpreting $\mathcal{Y}^0 = \{()\}$) with the property that $\forall y_1, \ldots, y_n \in \mathcal{Y}$, $\exists h \in \mathbb{C}$ with $(h(x_{()}), h(x_{y_1}), h(x_{y_{1:2}}), \ldots, h(x_{y_{1:(n-1)}})) = (y_1, \ldots, y_n)$. If no such largest n exists, define $\mathrm{L}(\mathbb{C}) = \infty$. Also define $\mathrm{L}(\emptyset) = -1$. When $\mathrm{L}(\mathbb{C}) < \infty$, one can show that L can equivalently be defined inductively as $\max_x \min_y \mathrm{L}(\mathbb{C}_{(x,y)}) + 1$, with $\mathrm{L}(\emptyset) = -1$ as the base case.

Another important quantity appearing in our results is the *VC dimension* (Vapnik and Chervonenkis, 1971, 1974). For any set $\mathcal{Z}$ and any non-empty set $\mathcal{F}$ of functions $\mathcal{Z} \to \{0,1\}$, the VC dimension $\mathrm{V}(\mathcal{F})$ is defined as the largest $n \in \mathbb{N} \cup \{0\}$ for which $\exists z_1, \ldots, z_n \in \mathcal{Z}$ with $\{(f(z_1), \ldots, f(z_n)) : f \in \mathcal{F}\} = \{0,1\}^n$: that is, every one of the 2^n possible binary patterns of length n can be realized by evaluating some function in $\mathcal{F}$ on the sequence $z_1, \ldots, z_n$. Define $\mathrm{V}(\mathcal{F}) = \infty$ if no such largest n exists. In particular, we will be interested both in $\mathrm{V}(\mathbb{C})$, the VC dimension of the concept class, and also in the *dual VC dimension*, $\mathrm{V}^*(\mathbb{C})$, defined as the largest $n \in \mathbb{N} \cup \{0\}$ for which $\exists h_1, \ldots, h_n \in \mathbb{C}$ with $\{(h_1(x), \ldots, h_n(x)) : x \in \mathcal{X}\} = \{0,1\}^n$, or $\mathrm{V}^*(\mathbb{C}) = \infty$ if no such largest n exists.

A final complexity measure appearing in our results is the *dual Helly number* (Bousquet, Hanneke, Moran, and Zhivotovskiy, 2020). For any non-empty sets $\mathbb{C}$ and $\mathbb{H}$ of concepts, define the *dual Helly number*, denoted $\mathrm{K}(\mathbb{C}, \mathbb{H})$, as the minimum $k \geq 2$ such that, for any $\mathbb{H}$-unrealizable set $S \subseteq (\mathcal{X} \times \mathcal{Y})$, there exists a $\mathbb{C}$-unrealizable $S' \subseteq S$ with $|S'| \leq k$. If no such k exists, define $\mathrm{K}(\mathbb{C}, \mathbb{H}) = \infty$. The dual Helly number was used to characterize the sample complexity of proper PAC learning by Bousquet, Hanneke, Moran, and Zhivotovskiy (2020). However, it also previously appeared in the literature on learning from equivalence queries, where it is known as the *strong consistency dimension* (Balcázar, Castro, Guijarro, and Simon, 2002b; Chase and Freitag, 2020).[4] It has also appeared in the literature on distributed learning under the name *co-VC dimension* (Kane, Livni, Moran, and Yehudayoff, 2019).

When $\mathbb{C}$ is clear from the context, we omit the argument $\mathbb{C}$, writing V, V^*, and L for $\mathrm{V}(\mathbb{C})$, $\mathrm{V}^*(\mathbb{C})$, and $\mathrm{L}(\mathbb{C})$, respectively, and when $\mathbb{H}$ is also clear from the context, we write K for $\mathrm{K}(\mathbb{C}, \mathbb{H})$. To avoid trivial cases, we always suppose $\mathrm{V} \geq 1$ and $\mathrm{V}^* \geq 1$ in all results.

The SOA. We will make use of an online learning algorithm originally introduced by Littlestone (1988), known as the *standard optimal algorithm*, denoted SOA. Specifically, for any hypothesis class $\mathbb{H}$, define a deterministic predictor $\mathrm{SOA}_{\mathbb{H}} : \mathcal{X} \to \mathcal{Y}$, which, for every $x \in \mathcal{X}$, predicts $\mathrm{SOA}_{\mathbb{H}}(x) = \mathrm{argmax}_{y \in \mathcal{Y}} \mathrm{L}(\mathbb{H}_{(x,y)})$. In particular, Littlestone (1988) proved that if $\mathrm{L}(\mathbb{H}) < \infty$, then every $x \in \mathcal{X}$ has at least one $y \in \mathcal{Y}$ with $\mathrm{L}(\mathbb{H}_{(x,y)}) < \mathrm{L}(\mathbb{H})$, and noted that this immediately implies that the online learning algorithm $\mathbb{A}_{\mathbb{C},\mathrm{SOA}}$ which predicts $\mathrm{SOA}_{\mathbb{C}_{\{(x_i,y_i)\}_{i=1}^{t-1}}}(x_t)$, for each $t \leq T$, has mistake bound $\mathrm{L}(\mathbb{C})$.

The result of Littlestone (1988) implies $\mathrm{MB}(\mathbb{C}, \mathcal{Y}^{\mathcal{X}}) \leq \mathrm{L}(\mathbb{C})$. However, at present there is no known simple description of the hypothesis class $\mathbb{H}(\mathbb{C}, \mathbb{A}_{\mathbb{C},\mathrm{SOA}})$ of the SOA predictor. One of the main contributions of the present work is to argue that there exists a learning algorithm $\mathbb{A}$ with a *simple* hypothesis class $\mathbb{H}(\mathbb{C}, \mathbb{A})$ which still achieves $\mathrm{MB}(\mathbb{C}, \mathbb{A}) = O(\mathrm{L}(\mathbb{C}))$. In particular, we find

4. Technically, the strong consistency dimension requires the unrealizable set to be a partial function. The two definitions only differ in the case the strong consistency dimension is 1, as the dual Helly number for $|\mathbb{C}| > 1$ is never smaller than 2 due to sets of the type $\{(x,0),(x,1)\}$.

such an algorithm with $\mathbb{H}(\mathbb{C}, \mathbb{A})$ contained in the set of *majority votes* of $O(\mathrm{V}^*)$ elements of $\mathbb{C}$. For instance, this has an important implication that the Littlestone and VC dimensions of $\mathbb{H}(\mathbb{C}, \mathbb{A})$ can be bounded in terms of $\mathrm{L}(\mathbb{C})$ and $\mathrm{V}(\mathbb{C})$: namely,

$$\mathrm{V}(\mathbb{H}(\mathbb{C}, \mathbb{A})) = O(\mathrm{V}\mathrm{V}^* \log(\mathrm{V}^*)) = O\big(\mathrm{V}^2 2^{\mathrm{V}}\big)$$

and

$$\mathrm{L}(\mathbb{H}(\mathbb{C}, \mathbb{A})) = O(\mathrm{L}\mathrm{V}^* \log(\mathrm{V}^*)) = O\big(\mathrm{L}\mathrm{V} 2^{\mathrm{V}}\big) .$$

These follow from composition theorems for the VC dimension (see Theorem 4.5 of Vidyasagar, 2003) and Littlestone dimension (Alon, Beimel, Moran, and Stemmer, 2020; Ghazi, Golowich, Kumar, and Manurangsi, 2021), together with the relation $\mathrm{V}^* < 2^{\mathrm{V}+1}$ due to Assouad (1983).

2.2. Summary of Main Results for Online Learning

We briefly summarize the main results of this work. Their detailed statements, and proofs, are presented in the sections below.

Theorem 1 *Every pair of classes* $\mathbb{C}, \mathbb{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ *satisfy* $\mathrm{MB}(\mathbb{C}, \mathbb{H}) = O(\mathrm{L}\mathrm{K} \log(\mathrm{K}))$.

Remark 2 *(Angluin, 1987; Littlestone, 1988; Balcázar, Castro, Guijarro, and Simon, 2002b) It is known that we always have* $\mathrm{MB}(\mathbb{C}, \mathbb{H}) \geq \max\{\mathrm{L}, \mathrm{K} - 1\}$*, and that there exist classes with* $\mathrm{MB}(\mathbb{C}, \mathbb{H}) = \Omega(\mathrm{L}\mathrm{K})$ *and other classes with* $\mathrm{MB}(\mathbb{C}, \mathbb{H}) = O(\max\{\mathrm{L}, \mathrm{K}\})$.

For any classifiers $h_1, \ldots, h_k$, define

$$\mathrm{Maj}(h_1, \ldots, h_k)(x) = \mathbb{1}\left[\sum_{i \leq k} h_i(x) \geq \sum_{i \leq k} (1 - h_i(x))\right] .$$

Define a hypothesis class $\mathrm{Maj}(\mathbb{C}^k) = \{\mathrm{Maj}(h_1, \ldots, h_{k'}) : 1 \leq k' \leq k, h_1, \ldots, h_{k'} \in \mathbb{C}\}$.

Theorem 3 *Every class* $\mathbb{C}$ *satisfies* $\mathrm{MB}(\mathbb{C}, \mathrm{Maj}(\mathbb{C}^{c\mathrm{V}^*})) = O(\mathrm{L})$*, where* c *is a universal constant.*

In fact, we prove a stronger result with *margins*. This can also be interpreted as a result about *randomized* predictors based on a distribution over $\mathbb{C}$. For any $k \in \mathbb{N}$ define

$$\mathrm{Vote}(h_1, \ldots, h_k)(x) = \frac{1}{k} \sum_{i \leq k} h_i(x),$$

and $\mathrm{Vote}(\mathbb{C}^k) = \{\mathrm{Vote}(h_1, \ldots, h_{k'}) : 1 \leq k' \leq k, h_1, \ldots, h_{k'} \in \mathbb{C}\}$.

Theorem 4 *For any* $\varepsilon \in (0, 1/2)$*, there is an algorithm* $\mathbb{A}$ *with* $\mathbb{H}(\mathbb{C}, \mathbb{A}) \subseteq \mathrm{Vote}(\mathbb{C}^{c\mathrm{V}^*})$ *(for a universal constant* c*) such that, for any* $T \in \mathbb{N}$*, running* $\mathbb{A}$ *on any* $\mathbb{C}$*-realizable sequence* $\{(x_1, y_1)\}_{t=1}^{T}$*, there are at most* $O\big(\frac{\mathrm{L}}{\varepsilon} \log \frac{1}{\varepsilon}\big)$ *times* $t \leq T$ *where* $|\bar{h}_t(x_t) - y_t| > \varepsilon$.

We also prove a result for *agnostic* online learning. In this case, rather than bounding the number of mistakes, we are interested in the *difference* of the number of mistakes made by $\mathbb{A}$ and the *minimum* number of mistakes made by any single $h \in \mathbb{C}$. This is known as the *regret* of the algorithm $\mathbb{A}$. It is known that any algorithm whose regret is $o(T)$ as $T \to \infty$ *must* be capable of using randomized predictors. We establish the following result.

