Proceedings of Machine Learning Research vol 247:1–52, 2024 37th Annual Conference on Learning Theory

The Star Number and Eluder Dimension: Elementary Observations About the Dimensions of Disagreement

Steve Hanneke STEVE.HANNEKE@GMAIL.COM
Purdue University

Editors: Shipra Agrawal and Aaron Roth

Abstract

This article presents a number of elementary observations and relations concerning commonly-studied combinatorial dimensions from the learning theory literature on classification and reinforcement learning: namely, the star number, eluder dimension, VC dimension, Littlestone dimension, threshold dimension, and cardinality of the class. One theme of the work is understanding how these dimensions may be re-expressed as natural dimensions of the convexity space of version spaces. Specifically, we find that the star number is precisely the VC dimension of version spaces (and of their disagreement regions), whereas the eluder dimension is precisely the threshold dimension of version spaces (and of their disagreement regions). We are also interested in understanding direct relations among these dimensions. For instance, we show that there is no infinite concept class with both finite Littlestone dimension and finite star number. Moreover, any infinite concept class must have infinite eluder dimension. In both cases, we also provide quantitative relations to the cardinality of the class. For the latter result, we also show an analogous relation for real-valued functions, where the cardinality of the class is replaced by the L_∞ covering number. As another relation between star numbers and VC dimension, we provide a simple, precise, and general characterization of the VC dimension of the minimal intersection-closed class containing a given concept class: namely, the 1-centered star number of the original class. Moreover, we generalize this result to provide a unifying approach to the design of certain sample compression schemes, along with a simple combinatorial dimension characterizing its compression size: the minimum star number. We also discuss a number of implications of many of these observations. Though the proofs of the above observations are actually all incredibly simple, it is interesting that such fundamental relations among these well-known quantities appear to have heretofore gone unnoticed in the literature.

Keywords: Star number, VC dimension, Eluder dimension, Littlestone dimension, Threshold dimension, Sample compression schemes, Active learning, Online learning, Differentially private learning, Reinforcement learning, Version spaces, Convexity spaces

1. Introduction

One of the major themes of statistical learning theory is the study of abstract combinatorial dimensions, which characterize various aspects of any given learning problem. For instance, in classical supervised learning, perhaps the most well-studied fundamental quantities are the *VC dimension* (Vapnik and Chervonenkis, 1971, 1974) and *Littlestone dimension* (Littlestone, 1988). These provide precise characterizations of learnability in supervised learning, for statistical learning and online learning, respectively, for (binary) classification. In the case of the Littlestone dimension, it is known that this quantity is also fundamentally related to a combinatorial dimension known as the *threshold dimension* (Shelah, 1978), and in particular, finiteness of either implies finiteness of the other. Other combinatorial dimensions arise in the context of other learning settings, beyond traditional supervised learning. For instance, in the context of *active learning*, the advantages of

active label queries over traditional supervised (passive) learning are precisely characterized by a combinatorial dimension known as the *star number* (Hanneke and Yang, 2015). As another example, in the problem of (adversarial) *reinforcement learning* (or contextual bandits), a combinatorial dimension known as the *eluder dimension* plays a central role (Russo and Van Roy, 2013; Osband and Van Roy, 2014; Foster, Rakhlin, Simchi-Levi, and Xu, 2021).

While there are already a number of known relations among the above combinatorial dimensions (Shelah, 1978; Hanneke and Yang, 2015; Li, Kamath, Foster, and Srebro, 2022), in this article we present several new general observations about these commonly-studied quantities. Our results will be valid for any concept class $\mathbb{C}$ of functions $\mathcal{X} \to \mathcal{Y}$ for discrete classification (i.e., where the dimensions are defined under the 0-1 loss, as discussed below). A common theme in several of the results is understanding these quantities in relation to the set of possible *version spaces* (Mitchell, 1977): that is, the collection of all subsets of $\mathbb{C}$ of the form

$$\{h \in \mathbb{C} : h(S) = h^\star(S)\},$$

where S ranges over all *data sets* (finite subsets of $\mathcal{X}$), and $h^\star$ is a fixed *target concept* in $\mathbb{C}$. We will also find interesting relations involving the *regions of disagreement of version spaces*: that is, the collection of all subsets of $\mathcal{X}$ of the form

$$\{x : \exists h \in \mathbb{C} \text{ with } h(S) = h^\star(S), h(x) \neq h^\star(x)\},$$

where S again ranges over all finite subsets of $\mathcal{X}$ and $h^\star$ is a fixed target concept in $\mathbb{C}$. Specifically, we establish the following elementary facts, stated *informally* for now (formal definitions and theorems will follow below); for simplicity, we only state results for finite $\mathcal{Y}$ for now (we discuss infinite $\mathcal{Y}$ where appropriate below).

