Proceedings of Machine Learning Research vol 125:1–28, 2020 33rd Annual Conference on Learning Theory

Proper Learning, Helly Number, and an Optimal SVM Bound

Olivier Bousquet OBOUSQUET@GOOGLE.COM
Google Research, Brain Team

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

Shay Moran SHAYMORAN@GOOGLE.COM
Google Research, Brain Team

Nikita Zhivotovskiy ZHIVOTOVSKIY@GOOGLE.COM
Google Research, Brain Team

Editors: Jacob Abernethy and Shivani Agarwal

Abstract

The classical PAC sample complexity bounds are stated for any Empirical Risk Minimizer (ERM) and contain an extra multiplicative logarithmic factor $\log \frac{1}{\varepsilon}$ which is known to be necessary for ERM in general. It has been recently shown by Hanneke (2016a) that the optimal sample complexity of PAC learning for any VC class $\mathbb{C}$ does not include this log factor and is achieved by a particular *improper learning algorithm*, which outputs a specific majority-vote of hypotheses in $\mathbb{C}$. This leaves the question of when this bound can be achieved by *proper learning algorithms*, which are restricted to always output a hypothesis from $\mathbb{C}$.

In this paper we aim to characterize the classes for which the optimal sample complexity can be achieved by a proper learning algorithm. We identify that these classes can be characterized by the *dual Helly number*, which is a combinatorial parameter that arises in discrete geometry and abstract convexity. In particular, under general conditions on $\mathbb{C}$, we show that the dual Helly number is bounded if and only if there is a proper learner that obtains the optimal dependence on ε.

As further implications of our techniques we resolve a long-standing open problem posed by Vapnik and Chervonenkis (1974) on the performance of the *Support Vector Machine* in $\mathbb{R}^n$ by proving that the sample complexity of SVM in the realizable case is

$$\Theta\left(\frac{n}{\varepsilon} + \frac{1}{\varepsilon}\log\frac{1}{\delta}\right).$$

This gives the first optimal PAC bound for Halfspaces in $\mathbb{R}^n$ achieved by a proper learning algorithm, and moreover is computationally efficient.

Keywords: Statistical Learning Theory, PAC Learning, Sample Complexity, Proper Learning, SVM.

1. Introduction

In the literature on the theory of PAC learning, there has been much work discussing the important distinction between *proper* vs *improper* learning algorithms, where a proper learner is required to output a hypothesis from the concept class being learned, while an improper learner may output any classifier, not necessarily in the class. Most of this literature has focused on the *computational* separations between proper and improper learning (see e.g., Kearns and Vazirani, 1994). However,

it is also interesting to consider the effect on *sample complexity* of proper vs improper learning. While the optimal sample complexity of PAC learning was recently resolved by Hanneke (2016a), the proposed learning algorithm is *improper*: constructing its classifier based on a majority vote of well-chosen classifiers from the concept class. Furthermore, it follows from arguments analogous to the work of Daniely and Shalev-Shwartz (2014) that the optimal sample complexity of PAC learning is sometimes *not achievable* by proper learners (see also our Theorem 11). While the question of characterizing the best sample complexity achievable by proper learners has been resolved for several special-case concept classes (e.g., Auer and Ortner, 2007; Darnstädt, 2015; Hanneke, 2016b), the best known *general* results on the sample complexity of proper learning in the prior literature are still only the results that hold for *all* empirical risk minimization (ERM) algorithms (Vapnik and Chervonenkis, 1974; Blumer, Ehrenfeucht, Haussler, and Warmuth, 1989; Hanneke, 2016b; Zhivotovskiy and Hanneke, 2018). However, it is well known that there are many classes where specific proper learners can achieve better sample complexities (by a log factor) than the worst ERM learner (Auer and Ortner, 2007). Thus, it is important to go beyond the analysis of general ERM learners if we are to understand the best sample complexity achievable by proper learners in general.

In the present work, we aim to provide such a general analysis of the sample complexity of proper learning, applicable to *every* concept class, by identifying the relevant combinatorial complexity measures of the class. We specifically find that a quantity called the *dual Helly number* (previously proposed by Kane, Livni, Moran, and Yehudayoff, 2019 under the name *coVC dimension*) is of critical importance. In particular, when the dual Helly number is finite, the logarithmic factor in the well-known sample complexity bounds for ERM (Vapnik and Chervonenkis, 1974) may be replaced by a bounded quantity. The proper learner achieving this bound is a variant of the optimal PAC learner of Hanneke (2016a), but modified in several steps so that it remains proper.

As a further implication of the techniques we develop, we find that in the case of learning Halfspaces in $\mathbb{R}^n$, the well-known *support vector machine* (SVM) learning algorithm achieves the optimal sample complexity $\Theta\left(\frac{n}{\varepsilon} + \frac{1}{\varepsilon}\log\frac{1}{\delta}\right)$. This resolves a question that appeared in the seminal work of Vapnik and Chervonenkis (1974). Moreover, this also provides the first proof that Halfspaces are properly learnable with the optimal sample complexity: that is, sample complexity of the form $\frac{n}{\varepsilon} + \frac{1}{\varepsilon}\log\frac{1}{\delta}$. As a further implication, we find that Maximum classes of any given VC dimension d are also properly learnable with optimal sample complexity $\Theta\left(\frac{\mathrm{d}}{\varepsilon} + \frac{1}{\varepsilon}\log\frac{1}{\delta}\right)$.

The known results on sample complexity are summarized in Figure 1 along with a (rough) statement of our new results.

Notation. To begin the formal discussion, we introduce some basic notation. Fix a space $\mathcal{X}$ equipped with a σ-algebra specifying the measurable subsets. Let $\mathcal{Y} = \{-1, 1\}$ denote the *label space*. A *classifier* is any measurable function $h : \mathcal{X} \to \mathcal{Y}$, and a *concept class* is any set $\mathbb{C}$ of classifiers. To focus on nontrivial cases here, we will always suppose $|\mathbb{C}| \geq 3$. A *learning algorithm* $\mathcal{A}$ maps any sequence (data set) $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ in $\mathcal{X} \times \mathcal{Y}$, of any length n, to a classifier $\hat{h}_n$; the map $\mathcal{A}$ may include randomization. A learning algorithm $\mathcal{A}$ is called *proper* (for $\mathbb{C}$) if $\hat{h}_n$ is always an element of $\mathbb{C}$, for all possible data sets. Otherwise $\mathcal{A}$ is called *improper*.

In the PAC learning problem, there is a *data distribution* $\mathcal{P}$ (a probability measure on $\mathcal{X}$), and a *target concept* $f^\star \in \mathbb{C}$. For any classifier h, define $\mathrm{er}_{\mathcal{P}}(h; f^\star) = \mathcal{P}(x : h(x) \neq f^\star(x))$. When $\mathcal{P}$ and $f^\star$ are clear from the context, we simply write $\mathrm{er}(h)$. In contexts where $\mathcal{P}$ is specified, we let $X_1, X_2, \ldots$ denote an i.i.d. sequence of $\mathcal{P}$-distributed random variables. Define $a \wedge b = \min\{a, b\}$ for $a, b \in \mathbb{R}$. Generally, for any sequence $x_1, x_2, \ldots$ and any $f : \mathcal{X} \to \mathcal{Y}$, we use the notation

Bounds on the sample complexity of PAC learning		
Improper Learning	$\Theta\left(\frac{\mathrm{d}}{\varepsilon} + \frac{1}{\varepsilon}\log\frac{1}{\delta}\right)$	Hanneke, 2016a Ehrenfeucht et al., 1989
Any ERM	$O\left(\frac{\mathrm{d}}{\varepsilon}\log(\frac{1}{\varepsilon} \wedge \frac{\mathfrak{s}}{\mathrm{d}}) + \frac{1}{\varepsilon}\log\frac{1}{\delta}\right)$ $\Omega\left(\frac{\mathrm{d}}{\varepsilon} + \frac{1}{\varepsilon}\log(\frac{1}{\varepsilon} \wedge \mathfrak{s}) + \frac{1}{\varepsilon}\log\frac{1}{\delta}\right)$	Hanneke, 2016b Vapnik and Chervonenkis, 1974
Proper Learning	$O\left(\frac{\mathrm{dk}^2}{\varepsilon}\log(\mathrm{k}) + \frac{\mathrm{k}^2}{\varepsilon}\log\frac{1}{\delta}\right)$ $\Omega\left(\frac{\mathrm{d}}{\varepsilon} + \frac{1}{\varepsilon}\log(\mathrm{k}) + \frac{1}{\varepsilon}\log\frac{1}{\delta}\right)$	New results in this work.
SVM / Halfspaces in $\mathbb{R}^n$	$\Theta\left(\frac{n}{\varepsilon} + \frac{1}{\varepsilon}\log\frac{1}{\delta}\right)$	New result in this work.
Maximum Class (Proper)	$\Theta\left(\frac{\mathrm{d}}{\varepsilon} + \frac{1}{\varepsilon}\log\frac{1}{\delta}\right)$	New result in this work.

Figure 1: Summary of results on the sample complexity of (ε, δ)-PAC learning, along with our new results. d denotes the *VC dimension* (Vapnik and Chervonenkis, 1971), $\mathfrak{s}$ the *star number* (Hanneke and Yang, 2015), and k the *dual Helly number* (Kane, Livni, Moran, and Yehudayoff, 2019) discussed in this article. Specific definitions, conditions, and ranges of parameters for which the results hold are discussed below.

$1:n = \{1, \ldots, n\}$, $x_{1:n} = \{x_1, \ldots, x_n\}$, and $(x_{1:n}, f(x_{1:n})) = \{(x_1, f(x_1)), \ldots, (x_n, f(x_n))\}$. The *sample complexity*, the central quantity of study in this work, is defined as follows.

Definition 1 *For any $\varepsilon, \delta \in (0,1)$, the* sample complexity *of (ε, δ)-PAC learning, denoted $\mathcal{M}(\varepsilon, \delta)$, is defined as the smallest $n \in \mathbb{N}$ for which there exists a learning algorithm $\mathcal{A}$ such that, for every data distribution $\mathcal{P}$ and every $f^\star \in \mathbb{C}$, the (random) classifier $\hat{h}_n = \mathcal{A}((X_{1:n}, f^\star(X_{1:n})))$ satisfies*

$$\Pr\Big(\mathrm{er}(\hat{h}_n) \leq \varepsilon\Big) \geq 1 - \delta.$$

The sample complexity of (ε, δ)-PAC proper *learning, denoted by $\mathcal{M}_{\mathrm{prop}}(\varepsilon, \delta)$, is defined identically, except that the learning algorithm $\mathcal{A}$ is required to be* proper*: it always outputs an element of $\mathbb{C}$.*

A fundamental quantity in characterizing the sample complexity is the *VC dimension* (Vapnik and Chervonenkis, 1971). We say $\mathbb{C}$ *shatters* a sequence of points $x_{1:n} \in \mathcal{X}^n$ if $\forall y_{1:n} \in \mathcal{Y}^n$, $\exists h \in \mathbb{C}$ with $h(x_{1:n}) = y_{1:n}$. The VC dimension of $\mathbb{C}$, denoted by d, is the largest $n \in \mathbb{N}$ for which there exists a sequence $x_{1:n}$ shattered by $\mathbb{C}$; otherwise if no such largest n exists, define $\mathrm{d} = \infty$.

The sample complexity of (unrestricted) PAC learning was recently proven by Hanneke (2016a) to satisfy $\mathcal{M}(\varepsilon, \delta) = \Theta\left(\frac{\mathrm{d}}{\varepsilon} + \frac{1}{\varepsilon}\log\left(\frac{1}{\delta}\right)\right)$, resolving a gap between the previously-known lower bound of this form (from Vapnik and Chervonenkis, 1974; Ehrenfeucht, Haussler, Kearns, and Valiant, 1989) and previous suboptimal upper bounds (Vapnik and Chervonenkis, 1974; Blumer, Ehrenfeucht, Haussler, and Warmuth, 1989; Haussler, Littlestone, and Warmuth, 1994; Simon, 2015). However, the optimal learning algorithm proposed by Hanneke (2016a) is *improper*.

Most of the work on the sample complexity of proper learning is based on the fact (due to Vapnik and Chervonenkis, 1974, Blumer, Ehrenfeucht, Haussler, and Warmuth, 1989) that any learning algorithm that outputs any $\hat{h} \in \mathbb{C}$ making no mistakes on the training data (called *empirical risk*

minimization, or ERM) is guaranteed to achieve a sample complexity $O\big(\frac{\mathrm{d}}{\varepsilon}\log\big(\frac{1}{\varepsilon}\big)+\frac{1}{\varepsilon}\log\big(\frac{1}{\delta}\big)\big)$. This bound has been refined in some special cases (Hanneke, 2016b; Zhivotovskiy and Hanneke, 2018), but it is known that it cannot generally be improved while still holding for all ERM learners (Blumer and Littlestone, 1989; Haussler, Littlestone, and Warmuth, 1994; Auer and Ortner, 2007). On the other hand, for some special types of concept classes, it was observed that $\mathcal{M}_{\mathrm{prop}}(\varepsilon,\delta)=\Theta\big(\frac{\mathrm{d}}{\varepsilon}+\frac{1}{\varepsilon}\log\big(\frac{1}{\delta}\big)\big)$: that is, proper learning is sometimes optimal. For instance, this was shown for any *intersection-closed* concept class, where the optimal sample complexity is achieved by a proper learner known as the *Closure* algorithm (Auer and Ortner, 2007; Darnstädt, 2015; Hanneke, 2016b). For the class of Halfspaces on $\mathbb{R}^n$, Vapnik and Chervonenkis (1974) found that the support vector machine (SVM) classifier (which is a proper learner) achieves the optimal dependence on d and ε for obtaining expected error ε, and they essentially asked the question of whether the exact optimal form $\Theta\big(\frac{\mathrm{d}}{\varepsilon}+\frac{1}{\varepsilon}\log\big(\frac{1}{\delta}\big)\big)$ for PAC learning is achieved by SVM. This question has remained open since then, with a number of works investigating the question (e.g., Blumer and Littlestone, 1989; Balcan and Long, 2013; Zhivotovskiy, 2017; Hanneke and Kontorovich, 2019; Long and Long, 2020). We answer this question affirmatively, finding that indeed the SVM classifier achieves the optimal sample complexity for learning Halfspaces.

Related results are known for the multi-class setting (where we may have $|\mathcal{Y}|>2$, or even infinite $\mathcal{Y}$). In this case, Daniely, Sabato, Ben-David, and Shalev-Shwartz (2015) showed that different ERMs may have strikingly different sample complexities, and Daniely and Shalev-Shwartz (2014) showed that there exist classes (with $|\mathcal{Y}|=\infty$) that are learnable but *not* properly learnable (i.e., $\mathcal{M}(\varepsilon,\delta)<\infty$ but $\mathcal{M}_{\mathrm{prop}}(\varepsilon,\delta)=\infty$). A similar separation between proper and improper learnability was also recently shown by Montasser, Hanneke, and Srebro (2019) for the problem of learning with adversarial robustness guarantees. Of course, these kinds of striking separations cannot happen in the binary classification setting ($|\mathcal{Y}|=2$) studied here, since the aforementioned result of Vapnik and Chervonenkis (1974) shows that ERM learners obtain sample complexities that are at most suboptimal by a factor $O(\log\frac{1}{\varepsilon})$. However, the argument used in the proofs of Daniely and Shalev-Shwartz (2014) and Montasser, Hanneke, and Srebro (2019) can be adapted to show that this logarithmic factor is sometimes *necessary*: that is, that there exist classes of any given VC dimension d for which $\mathcal{M}_{\mathrm{prop}}(\varepsilon,\delta)=\Omega(\frac{\mathrm{d}}{\varepsilon}\log\frac{1}{\varepsilon}+\frac{1}{\varepsilon}\log\frac{1}{\delta})$.[1]

One of the aims of this paper is to connect these scattered observations and explain the properties of the class $\mathbb{C}$ responsible for the optimality or sub-optimality of proper PAC learners.

Structure of the paper and main contributions

- Section 2 contains the definition of the dual Helly number and two variants of it: the *hollow star* number and the *projection number*. We show that these three parameters coincide whenever they are finite, and present some (infinite) classes where their values can differ.
- Section 3 contains an upper bound for proper learning classes with bounded dual Helly number. Specifically, we show that if the *projection number* is bounded, there is a particular ERM having a better sample complexity than arbitrary ERM. We also provide a proper learning algorithm achieving an even more-improved sample complexity, but this proper algorithm is not necessarily an ERM.
- Section 4 contains a lower bound on the sample complexity of proper algorithms when the hollow star number is large.

1. This result is unpublished, and essentially folklore, discovered independently by several different people familiar with the argument of Daniely and Shalev-Shwartz (2014).