Theorem 5 *For any $T \in \mathbb{N}$, there is an algorithm $\mathbb{A}$ with $\mathbb{H}(\mathbb{A}, T) \subseteq \mathrm{Vote}(\mathbb{C}^m)$, where $m = O\left(\frac{\mathrm{V}^* T}{\mathrm{L}\log(T/\mathrm{L})}\right)$, such that for any sequence $(x_1, y_1), \ldots, (x_T, y_T) \in \mathcal{X} \times \mathcal{Y}$,*

$$\sum_{t=1}^{T} \left|\bar{h}_t(x_t) - y_t\right| - \min_{h \in \mathbb{C}} \sum_{t=1}^{T} \mathbb{1}[h(x_t) \neq y_t] = O\left(\sqrt{\mathrm{L}T \log(T/\mathrm{L})}\right).$$

Note that $\mathbb{A}$ induces a randomized algorithm that on each round t interprets the hypothesis in $\mathrm{Vote}(\mathbb{C}^m)$ which $\mathbb{A}$ uses as a probability distribution π_t on $\mathbb{C}$ having support size at most m, and predicts $h_t(x_t)$ for a randomly drawn $h_t \sim \pi_t$. Then, the above theorem implies that the expected regret of this randomized algorithm is at most $O\left(\sqrt{\mathrm{L}T \log(T/\mathrm{L})}\right)$. This strengthen a similar result by Rakhlin, Sridharan, and Tewari (2015) who gave a non-constructive proof under further restrictions on the class $\mathbb{C}$. It also matches the bounds by Ben-David, Pál, and Shalev-Shwartz (2009); Rakhlin, Shamir, and Sridharan (2012) which were achieved by improper algorithms. Further, the above bound is tight up to the log factor. The recent work of Alon, Ben-Eliezer, Dagan, Moran, Naor, and Yogev (2021) used the non-constructive framework of Rakhlin, Sridharan, and Tewari (2015) to get an optimal bound without this log. It remains open to prove the optimal bound constructively.

2.3. When Does the Minimax Theorem Hold for VC Games?

We define a general binary-valued zero-sum game as follows. Let $\mathcal{A}$ and $\mathcal{B}$ be nonempty sets, called the *action sets*. We suppose they are each equipped with a σ-algebra defining the measurable subsets; in particular, we suppose all singleton sets $\{a\} \subseteq \mathcal{A}$, $\{b\} \subseteq \mathcal{B}$, are measurable. Let $\mathrm{val} : \mathcal{A} \times \mathcal{B} \to \{0, 1\}$ be a *value function*, assumed to be measurable in the product σ-algebra. For each $a \in \mathcal{A}$, we can interpret $\mathrm{val}(a, \cdot)$ as a function $\mathcal{B} \to \{0, 1\}$. We call $(\mathcal{A}, \mathcal{B}, \mathrm{val})$ a *VC game* if the VC dimension of $\{\mathrm{val}(a, \cdot) : a \in \mathcal{A}\}$ is finite.[5] A *subgame* of $(\mathcal{A}, \mathcal{B}, \mathrm{val})$ is any game $(\mathcal{A}', \mathcal{B}', \mathrm{val})$ where $\mathcal{A}' \subseteq \mathcal{A}$ and $\mathcal{B}' \subseteq \mathcal{B}$ are nonempty measurable subsets, and val here is interpreted as the restriction of val to $\mathcal{A}' \times \mathcal{B}'$. For any countable sequences $\{a_i\}_{i \in \mathbb{N}}$ in $\mathcal{A}$ and $\{b_i\}_{i \in \mathbb{N}}$ in $\mathcal{B}$, we say $(\{a_i : i \in \mathbb{N}\}, \{b_i : i \in \mathbb{N}\}, \mathrm{val})$ is an infinite triangular subgame if $\forall i, j \in \mathbb{N}$, $\mathrm{val}(a_i, b_j) = \mathbb{1}[i \leq j]$. Generally, for a set $\mathcal{S}$ (equipped with a σ-algebra defining the measurable subsets), denote by $\Pi(\mathcal{S})$ the set of all probability measures on $\mathcal{S}$.

Theorem 6 *A binary-valued VC game $(\mathcal{A}, \mathcal{B}, \mathrm{val})$ satisfies*

$$\inf_{P_A \in \Pi(\mathcal{A}')} \sup_{P_B \in \Pi(\mathcal{B}')} \mathop{\mathbb{E}}_{(a,b) \sim P_A \times P_B} [\mathrm{val}(a, b)] = \sup_{P_B \in \Pi(\mathcal{B}')} \inf_{P_A \in \Pi(\mathcal{A}')} \mathop{\mathbb{E}}_{(a,b) \sim P_A \times P_B} [\mathrm{val}(a, b)]$$

for all subgames $(\mathcal{A}', \mathcal{B}', \mathrm{val})$ if and only if it has no infinite triangular subgame. Moreover, this remains true even if $\Pi(\mathcal{A}')$, $\Pi(\mathcal{B}')$ are restricted to be just the probability measures having finite support.

We note that, in general, one cannot strengthen this result by replacing the $\inf$ and $\sup$ by $\min$ and $\max$. In other words, there are games for which the above result applies, but where optimal maximin and minimax strategies do not exist, as the optimal value is witnessed only in the limit.

5. We note that it follows from a known relation of Assouad (1983) between VC dimension and dual VC dimension that the VC dimension of $\{\mathrm{val}(a, \cdot) : a \in \mathcal{A}\}$ is finite if and only if the VC dimension of $\{\mathrm{val}(\cdot, b) : b \in \mathcal{B}\}$ is also finite.

One such simple example is the game *"Guess My Number"*. This game is played between two players whom we call Alice and Bob. Each of Alice and Bob privately picks a natural number, and Bob's goal is to pick the same number Alice picked. Bob wins the game if and only if they picked the same number. It is easy to see that Alice can win this game with probability arbitrarily close to 1: indeed, if Alice picks a uniform distribution over $\{1, \ldots, N\}$, then she wins with probability at least $1 - 1/N$. However, since every distribution over $\mathbb{N}$ must give a positive measure to at least one number, there is no single strategy for Alice with which she wins with probability 1.

Littlestone games. As a special case of particular importance in this work, we say $(\mathcal{A}, \mathcal{B}, \mathrm{val})$ is a *Littlestone game* if $\{\mathrm{val}(a, \cdot) : a \in \mathcal{A}\}$ has finite Littlestone dimension. Due to a well-known connection between the Littlestone dimension and the so-called *threshold dimension* it is clear that any Littlestone game has no infinite triangular subgame (Shelah, 1978; Hodges, 1997, see also Alon, Livni, Malliaris, and Moran, 2019). Thus, the following corollary is immediately entailed by Theorem 6.

Corollary 7 *Any binary-valued Littlestone game* $(\mathcal{A}, \mathcal{B}, \mathrm{val})$ *satisfies*

$$\inf_{P_A \in \Pi(\mathcal{A})} \sup_{P_B \in \Pi(\mathcal{B})} \mathop{\mathbb{E}}_{(a,b) \sim P_A \times P_B} [\mathrm{val}(a, b)] = \sup_{P_B \in \Pi(\mathcal{B})} \inf_{P_A \in \Pi(\mathcal{A})} \mathop{\mathbb{E}}_{(a,b) \sim P_A \times P_B} [\mathrm{val}(a, b)].$$

This remains true even if $\Pi(\mathcal{A})$ *and* $\Pi(\mathcal{B})$ *are restricted to just the probability measures having finite support.*

2.4. Expression of the Results in Terms of Equivalence Queries

There is a well-known correspondence between the online learning setting of Littlestone (1988) and the setting of Exact learning from Equivalence Queries introduced by Angluin (1987). In the problem of learning $\mathbb{C}$ using Equivalence Queries for $\mathbb{H} \supseteq \mathbb{C}$, there is some unknown target concept $h^* \in \mathbb{C}$, and proceeding in rounds, on each round an algorithm proposes a hypothesis $h \in \mathbb{H}$ as a query to an oracle, which then either certifies that $h = h^*$ or else returns a *counterexample* $x \in \mathcal{X}$ such that $h(x) \neq h^*(x)$. The *query complexity* $\mathrm{QC}_{\mathrm{EQ}}(\mathbb{C}, \mathbb{H})$ is defined as the minimum number q such that there is an algorithm that, for any $h^* \in \mathbb{C}$, regardless of the oracle's responses (as long as they are valid), the algorithm is guaranteed to query the oracle with $h = h^*$ within at most q queries; if no such number q exists, $\mathrm{QC}_{\mathrm{EQ}}(\mathbb{C}, \mathbb{H})$ is defined to be infinite.

It is easy to observe that $\mathrm{QC}_{\mathrm{EQ}}(\mathbb{C}, \mathbb{H}) = \mathrm{MB}(\mathbb{C}, \mathbb{H}) + 1$. To see this, note that we can always use an online learning algorithm to propose the hypotheses $h \in \mathbb{H}$, and update the learner using $(x, 1 - h(x))$ for the returned point x on rounds where the oracle returns an x. Since each such x is a mistake for the online learner, an optimal learner will need at most $\mathrm{MB}(\mathbb{C}, \mathbb{H})$ such rounds before its next hypothesis h equals h^*, in which case one final query suffices for the oracle to certify that $h = h^*$. In the other direction, given any optimal algorithm for Exact learning $\mathbb{C}$ with Equivalence Queries for $\mathbb{H}$, we can use the proposed hypothesis h as an online learner's predictor until the first time when it makes a mistake (x_t, y_t); that x_t would be a valid response from the oracle, so we may feed this into the Exact learning algorithm, which then produces its next hypothesis h, which becomes the new hypothesis for the online learner for its prediction on the point x_{t+1}, and so on until the next mistake. If the sequence $(x_1, y_1), (x_2, y_2), \ldots$ is $\mathbb{C}$-realizable, then this can continue for at most $\mathrm{QC}_{\mathrm{EQ}}(\mathbb{C}, \mathbb{H}) - 1$ rounds before the algorithm produces a hypothesis h that never makes another mistake on the sequence (it need not be equal the target concept h^* if the number of mistakes is strictly smaller than $\mathrm{QC_{EQ}}(\mathbb{C},\mathbb{H})-1$, but merely never encounters another counterexample to update on).

The problem of characterizing $\mathrm{QC_{EQ}}(\mathbb{C},\mathbb{H})$ is a classic question in the learning theory literature (e.g., Angluin, 1987, 1990; Hellerstein, Pillaipakkamnatt, Raghavan, and Wilkins, 1996; Balcázar, Castro, Guijarro, and Simon, 2002b; Chase and Freitag, 2020). For finite $\mathbb{C}$, it was shown by Balcázar, Castro, Guijarro, and Simon (2002b) that $\mathrm{K} \leq \mathrm{QC_{EQ}}(\mathbb{C},\mathbb{H}) \leq \lceil \mathrm{K}\ln(|\mathbb{C}|)\rceil$. The lower bound of Littlestone (1988) for $\mathrm{MB}(\mathbb{C},\mathcal{Y}^{\mathcal{X}})$ immediately implies a lower bound in terms of the Littlestone dimension L: namely, $\mathrm{L}+1 \leq \mathrm{QC_{EQ}}(\mathbb{C},\mathbb{H})$. Recently, Chase and Freitag (2020) established an upper bound expressed in terms of the Littlestone dimension L, which also holds for infinite classes: namely, $\mathrm{QC_{EQ}}(\mathbb{C},\mathbb{H}) \leq \mathrm{K}^{\mathrm{L}}$. They in fact show a bound where K is replaced by the sometimes-smaller *consistency dimension*; we refer the reader to that work for the details. This was the best known bound holding for all classes of a given L, with no dependence on $|\mathbb{C}|$. Thus, our Theorem 1 immediately implies a new near-optimal bound: $\mathrm{QC_{EQ}}(\mathbb{C},H) = O(\mathrm{LK}\log(\mathrm{K}))$. This resolves the optimal query complexity, up to a factor $\log(\mathrm{K})$ and unavoidable gaps, a problem which has been studied for several decades, and moreover this solution represents an exponential improvement in the dependence on L compared to the previous result of Chase and Freitag (2020).