1. The star number of $\mathbb{C}$ is equal the VC dimension of its version spaces, and also equal the VC dimension of regions of disagreement of its version spaces.

2. The eluder dimension of $\mathbb{C}$ is equal the threshold dimension of its version spaces, and also equal the threshold dimension of regions of disagreement of its version spaces.

3. There is no infinite concept class with *both* finite Littlestone dimension L *and* finite star number $\mathfrak{s}$. Moreover, for finite classes $\mathbb{C}$, we prove $\mathfrak{s}\mathrm{L} = \Omega(\log(|\mathbb{C}|))$.

4. The eluder dimension is never smaller than $\Omega(\log(|\mathbb{C}|))$. Moreover, any infinite concept class admits an infinite eluder sequence. The result also extends to $\mathbb{R}$-valued functions with ε-approximate eluder dimension, which is then never smaller than $\tilde{\Omega}(\log(\mathcal{N}(3\varepsilon, \mathbb{C}, \mathrm{L}_\infty)))$, where $\mathcal{N}(\cdot, \mathbb{C}, \mathrm{L}_\infty)$ denotes the L_∞ covering number.

5. The (unique) minimal *intersection-closed* concept class $\bar{\mathbb{C}}$ containing $\mathbb{C}$ has VC dimension equal the 1-centered star number of $\mathbb{C}$.

6. There exists a (stable, unlabeled) sample compression scheme of size equal the minimum (over h) of the h-centered star number of $\mathbb{C}$. This compression scheme is based on a new general principle for sample compression called the Generalized Closure Algorithm, which unifies certain existing sample compression schemes in the literature.

7. The $\emptyset$-centered star number of regions of disagreement of version spaces equals the star number of the concept class. A similar claim is also true for the eluder dimension.

Though the proofs of these observations are all actually incredibly simple, it appears that such fundamental relations among these well-studied quantities have heretofore gone unnoticed in the literature.

These observations have immediate implications when combined with the known roles of these quantities in characterizing various learning settings. For instance, the fact that every infinite class has either infinite Littlestone dimension or infinite star number (contribution 3) implies that approximate *differentially private* learning and *active learning* are fundamentally *incompatible* for infinite concept classes, since private learning requires finite Littlestone dimension (Alon, Livni, Malliaris, and Moran, 2019; Bun, Livni, and Moran, 2020; Alon, Bun, Livni, Malliaris, and Moran, 2022) while any significant advantages of active learning over passive supervised learning would require finite star number (Hanneke and Yang, 2015). As another implication of the above results, the result giving a new bound on the size of sample compression schemes (contribution 6) provides a unified approach to defining bounded-size sample compression schemes for several families of concept classes, for which previous works presented specialized constructions for each case (e.g., it unifies the compression schemes for classes of VC dimension 1 and for intersection-closed classes). We state a number of other implications of the above results in Section E, including a new proof of a result of Hanneke (2016) giving a high-probability bound on the probability in the region of disagreement of a version space. We also provide a number of new tangential related results. As one example, along the way toward establishing contribution 3, we also prove a new result for the query complexity of Exact learning with membership queries in terms of the Littlestone dimension (via a slight modification of an existing proof).

Going beyond discrete classification settings, the above dimensions all have known natural extensions to the *regression* problem ($\mathcal{Y} = \mathbb{R}$) under the squared loss $(y, y') \mapsto (y - y')^2$. In this case, we extend the $\log(|\mathbb{C}|)$ lower bound on the eluder dimension to this alternative definition (which is, in fact, the more-commonly studied variant of the eluder dimension in this literature), where $|\mathbb{C}|$ is replaced by the L_∞ covering numbers of $\mathbb{C}$.