- Section 5 presents a new upper bound for *stable compression schemes*: compression schemes whose choice of compression set is unaffected by removing points not in the compression set. In particular, as SVM can be viewed as a stable compression scheme, our result implies that SVM requires only $O\left(\frac{n}{\varepsilon}+\frac{1}{\varepsilon}\log\frac{1}{\delta}\right)$ examples in order to (ε,δ)-PAC learn n-dimensional halfspaces. This resolves a long-standing open problem from Vapnik and Chervonenkis (1974). As a further implication of our general result for stable compression schemes, we also find that all Maximum classes are properly learnable with optimal sample complexity.

2. The dual Helly number

Helly's Theorem is a fundamental result about convex sets (Helly, 1923). It asserts that for any finite sequence of convex sets $C_1 \dots C_m \subseteq \mathbb{R}^n$ such that $\cap_{i=1}^m C_i = \emptyset$ there is a subsequence $C_{i_1} \dots C_{i_k}$, with $k \leq n+1$ such that $\cap_{j=1}^k C_{i_j} = \emptyset$. This notion has been studied more abstractly in various settings (see, e.g., Levi, 1951; Danzer, Grünbaum, and Klee, 1963); it is defined in an abstract manner as follows: let $\mathcal{F}$ be a family of subsets over a domain $\mathcal{X}$. The *Helly number* of $\mathcal{F}$ is the minimum integer k such that whenever $\mathcal{C} \subseteq \mathcal{F}$ is a collection of sets whose intersection is empty then there is a subset $\mathcal{C}' \subseteq \mathcal{C}$ of size at most k whose intersection is empty. That is, the empty intersection of the entire collection $\mathcal{C}$ is witnessed by a subset of size at most k.

We adapt the Helly number (in a dual form) to our context, obtaining a parameter[2] critical to the proper sample complexity. For any $S \subseteq \mathcal{X} \times \mathcal{Y}$, define $\mathbb{C}[S] = \{h \in \mathbb{C} : \forall (x,y) \in S, h(x) = y\}$.

Definition 2 (The dual Helly number) *Define the dual Helly number of $\mathbb{C}$, denoted by k_w, as the smallest integer k such that, for any $S \subseteq \mathcal{X} \times \mathcal{Y}$ such that $\mathbb{C}[S] = \emptyset$, there is a set $W \subseteq S$ with $|W| \leq k$ such that $\mathbb{C}[W] = \emptyset$. That is, for any unrealizable set of examples, there is an unrealizable subset of size at most k. If no such k exists, we define $\mathrm{k}_w = \infty$.*

Observe that the dual Helly number is precisely the Helly number of the following family $\mathcal{F}$: for each $(x,y) \in \mathcal{X} \times \mathcal{Y}$ let $\mathbb{C}_{(x,y)} := \{h \in \mathbb{C} : h(x) = y\}$, and let $\mathcal{F} := \{\mathbb{C}_{(x,y)} : (x,y) \in \mathcal{X} \times \mathcal{Y}\}$. This definition is also related to the notion of a *teaching set* (Goldman and Kearns, 1995): recall that $W \subseteq X$ is a teaching set for h with respect to $\mathbb{C}$ if there exists no $h' \in \mathbb{C} \setminus \{h\}$ which agrees with h on W. In particular, observe that any $h \notin \mathbb{C}$ has a teaching set with respect to $\mathbb{C}$ of size at most k_w.

We proceed with the second definition. We will need the following notation: two sequences (or samples) $(x_1,y_1),\dots,(x_k,y_k) \in \mathcal{X} \times \mathcal{Y}$ and $(x'_1,y'_1),\dots,(x'_k,y'_k) \in \mathcal{X} \times \mathcal{Y}$ are said to be *neighbors* if $x_i = x'_i$ for all $i \in 1:k$ and there exists exactly one $j \in 1:k$ such that $y_j \neq y'_j$.

Definition 3 (The hollow star number) *Define the* hollow star number *of $\mathbb{C}$, denoted by k_o, as the largest integer k such that there is a sequence $S = ((x_1,y_1),\dots(x_k,y_k)) \in (\mathcal{X} \times \mathcal{Y})^k$ which is <u>not</u> realizable by $\mathbb{C}$ (i.e. $\mathbb{C}[S] = \emptyset$), however every sequence S' which is a neighbor of S is realizable by $\mathbb{C}$ (i.e., $\mathbb{C}[S'] \neq \emptyset$). If no such largest k exists, define $\mathrm{k}_o = \infty$.*

We refer to any unrealizable sequence $S \in (\mathcal{X} \times \mathcal{Y})^*$, such that every neighbor S' of S is realizable, as a *hollow star set*. Thus, k_o is the size of the largest finite hollow star set, or ∞

2. We note that an equivalent parameter was introduced by the name "coVC dimension" by Kane, Livni, Moran, and Yehudayoff (2019), where it was used to characterize proper learning in a distributed setting.

if there exist hollow star sets of unbounded finite sizes. Equivalently, a hollow star set $S = \{(x_1, y_1), \ldots, (x_k, y_k)\}$ satisfies $\mathbb{C}[S] = \emptyset$ and $\forall i \leq k, \exists h_i \in \mathbb{C}$ s.t. $\{j : h_i(x_j) \neq y_j\} = \{i\}$.

The name *hollow star number* is chosen to stress the similarity with the *star number* defined by Hanneke and Yang (2015). The *star number* $\mathfrak{s}$ is the maximum size of a realizable sequence S such that every sequence S' which is a neighbor of S is also realizable. Thus, the definitions of star number and hollow star number differ *only* in whether S is required to be realizable or unrealizable.

The star number $\mathfrak{s}$ was defined by Hanneke and Yang (2015) to characterize the PAC sample complexity of active learning, and was shown by Hanneke (2016b) to also characterize the general rate of convergence of ERM (see Section 2.2 below). Interestingly, it can be shown (Hanneke and Yang, 2015) that the star number upper bounds the size of the *teaching set* of any classifier. It also has many other connections to various quantities arising in the learning theory literature (Hanneke and Yang, 2015; Hanneke, 2016b). From the definitions, we immediately have the simple inequality:

$$\mathrm{k}_o - 1 \leq \mathfrak{s}.$$

However, as we discuss below, while the classes $\mathbb{C}$ having $\mathfrak{s} < \infty$ are very limited, classes with $\mathrm{k}_o < \infty$ are far more common (e.g., Halfspaces). Thus, this one small difference in the definition, requiring the star's *center* S to be unrealizable, significantly impacts the value of the quantity.

Our final definition is slightly more involved but it will play a key role in our upper bound. For a finite (multiset) $\mathbb{C}' \subseteq \mathbb{C}$ let $\mathrm{Majority}(\mathbb{C}') : \mathcal{X} \to \{0, 1, ?\}$ denote the majority-vote classifier:

$$\mathrm{Majority}(\mathbb{C}')(\mathrm{x}) = \begin{cases} 0 & \left|\{c \in \mathbb{C}' : c(x) = 0\}\right| > \frac{|\mathbb{C}'|}{2}, \\ 1 & \left|\{c \in \mathbb{C}' : c(x) = 1\}\right| > \frac{|\mathbb{C}'|}{2}, \\ ? & \text{else.} \end{cases}$$

For $\ell \geq 2$, define the set $\mathcal{X}_{\mathbb{C}',\ell} \subseteq \mathcal{X}$ of all the points x on which less than $\frac{1}{\ell}$-fraction of all classifiers in $\mathbb{C}'$ disagree with the majority. That is, letting $h_{\mathrm{maj}} = \mathrm{Majority}(\mathbb{C}')$,

$$\mathcal{X}_{\mathbb{C}',\ell} = \left\{ x \in \mathcal{X} : \sum_{h \in \mathbb{C}'} \mathbb{1}[h(x) \neq h_{\mathrm{maj}}(x)] < \frac{|\mathbb{C}'|}{\ell} \right\}, \tag{1}$$

Definition 4 (The projection number) *Define the* projection number *of* $\mathbb{C}$*, denoted by* k_p*, as the smallest integer* $k \geq 2$ *such that, for any finite multiset* $\mathbb{C}' \subseteq \mathbb{C}$ *there exists* $h \in \mathbb{C}$ *that agrees with* $\mathrm{Majority}(\mathbb{C}')$ *on the entire set* $\mathcal{X}_{\mathbb{C}',k}$*. If no such integer* k *exists, define* $\mathrm{k}_p = \infty$.

This definition allows us to "project" the majority vote of any classifiers in $\mathbb{C}$ to the class $\mathbb{C}$. Define

$$\mathrm{Proj}_{\mathbb{C}}(\mathbb{C}') \text{ is any element in } \left\{ h \in \mathbb{C} : h(x) = \mathrm{Majority}(\mathbb{C}'), \text{ for all } x \in \mathcal{X}_{\mathbb{C}',\mathrm{k}_p} \right\}.$$

The set used in this definition is always non-empty (by definition of k_p) when $\mathrm{k}_p < \infty$.

Following Kane, Livni, Moran, and Yehudayoff (2019), we say a class $\mathbb{C}$ is "closed" if every $S \subseteq \mathcal{X} \times \mathcal{Y}$ with $\mathbb{C}[S] = \emptyset$ has a finite subset $S' \subseteq S$ with $\mathbb{C}[S'] = \emptyset$. The following lemma connects these quantities. Its proof is included in Appendix A.

Lemma 5

- $\mathrm{k}_o \leq \mathrm{k}_p \leq \mathrm{k}_w$.
- *If* $\mathrm{k}_w < \infty$ *or* $\mathbb{C}$ *is closed, then* $\mathrm{k}_o = \mathrm{k}_p = \mathrm{k}_w$.

Remark 6 *Certainly if* $\mathcal{X}$ *or* $\mathbb{C}$ *is* finite *then* $\mathrm{k}_w < \infty$*, so that* $\mathrm{k}_o = \mathrm{k}_p = \mathrm{k}_w$*. However, in the general case, there are examples where each of these inequalities can be strict, due to one quantity being infinite and another finite. We discuss such examples in Section 2.1 below.*

2.1. Some Examples

Let us now argue that many classes of interest have finite values for these complexity measures. Where appropriate, details of the examples are provided in Appendix A. We start with Halfspaces.

Example 1 *Let $\mathbb{C}$ be a class induced by halfspaces in $\mathbb{R}^n$. Then $\mathrm{k}_o = \mathrm{k}_p = n + 2 = \mathrm{d} + 1$.*

Another simple example is the case of *intersection closed classes*. The class $\mathbb{C}$ is intersection closed if the set $\{\{x : h(x) = +1\} : h \in \mathbb{C}\}$ is closed under arbitrary intersections. Also, recall the notions of *Maximum* and *Extremal* classes (see e.g., Floyd and Warmuth, 1995; Lawrence, 1983; Bandelt, Chepoi, Dress, and Koolen, 2006; Moran and Warmuth, 2016). A class $\mathbb{C}$ is a *Maximum class* if, for every integer $m \geq \mathrm{d}$ and every distinct $x_1, \ldots, x_m \in \mathcal{X}$, we have $|\{(h(x_1), \ldots, h(x_m)) : h \in \mathbb{C}\}| = \sum_{i=0}^{\mathrm{d}} \binom{m}{i}$. A class $\mathbb{C}$ is an *Extremal class* if, for every sequence $x_1, \ldots, x_m$ shattered by $\mathbb{C}$, $x_1, \ldots, x_m$ is also shattered by a set of classifiers in $\mathbb{C}$ that agree on all of $\mathcal{X} \setminus \{x_1, \ldots, x_m\}$. It is known that every Maximum class is Extremal.

Example 2 *If $\mathbb{C}$ is intersection-closed or Extremal (e.g., any Maximum class), then $\mathrm{k}_o \leq \mathrm{d} + 1$.*

This also means any *closed* class that is intersection-closed or Extremal has $\mathrm{k}_o = \mathrm{k}_p = \mathrm{k}_w \leq \mathrm{d}+1$.

As mentioned, all of the inequalities in Lemma 5 can be strict, due to the larger quantity being infinite while the smaller one is finite. We now discuss this issue, and how it relates to whether a class is closed (in terms of the definition of "closed" above, from Kane, Livni, Moran, and Yehudayoff, 2019). In particular, in all these examples, merely adding the limit cases into the class brings the complexity measures into agreement. We begin with a simple example where $\mathrm{k}_o = 2$ but $\mathrm{k}_p = \infty$.

Example 3 *Consider $\mathcal{X} = \mathbb{N}$ and $\mathbb{C} = \{2\mathbb{1}_{\{t\}} - 1 : t \in \mathcal{X}\}$ the class of* singletons. *The only finite hollow star sets are of size 2: namely, sets $\{(x, 1), (x', 1)\}$ and $\{(x, 1), (x, -1)\}$. Thus, $\mathrm{k}_o = 2$. However, for any finite set $\mathbb{C}' \subset \mathbb{C}$ of size at least 3, $\mathrm{Majority}(\mathbb{C}')$ is -1 everywhere on $\mathcal{X}$, but any $\ell < |\mathbb{C}'|$ has $\mathcal{X}_{\mathbb{C}',\ell} = \mathcal{X}$, so that we must have $\mathrm{k}_p > \ell$; thus, $\mathrm{k}_p = \infty$.*

Note that the constant function $x \mapsto -1$ is a pointwise limit of functions in $\mathbb{C}$. By adding just this one extra function, the modified class $\mathbb{C} \cup \{x \mapsto -1\}$ is closed and has $\mathrm{k}_w = \mathrm{k}_p = \mathrm{k}_o = 2$.

Next we describe a simple example where $\mathrm{k}_p = 2$ but $\mathrm{k}_w = \infty$.

Example 4 *Consider $\mathcal{X} = [0, \infty)$ and $\mathbb{C} = \{2\mathbb{1}_{[t,\infty)} - 1 : t \in [0, \infty)\}$ the class of* threshold *functions. This class is Maximum with $\mathrm{d} = 1$. For any finite $\mathbb{C}' \subset \mathbb{C}$, the majority vote classifier is a median threshold from $\mathbb{C}'$, so there is always an $h \in \mathbb{C}$ that agrees with $\mathrm{Majority}(\mathbb{C}')$ on all of $\mathcal{X}$. Therefore, $\mathrm{k}_p = 2$, the smallest possible value of k_p. On the other hand, the set $S = \{(x, -1) : x \in \mathcal{X}\}$ is unrealizable by $\mathbb{C}$, but there is no finite set witnessing this fact, and therefore $\mathrm{k}_w = \infty$.*

In this case, there are an infinite number of functions that are pointwise limits of functions in $\mathbb{C}$: namely, $x \mapsto -1$, and every open threshold $2\mathbb{1}_{(t,\infty)} - 1$, $t \in \mathcal{X}$. By adding these functions, the class $\mathbb{C} \cup \{x \mapsto -1\} \cup \{2\mathbb{1}_{(t,\infty)} - 1 : t \in \mathcal{X}\}$ has $\mathrm{k}_w = \mathrm{k}_p = \mathrm{k}_o = 2$.

These distinctions can occur in more extreme forms as well, as the following example illustrates.

Example 5 *Consider $\mathcal{X} = \mathbb{R}$ and $\mathbb{C} = \{2\mathbb{1}_{[a,b]} - 1 : a, b \in \mathbb{R}\}$ the class of bounded closed intervals. This class is intersection-closed (and Maximum), with $\mathrm{d} = 2$, and has $\mathrm{k}_o = \mathrm{k}_p = 3$, but the set $S = \{(x, 1) : x \in \mathcal{X}\}$ has no finite subset witnessing its non-realizability, so $\mathrm{k}_w = \infty$.*

In this case, to bring the complexity measures into agreement, we must increase $\mathbb{C}$ to be the set of all *intervals (including unbounded intervals, closed intervals, open intervals, and half-open half-closed intervals); this expanded class is closed and has $\mathrm{k}_w = \mathrm{k}_p = \mathrm{k}_o = 3$.*

2.2. Star number and sample complexity of ERM

We finish this section by recalling the known relations between ERM and the star number. We use the following notation (also used in the results below). For $x \geq 0$, define $\mathrm{Log}(x) = \max\{\ln(x), 1\}$.