Acknowledgments

Roi Livni is supported by the ISF, grant no. 2188/20, and by a grant from Tel Aviv University Center for AI and Data Science (TAD) in collaboration with Google, as part of the initiative of AI and DS for social good. Shay Moran is a Robert J. Shillman Fellow and is supported by the ISF, grant no. 1225/20, by an Azrieli Faculty Fellowship, and by BSF grant 2018385.

Appendix A. The Optimal Mistake Bound for Learning $\mathbb{C}$ with $\mathbb{H}$

This section presents the details of the algorithm and proof of Theorem 1.

For a finite set Q of pairs (C_i, w_i), where $C_i \subseteq \mathbb{C}$ and $w_i \geq 0$, define

$$\mathrm{Vote}(Q)(x) = \frac{\sum_{(C_i,w_i)\in Q} w_i \mathrm{SOA}_{C_i}(x)}{\sum_{(C_j,w_j)\in Q} w_j},$$

and $\mathrm{Maj}(Q)(x) = \mathbb{1}[\mathrm{Vote}(Q)(x) \geq 1/2]$. Also, for $\varepsilon \in [0,1]$, define

$$\mathrm{HighVote}(Q,\varepsilon) = \{(x, \mathrm{Maj}(Q)(x)) : x \in \mathcal{X}, \mathrm{Vote}(Q)(x) \in [0,\varepsilon] \cup [1-\varepsilon, 1]\}.$$

Consider the following online learning algorithm, for any $\mathbb{C}$ and $\mathbb{H}$ with $\mathrm{K} = \mathrm{K}(\mathbb{C},\mathbb{H}) < \infty$, executed on any $\mathbb{C}$-realizable sequence $(x_1,y_1),\ldots,(x_T,y_T)$ in $\mathcal{X}\times\mathcal{Y}$.

Initialize $Q = \{(C_1, w_1)\} = \{(\mathbb{C}, 1)\}$, $\eta = 1/(2\mathrm{K})$, $t = 1$
Repeat while $t \leq T$
  If $\mathrm{HighVote}(Q, \eta)$ is $\mathbb{H}$-realizable
    Choose $\hat{h}_t \in \mathbb{H}$ correct on $\mathrm{HighVote}(Q, \eta)$, predict $\hat{y}_t = \hat{h}_t(x_t)$
    If $\hat{y}_t \neq y_t$ (i.e., mistake)
      For each $(C_i, w_i) \in Q$
        If $\mathrm{SOA}_{C_i}(x_t) \neq y_t$, set $w_i \leftarrow \eta \cdot w_i$
        $C_i \leftarrow \{h \in C_i : h(x_t) = y_t\}$
        If $C_i = \emptyset$, remove (C_i, w_i) from Q
    $t \leftarrow t + 1$
  Else let $\{(\tilde{x}_1, \tilde{y}_1), \ldots, (\tilde{x}_{\mathrm{K}}, \tilde{y}_{\mathrm{K}})\} \subseteq \mathrm{HighVote}(Q, \eta)$ be $\mathbb{C}$-unrealizable
    For each $(C_i, w_i) \in Q$
      For each $j \leq \mathrm{K}$
        If $\mathrm{SOA}_{C_i}(\tilde{x}_j) = \tilde{y}_j$, $w_{ij} \leftarrow \eta \cdot w_i$, else $w_{ij} \leftarrow w_i$
        Let $C_{ij} = \{h \in C_i : h(\tilde{x}_j) = 1 - \tilde{y}_j\}$
    $Q \leftarrow \{(C_{ij}, w_{ij}) : C_{ij} \neq \emptyset\}$

Theorem 8 $\mathrm{MB}(\mathbb{C}, \mathbb{H}) \leq 4\mathrm{LK}\ln(2\mathrm{K})$*, achieved by the above algorithm.*

Proof Suppose L and K are both finite. Let $h^* \in \mathbb{C}$ be a concept correct on $\{(x_t, y_t)\}_{t=1}^{T}$. On each round where $\mathrm{HighVote}(Q, \eta)$ is $\mathbb{H}$-realizable and $\hat{y}_t \neq y_t$, clearly h^* is correct on the (x_t, y_t) used to update the sets C_i. Moreover, on each round where $\mathrm{HighVote}(Q, \eta)$ is not $\mathbb{H}$-realizable, the sequence $\{(\tilde{x}_j, \tilde{y}_j)\}_{j=1}^{\mathrm{K}}$ is not $\mathbb{C}$-realizable, and therefore there is at least one $\tilde{x}_j$ for which $1 - \tilde{y}_j = 1 - \mathrm{Maj}(Q)(\tilde{x}_j) = h^*(\tilde{x}_j)$. So if we think of the set Q as developing like a tree (with each element at the end of the round being updated from the previous round using the point (x_t, y_t) on rounds of the first type, or else branching off of a previous element by updating with some $(\tilde{x}_j, 1 - \tilde{y}_j)$ on rounds of the second type), then there is a path in the tree where all of the updates are for h^*-consistent examples. To put this more formally, if at the beginning of round n there exists $(C, w) \in Q$ with $h^* \in C$, then (by the above observations) at the end of round n there will still exist some $(C', w') \in Q$ with $h^* \in C'$, so that by induction we maintain this property for all rounds n. Moreover, we note that, at the end of any round n, any $(C, w) \in Q$ having $h^* \in C$ must have $w \geq \eta^{\mathrm{L}}$, since on any sequence of labeled examples $(x_i', h^*(x_i'))$, there are at most L times i with $\mathrm{SOA}_{\mathbb{C}_{\{(x_j', h^*(x_j')) : j < i\}}}(x_i') \neq h^*(x_i')$, as established by Littlestone (1988). Thus, after n rounds of the outermost loop, $\exists (C_n^*, w_n^*) \in Q$ with

$$w_n^* \geq \eta^{\mathrm{L}} = (1/(2\mathrm{K}))^{\mathrm{L}}.$$

Now suppose the algorithm executes the outermost loop at least n times, and consider the total state of the algorithm after completing round n of the outermost loop. Let t_n denote the value of t after completing this round, and let M_n denote the number of $t \in \{1, \ldots, t_n - 1\}$ with $\hat{y}_t \neq y_t$: that is, the number of mistakes on the actual data sequence within the first n rounds. Let N_n denote the number of the first n rounds where $\mathrm{HighVote}(Q, \eta)$ is not $\mathbb{H}$-realizable. Let W_n denote the total weight in Q after completing n rounds.

On any round $n' \leq n$ where $\mathrm{HighVote}(Q, \eta)$ is $\mathbb{H}$-realizable and $\hat{y}_t \neq y_t$ (for $t = t_{n'} - 1$), since every $(x, y) \in \mathrm{HighVote}(Q, \eta)$ has $\hat{h}_t(x) = y$, yet we know that $\hat{h}_t(x_t) = \hat{y}_t \neq y_t$, it must be that $(x_t, y_t) \notin \mathrm{HighVote}(Q, \eta)$; therefore at least η fraction of the total weight is multiplied by η in Step 6: that is, $W_{n'} \leq \eta^2 W_{n'-1} + (1-\eta) W_{n'-1} = (1 - \eta(1-\eta)) W_{n'-1} = \left(1 - (2\mathrm{K}-1)/(2\mathrm{K})^2\right) W_{n'-1}$. On the other hand, on any round $n' \leq n$ where $\mathrm{HighVote}(Q, \eta)$ is *not* $\mathbb{H}$-realizable, each $j \leq \mathrm{K}$ has $(\tilde{x}_j, \tilde{y}_j)$ in $\mathrm{HighVote}(Q, \eta)$, so that at least $1 - \eta$ fraction of the total of weights w_i have $w_{ij} = \eta \cdot w_i$; thus, $W_{n'} \leq \mathrm{K}(\eta(1-\eta) W_{n'-1} + \eta W_{n'-1}) = (1 - 1/(4\mathrm{K})) W_{n'-1}$. By induction, we have

$$\begin{aligned} W_n &\leq (1 - (2\mathrm{K}-1)/(2\mathrm{K})^2)^{M_n} (1 - 1/(4\mathrm{K}))^{N_n} \\ &< \exp\{-M_n (2\mathrm{K}-1)/(2\mathrm{K})^2\} \exp\{-N_n/(4\mathrm{K})\} \leq \exp\{-M_n/(4\mathrm{K})\} \exp\{-N_n/(4\mathrm{K})\} \end{aligned}$$

since $\mathrm{K} \geq 1$. Thus, since $w_n^* \leq W_n$, we have

$$M_n/(4\mathrm{K}) + N_n/(4\mathrm{K}) < \mathrm{L} \ln(2\mathrm{K}).$$

This has two important implications. First, since we always have $n = N_n + t_n - 1$, and $t_n \leq T + 1$ while the above inequality implies $N_n < 4\mathrm{KL}\ln(2\mathrm{K})$, we have that the algorithm will terminate after a finite number of rounds. Second, the above inequality further implies that $M_n < 4\mathrm{KL}\ln(2\mathrm{K})$ for all rounds in the algorithm, so that this is also a bound on the total number of mistakes at the time of termination, after predicting for all T points in the sequence. This completes the proof. ■

Appendix B. A Characterization of Games Satisfying the Minimax Theorem for All Subgames

A key component of the proofs of our results on learning with majority votes and randomized proper predictors (Theorems 3, 4, 5) is a general characterization of games for which the minimax theorem holds: namely, Theorem 6 stated Section 2.2. We present the proof here, separating the two parts of the claim.

Proposition 9 *Any binary-valued VC game* $(\mathcal{A}, \mathcal{B}, \mathrm{val})$ *with no infinite triangular subgame satisfies*

$$\inf_{P_A \in \Pi(\mathcal{A})} \sup_{P_B \in \Pi(\mathcal{B})} \mathop{\mathbb{E}}_{(a,b) \sim P_A \times P_B} [\mathrm{val}(a,b)] = \sup_{P_B \in \Pi(\mathcal{B})} \inf_{P_A \in \Pi(\mathcal{A})} \mathop{\mathbb{E}}_{(a,b) \sim P_A \times P_B} [\mathrm{val}(a,b)].$$

Moreover, this remains true even if $\Pi(\mathcal{A})$, $\Pi(\mathcal{B})$ *are restricted to be just the probability measures having finite support.*

Proof We prove this result in the contrapositive. Suppose $(\mathcal{A}, \mathcal{B}, \mathrm{val})$ is a binary-valued VC game, let

$$\alpha = \sup_{P_B \in \Pi(\mathcal{B})} \inf_{P_A \in \Pi(\mathcal{A})} \mathop{\mathbb{E}}_{(a,b) \sim P_A \times P_B} [\mathrm{val}(a,b)]$$

and let

$$\beta = \inf_{P_A \in \Pi(\mathcal{A})} \sup_{P_B \in \Pi(\mathcal{B})} \mathop{\mathbb{E}}_{(a,b) \sim P_A \times P_B} [\mathrm{val}(a,b)],$$

and let us suppose $\alpha < \beta$. Let V_A denote the VC dimension of $\{\mathrm{val}(a, \cdot) : a \in \mathcal{A}\}$ and V_B denote the VC dimension of $\{\mathrm{val}(\cdot, b) : b \in \mathcal{B}\}$. Note that these are both finite by the assumption that $(\mathcal{A}, \mathcal{B}, \mathrm{val})$ is VC game, and the fact that $\mathrm{V}_B < 2^{\mathrm{V}_A + 1}$ (Assouad, 1983).