1.1. Notation

Throughout, we let $\mathcal{X}$ and $\mathcal{Y}$ be arbitrary non-empty sets, called the *instance space* and *label space*, respectively. Both $\mathcal{X}$ and $\mathcal{Y}$ may generally be infinite, though some results will be stated specifically for finite $\mathcal{Y}$. We always suppose $|\mathcal{Y}| \geq 2$ (to focus on non-trivial cases). We refer to any function $h : \mathcal{X} \to \mathcal{Y}$ as a *concept*. We let $\mathbb{C}$ be any non-empty set of concepts (possibly infinite), called the *concept class* (or *concept space*). A *data set* S is any finite sequence $\{(x_1, y_1), \ldots, (x_n, y_n)\} \in (\mathcal{X} \times \mathcal{Y})^n$, for any $n \in \mathbb{N} \cup \{0\}$. Such a sequence S is said to be *realizable* by $\mathbb{C}$ if $\exists h \in \mathbb{C}$ with $\forall i \leq n, h(x_i) = y_i$.

Outline of the paper: We summarize the main results of this work in the following sections, presenting the formal definitions of the associated combinatorial dimensions as we go. Section 2 presents the definitions and formal theorems relevant to contribution 1, regarding the star number and VC dimension. Section 3 presents the definitions and results relevant to contribution 3, establishing that the Littlestone dimension and star number cannot both be small for large concept classes. Section 4 presents contribution 2 relating the eluder dimension and threshold dimension,

and the formal theorems relevant to contribution 4 establishing a lower bound on the eluder dimension in terms of the cardinality of the concept class. Section 5 presents the extension of results to real-valued function classes. Section A presents additional results on the star number, including the first claim in contribution 7, and other observations about the centered star numbers, their duals, and relations to the dual VC dimension. Section B presents the definitions and results comprising contribution 5 on the minimal dimension of embedding into intersection-closed concept classes, followed by Section C presenting the definitions and formal statement of contribution 6, providing a new general sample compression scheme. Section D presents additional remarks and observations about the eluder dimension, including the second claim in contribution 7. Finally, Section E presents several implications of these results, including a new relation between the star number and a dimension from the literature on selective classification.

2. The Star Number and VC Dimension

We begin by introducing the classic *Vapnik-Chervonenkis (VC) dimension* (Vapnik and Chervonenkis, 1971, 1974).

Definition 1 (Vapnik and Chervonenkis, 1971, 1974) *For any non-empty set $\mathcal{Z}$, and any $\mathcal{D} \subseteq 2^{\mathcal{Z}}$ (where $2^{\mathcal{Z}}$ is the set of all subsets of $\mathcal{Z}$), the* VC dimension*, denoted by $\mathrm{VC}(\mathcal{D})$, is defined as the largest $n \in \mathbb{N} \cup \{0\}$ such that $\exists x_1, \ldots, x_n \in \mathcal{Z}$ for which*

$$\{D \cap \{x_1, \ldots, x_n\} : D \in \mathcal{D}\} = 2^{\{x_1, \ldots, x_n\}}.$$

Such a set $\{x_1, \ldots, x_n\}$ is said to be shattered *by $\mathcal{D}$. If no largest such n exists, define $\mathrm{VC}(\mathcal{D}) = \infty$.*

Additionally, in the case of $\mathcal{Y} = \{0, 1\}$, the definition naturally extends to concept classes $\mathbb{C}$, defining $\mathrm{VC}(\mathbb{C}) = \mathrm{VC}(\{\{x : h(x) = 1\} : h \in \mathbb{C}\})$ (i.e., equating indicator functions and sets in the natural way).

The VC dimension plays a fundamental role in determining which classes of binary functions satisfy *uniform convergence* (i.e., for i.i.d. samples, their empirical averages and expectations converge uniformly, at a distribution-independent rate; Vapnik and Chervonenkis, 1971). It also characterizes which binary-valued concept classes are PAC learnable or agnostically PAC learnable (Vapnik and Chervonenkis, 1974; Ehrenfeucht, Haussler, Kearns, and Valiant, 1989; Blumer, Ehrenfeucht, Haussler, and Warmuth, 1989; Haussler, 1992). These connections together are known as the *fundamental theorem of PAC learning* (Shalev-Shwartz and Ben-David, 2014).