Definition 7 (The worst-case sample complexity of ERM) *For any $\varepsilon, \delta \in (0,1)$, the worst-case sample complexity of ERM, denoted by $\mathcal{M}_{\mathrm{ERM}}(\varepsilon, \delta)$, is the smallest integer n such that for every possible data distribution $\mathcal{P}$ and every $f^\star \in \mathbb{C}$, for $S = (X_{1:n}, f^\star(X_{1:n}))$ with $X_{1:n} \sim \mathcal{P}^n$,*

$$\Pr\left(\sup_{h \in \mathbb{C}[S]} \mathrm{er}(h, f^\star) \leq \varepsilon\right) \geq 1 - \delta.$$

The following bounds were shown by (Hanneke, 2016b):

$$\frac{1}{\varepsilon}\left(\mathrm{d} + \mathrm{Log}\left(\mathfrak{s} \wedge \frac{1}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right) \lesssim \mathcal{M}_{\mathrm{ERM}}(\varepsilon, \delta) \lesssim \frac{1}{\varepsilon}\left(\mathrm{dLog}\left(\frac{\mathfrak{s}}{\mathrm{d}} \wedge \frac{1}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right).$$

In particular, having $\mathfrak{s} < \infty$ is necessary and sufficient for the existence of a distribution-free bound on the error rates of all ERMs converging at a rate inversely proportional to the sample size. Our results below show that a much weaker assumption $\mathrm{k}_p < \infty$ implies an analog of this property but only for *some* proper learners instead of *all* ERMs. It is important to notice that the projection number is finite for many expressive VC classes. In contrast, the star number, while implying an upper bound for k_o, is infinite except for some relatively simple classes: e.g., $\mathfrak{s} = \infty$ for Halfspaces in $\mathbb{R}^n$ if $n \geq 2$, and for many intersection-closed and maximum classes (Hanneke and Yang, 2015).

3. Upper bounds

Our upper bound will be established for the following algorithm, a modification of the optimal PAC learner of Hanneke (2016a). The main modifications compared to the original algorithm involve using $\mathrm{k}_p + 1$ recursive calls, rather than 3 calls, and using the projection operator (introduced above) to replace a majority vote classifier with an element of $\mathbb{C}$ so that the algorithm is a proper learner.

$\mathbb{A}(S; T)$
1. If $|S| < 4$, Return $\mathrm{ERM}(S \cup T)$
2. Let S_0 be the first $\lceil |S|/2 \rceil$ points in S
3. Let $S_1, \ldots, S_{\mathrm{k}_p+1}$ be independent uniform subsamples of $S \setminus S_0$ of size $\lfloor |S|/4 \rfloor$
4. Let $h_i = \mathbb{A}(S_0; T \cup S_i)$ for each $i = 1, \ldots, \mathrm{k}_p + 1$
5. Return $\hat{h} = \mathrm{Proj}_{\mathbb{C}}(h_1, \ldots, h_{\mathrm{k}_p+1})$.

To be precise, in Step 3, S_i is sampled without replacement. For this algorithm, we have the following theorem, representing one of the main results of this article.[3] We present its proof in Appendix B.

Theorem 8 *For any class $\mathbb{C}$, the proper sample complexity (achieved by $\mathbb{A}(S; \emptyset)$) satisfies*

$$\mathcal{M}_{\mathrm{prop}}(\varepsilon, \delta) = O\left(\frac{\mathrm{k}_p^2}{\varepsilon}\left(\mathrm{dLog}(\mathrm{k}_p) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right)\right).$$

3. We implicitly assume that $\mathbb{C}$ satisfies conditions so that all of the relevant random variables in the analysis are measurable. This is always the case if $\mathcal{X}$ is countable. For the uncountable case, we refer the reader to discussions by Blumer, Ehrenfeucht, Haussler, and Warmuth (1989); van der Vaart and Wellner (1996); van Handel (2013).

Interestingly, algorithm $\mathbb{A}$ above is not necessarily an ERM: it is not always consistent with the training sample. However, it is also possible to define a sample-consistent variant of the algorithm, only losing a factor of k_p in the sample complexity bound. Note that the standard VC bounds for ERM apply to *any* ERM classifier, whereas the bound we establish for this algorithm only holds for this *specific* choice of ERM classifier, which is therefore sometimes significantly better than the worst ERM. We proceed with the definition and theorem, the proof of which is given in Appendix B.

$\mathbb{A}_{\mathrm{ERM}}(S)$
1. If $|S| < \mathrm{k}_p + 1$, Return $\mathrm{ERM}(S)$
2. Split S into $\mathrm{k}_p + 1$ disjoint subsets S_i of size at least $\lfloor |S|/(\mathrm{k}_p + 1) \rfloor$
3. Set $h_i = \mathbb{A}_{\mathrm{ERM}}(\bigcup_{i' \neq i} S_{i'})$ for $i = 1, \ldots, \mathrm{k}_p + 1$.
4. Return $\hat{h} = \mathrm{Proj}_{\mathbb{C}}(h_1, \ldots, h_{\mathrm{k}_p+1})$.

To be precise, in Step 2, S is split into the subsets S_i based purely on the indices: for instance, take S_1 as the first $\lfloor |S|/(\mathrm{k}_p + 1) \rfloor$ data points in the sequence, S_2 as the next $\lfloor |S|/(\mathrm{k}_p + 1) \rfloor$ points in the sequence, and so on, with S_{k_p+1} as the last $|S| - \mathrm{k}_p \lfloor |S|/(\mathrm{k}_p+1) \rfloor$ data points in the sequence.

Theorem 9 *For any class $\mathbb{C}$, the sample complexity $\mathcal{M}_{\mathbb{A}_{ERM}}$ of $\mathbb{A}_{ERM}(S)$ satisfies*

$$\mathcal{M}_{\mathbb{A}_{ERM}}(\varepsilon, \delta) = O\left(\frac{\mathrm{k}_p^3}{\varepsilon}\left(\mathrm{dLog}(\mathrm{k}_p) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right)\right).$$

4. Lower Bounds

For the purpose of a lower bound, we will use the hollow star number k_o. From Lemma 5, for many classes we have $\mathrm{k}_o = \mathrm{k}_p = \mathrm{k}_w$, in which case this result indicates that the appearance of k_p in Theorem 8 is unavoidable (though there may be room to improve the specific dependence on k_p).

Theorem 10 *Every class with $\mathrm{k}_o < \infty$ has proper sample complexity*

$$\mathcal{M}_{\mathrm{prop}}(\varepsilon, \delta) = \Omega\left(\frac{\mathrm{d}}{\varepsilon} + \frac{1}{\varepsilon}\mathrm{Log}\left(\frac{1}{\delta}\right) + \frac{1}{\varepsilon}\mathrm{Log}(\mathrm{k}_o)\mathbb{1}[\varepsilon \leq 1/\mathrm{k}_o]\right).$$

Also, if $\mathrm{k}_o = \infty$ then

$$\mathcal{M}_{\mathrm{prop}}(\varepsilon, \delta) \neq o\left(\frac{1}{\varepsilon}\mathrm{Log}\left(\frac{1}{\varepsilon}\right)\right).$$

In other words, there is a sequence $\varepsilon_i \to 0$ with $\mathcal{M}_{\mathrm{prop}}(\varepsilon_i, \delta) \geq \frac{c}{\varepsilon_i}\log\left(\frac{1}{\varepsilon_i}\right)$ for a constant $c > 0$.

The proof, which is inspired by the arguments in (Daniely and Shalev-Shwartz, 2014), is deferred to Appendix C. Let us only provide some intuition behind the proof. Assume, for simplicity, that we want to lower bound the sample complexity in a particular regime where $\varepsilon = \frac{1}{2(\mathrm{k}_o - 1)}$, $\mathrm{k}_o \geq 2$. Let $S = \{(x_1, y_1), \ldots, (x_{\mathrm{k}_o}, y_{\mathrm{k}_o})\}$ be a hollow star set, and $\forall i \in 1 : \mathrm{k}_o$ let $h_i \in \mathbb{C}$ be such that $\{j : h_i(x_j) \neq y_j\} = \{i\}$. We set $f^\star = h_{i^*}$ for some $i^* \in 1 : \mathrm{k}_o$ and set $\mathcal{P}(\{x_{i^*}\}) = 0$ and $\mathcal{P}(\{x_i\}) = 2\varepsilon$ for $i \neq i^*$. Observe that the only way for the learner to output $\hat{h} \in \mathbb{C}$ having $\mathrm{er}_{\mathcal{P}}(\hat{h}, f^\star) \leq \varepsilon$ is for $\hat{h}$ to agree with h_{i^*} on *all* of S. However, the learner will not be able to identify

the corresponding point x_{i^*} having zero mass before it observes x_i for *every* $i \neq i^*$. By the standard coupon collector argument (see Lemma 19 in Appendix C) due to the fact that there will be some copies in the training sample we will need the sample size of order $\Omega(\mathrm{k}_o\mathrm{Log}(\mathrm{k}_o)) = \Omega\big(\frac{1}{\varepsilon}\mathrm{Log}(\mathrm{k}_o)\big)$. The formal application of this idea (including extension to any $\varepsilon \leq 1/\mathrm{k}_o$) is given in Appendix C.

Our second lower bound provides stronger guarantees. However, it is less general. In what follows, given d and k_w we present a *particular* class $\mathbb{C}$ and a space $\mathcal{X}$ such that the desired lower bound holds. The proof of this result uses the same logic and the technical details are deferred to Appendix C.

Theorem 11 *There is a numerical constant $c > 0$ such that, for any value of $\mathrm{d} \geq 1$ and $2 \leq \mathrm{k}_w < \infty$, there exists a space $\mathcal{X}$ and a class $\mathbb{C}$ with VC dimension d and dual Helly number k_w, for which the proper sample complexity satisfies*

$$\mathcal{M}_{\mathrm{prop}}(\varepsilon, \delta) \geq \frac{c}{\varepsilon}\left(\mathrm{dLog}\left(\frac{\mathrm{k}_w}{\mathrm{d}} \wedge \frac{1}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right),$$

for every $\varepsilon \in (0, 1/8)$ and $\delta \in (0, 1/100)$. Furthermore, for any $\mathrm{d} \geq 1$ there exists $\mathcal{X}$ and a space $\mathbb{C}$ with VC dimension d and hollow star number $\mathrm{k}_o = \infty$ and

$$\mathcal{M}_{\mathrm{prop}}(\varepsilon, \delta) \geq \frac{c}{\varepsilon}\left(\mathrm{dLog}\left(\frac{1}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right).$$

5. Stable compression schemes and optimality of SVM for PAC learning Halfspaces

In this section we establish a new generalization bound for a special type of compression scheme, referred to as a *stable* compression scheme, which removes a log factor (which is known to not be removable for general compression schemes). We apply the result to obtain new tighter bounds for several quantities of interest in the learning theory literature.

As our main application, we apply this new bound to resolve a long-standing open question: namely, showing that the well-known *support vector machine* (SVM) learning algorithm for Halfspaces achieves the optimal sample complexity. This resolves a question posed by Vapnik and Chervonenkis (1974), which has received considerable attention in the literature (e.g., Blumer and Littlestone, 1989; Balcan and Long, 2013; Zhivotovskiy, 2017; Hanneke and Kontorovich, 2019; Long and Long, 2020). According to a note by A. Chervonenkis (see Chapter I in Vovk, Papadopoulos, and Gammerman, 2015) the in-expectation version of the risk bound for SVM had been proven in 1966, even before the renowned uniform law of large numbers was announced (Vapnik and Chervonenkis, 1968). However, obtaining a high-probability version of the bound, without introducing additional log factors, remained an open problem since then. The following theorem resolves this problem. Furthermore, this is also the first proof that the optimal sample complexity for Halfspaces is achievable by *some* proper learner.

Theorem 12 *The PAC sample complexity of SVM in $\mathbb{R}^n$ is*

$$\mathcal{M}_{\mathrm{SVM}}(\varepsilon, \delta) = \Theta\left(\frac{n}{\varepsilon} + \frac{1}{\varepsilon}\log\frac{1}{\delta}\right).$$

The proof is presented below, after the abstract result that implies it. This bound improves the sample complexity bound $\mathcal{M}_{\mathrm{SVM}}(\varepsilon, \delta) = O\left(\frac{n}{\varepsilon}\log\frac{1}{\delta}\right)$ shown in (Zhivotovskiy, 2017) and the general bound $\mathcal{M}_{\mathrm{ERM}}(\varepsilon, \delta) = O\left(\frac{n}{\varepsilon}\log\frac{1}{\varepsilon} + \frac{1}{\varepsilon}\log\frac{1}{\delta}\right)$ holding for any ERM (Vapnik and Chervonenkis, 1974; Blumer, Ehrenfeucht, Haussler, and Warmuth, 1989).

5.1. Stable compression schemes

We present here our new generalization bound for stable compression schemes. Let us recall some standard definitions. A *sample compression scheme* consists of two functions: a *compression function* $\kappa : (\mathcal{X} \times \mathcal{Y})^* \to (\mathcal{X} \times \mathcal{Y})^*$ and a *reconstruction function* $\rho : (\mathcal{X} \times \mathcal{Y})^* \to \mathcal{Y}^{\mathcal{X}}$. The following definition is due to Littlestone and Warmuth (1986).

Definition 13 (Sample compression scheme) *The pair of functions (κ, ρ) define a sample compression scheme of size ℓ if for any $m \in \mathbb{N}$ and $h \in \mathbb{C}$ and any sample $x_{1:m} \in \mathcal{X}^m$ it holds that $\kappa((x_{1:m}, h(x_{1:m})) \subseteq (x_{1:m}, h(x_{1:m}))$ and $|\kappa((x_{1:m}, h(x_{1:m}))| \leq \ell$, and the classifier $\hat{h} = \rho(\kappa((x_{1:m}, h(x_{1:m}))))$ satisfies $\hat{h}(x_{1:m}) = h(x_{1:m})$: that is, it recovers h's classifications on $x_{1:m}$.*

We work with the following natural definition taking its roots in (Vapnik and Chervonenkis, 1974).

Definition 14 (Stable compression scheme) *A sample compression scheme (κ, ρ) is called* stable *if for any $m \in \mathbb{N}$, $h \in \mathbb{C}$, $x_{1:m} \in \mathcal{X}^m$ and any $(x, h(x)) \in (x_{1:m}, h(x_{1:m})) \setminus \kappa((x_{1:m}, h(x_{1:m}))$,*

$$\kappa((x_{1:m}, h(x_{1:m})) \setminus (x, h(x))) = \kappa((x_{1:m}, h(x_{1:m})).$$

This definition means that removing any $(x, h(x))$ *not* belonging to the compression set, the compression set of the sub-sample remains the same. Finally, we say that the sample compression scheme (κ, ρ) is *proper* if the image of the reconstruction function ρ is contained in $\mathbb{C}$.

It is known that for any stable compression scheme (κ, ρ) of any size ℓ, for any $\mathcal{P}$ and $f^\star \in \mathbb{C}$, for any $m \in \mathbb{N}$, $\mathbb{E}[\mathrm{er}(\rho(\kappa((X_{1:m}, f^\star(X_{1:m})))))] \leq \frac{\ell}{m+1}$. This follows from the leave-one-out analysis of Vapnik and Chervonenkis (1974) (see also Haussler, Littlestone, and Warmuth, 1994; Zhivotovskiy, 2017). The best known PAC generalization bound on $\mathrm{er}(\rho(\kappa(X_{1:m}, f^\star(X_{1:m}))))$ (holding with probability $1 - \delta$) valid for *any* sample compression scheme of a size ℓ is due to Littlestone and Warmuth (1986) (see also Floyd and Warmuth, 1995): $O\big(\frac{1}{m}\left(\ell \log(m) + \log\big(\frac{1}{\delta}\big)\right)\big)$, where the $\log(m)$ factor improves to $\mathrm{Log}\big(\frac{m}{\ell}\big)$ if ρ is permutation-invariant. Floyd and Warmuth (1995) showed that there exist spaces $\mathbb{C}$ and compression schemes for $\mathbb{C}$ for which this log factor cannot be improved. However, in the special case of *stable* compression schemes, Zhivotovskiy (2017) established a bound $O\big(\frac{\ell}{m} \log\big(\frac{1}{\delta}\big)\big)$, which is sometimes better.

As one of the main contributions of this work, the following result improves this PAC generalization bound by completely removing the log factor from the bound of Littlestone and Warmuth (1986). A simple proof of this result is provided in Appendix D.

Theorem 15 *Assume that $\mathbb{C}$ has a stable sample compression scheme (κ, ρ) of size ℓ. Then, for any $\mathcal{P}$ and any $f^\star \in \mathbb{C}$, for any integer $m > 2\ell$, given an i.i.d. sample $S = (X_{1:m}, f^\star(X_{1:m}))$ of size m, for any $\delta \in (0, 1)$, we have with probability at least $1 - \delta$,*

$$\mathrm{er}(\rho(\kappa(S))) < \frac{2}{m - 2\ell}\left(\ell \ln(4) + \ln\left(\frac{1}{\delta}\right)\right).$$

This result immediately yields the following proof of Theorem 12 establishing optimality of SVM for PAC learning Halfspaces.