We inductively define two sequences of mixed strategies $\{P_A^t\}_{t\in\mathbb{N}}$,$\{P_B^t\}_{t\in\mathbb{N}}$, where each $P_A^t \in \Pi(\mathcal{A})$ and each $P_B^t \in \Pi(\mathcal{B})$ such that, for a numerical constant c, each P_A^t is supported on at most $\frac{c\mathrm{V}_B}{(\beta-\alpha)^2}$ elements of $\mathcal{A}$ and has

$$\sup_{P_B \in \Pi\left(\bigcup_{i<t} \mathrm{supp}(P_B^i)\right)} \mathop{\mathbb{E}}_{(a,b)\sim P_A^t \times P_B}[\mathrm{val}(a,b)] \le \frac{2\alpha+\beta}{3} \tag{1}$$

while each P_B^t is supported on at most $\frac{c\mathrm{V}_A}{(\beta-\alpha)^2}$ elements of $\mathcal{B}$ and has

$$\inf_{P_A \in \Pi\left(\bigcup_{i\le t} \mathrm{supp}(P_A^i)\right)} \mathop{\mathbb{E}}_{(a,b)\sim P_A \times P_B^t}[\mathrm{val}(a,b)] \ge \frac{\alpha+2\beta}{3}. \tag{2}$$

As a base case, let $P_A^1 = \mathbb{1}_{\{a\}}$ for some $a \in \mathcal{A}$: that is, P_A^1 is any pure strategy. Now fix any $t \in \mathbb{N}$ and suppose there exist P_A^i, $i \in \{1,\ldots,t\}$ and P_B^i, $i \in \{1,\ldots,t-1\}$ satisfying the above properties. To complete the inductive construction, it suffices to specify P_A^t and P_B^{t+1} to extend the sequences. Letting $B_{<t} = \bigcup_{i<t} \mathrm{supp}(P_B^i)$, since this is a finite set (by the inductive hypothesis), there is a finite number of distinct sequences $\{\mathrm{val}(a,b)\}_{b\in B_{<t}}$ realized by elements $a \in \mathcal{A}$. Thus, there exists a finite set $\mathcal{A}' \subseteq \mathcal{A}$ such that every such sequence $\{\mathrm{val}(a,b)\}_{b\in B_{<t}}$ witnessed by an $a \in \mathcal{A}$ is also witnessed by $\{\mathrm{val}(a',b)\}_{b\in B_{<t}}$ for some $a' \in \mathcal{A}'$. In particular, by the classic minimax theorem for finite games (von Neumann and Morgenstern, 1944), $\exists P_A^* \in \Pi(\mathcal{A}')$ such that

$$\begin{aligned}
&\sup_{P_B\in\Pi(B_{<t})} \mathop{\mathbb{E}}_{(a,b)\sim P_A^*\times P_B}[\mathrm{val}(a,b)] = \sup_{P_B\in\Pi(B_{<t})} \inf_{P_A\in\Pi(\mathcal{A}')} \mathop{\mathbb{E}}_{(a,b)\sim P_A\times P_B}[\mathrm{val}(a,b)] \\
&= \sup_{P_B\in\Pi(B_{<t})} \inf_{P_A\in\Pi(\mathcal{A})} \mathop{\mathbb{E}}_{(a,b)\sim P_A\times P_B}[\mathrm{val}(a,b)] \le \sup_{P_B\in\Pi(\mathcal{B})} \inf_{P_A\in\Pi(\mathcal{A})} \mathop{\mathbb{E}}_{(a,b)\sim P_A\times P_B}[\mathrm{val}(a,b)] = \alpha,
\end{aligned}$$

where the second equality follows from the assumed property of $\mathcal{A}'$ and the subsequent inequality follows from the fact that every $P_B \in \Pi(B_{<t})$ is a restriction to $B_{<t}$ of some $P_B' \in \Pi(\mathcal{B})$ with $\mathrm{supp}(P_B') = \mathrm{supp}(P_B)$.

Now since $B_{<t}$ is a finite set, the classic uniform convergence property of VC classes holds (Lemma 19 of Appendix F), which in particular implies that there exists a sequence $\{a_i\}_{i\le m}$ in $\mathcal{A}'$ for some $m \le \frac{c\mathrm{V}_B}{(\beta-\alpha)^2}$ (for a universal constant c) such that, defining P_A^t as the empirical measure, i.e., $P_A^t(\cdot) = \frac{1}{m}\sum_{i=1}^m \mathbb{1}[a_i \in \cdot]$, it holds that every $b \in B_{<t}$ satisfies $\mathbb{E}_{a\sim P_A^t}[\mathrm{val}(a,b)] \le \mathbb{E}_{a\sim P_A^*}[\mathrm{val}(a,b)] + \frac{\beta-\alpha}{3}$. In particular, this implies

$$\sup_{P_B\in\Pi(B_{<t})} \mathop{\mathbb{E}}_{(a,b)\sim P_A^t\times P_B}[\mathrm{val}(a,b)] \le \alpha + \frac{\beta-\alpha}{3} = \frac{2\alpha+\beta}{3}.$$

That is, (1) holds.

Applying the same argument to the set $\mathcal{A}_{\le t} = \bigcup_{i\le t} \mathrm{supp}(P_A^i)$, implies the existence of a finite set $\mathcal{B}' \subseteq \mathcal{B}$ with $\sup_{P_A\in\Pi(A_{\le t})} \mathbb{E}_{(a,b)\sim P_A\times P_B^*}[\mathrm{val}(a,b)] \ge \beta$, and a sequence $\{b_i\}_{i\le m'}$ in $\mathcal{B}'$ for some $m' \le \frac{c\mathrm{V}_A}{(\beta-\alpha)^2}$ such that, defining P_B^{t+1} as the empirical measure, $P_B^{t+1} = \frac{1}{m'}\sum_{i=1}^{m'} \mathbb{1}[b_i \in \cdot]$, it holds that every $a \in A_{\le t}$ satisfies $\mathbb{E}_{b\sim P_B^{t+1}}[\mathrm{val}(a,b)] \ge \mathbb{E}_{b\sim P_B^*}[\mathrm{val}(a,b)] - \frac{\beta-\alpha}{3}$, which implies

$$\inf_{P_A\in\Pi(A_{\le t})} \mathop{\mathbb{E}}_{(a,b)\sim P_A\times P_B^{t+1}}[\mathrm{val}(a,b)] \ge \beta - \frac{\beta-\alpha}{3} = \frac{\alpha+2\beta}{3},$$

so that (2) holds. By the principle of induction, we have established the claimed existence of the above infinite sequences P_A^t,P_B^t.

For each $i \in \mathbb{N}$, let $m_i^A = |\text{supp}(P_A^i)|$ and $\{a_{i,1}, \dots, a_{i,m_i^A}\} = \text{supp}(P_A^i)$, and let $m_i^B = |\text{supp}(P_B^i)|$ and $\{b_{i,1}, \dots, b_{i,m_i^B}\} = \text{supp}(P_B^i)$, Now for each $i, j \in \mathbb{N}$, define two matrices: C^{ij} is a $m_i^A \times m_j^B$ matrix with entries $C_{k\ell}^{ij} = \text{val}(a_{i,k}, b_{j,\ell})$, and D^{ij} is a $m_j^A \times m_i^B$ matrix with entries $D_{k\ell}^{ij} = \text{val}(a_{j,k}, b_{i,\ell})$. Note that every $i, j \in \mathbb{N}$ have C^{ij} and D^{ij} of sizes no larger than $\frac{c\text{V}_B}{(\beta-\alpha)^2} \times \frac{c\text{V}_A}{(\beta-\alpha)^2}$. In particular, this implies that there are only a finite number of possible (C^{ij}, D^{ij}) pairs witnessed among choices of $i, j \in \mathbb{N}$. If we consider each possible (C, D) pair as a *color* for the pairs $(i, j) \in \mathbb{N}^2$ with $i < j$, the infinite Ramsey theorem implies that there exists an infinite increasing sequence $i_1, i_2, \dots$ in $\mathbb{N}$ such that $\forall (s, t), (s', t') \in \mathbb{N}^2$ with $s < t$ and $s' < t'$, we have $(C^{i_s i_t}, D^{i_s i_t}) = (C^{i_{s'} i_{t'}}, D^{i_{s'} i_{t'}})$. Let (C^*, D^*) denote this common value for the pair $(C^{i_s i_t}, D^{i_s i_t})$.

Now we claim $\exists k^*, \ell^*$ with $C_{k^*\ell^*}^* = 1$ and $D_{k^*\ell^*}^* = 0$. This is because, for each $(s, t) \in \mathbb{N}$ with $s < t$, we have $i_s < i_t$, so that

$$\begin{aligned}
&\inf_{P_A \in \Pi(\text{supp}(P_A^{i_s}))} \sup_{P_B \in \Pi(\text{supp}(P_B^{i_t}))} \mathop{\mathbb{E}}_{(a,b)\sim P_A \times P_B} [\text{val}(a, b)] \\
&\geq \inf_{P_A \in \Pi(A_{\leq i_t})} \mathop{\mathbb{E}}_{(a,b)\sim P_A \times P_B^{i_t}} [\text{val}(a, b)] \geq \frac{\alpha + 2\beta}{3},
\end{aligned}$$

while

$$\begin{aligned}
&\inf_{P_A \in \Pi(\text{supp}(P_A^{i_t}))} \sup_{P_B \in \Pi(\text{supp}(P_B^{i_s}))} \mathop{\mathbb{E}}_{(a,b)\sim P_A \times P_B} [\text{val}(a, b)] \\
&\leq \sup_{P_B \in \Pi(B_{< i_t})} \mathop{\mathbb{E}}_{(a,b)\sim P_A^{i_t} \times P_B} [\text{val}(a, b)] \leq \frac{2\alpha + \beta}{3}.
\end{aligned}$$

In other words, the value of the finite game represented by C^* is strictly greater than the value of the finite game represented by D^*. For this to be true, there must exist at least one pair k^*, ℓ^* with $C_{k^*\ell^*}^* > D_{k^*\ell^*}^*$, which (since these are binary-valued games) implies $C_{k^*\ell^*}^* = 1$ and $D_{k^*\ell^*}^* = 0$. Fix any such pair k^*, ℓ^*.

To complete the proof, we use this fact to construct an infinite triangular subgame. Define $\tilde{a}_t = a_{i_{2t-1},k^*}$ and $\tilde{b}_t = b_{i_{2t},\ell^*}$ for all $t \in \mathbb{N}$. Note that, for any $s, t \in \mathbb{N}$, if $s \leq t$, then $i_{2s-1} < i_{2t}$, so that $\text{val}(\tilde{a}_s, \tilde{b}_t) = \text{val}(a_{i_{2s-1},k^*}, b_{i_{2t},\ell^*}) = C_{k^*\ell^*}^{i_{2s-1} i_{2t}} = C_{k^*\ell^*}^* = 1$. On the other hand, if $s > t$, then $i_{2s-1} > i_{2t}$, so that $\text{val}(\tilde{a}_s, \tilde{b}_t) = \text{val}(a_{i_{2s-1},k^*}, b_{i_{2t},\ell^*}) = D_{k^*\ell^*}^{i_{2t} i_{2s-1}} = D_{k^*\ell^*}^* = 0$. Together we have $\text{val}(\tilde{a}_s, \tilde{b}_t) = \mathbb{1}[s \leq t]$ for all $s, t \in \mathbb{N}$, so that $(\{\tilde{a}_t : t \in \mathbb{N}\}, \{\tilde{b}_t : t \in \mathbb{N}\}, \text{val})$ is an infinite triangular subgame.