A second combinatorial dimension of interest in this work is the *star number* (Hanneke and Yang, 2015).

Definition 2 (Hanneke and Yang, 2015) *For any concept class $\mathbb{C}$, for any concept h (not necessarily in $\mathbb{C}$), the star number of $\mathbb{C}$ centered at h, denoted by $\mathfrak{s}_h = \mathfrak{s}_h(\mathbb{C})$, is defined as the largest $n \in \mathbb{N} \cup \{0\}$ such that $\exists x_1, \ldots, x_n \in \mathcal{X}$ satisfying*

$$\forall i \in \{0, \ldots, n\}, \exists h_i \in \mathbb{C} \text{ with } \forall j \in \{1, \ldots, n\}, h_i(x_j) = h(x_j) \iff j \neq i.$$

Such a set $\{x_1, \ldots, x_n\}$ is called a star set *centered at h. If no largest such n exists, define $\mathfrak{s}_h = \infty$.*

In other words, a star set $x_1, \ldots, x_n$ centered at h satisfies that $\{(x_1, h(x_1)), \ldots, (x_n, h(x_n))\}$ is realizable by $\mathbb{C}$, and for any $i \in \{1, \ldots, n\}$, there exists $y_i \neq h(x_i)$ such that even if we replace $(x_i, h(x_i))$ by (x_i, y_i) the data set remains realizable by $\mathbb{C}$. Also define the star number of $\mathbb{C}$:

$$\mathfrak{s} = \mathfrak{s}(\mathbb{C}) := \sup_{h \in \mathbb{C}} \mathfrak{s}_h.$$

Equivalently, $\mathfrak{s}$ is the size of the largest star set for $\mathbb{C}$ (allowing *any* center concept). It will sometimes be useful to consider an extension of this definition to allow non-realizable centers. Namely, for any concept h, define the *extended star number* centered at h denoted by $\bar{\mathfrak{s}}_h = \bar{\mathfrak{s}}_h(\mathbb{C})$, identically to Definition 2 except replacing "$\forall i \in \{0, \ldots, n\}$" with "$\forall i \in \{1, \ldots, n\}$". In other words, the definition is the same, except that we drop the requirement that $\{(x_1, h(x_1)), \ldots, (x_n, h(x_n))\}$ be realizable by $\mathbb{C}$; we refer to such a $\{x_1, \ldots, x_n\}$ as an *extended* star set).[1] The extended star number of $\mathbb{C}$ is then defined as $\bar{\mathfrak{s}} = \bar{\mathfrak{s}}(\mathbb{C}) := \sup_h \bar{\mathfrak{s}}_h$. Defining $\bar{\mathfrak{s}}$ will allow us to state certain results more precisely. However, it is an easy exercise to verify that every concept h satisfies $\mathfrak{s}_h \leq \bar{\mathfrak{s}}_h \leq \mathfrak{s}_h + 1$, and every $h \in \mathbb{C}$ satisfies $\bar{\mathfrak{s}}_h = \mathfrak{s}_h$.

The star number was introduced by Hanneke and Yang (2015) who also proved that (in combination with the VC dimension) it characterizes the optimal query complexity of *active learning* in the PAC setting (for binary classification[2]): that is, the problem where there is an unknown *target concept* $h^\star \in \mathbb{C}$, and given i.i.d. unlabeled samples from a distribution P on $\mathcal{X}$, the algorithm may interactively request to observe labels of selected examples, and thereby aims to produce a concept $\hat{h}$ close to $h^\star$ (in $L_1(P)$) with high probability. Later works have since found that the star number plays fundamental roles in lower-order factors for traditional supervised learning with noisy labels (satisfying Massart noise; Hanneke, 2016; Zhivotovskiy and Hanneke, 2018). The star number can also be viewed as describing the maximum possible *degree* of the *one-inclusion graph* of Haussler, Littlestone, and Warmuth (1994); Daniely and Shalev-Shwartz (2014) (whereas the VC dimension is the maximum possible dimension of a *cube* in the one-inclusion graph).