Proof of Theorem 12 Since SVM may be expressed as a stable compression scheme of size $n + 1$ (see (Long and Long, 2020) for a transparent proof of this fact, originally proven by Vapnik and Chervonenkis 1974), the upper bound is immediate from Theorem 15. The lower bound follows from (Ehrenfeucht, Haussler, Kearns, and Valiant, 1989). ■

Maximum classes. As a second application of Theorem 15 to proper learning, consider any *maximum class* $\mathbb{C}$ (Floyd and Warmuth, 1995; recall the definition from Section 2.1). Every maximum class $\mathbb{C}$ is known to have a proper stable compression scheme of size d; this follows from Theorem 5.1 and condition (R2) of Theorem 6.1 from the paper of Chalopin, Chepoi, Moran, and Warmuth (2019) (specifically, for their (κ, ρ), every realizable data set contains exactly one compression set for a concept in $\mathbb{C}$ consistent with the data, and hence removing any point not in that compression set cannot change the identity of this unique compression set). Hence, we have the following corollary of Theorem 15, establishing (for the first time) that maximum classes are properly learnable with sample complexity of the same order as the optimal PAC sample complexity.

Corollary 16 *For $\mathbb{C}$ which is a Maximum class,*

$$\mathcal{M}_{\mathrm{prop}}(\varepsilon, \delta) = \Theta\left(\frac{\mathrm{d}}{\varepsilon} + \frac{1}{\varepsilon}\log\left(\frac{1}{\delta}\right)\right).$$

Remark 17 *We note that Theorem 15 is also able to recover the optimal PAC sample complexity for the* Closure *algorithm for* intersection-closed *classes (Helmbold, Sloan, and Warmuth, 1990); this sample complexity result was already known via different arguments (Darnstädt, 2015; Hanneke, 2016b; Auer and Ortner, 2007), though Theorem 15 offers improved numerical constants.*

Remark 18 *A further implication of Theorem 15 and Theorem 10 together is that any class $\mathbb{C}$ with $\mathrm{k}_o = \infty$ does* not *have a proper stable compression scheme of bounded size.*

6. Remaining Gaps and Open Questions

Although our upper and lower bounds give a description of proper PAC sample complexity in many cases, there are still situations not explained by our general bounds. Consider Example 3 (singletons $x \mapsto 2\mathbb{1}_{\{t\}}(x) - 1$). Since $\mathrm{k}_o = 2$ but $\mathrm{k}_p = \infty$, our upper bound (Theorem 8) does not match our lower bound (Theorem 10), and also does not match the optimal (improper) sample complexity $\mathcal{M}(\varepsilon, \delta) = \Theta\left(\frac{1}{\varepsilon}\log\frac{1}{\delta}\right)$. However, it turns out there *does* exist a proper stable compression scheme of size 1 for this $\mathbb{C}$, and therefore Theorem 15 implies $\mathcal{M}_{\mathrm{prop}}(\varepsilon, \delta) = \Theta\left(\frac{1}{\varepsilon}\log\frac{1}{\delta}\right)$, matching the optimal (improper) sample complexity. Specifically, for any data set S, if $\exists (x, 1) \in S$, define $\kappa(S) = \{(x, 1)\}$ and $\rho(\{(x, 1)\}) = 2\mathbb{1}_{\{x\}} - 1$; if S is all negative examples, define $\kappa(S) = \{(x, -1)\}$ for the *largest* x in S, and $\rho(\{(x, -1)\}) = 2\mathbb{1}_{\{x+1\}} - 1$. Alternatively, we may simply note that $\mathbb{C}$ is a *Maximum* class, hence Corollary 16 applies. It is therefore important to ask whether $\mathcal{M}_{\mathrm{prop}}(\varepsilon, \delta)$ is perhaps *always* characterized by k_o (and d, ε, δ). It is also interesting to consider whether the optimal size of a proper stable compression scheme is also always characterized by k_o (and d). Another interesting direction is sharpening the dependence on k_p in our upper bounds.

Finally, we remark that the result of Section 5 is related to the question of obtaining generalization bounds for learning algorithms that are stable with respect to small perturbations in the training sample. In the case of uniformly stable algorithms, recent results (Feldman and Vondrak, 2019; Bousquet, Klochkov, and Zhivotovskiy, 2019) provide sharp high-probability bounds, and the proofs are based on a sub-sampling argument: the learner is tested on carefully chosen parts of the training sample. Similarly to uniformly stable algorithms, stable compression schemes easily provide sharp in-expectation error bounds (Haussler, Littlestone, and Warmuth, 1994). The challenging part, already pointed out by Vapnik and Chervonenkis (1974), is to prove that these algorithms admit sharp high-probability bounds. Our result, based on related arguments, is the first to prove that an optimal high-probability error bound holds for stable compression schemes, including SVM.

References

P. Auer and R. Ortner. A new PAC bound for intersection-closed concept classes. *Machine Learning*, 66(2-3):151–163, 2007.

M.-F. Balcan and P. M. Long. Active and passive learning of linear separators under log-concave distributions. In *Proceedings of the* 26^{th} *Conference on Learning Theory*, 2013.

H.-J. Bandelt, V. Chepoi, A. W. M. Dress, and J. H. Koolen. Combinatorics of lopsided sets. *Eur. J. Comb.*, 27(5):669–689, 2006.

A. Blumer and N. Littlestone. Learning faster than promised by the Vapnik-Chervonenkis dimension. *Discrete Applied Mathematics*, 24(1-3):47–53, 1989.

A. Blumer, A. Ehrenfeucht, D. Haussler, and M. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the Association for Computing Machinery*, 36(4):929–965, 1989.

O. Bousquet, Y. Klochkov, and N. Zhivotovskiy. Sharper bounds for uniformly stable algorithms. *arXiv preprint arXiv:1910.07833*, 2019.

M. Braverman, G. Kol, S. Moran, and R. R. Saxena. Convex set disjointness, distributed learning of halfspaces, and LP feasibility. *arXiv:1909.03547*, 2019.

J. Chalopin, V. Chepoi, S. Moran, and M. K. Warmuth. Unlabeled sample compression schemes and corner peelings for ample and maximum classes. In *Proceedings of the* 46^{th} *International Colloquium on Automata, Languages and Programming*, 2019.

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

A. Daniely, S. Sabato, S. Ben-David, and S. Shalev-Shwartz. Multiclass learnability and the ERM principle. *The Journal of Machine Learning Research*, 16(1):2377–2404, 2015.

L. Danzer, B. Grünbaum, and V. Klee. *Helly's theorem and its relatives*. Proceedings of symposia in pure mathematics: Convexity. American Mathematical Society, 1963.

M. Darnstädt. The optimal PAC bound for intersection-closed concept classes. *Information Processing Letters*, 115(4):458–461, 2015.

A. Ehrenfeucht, D. Haussler, M. Kearns, and L. Valiant. A general lower bound on the number of examples needed for learning. *Information and Computation*, 82(3):247–261, 1989.

V. Feldman and J. Vondrak. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In *Proceedings of the* 32^{nd} *Conference on Learning Theory*, pages 1270–1279, 2019.

S. Floyd and M. Warmuth. Sample compression, learnability, and the Vapnik-Chervonenkis dimension. *Machine Learning*, 21(3):269–304, 1995.

S. A. Goldman and M. J. Kearns. On the complexity of teaching. *Journal of Computer and System Sciences*, 50:20–31, 1995.

S. Hanneke. *Theoretical Foundations of Active Learning*. PhD thesis, Machine Learning Department, School of Computer Science, Carnegie Mellon University, 2009.

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

S. Hanneke. Refined error bounds for several learning algorithms. *Journal of Machine Learning Research*, 17(135):1–55, 2016b.

S. Hanneke and A. Kontorovich. Optimality of SVM: Novel proofs and tighter bounds. *Theoretical Computer Science*, 796:99–113, 2019.

S. Hanneke and L. Yang. Minimax analysis of active learning. *Journal of Machine Learning Research*, 16(12):3487–3602, 2015.

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

E. Helly. Über mengen konvexer körper mit gemeinschaftlichen punkte. *Jahresbericht der Deutschen Mathematiker-Vereinigung*, 32:175–176, 1923.

D. Helmbold, R. Sloan, and M. Warmuth. Learning nested differences of intersection-closed concept classes. *Machine Learning*, 5(2):165–196, 1990.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

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.

M. Kearns and U. Vazirani. *An Introduction to Computational Learning Theory*. The MIT Press, 1994.

J. Lawrence. Lopsided sets and orthant-intersection of convex sets. *Pacific J. Math.*, 104:155–173, 1983.

F. W. Levi. On Helly's theorem and the axioms of convexity. *J. Indian Math. Soc*, 15:65–76, 1951.

N. Littlestone and M. Warmuth. Relating data compression and learnability. *Unpublished manuscript*, 1986.

P. Long and R. Long. On the complexity of proper distribution-free learning of linear classifiers. In *Proceedings of the* 31^{st} *International Conference on Algorithmic Learning Theory*, 2020.

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 M. Warmuth. Labeled compression schemes for extremal classes. In *Proceedings of the* 27^{th} *International Conference on Algorithmic Learning Theory*, pages 34–49. Springer, 2016.

R. Motwani and P. Raghavan. *Randomized algorithms*. Chapman & Hall/CRC, 2010.

H. Simon. An almost optimal PAC algorithm. In *Proceedings of the* 28^{th} *Conference on Learning Theory*, 2015.

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

R. van Handel. The universal Glivenko–Cantelli property. *Probability Theory and Related Fields*, 155(3-4):911–934, 2013.

V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In *Proc. USSR Acad. Sci.*, 1968.

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.

V. Vovk, H. Papadopoulos, and A. Gammerman. *Measures of Complexity*. Springer, 2015.

N. Zhivotovskiy. Optimal learning via local entropies and sample compression. In *Proceedings of the* 30^{th} *Conference on Learning Theory*, pages 2023–2065, 2017.

N. Zhivotovskiy and S. Hanneke. Localization of VC classes: Beyond local Rademacher complexities. *Theoretical Computer Science*, 742:27–49, 2018.

Appendix A. Omitted proofs of Section 2

Proof of Lemma 5 We note that many of these arguments were given in some form in the course of the proofs of Kane, Livni, Moran, and Yehudayoff (2019). However, we provide the proofs explicitly here, particularly since our context is slightly different.

First, on a technical note, we remark that because of our assumption $|\mathbb{C}| \geq 3$ stated initially, all of k_o,k_p,k_w are at least 2 (so that the restriction to $\mathrm{k}_p \geq 2$ in its definition does not affect the claims).

For the first claimed inequalities, given the k_o classifiers $\mathbb{C}'$ in $\mathbb{C}$ witnessing a hollow star, for any $\ell < \mathrm{k}_o$ the region $\mathcal{X}_{\mathbb{C}',\ell}$ contains the hollow star, and hence the majority vote of the $\mathbb{C}'$ classifiers is unrealizable on $\mathcal{X}_{\mathbb{C}',\ell}$. Therefore, $\mathrm{k}_p \geq \mathrm{k}_o$.

If $\mathrm{k}_w = \infty$, then trivially $\mathrm{k}_p \leq \mathrm{k}_w$, so suppose $\mathrm{k}_w < \infty$. Suppose some finite multiset $\mathbb{C}' \subseteq \mathbb{C}$ has no $h \in \mathbb{C}$ that coincides with $\mathrm{Majority}(\mathbb{C}')$ on $\mathcal{X}_{\mathbb{C}',\mathrm{k}_w}$. Then $S = \{(x, \mathrm{Majority}(\mathbb{C}')(x)) : x \in \mathcal{X}_{\mathbb{C}',\mathrm{k}_w}\}$ is an unrealizable set. Therefore, it contains a subset W of size at most k_w that is also unrealizable. But (by definition of $\mathcal{X}_{\mathbb{C}',\mathrm{k}_w}$) each point (x, y) in W contradicts strictly fewer than $|\mathbb{C}'|/\mathrm{k}_w$ elements in $\mathbb{C}'$, so that there must be at least one $h \in \mathbb{C}'$ that survives: a contradiction. Therefore, $\mathrm{k}_p \leq \mathrm{k}_w$.

It remains to show that these quantities are all equal when $\mathrm{k}_w < \infty$. Let S be an unrealizable set such that the smallest unrealizable subset W has size $\mathrm{k}_w < \infty$. If W is not a hollow star set, then there exists a point $(x, y) \in W$ such that $(W \setminus \{(x, y)\}) \cup \{(x, -y)\}$ is also not realizable, which implies $W \setminus \{(x, y)\}$ is also not realizable: a contradiction. Therefore, W is a hollow star (indeed, any unreducible unrealizable set is a hollow star), and hence $\mathrm{k}_w \leq \mathrm{k}_o$. The equalities then follow from the first claim, established above.

For the remaining claim, if $\mathbb{C}$ is closed and yet $\mathrm{k}_w = \infty$, it implies there is a sequence S_i of sets with $\mathbb{C}[S_i] = \emptyset$ for which the smallest $W_i \subseteq S_i$ with $\mathbb{C}[W_i] = \emptyset$ has $\lim_{i\to\infty} |W_i| = \infty$, yet each W_i is finite (due to the "closedness" assumption). As above, these minimum-size W_i sets must be hollow star sets, which implies there is no finite bound on the size of all finite hollow star sets. Therefore $\mathrm{k}_o = \infty$, and the inequality established above then implies $\mathrm{k}_p = \infty$ as well. ■

Proof for Example 1 This follows from Proposition 2.8 in (Braverman, Kol, Moran, and Saxena, 2019). Indeed, this implies that $\mathrm{k}_o \leq n + 2$ since if a finite unrealizable sample S is a hollow star set, then in particular every proper subsample of S must be realizable; however, by Proposition 2.8 in (Braverman, Kol, Moran, and Saxena, 2019) the set S must contain an unrealizable subsample of size at most $n + 2$, and hence it must be that $|S| \leq n + 2$. To see that $\mathrm{k}_o \geq n + 2$, pick $x_1, \ldots, x_{n+1} \in \mathbb{R}^n$ to be the vertices of a simplex, and choose

$$S = \left\{(x_1, +1), \ldots, (x_{n+1}, +1), \left(\frac{x_1 + \ldots + x_{n+1}}{n+1}, -1\right)\right\}.$$

We leave it to the reader to verify that the above S witnesses that $\mathrm{k}_o \geq |S| = n + 2$.

It remains to show that $\mathrm{k}_p = n + 2$. By Lemma 5 it suffices to show that $\mathrm{k}_p \leq n + 2$. Let $\mathbb{C}'$ be a finite collection of halfspaces in $\mathbb{R}^n$. We need to show that there exists a halfspace h which agrees with $\mathrm{Majority}(\mathbb{C}')$ on the set $\mathcal{X}_{\mathbb{C}',n+2}$. Let $\mathcal{X}_+ \subseteq \mathcal{X}_{\mathbb{C}',n+2}$ denote the set of all points $x \in \mathcal{X}_{\mathbb{C}',n+2}$ such that $\mathrm{Majority}(\mathbb{C}')(\mathrm{x}) = +1$ and similarly let $\mathcal{X}_- \subseteq \mathcal{X}_{\mathbb{C}',n+2}$ denote the set of all points $x \in \mathcal{X}_{\mathbb{C}',n+2}$ such that $\mathrm{Majority}(\mathbb{C}')(\mathrm{x}) = -1$. We first claim that the convex hulls $\mathsf{conv}(\mathcal{X}_+)$ and $\mathsf{conv}(\mathcal{X}_-)$ are disjoint; indeed, otherwise by Proposition 2.8 in (Braverman, Kol, Moran, and Saxena, 2019) there exist $S_+ \subseteq \mathcal{X}_+, S_- \subseteq \mathcal{X}_-$ such that $\mathsf{conv}(S_-) \cap \mathsf{conv}(S_+) \neq \emptyset$ and $|S_-| + |S_+| \leq n + 2$ However, every $x \in S_- \cup S_+$ is classified correctly by more than $1 - \frac{1}{n+2}$ fraction of the halfspaces in $\mathbb{C}'$ and hence there must be a halfspace in $\mathbb{C}'$ that classifies correctly all points in $S_+ \cup S_-$ and so $\mathsf{conv}(S_-) \cap \mathsf{conv}(S_+) = \emptyset$, which is a contradiction.

Having established that $\mathsf{conv}(S_-) \cap \mathsf{conv}(S_+) = \emptyset$, we are ready to finish the proof. Indeed, by the *Hyperplane Separation Theorem* there exists a linear function $L : \mathbb{R}^n \to \mathbb{R}$ and a value v such that $L(x) \leq v$ for every $v \in \mathcal{X}_-$ and $L(x) \geq v$ for every $v \in \mathcal{X}_+$. However, note that in fact $L(x) < v$ for every $x \in \mathcal{X}_-$: this follows because $\mathcal{X}_-$ is an open set (indeed, it can be written as a union of (finite) intersections of sets of the form $h^{-1}(-1)$, where $h \in \mathbb{C}'$, each of which is an open set). Thus, the halfspace $\{x : L(x) \geq v\}$ agrees with $\mathrm{Majority}(\mathbb{C}')$ on the set $\mathcal{X}_{\mathbb{C}',n+2}$, as required.[4] ■

Proof for Example 2 Let us sketch the proof. The bound $\mathrm{k}_o \leq \mathrm{d} + 1$ on the dual Helly number for extremal classes follows from the fact that: (i) the complement of an extremal class is extremal, and

4. Note that we use the definition of Halfspaces where the positive side of each halfspace is closed: i.e., every halfspace is of the form $\mathrm{sign}(L(x) - v)$, where $\mathrm{sign}(0) = +1$.