The final claim about restricting $\Pi(\mathcal{A})$ and $\Pi(\mathcal{B})$ to have finite support follows by noting that all probability measures P_A and P_B used in the above proof have finite support, so that if the claimed equality is violated for finite-support probability measures, then the above construction still implies the existence of an infinite triangular subgame. ∎

Proposition 10 *Any binary-valued VC game* $(\mathcal{A}, \mathcal{B}, \mathrm{val})$ *with an infinite triangular subgame* $(\mathcal{A}', \mathcal{B}', \mathrm{val})$ *satisfies*

$$1 = \inf_{P_A \in \Pi(\mathcal{A}')} \sup_{P_B \in \Pi(\mathcal{B}')} \mathop{\mathbb{E}}_{(a,b) \sim P_A \times P_B} [\mathrm{val}(a, b)] > \sup_{P_B \in \Pi(\mathcal{B}')} \inf_{P_A \in \Pi(\mathcal{A}')} \mathop{\mathbb{E}}_{(a,b) \sim P_A \times P_B} [\mathrm{val}(a, b)] = 0.$$

Proof Let $\{a_i : i \in \mathbb{N}\} = \mathcal{A}'$ and $\{b_i : i \in \mathbb{N}\} = \mathcal{B}'$ so that $\mathrm{val}(a_i, b_j) = \mathbb{1}[i \leq j]$, as guaranteed by the defining property of an infinite triangular subgame. Now note that, for any $P_A \in \Pi(\mathcal{A}')$, we have

$$\sup_{P_B \in \Pi(\mathcal{B}')} \mathop{\mathbb{E}}_{(a,b) \sim P_A \times P_B} [\mathrm{val}(a, b)] \geq \lim_{j \to \infty} \mathop{\mathbb{E}}_{a_i \sim P_A} [\mathrm{val}(a_i, b_j)] = \mathop{\mathbb{E}}_{a_i \sim P_A} \left[\lim_{j \to \infty} \mathrm{val}(a_i, b_j) \right] = 1,$$

where the first equality is due to the monotone convergence theorem, which applies since $\mathrm{val}(a_i, b_j)$ is nondecreasing in j, with limiting value 1 (achieved for all $j \geq i$). Since the values are bounded by 1, this implies the leftmost claimed equality.

Likewise, for any $P_B \in \Pi(\mathcal{B}')$, we have

$$\inf_{P_A \in \Pi(\mathcal{A}')} \mathop{\mathbb{E}}_{(a,b) \sim P_A \times P_B} [\mathrm{val}(a, b)] \leq \lim_{i \to \infty} \mathop{\mathbb{E}}_{b_j \sim P_B} [\mathrm{val}(a_i, b_j)] = \mathop{\mathbb{E}}_{b_j \sim P_B} \left[\lim_{i \to \infty} \mathrm{val}(a_i, b_j) \right] = 0,$$

where the first equality holds by the dominated convergence theorem, since $|\mathrm{val}(a_i, b_j)| \leq 1$ and $\mathrm{val}(a_i, b_j)$ has limit 0 as $i \to \infty$ (achieved for all $i > j$). Since the values are bounded below by 0, this implies the rightmost claimed equality, and this completes the proof. ■

Theorem 6 follows immediately from Propositions 9 and 10, as follows.

Proof of Theorem 6 If $(\mathcal{A}, \mathcal{B}, \mathrm{val})$ is a VC game for which some subgame $(\mathcal{A}', \mathcal{B}', \mathrm{val})$ satisfies

$$\inf_{P_A \in \Pi(\mathcal{A}')} \sup_{P_B \in \Pi(\mathcal{B}')} \mathop{\mathbb{E}}_{(a,b) \sim P_A \times P_B} [\mathrm{val}(a, b)] > \sup_{P_B \in \Pi(\mathcal{B}')} \inf_{P_A \in \Pi(\mathcal{A}')} \mathop{\mathbb{E}}_{(a,b) \sim P_A \times P_B} [\mathrm{val}(a, b)], \quad (3)$$

then Proposition 9 implies $(\mathcal{A}', \mathcal{B}', \mathrm{val})$ has an infinite triangular subgame, which is therefore also a subgame of $(\mathcal{A}, \mathcal{B}, \mathrm{val})$. By Proposition 9, this remains true even if we restrict to finite-support probability measures. For the other direction, any game $(\mathcal{A}, \mathcal{B}, \mathrm{val})$ with an infinite triangular subgame $(\mathcal{A}', \mathcal{B}', \mathrm{val})$ satisfies (3) for these $\mathcal{A}'$, $\mathcal{B}'$ by Proposition 10. This completes the proof. ■

Appendix C. Optimal Online Learning with Small Majorities

The question we address here is how simple of a hypothesis class $\mathbb{H}$ can we use while ensuring that an optimal mistake bound is still achievable (up to numerical constant factors). Here we find this is possible using $\mathbb{H}$ based on majority votes of $O(\mathrm{V}^*)$ classifiers from $\mathbb{C}$. We first prove a coarse bound achieved by the general algorithm above by showing that $\mathrm{K} = O(\mathrm{V})$ for this class $\mathbb{H}$. This yields a mistake bound $O(\mathrm{VL}\log(\mathrm{V}))$. We then refine this by a direct analysis, showing that an essentially-similar algorithm achieves a mistake bound of $O(\mathrm{L})$ with this same class $\mathbb{H}$.

C.1. Supporting Lemmas

We first establish the following helpful lemmas. The first is a simple application of the minimax theorem and the classic result on the size of ε-approximating sets for VC classes. The second is based on a technique for sparsifying majority votes having a margin, proposed by Moran and Yehudayoff (2016).

Lemma 11 *Suppose* $\mathrm{L} < \infty$. *For any* $\varepsilon \in (0, 1]$ *and any set* $S \subseteq \mathcal{X} \times \mathcal{Y}$, *if, for every finite-support probability measure* π *on* $\mathbb{C}$, *there exists* $(x, y) \in S$ *such that* $\pi(h : h(x) \neq y) > \varepsilon$, *then for every integer* $m \geq \frac{c_1 \mathrm{V}}{\varepsilon} \log\left(\frac{1}{\varepsilon}\right)$ *(for a universal constant* c_1*), there exists a sequence* S' *in* S *with* $|S'| = m$ *such that every* $h \in \mathbb{C}$ *has* $\frac{1}{m} \sum_{(x,y) \in S'} \mathbb{1}[h(x) \neq y] > \frac{\varepsilon}{2}$.

Proof By Corollary 7, there exists a finite-support probability measure P on S such that every $h \in \mathbb{C}$ has $P((x, y) : h(x) \neq y) \geq \frac{2}{3}\varepsilon$. Since P has finite support, the classic relative uniform convergence bounds of Vapnik and Chervonenkis (1974) hold (see Lemma 19 in Appendix F). This implies that for a universal constant c_1, for any $m \geq \frac{c_1 \mathrm{V}}{\varepsilon} \ln\left(\frac{1}{\varepsilon}\right)$, there exists a sequence S' of length m in S such that every $h \in \mathbb{C}$ satisfies $\frac{1}{m} \sum_{(x,y) \in S'} \mathbb{1}[h(x) \neq y] > \frac{\varepsilon}{2}$. ■

Lemma 12 *For any* $\varepsilon \in (0, 1/2)$ *and any set* $S \subseteq \mathcal{X} \times \mathcal{Y}$, *if there exists a finite-support probability measure* π *on* $\mathbb{C}$ *such that every* $(x, y) \in S$ *satisfies* $\pi(h : h(x) \neq y) \leq \varepsilon$, *then there exists a multiset* $\mathbb{C}' \subseteq \mathbb{C}$ *with* $|\mathbb{C}'| \leq \frac{c_2 \mathrm{V}^*}{\varepsilon^2}$ *(for* c_2 *a universal constant) such that every* $(x, y) \in S$ *satisfies* $\frac{1}{|\mathbb{C}'|} \sum_{h \in \mathbb{C}'} \mathbb{1}[h(x) \neq y] < 2\varepsilon$.

Proof To prove this, we apply the classic uniform convergence guarantees based on the chaining argument. Specifically, we apply Lemma 19 of Appendix F with $\mathcal{Z} = \mathbb{C}$ and $\mathcal{F}$ the set of functions $g_x : \mathbb{C} \to \{0, 1\}$, $x \in \mathcal{X}$, defined as $g_x(h) = h(x)$ for $h \in \mathbb{C}$. The existence of $\mathbb{C}'$ with the claimed properties then follows immediately from Lemma 19 by noting that $\mathrm{V}(\mathcal{F}) = \mathrm{V}^*(\mathbb{C})$. ■

C.2. The Dual Helly Number of Small Majorities

To start, we state a coarse bound based on a direct application of Theorem 8, by bounding the dual Helly number.

Proposition 13 *For a numerical constant* c *and* $\mathbb{H} = \mathrm{Maj}(\mathbb{C}^{c\mathrm{V}^*})$, *it holds that* $\mathrm{K} = O(\mathrm{V})$, *and hence* $\mathrm{MB}(\mathbb{C}, \mathbb{H}) = O(\mathrm{LV}\log(\mathrm{V}))$.

Proof The statement is vacuous for $\mathrm{L} = \infty$, so suppose $\mathrm{L} < \infty$. Let $c = 16c_2$, for c_2 from Lemma 12. Let S be a set not realizable by $\mathbb{H}$. In particular, applying Lemma 12 with $\varepsilon = 1/4$, we conclude that for every finite-support probability measure π on $\mathbb{C}$, there exists $(x, y) \in S$ such that $\pi(h : h(x) \neq y) > 1/4$. Lemma 11 then implies there exists $S' \subseteq S$ with $|S'| \leq \lceil 4c_1 \mathrm{V} \log(4) \rceil \leq 7c_1 \mathrm{V}$, such that every $h \in \mathbb{C}$ has $\frac{1}{|S'|} \sum_{(x,y) \in S'} \mathbb{1}[h(x) \neq y] > \frac{1}{8}$. In particular, this implies S' is not realizable by $\mathbb{C}$. Thus, $\mathrm{K} \leq 7c_1 \mathrm{V}$. A bound $\mathrm{MB}(\mathbb{C}, \mathbb{H}) \leq 28c_1 \mathrm{LV} \ln(14c_1 \mathrm{V})$ then follows from Theorem 8. ■

C.3. Optimal Mistake Bound via a Direct Analysis

Proposition 13 reveals that it is possible to learn any $\mathbb{C}$ using majorities of $O(\mathrm{V}^*)$ elements of $\mathbb{C}$, with mistake bound at most $O(\mathrm{LV}\log(\mathrm{V}))$. Here we find that this can be improved to an essentially *optimal* mistake bound: that is, $O(\mathrm{L})$. To put this another way, we find that the optimal SOA predictor of Littlestone (1988) can be approximated (in the appropriate sense) by small majority votes of classifiers from $\mathbb{C}$.

The following is a more-detailed form of Theorem 3, from which the original statement of Theorem 3 immediately follows.

Theorem 14 *For $c = 36c_2$ and $\mathbb{H} = \mathrm{Maj}(\mathbb{C}^{c\mathrm{V}^*})$, it holds that $\mathrm{MB}(\mathbb{C}, \mathbb{H}) \leq 80\mathrm{L}$.*

Theorem 14 will immediately follow from a result stated in the following section, which presents a stronger result where we are guaranteed not merely a small number of mistakes, but also a small number of points where the majority vote fails to have *high margin*: namely, Theorem 15. Specifically, Theorem 14 follows by plugging $\varepsilon = 1/3$ into Theorem 15.

Appendix D. Online Learning with Votes of Large Margin

Since $\mathrm{K}(\mathbb{C}, \mathbb{C})$ is sometimes large or infinite, proper learning isn't always viable; however, since Proposition 13 indicates $\mathrm{K}(\mathbb{C}, \mathbb{H})$ is small for $\mathbb{H}$ the set of majority votes of $O(\mathrm{V}^*)$ classifiers of $\mathbb{C}$, we can define predictors that, for each time t, *sample* a classifier $\hat{h}_t \sim \pi_t$ for some distribution π_t over $\mathbb{C}$, and will be correct with probability greater than $1/2$ against an adversary that only knows π_t before selecting the next (x_t, y_t), but does not know the specific $\hat{h}_t$ sampled from π_t. We may however be interested in having even greater probability of predicting correctly, defining a loss function that is 1 if $\pi_t(h : h(x_t) \neq y_t) > \varepsilon$ for some given $\varepsilon \in (0, 1/2)$. We would then be interested in bounding the number of times t for which this occurs.