2.1. The Star Number is the VC Dimension of Version Spaces

Our first result reveals an *equivalence* between these two fundamental dimensions, via a change in perspective to the set of *version spaces* and *disagreement regions* thereof. Formally, for any $n \in \mathbb{N} \cup \{0\}$ and any data set $S = \{(x_1, y_1), \ldots, (x_n, y_n)\} \in (\mathcal{X} \times \mathcal{Y})^n$, define

$$\mathbb{C}_S := \{h \in \mathbb{C} : \forall i \in \{1, \ldots, n\}, h(x_i) = y_i\},$$

the *version space* induced by S (Mitchell, 1977). In particular, $\mathbb{C}_S$ is non-empty if and only if S is realizable by $\mathbb{C}$.

1. This is related to a notion of *hollow star* set, studied by Bousquet, Hanneke, Moran, and Zhivotovskiy (2020a) (see Section A below). An extended star set may be either hollow or non-hollow, and $\bar{\mathfrak{s}}$ is the size of the largest such set.
2. It is a straightforward exercise to show that the proofs of Hanneke and Yang (2015); Hanneke (2016) also establish that the star number characterizes the query complexity of active learning for general $\mathcal{Y}$ spaces, replacing the VC dimension by the DS dimension (Brukhim, Carmon, Dinur, Moran, and Yehudayoff, 2022; Daniely and Shalev-Shwartz, 2014). Specifically, Hanneke (2016) shows that the CAL active learner identifies the target labels of n i.i.d. examples using a number of queries $\tilde{O}(\mathfrak{s} \log(n))$, in which case any passive supervised learner may be applied to these labeled examples. Applying the multiclass learner of Brukhim, Carmon, Dinur, Moran, and Yehudayoff (2022) then suffices to learn any class of finite DS dimension, with n polynomial in the relevant learning parameters.

For any $\mathbb{C}' \subseteq \mathbb{C}$, the *region of disagreement* of $\mathbb{C}'$ (Cohn, Atlas, and Ladner, 1994), denoted by $\mathrm{DIS}(\mathbb{C}')$, is defined as

$$\mathrm{DIS}(\mathbb{C}') := \{x \in \mathcal{X} : \exists h, h' \in \mathbb{C}' \text{ with } h(x) \neq h'(x)\}.$$

Define

$$\mathcal{V}(\mathbb{C}) := \left\{ \mathbb{C}_S : S \in \bigcup_{n \geq 0} (\mathcal{X} \times \mathcal{Y})^n \right\},$$

the set of all version spaces of $\mathbb{C}$, and define

$$\mathcal{D}(\mathbb{C}) := \{\mathrm{DIS}(V) : V \in \mathcal{V}(\mathbb{C})\},$$

the set of all regions of disagreement of version spaces of $\mathbb{C}$. Also, for any concept h, let

$$\mathcal{V}_h(\mathbb{C}) := \{\mathbb{C}_S : S = \{(x_1, h(x_1)), \ldots, (x_n, h(x_n)) \in (\mathcal{X} \times \mathcal{Y})^n, n \in \mathbb{N} \cup \{0\}\}$$
$$\textit{and} \qquad \mathcal{D}_h(\mathbb{C}) := \{\mathrm{DIS}(V) : V \in \mathcal{V}_h(\mathbb{C})\}$$

denote the set of version spaces of $\mathbb{C}$ and regions of disagreement of version spaces of $\mathbb{C}$ induced by data sets S consistent with the concept h.

Our first formal result of this work is the following simple observation.

Theorem 3 *For any concept class $\mathbb{C}$, we have $\mathrm{VC}(\mathcal{V}(\mathbb{C})) = \bar{\mathfrak{s}}$, and for any concept $h \in \mathbb{C}$, $\mathrm{VC}(\mathcal{D}_h(\mathbb{C})) = \mathrm{VC}(\mathcal{V}_h(\mathbb{C})) = \mathfrak{s}_h$. Moreover, for any concept $h \in \mathcal{Y}^{\mathcal{X}}$, $\mathrm{VC}(\mathcal{V}_h(\mathbb{C})) = \bar{\mathfrak{s}}_h$.*

We present a simple proof of this observation in Section F.1. While the proof itself is rather straightforward, it is noteworthy that this connection between such fundamental quantities was not previously noticed in the literature. Moreover, we present a number of implications below, in Section E, including a further relation between the star number and a dimension studied by El-Yaniv and Wiener (2010), as well as a new proof of a bound from Hanneke (2016) on the probability in the region of disagreement of a version space.