(ii) the *one-inclusion graph* on an extremal graph projected to any data set is connected (we refer the reader to Bandelt, Chepoi, Dress, and Koolen, 2006 and Moran and Warmuth, 2016 for these definitions and results). In particular, it follows that the one-inclusion graph of the complement of $\mathbb{C}$ projected to any data set is connected.

Since a hollow star corresponds to a one-inclusion graph where the star's center is an isolated vertex in the one-inclusion graph of the complement of $\mathbb{C}$ projected to the points, it must be the only vertex in the complement, which means any strict subset of the points is shattered by $\mathbb{C}$. This immediately implies $\mathrm{k}_o - 1 \leq \mathrm{d}$.

The case of intersection-closed classes is also straightforward. Let $S = \{(x_1, y_1), \ldots, (x_k, y_k)\}$ be a finite hollow star set, and let $h_1, \ldots, h_k$ be elements of $\mathbb{C}$ such that $\{j : h_i(x_j) \neq y_j\} = \{i\}$ for each i. Denote by m the number of y_i values equal -1. We will first show that $m \leq 1$. Indeed, if the are at least two values y_i and y_j ($i \neq j$) both equal to -1, then letting h_0 be the classifier in $\mathbb{C}$ with $\{x : h_0(x) = 1\} = \{x : h_i(x) = 1\} \cap \{x : h_j(x) = 1\}$ (which exists since $\mathbb{C}$ is intersection-closed), we would have h_0 correct on all of S: a contradiction to S being unrealizable. Next we argue that there are at most d values y_i equal 1: that is, $k - m \leq \mathrm{d}$. Suppose $y_{i_1}, \ldots, y_{i_{k-m}}$ are equal 1. Then for any $y'_{i_1}, \ldots, y'_{i_{k-m}} \in \mathcal{Y}$, there exists a classifier $h \in \mathbb{C}$ with $\{x : h(x) = 1\} = \bigcap\{\{x : h_{i_j}(x) = 1\} : y'_{i_j} = -1\}$, and this h has $(h(x_{i_1}), \ldots, h(x_{i_{k-m}})) = (y'_{i_1}, \ldots, y'_{i_{k-m}})$; thus, the set $\{x_{i_1}, \ldots, x_{i_{k-m}}\}$ is shattered by $\mathbb{C}$, and hence has size at most d. Altogether, we have that $k = (k-m) + m \leq \mathrm{d} + 1$. Since this applies to *any* finite hollow star set S, we conclude that $\mathrm{k}_o \leq \mathrm{d} + 1$. ■

Proof for Example 5 We provide the details for the final claim that the augmented class has $\mathrm{k}_w = \mathrm{k}_p = \mathrm{k}_o = 3$. Since the class is intersection-closed with $\mathrm{d} = 2$, Example 2 implies $\mathrm{k}_o \leq 3$. Thus, since $\mathrm{k}_o \geq 3$ (witnessed by the hollow star set $\{(1,1), (2,-1), (3,1)\}$), Lemma 5 implies it suffices to show the class is closed. Let S be an infinite unrealizable set; we aim to show it must contain a finite unrealizable subset. If S contains $(x, 1)$ and $(x, -1)$ for some x, then $\{(x,1), (x,-1)\}$ is a finite unrealizable subset. Otherwise, suppose no such x exists. Notice that S must contain at least one point having label 1, since otherwise S would be realizable by the constant classifier $h_{-1} = -1$, which is in the class. Let $\underline{x} = \inf\{x : (x,1) \in S\}$ and $\bar{x} = \sup\{x : (x,1) \in S\}$. Note that every $(x, y) \in S$ with $x > \bar{x}$ has $y = -1$ and similarly every $(x, y) \in S$ with $x < \underline{x}$ has $y = -1$. In particular, this implies that if every $(x, y) \in S$ with $\underline{x} < x < \bar{x}$ has $y = 1$, then S would be realizable by a classifier corresponding to one of the intervals $(\underline{x}, \bar{x}]$ (if $(\underline{x}, 1) \notin S$ and $(\bar{x}, 1) \in S$), $[\underline{x}, \bar{x})$ (if $(\underline{x}, 1) \in S$ and $(\bar{x}, 1) \notin S$), $[\underline{x}, \bar{x}]$ (if $(\underline{x}, 1) \in S$ and $(\bar{x}, 1) \in S$), or $(\underline{x}, \bar{x})$ (if $(\underline{x}, 1) \notin S$ and $(\bar{x}, 1) \notin S$): a contradiction. Therefore, there exists $(x_2, -1) \in S$ with $\underline{x} < x_2 < \bar{x}$. By the definitions of $\underline{x}$ and $\bar{x}$, this implies there exist finite x_1, x_3 with $x_1 < x_2 < x_3$ such that $(x_1, 1)$ and $(x_3, 1)$ are both in S. Then we have that the set $\{(x_1,1), (x_2,-1), (x_3,1)\} \subset S$ is a finite unrealizable subset. Since this applies to *any* unrealizable infinite set S, we conclude that the class is closed. ■

Appendix B. Proofs of the upper bounds

Proof of Theorem 8 Fix any target concept $f^\star \in \mathbb{C}$ and distribution $\mathcal{P}$. We first argue that, for any finite data sets S and T with labels consistent with $f^\star$, the proper learner $\mathbb{A}(S; T)$ outputs

$\hat{h} \in \mathbb{C}$ such that $\hat{h}$ is correct on T. Note that this is trivially true if $|S| < 4$, since Step 1 returns $\mathrm{ERM}(S \cup T)$, which is correct on T by definition. Now, for induction, suppose S is a correctly labeled finite data set such that, for any strict subset $S' \subset S$, and any correctly labeled finite data set T', $\mathbb{A}(S'; T')$ returns a classifier in $\mathbb{C}$ that is correct on T'. Now we extend this property to the full set S. Fix any correctly labeled finite data set T. Recalling the notation from the algorithm, define $h_{\mathrm{maj}}(x) = \mathrm{Majority}(h_1(x), \ldots, h_{\mathrm{k}_p+1}(x))$ (breaking ties to favor label -1, say), and recalling the notation (1) let

$$\mathcal{X}_0 = \mathcal{X}_{\{h_1,\ldots,h_{\mathrm{k}_p+1}\},\mathrm{k}_p}.$$

By definition (from Step 5 in $\mathbb{A}(S;T)$), the classifier $\hat{h} = \mathbb{A}(S;T) \in \mathbb{C}$ has $\hat{h}(x) = h_{\mathrm{maj}}(x)$ on every $x \in \mathcal{X}_0$. Furthermore, since $h_i = \mathbb{A}(S_0; T \cup S_i)$ for each i, where $S_0 \subset S$, the inductive hypothesis implies h_i is correct on T. Therefore, all h_i agree on the labels in T, and hence the set of points in T is contained in $\mathcal{X}_0$, which implies $\hat{h}$ is correct on T as well. By induction, this implies that for any correctly labeled finite data sets S and T, $\hat{h} = \mathbb{A}(S;T)$ is correct on T.

Next we argue that, for any $m_0 \in \mathbb{N}$ and any $\delta_0 \in (0,1)$, if T is any correctly labeled finite data set and S is an i.i.d. labeled data set of size m_0, with $\mathcal{P}$ marginal distribution and $f^\star$ labels, then with probability at least $1 - \delta_0$, we have that $\hat{h}$ satisfies

$$\mathrm{er}(\hat{h}) \leq \frac{c \cdot \mathrm{k}_p^2}{m_0}\left(\mathrm{dLog}(\mathrm{k}_p) + \mathrm{Log}\left(\frac{1}{\delta_0}\right)\right), \tag{2}$$

where $c \geq 1$ is an appropriate finite numerical constant. Note that this would imply Theorem 8, since setting $\delta_0 = \delta$ and m_0 of size proportional to the claimed bound on $\mathcal{M}_{\mathrm{prop}}(\varepsilon, \delta)$ from Theorem 8, the bound (2) is less than ε.

If $m_0 < 4$, the claim trivially holds, as the bound is greater than 1. In particular, this will be our base case in an inductive argument. Now, for induction, suppose $m \geq 4$, and that for any $\delta_0 \in (0,1)$ and any $m_0 < m$, if S is an i.i.d. data set of size m_0 (with $\mathcal{P}$ marginal distribution and $f^\star$ labels) and T is any finite data set with $f^\star$ labels, then with probability at least $1 - \delta_0$ the inequality (2) holds for $\hat{h} = \mathbb{A}(S;T)$.

Next we extend this claim to hold for $m_0 = m$. Fix any $\delta_0 \in (0,1)$. If $m < 160\mathrm{Log}\left(\frac{6\mathrm{k}_p^2}{\delta_0}\right)$, the inequality trivially holds since the bound is greater than 1 (for sufficiently large choice of c), so suppose $m \geq 160\mathrm{Log}\left(\frac{6\mathrm{k}_p^2}{\delta_0}\right)$. Consider the sets S_i and classifiers h_i as defined in the specification of the algorithm $\mathbb{A}(S;T)$ above Theorem 8, and with a slight abuse of notation we also use S_i to denote the *unlabeled* portion of the set S_i (i.e., the points x such that $(x, f^\star(x)) \in S_i$). As argued above, each h_i is correct on $T \cup S_i$.

For any h, define $\mathrm{ER}(h) = \{x : h(x) \neq f^\star(x)\}$. We claim that

$$\mathrm{ER}(\hat{h}) \subseteq \bigcup_{i,j:i\neq j} \mathrm{ER}(h_i) \cap \mathrm{ER}(h_j). \tag{3}$$

To see this, note that since $\hat{h}$ agrees with h_{maj} on $\mathcal{X}_0$, we have

$$\mathrm{ER}(\hat{h}) \subseteq (\mathcal{X} \setminus \mathcal{X}_0) \cup (\mathcal{X}_0 \cap \mathrm{ER}(h_{\mathrm{maj}})). \tag{4}$$

Furthermore, for any $x \in \mathcal{X} \setminus \mathcal{X}_0$, at least two values of i have $h_i(x)$ different from the majority of the values $h_1(x), \ldots, h_{\mathrm{k}_p+1}(x)$, which means there are at least two classifiers predicting each label

in $\mathcal{Y}$, and hence there are at least two classifiers h_i with $h_i(x) \neq f^\star(x)$. Furthermore, any x with $h_{\mathrm{maj}}(x) \neq f^\star(x)$ certainly also has at least two h_i classifiers with $h_i(x) \neq f^\star(x)$. Therefore, the set on the right hand side of (4) is contained within $\bigcup_{i,j:i\neq j} \mathrm{ER}(h_i) \cap \mathrm{ER}(h_j)$, and (3) follows.

In particular, (3) implies

$$\mathrm{er}(\hat{h}) = \mathcal{P}(\mathrm{ER}(\hat{h})) \leq \mathcal{P}\left(\bigcup_{i,j:i<j} \mathrm{ER}(h_i) \cap \mathrm{ER}(h_j)\right) \leq \sum_{i,j:i<j} \mathcal{P}(\mathrm{ER}(h_i) \cap \mathrm{ER}(h_j))\,. \tag{5}$$

The remainder of the proof will establish that each term $\mathcal{P}(\mathrm{ER}(h_i) \cap \mathrm{ER}(h_j))$ is small with high probability. This is achieved using a "Win-Win" argument showing that for every distinct i, j, either $\mathcal{P}(\mathrm{ER}(h_i)) = \mathrm{er}(h_i)$ is small, or else $\mathcal{P}(\mathrm{ER}(h_j)|\mathrm{ER}(h_i))$ is small. In either case, it will follow that $\mathcal{P}(\mathrm{ER}(h_i) \cap \mathrm{ER}(h_j))$ is small.

Specifically, we follow a "conditioning" argument.[5] We claim that, with probability at least $1 - \delta_0/3$, for every pair i, j with $i < j$, either

$$\mathrm{er}(h_i) < \frac{320}{m} \ln\left(\frac{6\mathrm{k}_p^2}{\delta_0}\right) \tag{6}$$

or else

$$|\mathrm{ER}(h_i) \cap (S_j \setminus S_i)| \geq \mathrm{er}(h_i)m/80, \tag{7}$$

where here the notation $S_j \setminus S_i$ denotes the set of samples from S that are in S_j but not S_i (distinguished by their *indices*, so that if S contains two copies of some $x \in \mathcal{X}$ and one is in S_i and the other in S_j, then the latter will still appear in $S_j \setminus S_i$).

Toward establishing the above claim, note that for each distinct i, j we have

$$\begin{aligned}
&\Pr\left(|\mathrm{ER}(h_i) \cap (S_j \setminus S_i)| < \mathrm{er}(h_i)m/80 \text{ and } \mathrm{er}(h_i) \geq \frac{320}{m} \ln\left(\frac{6\mathrm{k}_p^2}{\delta_0}\right)\right) \\
&\leq \Pr\left(|\mathrm{ER}(h_i) \cap (S_j \setminus S_i)| < (1/2)\mathrm{er}(h_i)|S_j \setminus S_i| \text{ and } \mathrm{er}(h_i) \geq \frac{320}{m} \ln\left(\frac{6\mathrm{k}_p^2}{\delta_0}\right) \text{ and } |S_j \setminus S_i| \geq \frac{m}{40}\right) \\
&\qquad + \Pr\left(|S_j \setminus S_i| < \frac{m}{40}\right).
\end{aligned} \tag{8}$$

We begin with bounding the second term. Note that $\mathbb{E}\left[|S_j \setminus S_i| \middle| S_i\right] = \left(1 - \frac{\lfloor m/4 \rfloor}{\lfloor m/2 \rfloor}\right) \lfloor m/4 \rfloor \geq m/20$ (noting that, since $m \geq 20$, we have $\lfloor m/4 \rfloor \geq m/5$ and $\lfloor m/2 \rfloor \geq m/3$). Therefore, by a multiplicative Chernoff bound[6] conditioned on S_i, and the law of total probability, we have $\Pr(|S_j \setminus S_i| < m/40) \leq e^{-m/160} \leq \frac{\delta_0}{6\mathrm{k}_p^2}$.

5. Conditioning arguments of this type originate in the work of Hanneke (2009) on ERM bounds and active learning, and were later used to analyze several different learning algorithms (e.g., Darnstädt, 2015; Hanneke, 2016b; Zhivotovskiy and Hanneke, 2018). Notably, the argument was used by Simon (2015) to analyze majority votes of independent ERMs, and was used by Hanneke (2016a) in the proof of the optimal PAC sample complexity. Its use in the present proof most closely follows this latter work.
6. As proven by Hoeffding, 1963, the moment generating function for sampling without replacement is upper bounded by the moment generating function for sampling with replacement, and hence the usual Chernoff bounds for sampling with replacement also hold for sampling without replacement.

Next we bound the first term in (8). We note that, conditioned on $|S_j \setminus S_i|$, the samples in $S_j \setminus S_i$ are i.i.d. (with distribution $\mathcal{P}$) and independent of h_i. Thus, a multiplicative Chernoff bound implies (almost surely)

$$\Pr\Big(|\mathrm{ER}(h_i) \cap (S_j \setminus S_i)| < (1/2)\mathrm{er}(h_i)|S_j \setminus S_i| \Big| h_i, |S_j \setminus S_i|\Big) \leq e^{-(1/8)\mathrm{er}(h_i)|S_j \setminus S_i|}.$$

Therefore, the first term in (8) is bounded by

$$\mathbb{E}\left[e^{-(1/8)\mathrm{er}(h_i)|S_j \setminus S_i|}\mathbb{1}\left[\mathrm{er}(h_i) \geq \frac{320}{m}\ln\left(\frac{6\mathrm{k}_p^2}{\delta_0}\right) \text{ and } |S_j \setminus S_i| \geq m/40\right]\right] \leq \frac{\delta_0}{6\mathrm{k}_p^2}.$$

Altogether, we have that (8) is at most $\frac{\delta_0}{3\mathrm{k}_p^2}$. Finally, by a union bound over all pairs i, j with $i < j$, we conclude that with probability at least $1 - \frac{\delta_0}{3}$, for every i, j with $i < j$, at least one of (6) or (7) holds, as claimed.