Equivalently, we can interpret this criterion in terms of the *margin* of the majority vote classifier: that is, we can think of the learning algorithm as outputting the *conditional mean* $\bar{h}_t(x) := \pi_t(h : h(x_t) = 1)$, and we are interested in bounding the number of times t where $|\bar{h}_t(x_t) - y_t| > \varepsilon$.

The following is a more-detailed restatement of Theorem 4, from which the statement of Theorem 4 immediately follows.

Theorem 15 *Let $c = 4c_2$ (for c_2 from Lemma 12). For any $\varepsilon \in (0, 1/2)$, there is an algorithm $\mathbb{A}$ with $\mathbb{H}(\mathbb{C}, \mathbb{A}) \subseteq \mathrm{Vote}(\mathbb{C}^{c\mathrm{V}^*/\varepsilon^2})$ such that, for any $T \in \mathbb{N}$, running $\mathbb{A}$ on any $\mathbb{C}$-realizable sequence $(x_1, y_1), \ldots, (x_T, y_T)$, there are at most $\frac{8\mathrm{L}}{\varepsilon(1-\varepsilon/8)} \ln\left(\frac{8}{\varepsilon}\right)$ times t where $|\bar{h}_t(x_t) - y_t| > \varepsilon$.*

This bound will be achieved by the following algorithm, which takes as input $\mathbb{C}$ and any value $\varepsilon \in (0, 1/2)$, and processes a $\mathbb{C}$-realizable adversarial sequence $(x_1, y_1), \ldots, (x_T, y_T)$.

0. Initialize $Q = \{(C_1, w_1)\} = \{(\mathbb{C}, 1)\}$, $\eta = \varepsilon/8$, $m = \lceil \frac{2c_1\mathrm{V}}{\varepsilon} \log(\frac{2}{\varepsilon}) \rceil$, $t = 1$
1. Repeat while $t \leq T$
2. If $\exists h \in \mathrm{Vote}(\mathbb{C}^{c\mathrm{V}^*/\varepsilon^2})$ with $\sup_{(x,y)\in\mathrm{HighVote}(Q,\varepsilon/8)} |h(x) - y| \leq \varepsilon$
3. Choose $\bar{h}_t = h$ for some such h, predict $\bar{y}_t = \bar{h}_t(x_t)$
4. If $|\bar{y}_t - y_t| > \varepsilon$
5. For each $(C_i, w_i) \in Q$
6. If $\mathrm{SOA}_{C_i}(x_t) \neq y_t$, set $w_i \leftarrow \eta \cdot w_i$
7. $C_i \leftarrow \{h \in C_i : h(x_t) = y_t\}$
8. If $C_i = \emptyset$, remove (C_i, w_i) from Q
9. $t \leftarrow t + 1$
10. Else let $\{(\tilde{x}_1, \tilde{y}_1), \ldots, (\tilde{x}_m, \tilde{y}_m)\} \subseteq \mathrm{HighVote}(Q, \varepsilon/8)$ be such that every $h \in \mathbb{C}$ has $\frac{1}{m}\sum_{i=1}^{m} \mathbb{1}[h(x_i) \neq y_i] > \varepsilon/4$
11. For each $(C_i, w_i) \in Q$
12. For each $j \leq m$
13. If $\mathrm{SOA}_{C_i}(\tilde{x}_j) = \tilde{y}_j$, set $w_{ij} \leftarrow \eta \cdot w_i$, else $w_{ij} \leftarrow w_i$
14. Let $C_{ij} = \{h \in C_i : h(\tilde{x}_j) = 1 - \tilde{y}_j\}$
15. $Q \leftarrow \{(C_{ij}, w_{ij}) : C_{ij} \neq \emptyset\}$

We now present the proof of Theorem 15.

Proof of Theorem 15 The proof is similar to the proof of Theorem 8, with a few important changes. Suppose $\mathrm{L} < \infty$, let $h^* \in \mathbb{C}$ be a concept correct on $\{(x_t, y_t)\}_{t=1}^{T}$, and let $\mathbb{H} = \mathrm{Vote}(\mathbb{C}^{c\mathrm{V}^*/\varepsilon^2})$. First note that, by Lemma 12, on any given round, if the condition in Step 2 fails, then for every finite-support probability measure π on $\mathbb{C}$, there exists $(x, y) \in \mathrm{HighVote}(Q, \varepsilon/8)$ such that $\pi(h : h(x) \neq \mathrm{Maj}(Q)(x)) > \varepsilon/2$. Thus, the existence of the sequence $(\tilde{x}_1, \tilde{y}_1), \ldots, (\tilde{x}_m, \tilde{y}_m)$ in Step 10 is guaranteed by Lemma 11.

On each round where the condition in Step 2 holds but $|\bar{y}_t - y_t| > \varepsilon$, we clearly have that h^* is correct on the (x_t, y_t) used to update the sets C_i. Moreover, on each round where the condition in Step 2 fails, since $h^* \in \mathbb{C}$, the defining property of $(\tilde{x}_1, \tilde{y}_1), \ldots, (\tilde{x}_m, \tilde{y}_m)$ in Step 10 guarantees that at least $(\varepsilon/4)m$ of these $(\tilde{x}_j, \tilde{y}_j)$ have $h^*(\tilde{x}_j) = 1 - \tilde{y}_j$. So if we think of the set Q as developing like a tree (with each element at the end of the round being updated from the previous round using the point (x_t, y_t) on rounds of the first type, or else branching off of a previous element by updating with one of the $(\tilde{x}_j, 1 - \tilde{y}_j)$ examples on rounds of the second type), then there are a number of paths in the tree where all of the corresponding examples (x, y) have $y = h^*(x)$. More formally, if at the beginning of a round there exist q elements $(C, w) \in Q$ with $h^* \in C$, then on rounds where the condition in Step 2 holds, at the end of the round there will still be exactly q such elements in Q (i.e., none of them are removed in Step 8, since all contain h^*); on the other hand, on rounds where the condition in Step 2 fails, then since at least $(\varepsilon/4)m$ of the examples $(\tilde{x}_j, \tilde{y}_j)$ have $h^*(\tilde{x}_j) = 1 - \tilde{y}_j$, at the end of the round there will be at least $(\varepsilon/4)mq$ elements $(C', w') \in Q$ with $h^* \in C'$.

Now suppose the algorithm executes the outermost loop at least n times, and consider the total state of the algorithm after completing round n of the outermost loop. Let t_n denote the value of t after completing this round, and let M_n denote the number of $t \in \{1, \ldots, t_n - 1\}$ with $|\bar{y}_t - y_t| > \varepsilon$. Let N_n denote the number of the first n rounds for which the condition in Step 2 fails. Let W_n denote the total of the weights in Q after round n, and define $W_0 = 1$.

Applying the above argument inductively, we have that after completing round n of the outermost loop, there are at least $((\varepsilon/4)m)^{N_n}$ elements $(C, w) \in Q$ with $h^* \in C$. Moreover, we note that any $(C, w) \in Q$ with $h^* \in C$ must have $w \geq \eta^{\mathrm{L}}$, since on any sequence of labeled examples $(x'_i, h^*(x'_i))$, there are at most L times i with $\mathrm{SOA}_{\mathbb{C}_{\{(x'_j, h^*(x'_j)): j<i\}}}(x'_i) \neq h^*(x'_i)$, as established by Littlestone (1988). It follows that after completing round n we have

$$W_n \geq ((\varepsilon/4)m)^{N_n} \cdot \eta^{\mathrm{L}}.$$

On the other hand, on every round $n' \leq n$ where the condition in Step 2 holds but $|\bar{y}_t - y_t| > \varepsilon$ in Step 4 (for $t = t_{n'} - 1$), note that it cannot be that $(x_t, y_t) \in \mathrm{HighVote}(Q, \varepsilon/8)$, since this would contradict either $|\bar{y}_t - y_t| > \varepsilon$ or the defining property of $\bar{h}_t$. Thus, since $(x_t, y_t) \notin \mathrm{HighVote}(Q, \varepsilon/8)$, at least $\varepsilon/8$ fraction of the total weight is multiplied by η: that is, $W_{n'} \leq W_{n'-1}\left(\eta(\varepsilon/8) + (1 - (\varepsilon/8))\right)$. Furthermore, on every round where the condition in Step 2 fails, for each $j \leq m$, since $(\tilde{x}_j, \tilde{y}_j) \in \mathrm{HighVote}(Q, \varepsilon/8)$, it must be that at least $1 - (\varepsilon/8)$ fraction of the total weights w_i from the previous round will have $w_{ij} = \eta \cdot w_i$. Together with the growth by a factor of m from branching, we have that on such rounds $n' \leq n$, $W_{n'} \leq W_{n'-1} m \left(\left(1 - \frac{\varepsilon}{8}\right)\eta + \frac{\varepsilon}{8}\right)$. By induction we have that, after round n,

$$W_n \leq \left(\eta\frac{\varepsilon}{8} + 1 - \frac{\varepsilon}{8}\right)^{M_n} m^{N_n} \left(\left(1 - \frac{\varepsilon}{8}\right)\eta + \frac{\varepsilon}{8}\right)^{N_n}.$$

Combining the upper and lower bounds, we have

$$\left(\frac{\varepsilon}{4}m\right)^{N_n} \cdot \eta^{\mathrm{L}} \leq \left(\eta\frac{\varepsilon}{8} + 1 - \frac{\varepsilon}{8}\right)^{M_n} m^{N_n} \left(\left(1 - \frac{\varepsilon}{8}\right)\eta + \frac{\varepsilon}{8}\right)^{N_n}.$$

Plugging in $\eta = \varepsilon/8$ we have

$$\begin{aligned}\left(\frac{\varepsilon}{4}m\right)^{N_n} \cdot \left(\frac{\varepsilon}{8}\right)^{\mathrm{L}} &\leq \left(1 - \frac{\varepsilon}{8}\left(1 - \frac{\varepsilon}{8}\right)\right)^{M_n} \left(\left(1 - \frac{\varepsilon}{16}\right)\frac{\varepsilon}{4}m\right)^{N_n} \\ &\leq \exp\left\{-\frac{\varepsilon}{8}\left(1 - \frac{\varepsilon}{8}\right)M_n\right\} \cdot \left(\left(1 - \frac{\varepsilon}{16}\right)\frac{\varepsilon}{4}m\right)^{N_n}.\end{aligned}$$

Taking logarithms of both sides and simplifying, we have

$$M_n + N_n \frac{8}{\varepsilon\left(1 - \varepsilon/8\right)} \ln\left(\frac{1}{1 - \varepsilon/16}\right) \leq \frac{8\mathrm{L}}{\varepsilon(1 - \varepsilon/8)} \ln\left(\frac{8}{\varepsilon}\right).$$

This has two important implications. First, since we always have $n = N_n + t_n - 1$, and $t_n \leq T + 1$ while the above inequality implies $N_n \leq \mathrm{L}\frac{\ln(8/\varepsilon)}{\ln(1/(1-\varepsilon/16))} < \infty$, we may conclude that the algorithm will terminate after a finite number of rounds. Second, the above inequality further implies $M_n \leq \frac{8\mathrm{L}}{\varepsilon(1-\varepsilon/8)} \ln\left(\frac{8}{\varepsilon}\right)$ for all rounds n in the algorithm, so that this also bounds the total number of times $t \leq T$ with $|\bar{y}_t - y_t| > \varepsilon$. This completes the proof. ■

Remark 16 *We also remark that the above algorithm can actually be executed with* any *online learning algorithm* $\mathbb{A}$ *in place of* SOA, *resulting in a conversion to a method using votes of* $O(\mathrm{V}^*/\varepsilon^2)$ *concepts in* $\mathbb{C}$, *and having a number of rounds with* $|\bar{h}_t(x_t) - y_t| > \varepsilon$ *at most* $O\big(\mathrm{MB}(\mathbb{C}, \mathbb{A}, T)\frac{1}{\varepsilon}\log\frac{1}{\varepsilon}\big)$.