Given that Theorem 3 establishes that $\mathrm{VC}(\mathcal{D}_h(\mathbb{C})) = \mathrm{VC}(\mathcal{V}_h(\mathbb{C})) = \mathfrak{s}_h$ and $\mathrm{VC}(\mathcal{V}(\mathbb{C})) = \bar{\mathfrak{s}}$, it is natural to ask whether the latter fact extends to $\mathrm{VC}(\mathcal{D}(\mathbb{C})) = \bar{\mathfrak{s}}$. However, the technique in the proof of Theorem 3 does not seem sufficient to establish this, and we leave open the question of whether $\mathrm{VC}(\mathcal{D}(\mathbb{C})) = \bar{\mathfrak{s}}$ in general. That said, we are able to establish the following weaker relation. Its proof is included in Section F.1.

Proposition 4 *For any concept class $\mathbb{C}$, we have $\mathrm{VC}(\mathcal{D}(\mathbb{C})) \leq 2\mathfrak{s} \log_2(e|\mathcal{Y}|)$.*

We present several useful additional results, relating the star number, dual VC dimension, and dual star number, in Section A, along with an interesting discussion of a relation to abstract convexity theory.

3. There Does Not Exist an Infinite Concept Class With Both Finite Littlestone Dimension and Finite Star Number

The results of this section reveal a kind of tension between the star number and the *Littlestone dimension*, another fundamental quantity of interest in learning theory. In particular, we find there cannot exist an infinite concept class with *both* finite star number *and* finite Littlestone dimension. Formally, the Littlestone dimension is defined as follows.

Definition 5 (Littlestone, 1988; Daniely, Sabato, Ben-David, and Shalev-Shwartz, 2015) *A* Littlestone tree *is a rooted binary tree, where each node is labeled by an associated point $x \in \mathcal{X}$, and the two edges to its children are each labeled by* distinct *values $y \in \mathcal{Y}$. The tree is said to be* shattered *by a concept class $\mathbb{C}$ if, for every branch in the tree, there exists a concept $h \in \mathbb{C}$ which is consistent with the edges of the branch (in the sense that for each edge on the branch, if y is its label and it eminates from a node labeled x, then $h(x) = y$; this should hold for all edges on the branch). The* Littlestone dimension *of a concept class $\mathbb{C}$, denoted by $\mathrm{L} = \mathrm{L}(\mathbb{C})$, is defined as the largest $n \in \mathbb{N} \cup \{0\}$ for which there exists a perfect Littlestone tree of depth n shattered by $\mathbb{C}$ (where* perfect *means that all internal nodes have 2 children and all leaves have depth n). If no such largest n exists, define $\mathrm{L} = \infty$.*

The Littlestone dimension was originally introduced by Littlestone (1988) as a characterization of the optimal mistake bound in realizable-case online learning for binary classification. It was later found by Ben-David, Pál, and Shalev-Shwartz (2009); Alon, Ben-Eliezer, Dagan, Moran, Naor, and Yogev (2021) to also characterize the optimal regret in *agnostic* online learning for binary classification. It was extended by Daniely, Sabato, Ben-David, and Shalev-Shwartz (2015) to the above definition for general $\mathcal{Y}$ spaces, which they show still provides the optimal mistake bound for realizable-case *multiclass* online learning. This was recently shown to again remain the case for *agnostic* multiclass online learning (Hanneke, Moran, Raman, Subedi, and Tewari, 2023a), where L characterizes agnostic learnability and the optimal regret. The Littlestone dimension has also been found to play a fundamental role in several other learning settings. Notably, finiteness of Littlestone dimension was shown to be both necessary and sufficient for approximately *differentially private* learnability (Alon, Livni, Malliaris, and Moran, 2019; Bun, Livni, and Moran, 2020; Alon, Bun, Livni, Malliaris, and Moran, 2022). The Littlestone dimension also plays fundamental roles in characterizing query learning (Chase and Freitag, 2020), transductive online learning (Ben-David, Kushilevitz, and Mansour, 1997; Hanneke, Moran, and Shafer, 2023b), adversarially robust learning with an attack oracle (Montasser, Hanneke, and Srebro, 2021), and (in an extended definition allowing infinite ordinal values) universal learning rates (Bousquet, Hanneke, Moran, van Handel, and Yehudayoff, 2021).