Since h_j is correct on S_j, it is certainly correct on $\mathrm{ER}(h_i) \cap (S_j \setminus S_i)$. Also note that the samples in $\mathrm{ER}(h_i) \cap (S_j \setminus S_i)$ are conditionally i.i.d. given h_i and $|\mathrm{ER}(h_i) \cap (S_j \setminus S_i)|$, with conditional distribution $\mathcal{P}(\cdot|\mathrm{ER}(h_i))$. We can therefore apply the classic PAC bound for ERM (Vapnik and Chervonenkis, 1974; Blumer, Ehrenfeucht, Haussler, and Warmuth, 1989), under the conditional distribution given h_i and $|\mathrm{ER}(h_i) \cap (S_j \setminus S_i)|$, together with the law of total probability, to obtain that, for any distinct i, j, with probability at least $1 - \frac{\delta_0}{3\mathrm{k}_p^2}$,

$$\mathcal{P}(\mathrm{ER}(h_j)|\mathrm{ER}(h_i)) \leq \frac{2/\ln(2)}{|\mathrm{ER}(h_i) \cap (S_j \setminus S_i)|}\left(\mathrm{dLog}\left(\frac{2e|\mathrm{ER}(h_i) \cap (S_j \setminus S_i)|}{\mathrm{d}}\right) + \mathrm{Log}\left(\frac{6\mathrm{k}_p^2}{\delta_0}\right)\right).$$

(interpreting the bound to be infinite in the case $|\mathrm{ER}(h_i) \cap (S_j \setminus S_i)| = 0$). By the union bound, this holds simultaneously for all i, j with $i < j$ with probability at least $1 - \frac{\delta_0}{3}$. Combining this with the event established above, by the union bound and monotonicity of $x \mapsto (1/x)\mathrm{Log}(ax)$, with probability at least $1 - \frac{2}{3}\delta_0$, every i, j with $i < j$ either satisfy (6) or

$$\mathcal{P}(\mathrm{ER}(h_j)|\mathrm{ER}(h_i)) \leq \frac{160/\ln(2)}{\mathrm{er}(h_i)m}\left(\mathrm{dLog}\left(\frac{\mathrm{er}(h_i)me}{40\mathrm{d}}\right) + \mathrm{Log}\left(\frac{6\mathrm{k}_p^2}{\delta_0}\right)\right). \tag{9}$$

Since S_0 is i.i.d. (with marginal distribution $\mathcal{P}$ and $f^\star$ labels) with $m/2 \leq |S_0| < m$, the inductive hypothesis and the union bound imply that, with probability at least $1 - \frac{\delta_0}{3}$, every h_i has

$$\mathrm{er}(h_i) \leq \frac{c\mathrm{k}_p^2}{m/2}\left(\mathrm{dLog}(\mathrm{k}_p) + \mathrm{Log}\left(\frac{3(\mathrm{k}_p + 1)}{\delta_0}\right)\right).$$

Plugging this into the log in (9), by the union bound we have that, with probability at least $1 - \delta_0$, every i, j with $i < j$ either satisfy (6) or $\mathcal{P}(\mathrm{ER}(h_j)|\mathrm{ER}(h_i))$ is upper bounded by

$$\frac{160/\ln(2)}{\mathrm{er}(h_i)m}\left(\mathrm{dLog}\left(c\mathrm{k}_p^2\left(\mathrm{Log}(\mathrm{k}_p) + \frac{1}{\mathrm{d}}\mathrm{Log}\left(\frac{3(\mathrm{k}_p + 1)}{\delta_0}\right)\right)\right) + \mathrm{Log}\left(\frac{6\mathrm{k}_p^2}{\delta_0}\right)\right).$$

In either case (i.e., whether (6) holds or not), on this event

$$\mathcal{P}(\mathrm{ER}(h_i) \cap \mathrm{ER}(h_j)) = \mathrm{er}(h_i)\mathcal{P}(\mathrm{ER}(h_j)|\mathrm{ER}(h_i))$$
$$\leq \frac{320}{m}\left(\mathrm{dLog}\left(c\mathrm{k}_p^2\left(\mathrm{Log}(\mathrm{k}_p) + \frac{1}{\mathrm{d}}\mathrm{Log}\left(\frac{3(\mathrm{k}_p+1)}{\delta_0}\right)\right)\right) + \mathrm{Log}\left(\frac{6\mathrm{k}_p^2}{\delta_0}\right)\right).$$

Combining this with (5) we have that, with probability at least $1-\delta_0$,

$$\mathrm{er}(\hat{h}) \leq \binom{\mathrm{k}_p+1}{2}\frac{320}{m}\left(\mathrm{dLog}\left(c\mathrm{k}_p^2\left(\mathrm{Log}(\mathrm{k}_p) + \frac{1}{\mathrm{d}}\mathrm{Log}\left(\frac{3(\mathrm{k}_p+1)}{\delta_0}\right)\right)\right) + \mathrm{Log}\left(\frac{6\mathrm{k}_p^2}{\delta_0}\right)\right).$$

Finally, simplifying the expression, and noting that the constant c only appears in a logarithmic term, one can verify that for a sufficiently large choice of numerical constant c (e.g., any $c \geq 2^{17}$ would suffice), the right hand side is at most

$$\frac{c\mathrm{k}_p^2}{m}\left(\mathrm{dLog}(\mathrm{k}_p) + \mathrm{Log}\left(\frac{1}{\delta_0}\right)\right),$$

which extends the inductive hypothesis to $m_0 = m$. The result now follows by the principle of induction. ■

Proof of Theorem 9 Fix any target concept $f^\star \in \mathbb{C}$ and any distribution $\mathcal{P}$ on $\mathcal{X}$. We will argue that, for any finite labeled data set S with $f^\star$ labels, the proper learner $\mathbb{A}_{\mathrm{ERM}}(S)$ outputs $\hat{h} \in \mathbb{C}$ correct on S, and in the case that S is m_0 i.i.d. training examples (with $\mathcal{P}$ marginal distribution and $f^\star$ labels), then for any $\delta_0 \in (0,1)$, with probability at least $1-\delta_0$, the classifier $\hat{h} = \mathbb{A}(S)$ satisfies

$$\mathrm{er}(\hat{h}) \leq \frac{c\mathrm{k}_p^3}{m_0}\left(\mathrm{dLog}(\mathrm{k}_p) + \mathrm{Log}\left(\frac{1}{\delta_0}\right)\right), \tag{10}$$

where $c \geq 1$ is an appropriate finite numerical constant. Note that Theorem 9 would immediately follow from this (since it holds for any $\mathcal{P}$ and any $f^\star \in \mathbb{C}$), taking $\delta_0 = \delta$, and noting that m_0 of size proportional to the stated bound on $\mathcal{M}_{\mathbb{A}_{\mathrm{ERM}}}(\varepsilon,\delta)$ makes the right hand side of (10) less than ε.

If $m_0 < \mathrm{k}_p + 1$, the algorithm returns $\hat{h} = \mathrm{ERM}(S)$ in Step 1, which is an element of $\mathbb{C}$ that is correct on S by definition; furthermore, the inequality trivially holds in this case, as the right hand side is greater than 1. These values of m_0 will serve as our base case in an inductive argument. Now, for induction, suppose $m \geq \mathrm{k}_p + 1$, and that for any $\delta_0 \in (0,1)$ and $m_0 < m$, for any correctly labeled data set S of size m_0, the classifier $\hat{h}$ returned by $\mathbb{A}_{\mathrm{ERM}}(S)$ is in $\mathbb{C}$ and is correct on S, and in the case that S is i.i.d. (with $\mathcal{P}$ marginal distribution and $f^\star$ labels), then with probability at least $1-\delta_0$ (10) holds.

Next we extend this claim to $m_0 = m$. Consider a run of $\mathbb{A}_{\mathrm{ERM}}(S)$ with a correctly labeled finite data set S of size m. Consider the sets S_i and classifiers h_i as defined in the algorithm, and with a slight abuse of notation we also use S_i to denote the unlabeled portion of S_i (i.e., the points x such that $(x, f^\star(x)) \in S_i$). Define $h_{\mathrm{maj}}(x) = \mathrm{Majority}(h_1(x), \ldots, h_{\mathrm{k}_p+1}(x))$ (breaking ties to favor label -1, say), and as before using the notation (1),

$$\mathcal{X}_0 = \mathcal{X}_{\{h_1,\ldots,h_{\mathrm{k}_p+1}\},\mathrm{k}_p}.$$

Since each h_i is in $\mathbb{C}$ (by the inductive hypothesis), the classifier $\hat{h} = \mathrm{Proj}_{\mathbb{C}}(h_1, \ldots, h_{\mathrm{k}_p+1})$ in Step 4 is well defined, and by definition, $\hat{h} \in \mathbb{C}$ and has $\hat{h}(x) = h_{\mathrm{maj}}(x)$ on every $x \in \mathcal{X}_0$. Note that since every $(x, y) \in S$ is included in just one set S_i, and hence is in every set $\bigcup_{j' \neq j} S_{j'}$ except $j = i$, and by the inductive hypothesis every $j \neq i$ has h_j correct on $\bigcup_{j' \neq j} S_{j'}$, we see that every $(x, y) \in S$ has $x \in \mathcal{X}_0$, and $h_{\mathrm{maj}}(x) = y$, so that this extends the claim that $\hat{h}$ is in $\mathbb{C}$ and is correct on S for the inductive proof, and all that remains is to extend the bound on the error rate to hold for $m_0 = m$.

Toward this end, consider the case that S is an i.i.d. data set of size m (with marginal distribution $\mathcal{P}$ and $f^\star$ labels). Fix any $\delta_0 \in (0, 1)$. By the inductive hypothesis each h_i is correct on $\bigcup_{j \neq i} S_j$ and has $h_i \in \mathbb{C}$. We will follow a similar "conditioning" argument to that used in the proof of Theorem 8. As in that proof, define $\mathrm{ER}(h) = \{x : h(x) \neq f^\star(x)\}$ for any classifier h. Since $\hat{h}$ agrees with h_{maj} on $\mathcal{X}_0$, we have

$$\mathrm{ER}(\hat{h}) \subseteq (\mathcal{X} \setminus \mathcal{X}_0) \cup (\mathcal{X}_0 \cap \mathrm{ER}(h_{\mathrm{maj}})). \tag{11}$$

Furthermore, for any $x \in \mathcal{X} \setminus \mathcal{X}_0$, at least two values of i have $h_i(x)$ different from the majority of the values $h_1(x), \ldots, h_{\mathrm{k}_p+1}(x)$, which means there are at least two classifiers predicting each label in $\mathcal{Y}$, and hence there are at least two classifiers h_i with $h_i(x) \neq f^\star(x)$. Furthermore, any x with $h_{\mathrm{maj}}(x) \neq f^\star(x)$ certainly also has at least two h_i classifiers with $h_i(x) \neq f^\star(x)$. Therefore, the set on the right hand side of (11) is contained within $\bigcup_{i,j:i\neq j} \mathrm{ER}(h_i) \cap \mathrm{ER}(h_j)$. In particular, this implies

$$\mathrm{er}(\hat{h}) = \mathcal{P}(\mathrm{ER}(\hat{h})) \leq \mathcal{P}\left(\bigcup_{i,j:i<j} \mathrm{ER}(h_i) \cap \mathrm{ER}(h_j)\right) \leq \sum_{i,j:i<j} \mathcal{P}(\mathrm{ER}(h_i) \cap \mathrm{ER}(h_j)). \tag{12}$$

The remainder of the proof will establish that each term $\mathcal{P}(\mathrm{ER}(h_i) \cap \mathrm{ER}(h_j))$ is small with high probability.

For any i, we have

$$\begin{aligned}
&\Pr\left(|\mathrm{ER}(h_i) \cap S_i| < (1/2)\mathrm{er}(h_i)|S_i| \text{ and } \mathrm{er}(h_i) \geq \frac{8}{|S_i|}\ln\left(\frac{3(\mathrm{k}_p+1)}{\delta_0}\right)\right) \\
&= \mathbb{E}\left[\Pr\left(|\mathrm{ER}(h_i) \cap S_i| < (1/2)\mathrm{er}(h_i)|S_i| \middle| \mathrm{er}(h_i)\right)\mathbb{1}\left[\mathrm{er}(h_i) \geq \frac{8}{|S_i|}\ln\left(\frac{3(\mathrm{k}_p+1)}{\delta_0}\right)\right]\right].
\end{aligned} \tag{13}$$

Since S_i is excluded from the training set $\bigcup_{j\neq i} S_j$ producing h_i, we have that S_i and h_i are independent random variables. Therefore, a multiplicative Chernoff bound implies that (almost surely)

$$\Pr\left(|\mathrm{ER}(h_i) \cap S_i| < (1/2)\mathrm{er}(h_i)|S_i| \middle| \mathrm{er}(h_i)\right) \leq \exp(-\mathrm{er}(h_i)|S_i|/8),$$

so that (13) is at most $\frac{\delta_0}{3(\mathrm{k}_p+1)}$. In other words, with probability at least $1 - \frac{\delta_0}{3(\mathrm{k}_p+1)}$, either

$$\mathrm{er}(h_i) < \frac{8}{|S_i|}\log\left(\frac{3(\mathrm{k}_p+1)}{\delta_0}\right) \tag{14}$$

or else

$$|\mathrm{ER}(h_i) \cap S_i| \geq (1/2)\mathrm{er}(h_i)|S_i|. \tag{15}$$

By the union bound, with probability at least $1 - \frac{\delta_0}{3}$, every i satisfies at least one of (14) or (15).

For distinct j, i, since h_j is correct on $\bigcup_{i' \neq j} S_{i'} \supseteq S_i$ it is certainly correct on $\mathrm{ER}(h_i) \cap S_i$. Also note that the samples in $\mathrm{ER}(h_i) \cap S_i$ are conditionally i.i.d. given h_i and $|\mathrm{ER}(h_i) \cap S_i|$, with conditional distribution $\mathcal{P}(\cdot|\mathrm{ER}(h_i))$. Thus, by the classic PAC bound for ERM (Vapnik and Chervonenkis, 1974; Blumer, Ehrenfeucht, Haussler, and Warmuth, 1989) (applied under the conditional distribution given h_i and $|\mathrm{ER}(h_i) \cap S_i|$) and the law of total probability, with probability at least $1 - \frac{\delta_0}{3\mathrm{k}_p^2}$,

$$\mathcal{P}(\mathrm{ER}(h_j)|\mathrm{ER}(h_i)) \leq \frac{2/\ln(2)}{|\mathrm{ER}(h_i) \cap S_i|}\left(\mathrm{dLog}\left(\frac{2e|\mathrm{ER}(h_i) \cap S_i|}{\mathrm{d}}\right) + \mathrm{Log}\left(\frac{6\mathrm{k}_p^2}{\delta_0}\right)\right).$$

By the union bound (over i, j pairs with $i < j$, and combining with the event above) and monotonicity of $x \mapsto (1/x)\mathrm{Log}(ax)$, with probability at least $1 - \frac{2}{3}\delta_0$, every pair i, j with $i < j$ have either (14) or

$$\mathcal{P}(\mathrm{ER}(h_j)|\mathrm{ER}(h_i)) \leq \frac{4/\ln(2)}{\mathrm{er}(h_i)|S_i|}\left(\mathrm{dLog}\left(\frac{e\,\mathrm{er}(h_i)|S_i|}{\mathrm{d}}\right) + \mathrm{Log}\left(\frac{6\mathrm{k}_p^2}{\delta_0}\right)\right). \tag{16}$$

Since each h_i is correct on $\bigcup_{j \neq i} S_j$, which is itself an i.i.d. data set of size strictly smaller than m and no smaller than $\mathrm{k}_p\lfloor m/(\mathrm{k}_p+1)\rfloor \geq \frac{m}{3}$ (using the fact that $\mathrm{k}_p \geq 2$, from the definition), the inductive hypothesis and the union bound imply that with probability at least $1 - \frac{\delta_0}{3}$, every h_i has

$$\mathrm{er}(h_i) \leq \frac{3c\mathrm{k}_p^3}{m}\left(\mathrm{dLog}(\mathrm{k}_p) + \mathrm{Log}\left(\frac{3(\mathrm{k}_p+1)}{\delta_0}\right)\right).$$

Plugging this into the log in (16) above, together with the fact that $|S_i| \geq \lfloor m/(\mathrm{k}_p+1)\rfloor \geq (1/2)m/(\mathrm{k}_p+1)$, we have by the union bound that with probability at least $1 - \delta_0$, every pair i, j with $i < j$ either have (14) or have that $\mathcal{P}(\mathrm{ER}(h_j)|\mathrm{ER}(h_i))$ is upper bounded by

$$\frac{(8/\ln(2))(\mathrm{k}_p+1)}{\mathrm{er}(h_i)m}\left(\mathrm{dLog}\left(\frac{3}{2}ec\mathrm{k}_p^2\left(\mathrm{Log}(\mathrm{k}_p) + \frac{1}{\mathrm{d}}\mathrm{Log}\left(\frac{6\mathrm{k}_p^2}{\delta_0}\right)\right)\right) + \mathrm{Log}\left(\frac{6\mathrm{k}_p^2}{\delta_0}\right)\right).$$