Appendix E. Near-Optimal Agnostic Online Learning with Randomized Proper Predictors

This section presents the details of the algorithm and proof for Theorem 5. Recall the statement of the theorem, as follows.

Theorem 5 (Restated) For any $T \in \mathbb{N}$, there is an algorithm $\mathbb{A}$ with $\mathbb{H}(\mathbb{A}, T) \subseteq \mathrm{Vote}(\mathbb{C}^m)$, where $m = O\left(\frac{\mathrm{V}^* T}{\mathrm{L} \log(T/\mathrm{L})}\right)$, such that for any sequence $(x_1, y_1), \ldots, (x_T, y_T) \in \mathcal{X} \times \mathcal{Y}$,

$$\sum_{t=1}^{T} \left|\bar{h}_t(x_t) - y_t\right| - \min_{h \in \mathbb{C}} \sum_{t=1}^{T} \mathbb{1}[h(x_t) \neq y_t] = O\left(\sqrt{\mathrm{L} T \log(T/\mathrm{L})}\right).$$

Before presenting the proof, let us first note that the algorithm we propose is, to a large extent, constructive, in the sense that on each round t it constructs a probability measure p_t on $\mathbb{C}$, based on $(x_1, y_1), \ldots, (x_{t-1}, y_{t-1})$, so that $\bar{h}_t(x_t) = p_t(h : h(x_t) = 1)$. Thus, we may regard this method as a *randomized proper* learning algorithm, in the sense that on each round, if we draw a random $h_t \sim p_t$ (independently, given the data sequence) then $\mathbb{E} \sum_{t=1}^{T} \mathbb{1}[h_t(x_t) \neq y_t] = \sum_{t=1}^{T} \left|\bar{h}_t(x_t) - y_t\right|$, so that the *expected regret* of these sampled classifiers is at most $O\left(\sqrt{\mathrm{L} T \log(T/\mathrm{L})}\right)$.

As discussed above, this result can be interpreted as a more constructive version of a result of Rakhlin, Sridharan, and Tewari (2015). The result in Theorem 5 is also more general, as it removes additional restrictions on $\mathbb{C}$ required by Rakhlin, Sridharan, and Tewari (2015).

Turning now to the task of proving Theorem 5, we will rely on the following Lemma from Ben-David, Pál, and Shalev-Shwartz (2009). For any sequence $x_1, \ldots, x_t$, abbreviate $x_{1:t} = (x_1, \ldots, x_t)$.

Lemma 17 (Ben-David, Pál, and Shalev-Shwartz, 2009, Lemma 12) *For any concept class $\mathbb{C}$ of Littlestone dimension L, for any $T \in \mathbb{N}$, there exists a family $\mathbb{G}_T := \{g_I : I \subseteq \{1, \ldots, T\}, |I| \leq \mathrm{L}\}$ of functions $\mathcal{X}^* \to \mathcal{Y}$ such that, for every $x_1, \ldots, x_T \in \mathcal{X}$,*

$$\{(h(x_1), \ldots, h(x_T)) : h \in \mathbb{C}\} = \{(g_I(x_1), g_I(x_{1:2}), \ldots, g_I(x_{1:T})) : g_I \in \mathbb{G}_T\}.$$

Specifically, Ben-David, Pál, and Shalev-Shwartz (2009) construct this family $\mathbb{G}_T$ by letting $g_I(x_{1:t})$ be the prediction of $\mathrm{SOA}_{H(t,I)}(x_t)$, where $H(0, I) = \mathbb{C}$, and $H(t, I)$ is inductively defined as $H(t, I) = \{h \in H(t-1, I) : h(x_t) = 1 - \mathrm{SOA}_{H(t-1,I)}(x_t))\}$ if $t \in I$ and $\{h \in H(t-1, I) : h(x_t) = 1 - \mathrm{SOA}_{H(t-1,I)}(x_t))\} \neq \emptyset$, and $H(t, I) = H(t-1, I)$ otherwise. The property guaranteed by Lemma 17 then follows from the L mistake bound of Littlestone (1988) for SOA: that is, the g_I that agrees with h on $x_{1:T}$ simply takes I as the (at most L) times when SOA_H would make mistakes on $(x_1, h(x_1)), \ldots, (x_T, h(x_T))$, given that it updates $H \leftarrow \{h' \in H : h'(x_t) = h(x_t)\}$ after each mistake $(x_t, h(x_t))$.

We will also make use of a classic result for learning from expert advice for the absolute loss (Vovk, 1990, 1992; Littlestone and Warmuth, 1994; Cesa-Bianchi, Freund, Haussler, Helmbold, Schapire, and Warmuth, 1997; Kivinen and Warmuth, 1999; Singer and Feder, 1999); see Theorem 2.2 of Cesa-Bianchi and Lugosi (2006).

Lemma 18 *(Cesa-Bianchi and Lugosi, 2006, Theorem 2.2) For any $N, T \in \mathbb{N}$ and $f_1, \ldots, f_N$ functions $\mathcal{X}^* \to [0, 1]$, letting $\eta = \sqrt{(8/T)\ln(N)}$, for any $(x_1, y_1), \ldots, (x_T, y_T) \in \mathcal{X} \times [0, 1]$,* *letting* $w_{0,i} = 1$ *and* $w_{t,i} = e^{-\eta \sum_{s \le t} |f_i(x_{1:s}) - y_s|}$ *for each* $t \le T$, $i \le N$, *letting* $\bar{f}_t(x_{1:t}, y_{1:(t-1)}) = \sum_i w_{t-1,i} f_i(x_{1:t}) / \sum_{i'} w_{t-1,i'}$, *it holds that*

$$\sum_{t=1}^{T} \left| \bar{f}_t(x_{1:t}, y_{1:(t-1)}) - y_t \right| - \min_{1 \le i \le N} \sum_{t=1}^{T} |f_i(x_{1:t}) - y_t| \le \sqrt{(T/2)\ln(N)}.$$

We are now ready for the proof of Theorem 5.

Proof of Theorem 5 If $T < 10\mathrm{L}$, the regret bound is $\ge T$ (for appropriate constants), which trivially holds, so let us suppose $T \ge 10\mathrm{L}$. Fix a value $\varepsilon = \sqrt{(\mathrm{L}/T)\log(eT/\mathrm{L})}$ and let $\mathbb{A}$ be the algorithm guaranteed by Theorem 15: that is, for any $\mathbb{C}$-realizable sequence $(x_1, y_1), \ldots, (x_T, y_T)$, for each $t \le T$, $\mathbb{A}$ proposes a finite-support probability measure π_t on $\mathbb{C}$ (namely, a uniform distribution on at most $O(\mathrm{V}^*/\varepsilon^2)$ elements of $\mathbb{C}$) defined based solely on $(x_1, y_1), \ldots, (x_{t-1}, y_{t-1})$, and there are at most $O\big(\frac{\mathrm{L}}{\varepsilon}\log\frac{1}{\varepsilon}\big)$ times t for which $|\pi_t(h : h(x_t) = 1) - y_t| > \varepsilon$.

Now consider running $\mathbb{A}$ for the sequence $(x_1, g_I(x_1)), (x_2, g_I(x_{1:2})), \ldots, (x_T, g_I(x_{1:T}))$ for $g_I \in \mathbb{G}_T$, where $\mathbb{G}_T$ is from Lemma 17. In particular, Lemma 17 guarantees this is a $\mathbb{C}$-realizable sequence. Let π_t^I denote the corresponding probability measure that would be proposed on each round t, and let $\bar{h}_t^I(x) = \pi_t^I(h : h(x) = 1)$.

Let $h^* = \mathrm{argmin}_{h \in \mathbb{C}} \sum_{t=1}^{T} \mathbb{1}[h(x_t) \ne y_t]$. By Lemma 17, there exists $g_{I^*} \in \mathbb{G}_T$ with $g_{I^*}(x_{1:t}) = h^*(x_t)$ for every $t \le T$. We therefore have

$$\sum_{t=1}^{T} \mathbb{1}\left[\left|\bar{h}_t^{I^*}(x_t) - h^*(x_t)\right| > \varepsilon\right] = O\left(\frac{\mathrm{L}}{\varepsilon}\log\frac{1}{\varepsilon}\right).$$

In particular, this immediately implies

$$\sum_{t=1}^{T} \left|\bar{h}_t^{I^*}(x_t) - h^*(x_t)\right| \le \varepsilon T + O\left(\frac{\mathrm{L}}{\varepsilon}\log\frac{1}{\varepsilon}\right). \tag{4}$$

Let $\mathcal{I}_T = \{I \subseteq \{1, \ldots, T\} : |I| \le \mathrm{L}\}$. We will now treat the predictors $\bar{h}_t^I$, $I \in \mathcal{I}_T$, as *experts* in the traditional framework of learning from expert advice for the absolute loss. Let $N = \binom{T}{\le \mathrm{L}} = \sum_{j=0}^{\mathrm{L}} \binom{T}{j}$. Following Lemma 18, let $\eta = \sqrt{(8/T)\ln(N)}$, and for each $I \in \mathcal{I}_T$ let $w_{0,I} = 1$ and $w_{t,I} = e^{-\eta \sum_{s \le t} |\bar{h}_s^I(x_s) - y_s|}$ for each $t \le T$. Finally, define $\bar{h}'_t(x) = \sum_{I \in \mathcal{I}_T} w_{t-1,I} \bar{h}_t^I(x) / \sum_{I \in \mathcal{I}_T} w_{t-1,I}$.

By Lemma 18, we have

$$\begin{aligned} \sum_{t=1}^{T} \left|\bar{h}'_t(x_t) - y_t\right| &\le \sqrt{(T/2)\ln(N)} + \min_{I \in \mathcal{I}_T} \sum_{t=1}^{T} \left|\bar{h}_t^I(x_t) - y_t\right| \\ &\le \sqrt{\mathrm{L}(T/2)\ln\left(\frac{eT}{\mathrm{L}}\right)} + \sum_{t=1}^{T} \left|\bar{h}_t^{I^*}(x_t) - y_t\right|. \end{aligned}$$

By the triangle inequality, the rightmost expression is at most

$$\sqrt{\mathrm{L}(T/2)\ln\left(\frac{eT}{\mathrm{L}}\right)} + \sum_{t=1}^{T} \left|\bar{h}_t^{I^*}(x_t) - h^*(x_t)\right| + \sum_{t=1}^{T} |h^*(x_t) - y_t|.$$

Combining this with (4) and plugging in $\varepsilon = \sqrt{(\mathrm{L}/T)\log(eT/\mathrm{L})}$ yields

$$\sum_{t=1}^{T} \left|\bar{h}'_t(x_t) - y_t\right| - \sum_{t=1}^{T} |h^*(x_t) - y_t| = O\left(\sqrt{\mathrm{L}T \log\left(\frac{T}{\mathrm{L}}\right)}\right).$$

The claimed regret now follows by noting that $\bar{h}'_t(x) = p'_t(h : h(x) = 1)$ by choosing $p'_t = \sum_{I \in \mathcal{I}_T} \frac{w_{t-1,I}}{\sum_{I' \in \mathcal{I}_T} w_{t-1,I'}} \pi_t^I$. Since each π_t^I is a finite-support probability measure on $\mathbb{C}$, p'_t is also a finite-support probability measure on $\mathbb{C}$.