In either case (i.e., whether (14) holds or not), on this event we have

$$\begin{aligned}&\mathcal{P}(\mathrm{ER}(h_i) \cap \mathrm{ER}(h_j)) = \mathrm{er}(h_i)\mathcal{P}(\mathrm{ER}(h_j)|\mathrm{ER}(h_i)) \\ &\leq \frac{16(\mathrm{k}_p+1)}{m}\left(\mathrm{dLog}\left(\frac{3}{2}ec\mathrm{k}_p^2\left(\mathrm{Log}(\mathrm{k}_p) + \frac{1}{\mathrm{d}}\mathrm{Log}\left(\frac{6\mathrm{k}_p^2}{\delta_0}\right)\right)\right) + \mathrm{Log}\left(\frac{6\mathrm{k}_p^2}{\delta_0}\right)\right).\end{aligned}$$

Combining this with (12), we conclude that on the above event of probability at least $1 - \delta_0$,

$$\mathrm{er}(\hat{h}) \leq \binom{\mathrm{k}_p+1}{2}\frac{16(\mathrm{k}_p+1)}{m}\left(\mathrm{dLog}\left(\frac{3}{2}ec\mathrm{k}_p^2\left(\mathrm{Log}(\mathrm{k}_p) + \frac{1}{\mathrm{d}}\mathrm{Log}\left(\frac{6\mathrm{k}_p^2}{\delta_0}\right)\right)\right) + \mathrm{Log}\left(\frac{6\mathrm{k}_p^2}{\delta_0}\right)\right).$$

By simplifying the expression on the right hand side and noting that the constant c only appears in a logarithm, one can verify that for a sufficiently large choice of numerical constant c (e.g., any $c \geq e^8$ would suffice), the right hand side is at most

$$\frac{c\mathrm{k}_p^3}{m}\left(\mathrm{dLog}(\mathrm{k}_p) + \mathrm{Log}\left(\frac{1}{\delta_0}\right)\right),$$

which extends the inductive hypothesis to $m_0 = m$. The result now follows by the principle of induction. ■

Appendix C. Proofs of lower bounds

We start with the analysis of the *coupon collector's problem* and corresponding lower bounds (see e.g., (Motwani and Raghavan, 2010)). Since we need a slightly more general result (not appearing in the standard textbooks to the best of our knowledge), we present a short proof for the sake of completeness. We remark that a similar argument was used in (Simon, 2015).

Lemma 19 (Generalized coupon collector's problem) *Let $m \leq k \in \mathbb{N}$. Consider a sequence $x_1, x_2, \ldots$ of independent uniform draws from a set of size k. Assume $z \in \mathbb{N}$ satisfies that with probability at least $1/2$, the number of distinct elements among $x_1, \ldots, x_z$ is at least $k - m$. Then,*

$$z \geq k \left(\ln \frac{k}{m} - 1 - \sqrt{\frac{2}{m}} \right).$$

Proof Let Z denote the random variable that counts the number of independent draws until at least $k - m$ distinct elements are present among $x_1, x_2, \ldots$. We may write $Z = \sum_{i=1}^{k-m} Z_i$, where Z_i represents the (random) number of draws after $i - 1$ distinct elements were observed and up to and including the first draw when i distinct elements have been observed. Observe that Z_i-s are independent, each having the geometric distribution with parameter $p_i = \frac{k-i+1}{k}$. Thus, $\mathbb{E}Z_i = \frac{1}{p_i}$ and $\mathrm{Var}(Z_i) = \frac{1-p_i}{p_i^2}$, which implies

$$\mathbb{E}Z = \sum_{i=1}^{k-m} \mathbb{E}Z_i = \sum_{i=1}^{k-m} \frac{k}{k-i+1} = k(H_k - H_m),$$

where $H_p = \sum_{i=1}^{p} \frac{1}{i}$ stands for the p-th Harmonic number. Further, we have

$$\mathrm{Var}(Z) = \sum_{i=1}^{k-m} \frac{k(i-1)}{(k-i+1)^2} = \sum_{j=m+1}^{k} \frac{k(k-j)}{j^2} \leq k^2 \sum_{j=m+1}^{k} \frac{1}{j^2} \leq \frac{k^2}{m}.$$

Finally, by Chebyshev's inequality and the relation $\ln p \leq H_p \leq \ln p + 1$ we have, with probability at least $\frac{1}{2}$,

$$Z \geq \mathbb{E}Z - k\sqrt{\frac{2}{m}} \geq k \left(\ln \frac{k}{m} - 1 - \sqrt{\frac{2}{m}} \right).$$

The claim follows. ■

Remark 20 *We will often use the following handy corollary of Lemma 19: under the conditions of this result, $z \geq \frac{k}{2} \ln \frac{k}{m}$, provided that $1 + \sqrt{\frac{2}{m}} \leq \frac{1}{2} \ln \frac{k}{m}$.*

Proof of Theorem 10 A lower bound $\Omega\big(\frac{\mathrm{d}}{\varepsilon}+\frac{1}{\varepsilon}\log\big(\frac{1}{\delta}\big)\big)$ holds for all learning algorithms (Vapnik and Chervonenkis, 1974; Ehrenfeucht, Haussler, Kearns, and Valiant, 1989), so we focus only on establishing a lower bound $\Omega\big(\frac{1}{\varepsilon}\log(\mathrm{k}_o)\mathbb{1}[\varepsilon\le 1/\mathrm{k}_o]\big)$ for proper learners. Without loss of generality we may assume that $\mathrm{k}_o \ge 128$ since for smaller values of k_o the lower bound is automatically established by choosing a small enough numerical constant factor. We in fact establish a stronger result, which also implies the second claim: namely, for any $k \ge 128$ such that there exists a hollow star of size k, any $\varepsilon \le 1/k$ has $\mathcal{M}_{\mathrm{prop}}(\varepsilon,\delta) \ge \frac{c}{\varepsilon}\log(k)$ (for a numerical constant $c>0$). Note that both of the claimed lower bounds will follow from this, since when $\mathrm{k}_o < \infty$ there exists a hollow star of size k_o, and when $\mathrm{k}_o = \infty$ there exists a sequence $k_i \to \infty$ for which there exist hollow stars of each size k_i, so that choosing $\varepsilon_i = 1/k_i$ the lower bound $\frac{c}{\varepsilon_i}\log\Big(\frac{1}{\varepsilon_i}\Big)$ holds for each ε_i.

Fix any $k \ge 128$ such that there exists a hollow star set $S = \{(x_1,y_1),\ldots,(x_k,y_k)\}$, and for each $i \in \{1,\ldots,k\}$ let $h_i \in \mathbb{C}$ be such that $\{j : h_i(x_j) \ne y_j\} = \{i\}$. Fix any proper learning algorithm $\mathbb{A}'$. We construct a target function $f^\star$ and distribution $\mathcal{P}$ to witness the lower bound via the probabilistic method. Let $\varepsilon \le 1/k$ and choose $i^* \sim \mathrm{Uniform}(\{2,\ldots,k\})$, and set $f^\star = h_{i^*}$ and $\mathcal{P}(\{x_i\}) = \varepsilon/(1-\varepsilon)$ for $i \in \{2,\ldots,k\}\setminus\{i^*\}$, $\mathcal{P}(\{x_1\}) = 1-(k-2)\varepsilon/(1-\varepsilon)$ (which is greater than ε), and $\mathcal{P}(\{x_{i^*}\}) = 0$. Consider running $\mathbb{A}'$ with a data set D_n (conditionally i.i.d. given i^*, with each point (X,Y) having $X \sim \mathcal{P}$ and $Y = f^\star(X)$), of some size $n < \frac{1}{8}\frac{1-\varepsilon}{\varepsilon}\ln(k-2)$, and let $\hat{h}$ be the classifier it outputs. Since $\mathbb{A}'$ is proper ($\hat{h} \in \mathbb{C}$) and $\mathbb{C}[S] = \emptyset$ (S being a hollow star), we know that $\hat{h}$ cannot realize the y_i classification of every x_i, so there must be a non-empty set $\hat{I} = \{i : \hat{h}(x_i) \ne y_i\} \ne \emptyset$. If any of these $\hat{i} \in \hat{I}$ are not equal i^*, then $\mathrm{er}(\hat{h}) > \varepsilon$. Denote by $\hat{n}$ the number of the n data points in D_n falling in $\{(x_2,y_2),\ldots,(x_k,y_k)\}\setminus\{(x_{i^*},y_{i^*})\}$, and denote by $\hat{n}_1$ the number of *distinct* elements of $\{(x_2,y_2),\ldots,(x_k,y_k)\}\setminus\{(x_{i^*},y_{i^*})\}$ observed in the data set D_n. By Markov's inequality, with probability at least $\frac{3}{4}$, we have $\hat{n} \le 4\frac{n(k-2)\varepsilon}{1-\varepsilon} < \frac{1}{2}(k-2)\ln(k-2)$. Furthermore, note that conditioned on $\hat{n}$ and i^*, the $\hat{n}$ samples in D_n falling in $\{(x_2,y_2),\ldots,(x_k,y_k)\}\setminus\{(x_{i^*},y_{i^*})\}$ are conditionally independent, with conditional distribution uniform on this set. Therefore, on the event that $\hat{n} < \frac{1}{2}(k-2)\ln(k-2)$, Lemma 19 implies that $\Pr(\hat{n}_1 < k-3 | \hat{n}, i^*) > \frac{1}{2}$ (noting that $k \ge 128$ implies $1+\sqrt{2} \le \frac{1}{2}\ln(k-2)$). Finally, note that conditioned on D_n, the variable i^* has conditional distribution uniform on the $k-1-\hat{n}_1$ values $i \in \{2,\ldots,k\}$ with $(x_i,y_i) \notin D_n$. Thus, on the event that $\hat{n}_1 < k-3$, we have that $\Pr(\hat{I} \ne \{i^*\} | \hat{I}, D_n) \ge \frac{(k-1-\hat{n}_1)-1}{k-1-\hat{n}_1} \ge \frac{1}{2}$. Altogether, we have

$$\Pr\Big(\hat{I} \ne \{i^*\}\Big) \ge \mathbb{E}\Big[\Pr\Big(\hat{I} \ne \{i^*\}\Big|\hat{I}, D_n\Big)\,\mathbb{1}[\hat{n}_1 < k-3]\Big] \ge \tfrac{1}{2}\Pr(\hat{n}_1 < k-3)$$
$$\ge \tfrac{1}{2}\mathbb{E}\big[\Pr(\hat{n}_1 < k-3|\hat{n},i^*)\,\mathbb{1}\big[\hat{n} < \tfrac{1}{2}(k-2)\ln(k-2)\big]\big] \ge \tfrac{1}{4}\Pr\big(\hat{n} < \tfrac{1}{2}(k-2)\ln(k-2)\big) \ge \tfrac{3}{16}.$$

In particular, this implies that for any proper learning algorithm $\mathbb{A}'$, if $n < \frac{1}{8}\frac{1-\varepsilon}{\varepsilon}\ln(k-2)$, there exist fixed choices of $f^\star \in \mathbb{C}$ and $\mathcal{P}$ such that, with probability at least $3/16$, the classifier $\hat{h}$ returned by $\mathbb{A}'$ has $\mathrm{er}(\hat{h}) > \varepsilon$. The claim follows. ■

Proof of Theorem 11 Fix any d, k_w. Since there is already a known lower bound $\frac{c'}{\varepsilon}\big(\mathrm{d}+\mathrm{Log}\big(\frac{1}{\delta}\big)\big)$ from (Ehrenfeucht, Haussler, Kearns, and Valiant, 1989; Blumer, Ehrenfeucht, Haussler, and Warmuth, 1989; Vapnik and Chervonenkis, 1974), for PAC learning in general (for some numerical constant $c'>0$), we focus on showing a lower bound $\frac{c\mathrm{d}}{\varepsilon}\mathrm{Log}\big(\frac{\mathrm{k}_w}{\mathrm{d}} \wedge \frac{1}{\varepsilon}\big)$ for some numerical constant $c>0$. Furthermore, since $\mathrm{Log}(x) \ge 1$ (from its definition above), this lower bound again already

follows from the lower bound of Ehrenfeucht, Haussler, Kearns, and Valiant (1989) in the case that $\mathrm{k}_w < 126\mathrm{d}$ (for instance, taking $\mathcal{X} = \{1, \ldots, \mathrm{d}-1+\mathrm{k}_w\}$ and $\mathbb{C} = \{x \mapsto 2\mathbb{1}[x \in I] - 1 : I \subseteq \mathcal{X}, |I \cap \{1, \ldots, \mathrm{k}_w\}| = 1\}$, which one can easily verify has VC dimension d and dual Helly number k_w). To address the remaining case, suppose $\mathrm{k}_w \geq 126\mathrm{d}$.

In the special case $\mathrm{d} = 1$, simply take $\mathcal{X} = \{1, \ldots, \mathrm{k}_w\}$ and $\mathbb{C} = \{x \mapsto 2\mathbb{1}[x = t] - 1 : t \in \mathcal{X}\}$ the singleton classifiers. It is an easy exercise to verify that the VC dimension of $\mathbb{C}$ is indeed 1, and that $\{(x, -1) : x \in \mathcal{X}\}$ is a hollow star set of size k_w, so that Lemma 5 (together with the fact that this is clearly the largest possible hollow star set, and that $\mathbb{C}$ is finite and therefore closed) implies the dual Helly number is indeed k_w. The claimed lower bound for this case then follows from Theorem 10. To address the remaining case, for the rest of the proof suppose $\mathrm{d} \geq 2$.

Let $\mathcal{X} = \{(i, j) : i \in \{\mathrm{d}-1, \ldots, \mathrm{k}_w+\mathrm{d}-2\}, j \in \{1, \ldots, i\}\}$. For each $i \in \{\mathrm{d}-1, \ldots, \mathrm{k}_w+\mathrm{d}-2\}$ and each $J \subseteq \{1, \ldots, i\}$ with $|J| = \mathrm{d}-1$, define a classifier $h_{i,J}(i', j) = 1 - 2\mathbb{1}[i' = i]\mathbb{1}[j \notin J]$: that is, $h_{i,J}$ classifies as 1 everything that does not have first coordinate equal i, and exactly $\mathrm{d}-1$ of the points that do have first coordinate equal i. Set

$$\mathbb{C} = \{h_{i,J} : i \in \{\mathrm{d}-1, \ldots, \mathrm{k}_w + \mathrm{d} - 2\}, J \subseteq \{1, \ldots, i\}, |J| = \mathrm{d}-1\}.$$

We first show that the VC dimension of $\mathbb{C}$ is indeed d. To see that d points can be shattered, simply take $i = 2\mathrm{d}-1$ (which has $i \leq \mathrm{k}_w + \mathrm{d} - 2$ since $\mathrm{k}_w \geq \mathrm{d}+1$) and we claim that the points $(i, 1), \ldots, (i, \mathrm{d})$ are shattered: for any strict subset $J \subset \{1, \ldots, \mathrm{d}\}$, we can realize a labeling with $\{(i, j) : j \in J\}$ positive and the other $\mathrm{d} - |J|$ negative with $h_{i,J \cup J'}$ where J' is any subset of $\{\mathrm{d}+1, \ldots, 2\mathrm{d}-1\}$ with $|J'| = \mathrm{d} - 1 - |J|$; also, we can realize the all-positive labeling of these d points with $h_{\mathrm{d}-1,1:(\mathrm{d}-1)}$. To show no set of $\mathrm{d}+1$ points can be shattered, note that if $\{((i_1, j_1), -1), \ldots, ((i_{\mathrm{d}+1}, j_{\mathrm{d}+1}), -1)\}$ is realizable, then $i_1 = \cdots = i_{\mathrm{d}+1}$; but in this case, $\{((i_1, j_1), 1), \ldots, ((i_{\mathrm{d}}, j_{\mathrm{d}}), 1), ((i_{\mathrm{d}+1}, j_{\mathrm{d}+1}), -1)\}$ is not realizable, and hence no set of size $\mathrm{d}+1$ is shattered.

Next we argue that the dual Helly numer of $\mathbb{C}$ is indeed k_w. To see that it is at least k_w, note that $\{((\mathrm{k}_w + \mathrm{d} - 2, 1), -1), \ldots, ((\mathrm{k}_w + \mathrm{d} - 2, \mathrm{k}_w), -1)\}$ is a hollow star set of size k_w, so that Lemma 5 implies the dual Helly number is at least k_w. To see that it is also at most k_w, consider any unrealizable set S. If some x has $\{(x, -1), (x, 1)\} \subseteq S$, this is clearly an unrealizable subset of size $2 \leq \mathrm{k}_w$. Otherwise if no such x exists, then note that since $h_{\mathrm{d}-1,1:(\mathrm{d}-1)}$ is positive on all of $\mathcal{X}$, there must be some (i, j_-) with $((i, j_-), -1) \in S$. If there are in fact *two* points (i, j), (i', j') with $i \neq i'$ and $\{((i, j), -1), ((i', j'), -1)\} \subseteq S$, then again this is an unrealizable subset of size $2 \leq \mathrm{k}_w$. Otherwise, if every (i, j) with $((i, j), -1) \in S$ has the *same* i, then it must be that either there exist $j_1, \ldots, j_{\mathrm{d}}$ with $\{((i, j_1), 1), \ldots, ((i, j_{\mathrm{d}}), 1)\} \subset S$, in which case $\{((i, j_1), 1), \ldots, ((i, j_{\mathrm{d}}), 1), ((i, j_-), -1)\}$ is an unrealizable subset of size $\mathrm{d}+1 \leq \mathrm{k}_w$, or else there exist $j_1, \ldots, j_{i-(\mathrm{d}-2)}$ with $\{((i, j_1), -1), \ldots, ((i, j_{i-(\mathrm{d}-2)}), -1)\} \subseteq S$, in which case this is an unrealizable subset of size $i - (\mathrm{d}-2) \leq \mathrm{k}_w$. Since this covers all possible cases for the set S, we conclude that the dual Helly number is equal k_w.

Fix any $\delta \in (0, 1/100)$. For any $\varepsilon \in (1/504, 1/8)$, a lower bound $\frac{\mathrm{cd}}{\varepsilon}\mathrm{Log}\left(\frac{\mathrm{k}_w}{\mathrm{d}} \wedge \frac{1}{\varepsilon}\right)$ follows from the lower bound $\frac{c'\mathrm{d}}{\varepsilon}$ of Ehrenfeucht, Haussler, Kearns, and Valiant (1989) (for c a sufficiently small numerical constant). To address the remaining case, fix any $\varepsilon \in (0, 1/504]$. If $\varepsilon \geq \frac{\mathrm{d}-1}{4(\mathrm{k}_w-1)}$, let $i_\varepsilon = \lfloor(\mathrm{d}-1)/(4\varepsilon)\rfloor + \mathrm{d} - 1$, and otherwise let $i_\varepsilon = \mathrm{k}_w + \mathrm{d} - 2$. We prove the lower bound via the probabilistic method. Let J^* be a subset of $\{1, \ldots, i_\varepsilon\}$ with $|J^*| = \mathrm{d}-1$ chosen uniformly at random (without replacement). Let $\mathcal{P}(\{(i_\varepsilon, j)\}) = \frac{4\varepsilon}{\mathrm{d}-1}$ for every $j \in \{1, \ldots, i_\varepsilon\} \setminus J^*$, and let

$\mathcal{P}(\{(\mathrm{d}-1,1)\}) = 1-(i_\varepsilon-(\mathrm{d}-1))\frac{4\varepsilon}{\mathrm{d}-1}$, and define the target concept $f^\star = h_{i_\varepsilon,J^*}$: in particular, $f^\star((i_\varepsilon,j)) = 2\mathbb{1}[j \in J^*]-1$, and hence $\mathcal{P}$ has zero mass on the set of all points (i_ε,j) where $f^\star((i_\varepsilon,j)) = 1$ and has mass $\frac{4\varepsilon}{\mathrm{d}-1}$ on every point (i_ε,j) where $f^\star((i_\varepsilon,j)) = -1$; any remaining probability mass is placed on $(\mathrm{d}-1,1)$, which is an uninformative point (since every $h_{i,J}$ classifies it 1).

Fix any sample size $n \in \mathbb{N}$ with $n < \frac{\mathrm{d}-1}{32e\varepsilon}\ln\Big(\frac{1}{\mathrm{d}-1}\min\big\{\big\lfloor\frac{\mathrm{d}-1}{4\varepsilon}\big\rfloor, \mathrm{k}_w-1\big\}\Big)$, fix any proper learning algorithm $\mathbb{A}'$, and let $\hat{h}$ be the classifier returned by running $\mathbb{A}'$ on a conditionally i.i.d. (given J^*) training set D_n of size n (with each $(X,Y) \in D_n$ having $X \sim \mathcal{P}$ and $Y = f^\star(X)$ given J^*). Let $\mathcal{Z}_{i_\varepsilon} = \{((i_\varepsilon,j),-1) : j \in \{1,\ldots,i_\varepsilon\} \setminus J^*\}$ and $\hat{n} = |D_n \cap \mathcal{Z}_{i_\varepsilon}|$, and note that we have $\mathbb{E}[\hat{n}|J^*] = \frac{4\varepsilon}{\mathrm{d}-1}(i_\varepsilon-(\mathrm{d}-1))n$. Thus, by a Chernoff bound and the law of total probability, with probability at least $1/2$, it holds that $\hat{n} \le 1+2e\mathbb{E}[\hat{n}|J^*] = 1+\frac{8e\varepsilon}{\mathrm{d}-1}(i_\varepsilon-(\mathrm{d}-1))n$ (see Motwani and Raghavan, 2010). Combining this with the constraint on n, on this event we have $\hat{n} < 1+\frac{i_\varepsilon-(\mathrm{d}-1)}{4}\ln\Big(\frac{1}{\mathrm{d}-1}\min\big\{\big\lfloor\frac{\mathrm{d}-1}{4\varepsilon}\big\rfloor, \mathrm{k}_w-1\big\}\Big) \le \frac{i_\varepsilon-(\mathrm{d}-1)}{2}\ln\Big(\frac{i_\varepsilon-(\mathrm{d}-1)}{\mathrm{d}-1}\Big)$ (using the fact that $\varepsilon \le 1/504$). Furthermore, note that the samples in $D_n \cap \mathcal{Z}_{i_\varepsilon}$ are conditionally i.i.d. $\mathrm{Uniform}(\mathcal{Z}_{i_\varepsilon})$ given $\hat{n}$. Also note that the assumptions that $\mathrm{k}_w \ge 126\mathrm{d}$ and $\varepsilon \le 1/504$ imply $i_\varepsilon-(\mathrm{d}-1) \ge 126(\mathrm{d}-1)$, so that $1+\sqrt{\frac{2}{\mathrm{d}-1}} \le \frac{1}{2}\ln\frac{i_\varepsilon-(\mathrm{d}-1)}{\mathrm{d}-1}$. Therefore, denoting by $\hat{n}_1$ the number of *distinct* elements of $\mathcal{Z}_{i_\varepsilon}$ present in D_n, we have that, on the event that $\hat{n} < \frac{i_\varepsilon-(\mathrm{d}-1)}{2}\ln\Big(\frac{i_\varepsilon-(\mathrm{d}-1)}{\mathrm{d}-1}\Big)$, Lemma 19 implies $\Pr(\hat{n}_1 < i_\varepsilon-2(\mathrm{d}-1)|\hat{n},J^*) > \frac{1}{2}$.

Since $\mathbb{A}'$ is a proper learning algorithm, it must be that $\hat{h} = h_{\hat{i},\hat{J}}$ for some $\hat{i} \in \{\mathrm{d}-1,\ldots,\mathrm{k}_w+\mathrm{d}-2\}$ and $\hat{J} \subseteq \{1,\ldots,\hat{i}\}$ with $|\hat{J}| = \mathrm{d}-1$. If $\hat{i} \ne i_\varepsilon$, then $\mathrm{er}(\hat{h}) = (i_\varepsilon-(\mathrm{d}-1))\frac{4\varepsilon}{\mathrm{d}-1} > \varepsilon$. Otherwise, suppose $\hat{i} = i_\varepsilon$. Then $\mathrm{er}(\hat{h}) = |\hat{J} \setminus J^*|\frac{4\varepsilon}{\mathrm{d}-1}$. Note that, conditioned on D_n, the variable J^* has conditional distribution uniform on the subsets of the (size $i_\varepsilon-\hat{n}_1$) set $\{j \in \{1,\ldots,i_\varepsilon\} : ((i_\varepsilon,j),-1) \notin D_n\}$ of size $\mathrm{d}-1$. In particular, on the event $\hat{i} = i_\varepsilon$, we have $\mathbb{E}\Big[|\hat{J} \setminus J^*|\Big|D_n,\hat{J},\hat{i}\Big] \ge (\mathrm{d}-1)\frac{i_\varepsilon-\hat{n}_1-(\mathrm{d}-1)}{i_\varepsilon-\hat{n}_1}$. On the event that $\hat{n}_1 < i_\varepsilon-2(\mathrm{d}-1)$, this implies $\mathbb{E}\Big[|\hat{J} \setminus J^*|\Big|D_n,\hat{J},\hat{i}\Big] > \frac{\mathrm{d}-1}{2}$. Therefore, a Chernoff bound (for sampling without replacement; see Hoeffding, 1963) implies that, on the events that $\hat{n}_1 < i_\varepsilon-2(\mathrm{d}-1)$ and $\hat{i} = i_\varepsilon$, we have $\Pr\Big(|\hat{J} \setminus J^*| \le \frac{\mathrm{d}-1}{4}\Big|D_n,\hat{J},\hat{i}\Big) \le \exp\big\{-\frac{\mathrm{d}-1}{16}\big\} \le e^{-1/16} < 1-\frac{1}{17}$. In particular, note that if $\hat{i} = i_\varepsilon$ and $|\hat{J} \setminus J^*| > \frac{\mathrm{d}-1}{4}$, then $\mathrm{er}(\hat{h}) > \varepsilon$. Altogether, we have that

$$\begin{aligned}
\Pr(\mathrm{er}(\hat{h}) > \varepsilon) &\ge \Pr(\hat{i} \ne i_\varepsilon) + \mathbb{E}\left[\Pr\Big(|\hat{J} \setminus J^*| > \frac{\mathrm{d}-1}{4}\Big|D_n,\hat{J},\hat{i}\Big)\mathbb{1}[\hat{n}_1 < i_\varepsilon-2(\mathrm{d}-1)]\mathbb{1}[\hat{i} = i_\varepsilon]\right] \\
&\ge \frac{1}{17}\mathbb{E}\left[\Pr\Big(\hat{n}_1 < i_\varepsilon-2(\mathrm{d}-1)\Big|\hat{n},J^*\Big)\mathbb{1}\left[\hat{n} < \frac{i_\varepsilon-(\mathrm{d}-1)}{2}\ln\Big(\frac{i_\varepsilon-(\mathrm{d}-1)}{\mathrm{d}-1}\Big)\right]\right] \\
&\ge \frac{1}{34}\Pr\left(\hat{n} < \frac{i_\varepsilon-(\mathrm{d}-1)}{2}\ln\Big(\frac{i_\varepsilon-(\mathrm{d}-1)}{\mathrm{d}-1}\Big)\right) \ge \frac{1}{68} > \delta.
\end{aligned}$$

In particular, this implies that there exists a non-random choice of $f^\star \in \mathbb{C}$ and $\mathcal{P}$ such that, with probability strictly greater than δ, it holds that $\mathrm{er}(\hat{h}) > \varepsilon$. The claimed lower bound on $\mathcal{M}_{\mathrm{prop}}(\varepsilon,\delta)$ follows by simplifying the expression of the constraint on n above (which is lower-bounded by the expression in the theorem, for a sufficiently small choice of the numerical constant c).

For the final claim in the theorem, it is clear that we can extend the above construction to an infinite space by allowing all $i \in \mathbb{N}$ with $i \ge \mathrm{d}-1$, in which case $\mathrm{k}_o = \mathrm{k}_w = \infty$ (since

there exist hollow star sets of unbounded sizes, following the same argument given above), and the $\mathrm{dLog}\big(\frac{\mathrm{k}_w}{\mathrm{d}} \wedge \frac{1}{\varepsilon}\big)$ term simplifies to $\mathrm{dLog}\big(\frac{1}{\varepsilon}\big)$. ■

Appendix D. Proof of Theorem 15

The essence of the proof of this result is in fact very simple, relying only on one technical construction: a set system on the data indices. Specifically, for any $m \in \mathbb{N}$, consider a family $\mathcal{I}_m$ of subsets of $\{1, \ldots, m\}$ satisfying the following two properties, for some $T_m \in \{1, \ldots, m\}$:

(i) each $I \in \mathcal{I}_m$ has size $|I| \leq m - T_m$,

(ii) for every $i_1, i_2, \ldots i_\ell \in \{1, \ldots, m\}$ there exists $I \in \mathcal{I}_m$ such that $\{i_1, i_2, \ldots i_\ell\} \subseteq I$.

Let (κ, ρ) be a stable compression scheme of size ℓ. Fix any distribution $\mathcal{P}$, any $f^\star \in \mathbb{C}$, and any $\delta \in (0,1)$, and let $S = (X_{1:m}, f^\star(X_{1:m}))$ be such that $X_{1:m} \sim \mathcal{P}^m$. Given any family $\mathcal{I}_m$ satisfying (i) and (ii), we will establish that with probability at least $1 - \delta$,

$$\mathrm{er}(\rho(\kappa(S))) \leq \frac{1}{T_m}\left(\ln(|\mathcal{I}_m|) + \ln\left(\frac{1}{\delta}\right)\right). \tag{17}$$

As a simple example of such a family $\mathcal{I}_m$ that yields Theorem 15, consider any partition of $\{1, \ldots, m\}$ into disjoint blocks $I_1, \ldots, I_{2\ell}$, each of size either $\lceil m/(2\ell) \rceil$ or $\lfloor m/(2\ell) \rfloor$. Then we can define

$$\mathcal{I}_m = \left\{\bigcup\{I_j : j \in \mathcal{J}\} : \mathcal{J} \subseteq \{1, \ldots, 2\ell\}, |\mathcal{J}| = \ell\right\}. \tag{18}$$

This clearly satisfies the above properties, with $T_m = \ell\lfloor m/(2\ell) \rfloor$, and has size $|\mathcal{I}_m| = \binom{2\ell}{\ell} < 4^\ell$, and hence plugging this into (17) yields the bound stated in Theorem 15. We now finish the proof of Theorem 15 by establishing the bound (17).

Fix any family $\mathcal{I}_m$ satisfying (i) and (ii) above. For the set S as introduced above, for any $I \subseteq \{1, \ldots, m\}$ define $S_I = \{(X_i, f^\star(X_i)) : i \in I\}$. For any $I \in \mathcal{I}_m$, since $S_{(1:m)\setminus I}$ is independent of S_I, and property (i) implies $|S_{(1:m)\setminus I}| \geq T_m$, we have

$$\Pr\Big(\rho(\kappa(S_I)) \text{ is correct on } S_{(1:m)\setminus I} \text{ and } \mathrm{er}(\rho(\kappa(S_I))) > \varepsilon\Big) \leq (1-\varepsilon)^{T_m}.$$

However, by property (ii) there must exist at least one $I^* \in \mathcal{I}$ with $\kappa(S) \subseteq S_{I^*}$, which means $\rho(\kappa(S_{I^*})) = \rho(\kappa(S))$ (by the stability property). Thus, since $\rho(\kappa(S))$ is correct on all of S (because (κ, ρ) is a valid compression scheme), including $S_{(1:m)\setminus I^*}$, we have for this (data-dependent) choice of I^* that $\rho(\kappa(S_{I^*}))$ is correct on $S_{(1:m)\setminus I^*}$. Therefore, by basic inequalities and a union bound,

$$\begin{aligned}
\Pr(\mathrm{er}(\rho(\kappa(S))) > \varepsilon) &= \Pr(\mathrm{er}(\rho(\kappa(S_{I^*}))) > \varepsilon) \\
&\leq \Pr\big(\exists I \in \mathcal{I}_m : \rho(\kappa(S_I)) \text{ is correct on } S_{(1:m)\setminus I} \text{ and } \mathrm{er}(\rho(\kappa(S_I))) > \varepsilon\big) \\
&\leq |\mathcal{I}_m|(1-\varepsilon)^{T_m} \leq |\mathcal{I}_m| e^{-\varepsilon T_m}.
\end{aligned} \tag{19}$$

In particular, for any $\delta \in (0,1)$, choosing ε equal the expression on the right hand side of (17) makes the rightmost expression in (19) equal δ, which therefore completes the proof of the abstract bound (17). The bound in Theorem 15 follows by plugging in the family $\mathcal{I}_m$ from (18), which has $|\mathcal{I}_m| < 4^\ell$ and $T_m = \ell\lfloor m/(2\ell) \rfloor > (m - 2\ell)/2$. ■