To conclude, we note that the probability measures p'_t in the above proof may have supports of size up to $\tilde{O}\left(\mathrm{V}^* \left(\frac{eT}{\mathrm{L}}\right)^{\mathrm{L}+1}\right)$. However, we can apply the same sparsification argument used in previous sections: that is, since p'_t has finite support, we can apply the classic result on the size of ε-approximating sets (Lemma 19 of Appendix F), which implies there exists $\bar{h}_t \in \mathrm{Vote}(\mathbb{C}^m)$ with $m = O\left(\frac{\mathrm{V}^*}{\varepsilon^2}\right)$ such that $\sup_x |\bar{h}_t(x) - p'_t(h : h(x) = 1)| \leq \varepsilon$, so that using $\bar{h}_t$ on each round, rather than $\bar{h}'_t$, by the triangle inequality we have

$$\begin{aligned}
&\sum_{t=1}^{T} \left|\bar{h}_t(x_t) - y_t\right| - \min_{h \in \mathbb{C}} \sum_{t=1}^{T} |h(x_t) - y_t| \\
&\leq \varepsilon T + \sum_{t=1}^{T} \left|\bar{h}'_t(x_t) - y_t\right| - \min_{h \in \mathbb{C}} \sum_{t=1}^{T} |h(x_t) - y_t| = \varepsilon T + O\left(\sqrt{\mathrm{L}T \log(T/\mathrm{L})}\right).
\end{aligned}$$

Taking $\varepsilon = \sqrt{(\mathrm{L}/T)\log(T/\mathrm{L})}$ then retains the regret bound $O\left(\sqrt{\mathrm{L}T \log(T/\mathrm{L})}\right)$ required by Theorem 5, with $\bar{h}_t \in \mathrm{Vote}(\mathbb{C}^m)$ for $m = O\left(\frac{\mathrm{V}^* T}{\mathrm{L}\log(T/\mathrm{L})}\right)$. ■

Ackgnoweledgements This work has received support from the Israeli Science Foundation (ISF) grant no. 2188/20, and by a grant from Tel Aviv University Center for AI and Data Science (TAD) in collaboration with Google, as part of the initiative of AI and DS for social good.

References

N. Agarwal, B. Bullins, E. Hazan, S. M. Kakade, and K. Singh. Online control with adversarial disturbances. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

N. Alon, R. Livni, M. Malliaris, and S. Moran. Private PAC learning implies finite Littlestone dimension. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, 2019.

N. Alon, A. Beimel, S. Moran, and U. Stemmer. Closure properties for private classification and online prediction. In *Proceedings of the 33rd Conference on Learning Theory*, 2020.

N. Alon, O. Ben-Eliezer, Y. Dagan, S. Moran, M. Naor, and E. Yogev. Adversarial laws of large numbers and optimal regret in online classification. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, 2021.

O. Anava, E. Hazan, S. Mannor, and O. Shamir. Online learning for time series prediction. In *Proceedings of the 26th Conference on Learning Theory*, 2013.

D. Angluin. Queries and concept learning. *Machine Learning*, 2(4):319–342, 1987.

D. Angluin. Negative results for equivalence queries. *Machine Learning*, 5:121–150, 1990.

D. Angluin. Queries revisited. *Theoretical Computer Science*, 313:175–194, 2004.

D. Angluin and T. Dohrn. The power of random counterexamples. *Theor. Comput. Sci.*, 808:2–13, 2020.

P. Assouad. Densité et dimension. *Annales de l'Institut Fourier (Grenoble)*, 33(3):233–282, 1983.

J. L. Balcázar, J. Castro, and D. Guijarro. A new abstract combinatorial dimension for exact learning via queries. *Journal of Computer and System Sciences*, 64:2–21, 2002a.

J. L. Balcázar, J. Castro, D. Guijarro, and H. Simon. The consistency dimension and distribution-dependent learning from queries. *Theoretical Computer Science*, 288(2):197–215, 2002b.

José L. Balcázar, Jorge Castro, and David Guijarro. A general dimension for exact learning. In *14th Conference on Learning Theory*, 2001.

S. Ben-David, D. Pál, and S. Shalev-Shwartz. Agnostic online learning. In *Proceedings of the 22nd Conference on Learning Theory*, 2009.

O. Bousquet, S. Hanneke, S. Moran, and N. Zhivotovskiy. Proper learning, Helly number, and an optimal SVM bound. In *Proceedings of the* $33^{\rm rd}$ *Conference on Learning Theory*, 2020.

E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

N. Cesa-Bianchi, Y. Freund, D. Haussler, D. P. Helmbold, R. E. Schapire, and M. K. Warmuth. How to use expert advice. *Journal of the Association for Computing Machinery*, 44(3):427–485, 1997.

H. Chase and J. Freitag. Bounds in query learning. In *Proceedings of the 33rd Conference on Learning Theory*, 2020.

A. Daniely and S. Shalev-Shwartz. Optimal learners for multiclass problems. In *Proceedings of the* $27^{\rm th}$ *Conference on Learning Theory*, 2014.

A. Daniely, S. Sabato, S. Ben-David, and S. Shalev-Shwartz. Multiclass learnability and the ERM principle. *Journal of Machine Learning Research*, 16(12):2377–2404, 2015.

B. Ghazi, N. Golowich, R. Kumar, and P. Manurangsi. Near-tight closure bounds for Littlestone and threshold dimensions. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, 2021.

S. Hanneke. The optimal sample complexity of PAC learning. *Journal of Machine Learning Research*, 17(38):1–15, 2016.

D. Haussler, N. Littlestone, and M. Warmuth. Predicting $\{0,1\}$-functions on randomly drawn points. *Information and Computation*, 115(2):248–292, 1994.

Y. Hayashi, S. Matsumoto, A. Shinohara, and M. Takeda. Uniform characterizations of polynomial-query learnabilities. *Theor. Comput. Sci.*, 292(2):377–385, 2003.

E. Hazan and T. Ma. A non-generative framework and convex relaxations for unsupervised learning. In *Advances in Neural Information Processing Systems 29*, 2016.

E. Hazan, R. Livni, and Y. Mansour. Classification with low rank and missing data. In *Proceedings of the 32nd International Conference on Machine Learning*, 2015.

E. Hazan, S. Kale, and S. Shalev-Shwartz. Near-optimal algorithms for online matrix prediction. *SIAM J. Comput.*, 46(2):744–773, 2017.

L. Hellerstein, K. Pillaipakkamnatt, V. Raghavan, and D. Wilkins. How many queries are needed to learn? *Journal of the Association for Computing Machinery*, 43(5):840–862, 1996.

W. Hodges. *A Shorter Model Theory*. Cambridge University Press, New York, NY, USA, 1997.

D. Kane, R. Livni, S. Moran, and A. Yehudayoff. On communication complexity of classification problems. In *Proceedings of the* 32^{nd} *Conference on Learning Theory*, pages 1903–1943, 2019.

J. Kivinen and M. K. Warmuth. Averaging expert predictions. In *Proceedings of the* 4^{th} *European Conference on Computational Learning Theory*, 1999.

N. Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2:285–318, 1988.

N. Littlestone and M. K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2):212–261, 1994.

O. Montasser, S. Hanneke, and N. Srebro. VC classes are adversarially robustly learnable, but only improperly. In *Proceedings of the* 32^{nd} *Conference on Learning Theory*, 2019.

S. Moran and A. Yehudayoff. Sample compression schemes for VC classes. *Journal of the ACM*, 63(3):1–10, 2016.

A. Rakhlin, O. Shamir, and K. Sridharan. Relax and randomize: from value to algorithms. In *Advances in Neural Information Processing Systems 25*, 2012.

A. Rakhlin, K. Sridharan, and A. Tewari. Online learning via sequential complexities. *Journal of Machine Learning Research*, 16(1):155–186, 2015.

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

A. Singer and M. Feder. Universal linear prediction by model order weighting. *IEEE Transactions on Signal Processing*, 47(10):2685–2699, 1999.

N. Srebro, J. Rennie, and T. S. Jaakkola. Maximum-margin matrix factorization. In *Advances in Neural Information Processing Systems 18*, pages 1329–1336, 2005.

M. Talagrand. Sharper bounds for gaussian and empirical processes. *The Annals of Probability*, 22: 28–76, 1994.

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996.

V. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, Inc., 1998.

V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.

V. Vapnik and A. Chervonenkis. *Theory of Pattern Recognition*. Nauka, Moscow, 1974.

M. Vidyasagar. *Learning and Generalization with Applications to Neural Networks*. Springer-Verlag, 2$^{\text{nd}}$ edition, 2003.

J. von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, 1944.

V. Vovk. Aggregating strategies. In *Proceedings of the* 3$^{\text{rd}}$ *Annual Workshop on Computational Learning Theory*, 1990.

V. Vovk. Universal forecasting algorithms. *Information and Computation*, 96(2):245–277, 1992.

Appendix F. Uniform Convergence Bounds for VC Classes

Lemma 19 *There is a universal constant c_0 such that, for any set $\mathcal{Z}$ and any finitely-supported probability measure P on $\mathcal{Z}$, for any set $\mathcal{F}$ of functions $\mathcal{Z} \to \{0,1\}$, for $m \in \mathbb{N}$ and $\varepsilon \in (0,1]$, if $m \geq \frac{c_0 \mathrm{V}(\mathcal{F})}{\varepsilon} \log \frac{1}{\varepsilon}$, then there exists $z_1, \ldots, z_m \in \mathcal{Z}$ such that every $f \in \mathcal{F}$ with $P(z : f(z) = 1) \geq \frac{2}{3}\varepsilon$ has $\frac{1}{m}\sum_{t=1}^{m} f(z_t) > \frac{\varepsilon}{2}$. Moreover, if $m \geq \frac{c_0 \mathrm{V}(\mathcal{F})}{\varepsilon^2}$, then there exists $z_1, \ldots, z_m \in \mathcal{Z}$ such that every $f \in \mathcal{F}$ satisfies $\left|P(z : f(z) = 1) - \frac{1}{m}\sum_{t=1}^{m} f(z_t)\right| < \varepsilon$.*

Proof The first claim follows from the relative uniform convergence bounds of Vapnik and Chervonenkis (1974), which hold without further restrictions on $\mathcal{F}$ since P has finite support; see Theorem 4.4 of Vapnik (1998). Specifically, Theorem 4.4 of Vapnik (1998) guarantees that (for an appropriate constant c_0), for $(Z_1, \ldots, Z_m) \sim P^m$, with nonzero probability, every $f \in \mathcal{F}$ with $P(z : f(z) = 1) \geq \frac{2}{3}\varepsilon$ has $\frac{1}{m}\sum_{t=1}^{m} f(Z_t) > \frac{\varepsilon}{2}$, which then implies there exist at least one such sequence $z_1, \ldots, z_m$ satisfying the claim. Similarly, for the second claim, by the classic uniform convergence guarantees based on the chaining argument (see Talagrand, 1994; van der Vaart and Wellner, 1996), which again hold without further restrictions to $\mathcal{F}$ because P has finite support, it holds that, if $m \geq \frac{c_0 \mathrm{V}(\mathcal{F})}{\varepsilon^2}$ (for an appropriate constant c_0), for $(Z_1, \ldots, Z_m) \sim P^m$, with nonzero probability, every $f \in \mathcal{F}$ satisfies $\left|P(z : f(z) = 1) - \frac{1}{m}\sum_{t=1}^{m} f(z_t)\right| < \varepsilon$. In particular, this implies there must exist at least one such sequence $z_1, \ldots, z_m$ satisfying the claim. ■