As our next result, we establish a fundamental relation between the Littlestone dimension, star number, and cardinality of the concept class. In particular, the result implies that there *cannot exist* an infinite concept class with *both* finite Littlestone dimension *and* finite star number. As discussed above, both the Littlestone dimension and star number play fundamental roles in characterizing various learning settings. In light of this, this result is quite interesting, as it shows that these settings are in some sense *incompatible*. For instance, it reveals that for infinite concept classes, we should not expect significant savings in the number of labeled examples sufficient for *active learning* compared to traditional supervised learning if we require that the learning algorithm be approximately differentially private (since such savings in were shown by Hanneke and Yang, 2015

to require finite star number, whereas approximate differentially private learning was shown by Alon, Bun, Livni, Malliaris, and Moran, 2022 to require finite Littlestone dimension). Formally:

Theorem 6 *For finite $\mathcal{Y}$ and any concept class $\mathbb{C}$, if $|\mathbb{C}| = \infty$, then either $\mathrm{L} = \infty$ or $\mathfrak{s} = \infty$. This also holds for infinite $\mathcal{Y}$ if every $x \in \mathcal{X}$ has $|\{h(x) : h \in \mathbb{C}\}| < \infty$ (*not *necessarily of uniformly bounded size). Moreover, if $|\mathcal{Y}| < \infty$ and $|\mathbb{C}| < \infty$, then $\mathfrak{s}\mathrm{L} \geq \log_{|\mathcal{Y}|}(|\mathbb{C}|)$.*

The proof of this result follows from new upper and lower bounds for the problem of *Exact Learning with Membership Queries* (Angluin, 1987, 2004; Hegedüs, 1995; Hellerstein, Pillaipakkamnatt, Raghavan, and Wilkins, 1996). In this setting , there is an unknown target concept $h^\star \in \mathbb{C}$, and a learning algorithm may choose any $x_1 \in \mathcal{X}$, *query* to observe $h^\star(x_1)$ (called a *membership query*), then choose another x_2, query for $h^\star(x_2)$, and so on, up to some Q times total, at which point it must return $h^\star$. For simplicity, we will require that the algorithm be deterministic. Define $\mathrm{QC}_{\mathrm{MQ}}(\mathbb{C})$, the *query complexity* of Exact Learning $\mathbb{C}$ with Membership Queries, as the minimum $Q \in \mathbb{N}$ such that there exists a learning algorithm (as described above) such that, for every $h^\star \in \mathbb{C}$, the algorithm successfully returns $h^\star$ while making at most Q membership queries. Define $\mathrm{QC}_{\mathrm{MQ}}(\mathbb{C}) = \infty$ if no such Q exists.

For the case $|\mathcal{Y}| = 2$, Hegedüs (1995) proves $\log_2(|\mathbb{C}|) \leq \mathrm{QC}_{\mathrm{MQ}}(\mathbb{C}) \leq \mathrm{XTD}(\mathbb{C}) \log_2(|\mathbb{C}|)$, where $\mathrm{XTD}(\mathbb{C})$ is a combinatorial dimension defined by Hegedüs (1995) called the *extended teaching dimension* (a variant of the *teaching dimension* of Goldman and Kearns, 1995): namely, $\mathrm{XTD}(\mathbb{C})$ is the minimum $t \in \mathbb{N}$ such that, for every $f : \mathcal{X} \to \mathcal{Y}$ (not necessarily in $\mathbb{C}$), there exists $S \in \mathcal{X}^t$ for which $|\{h \in \mathbb{C} : h(S) = f(S)\}| \leq 1$. Define $\mathrm{XTD}(\mathbb{C}) = \infty$ if no such t exists. In words, $\mathrm{XTD}(\mathbb{C})$ is the number of examples needed to whittle down the version space to at most one function (which must be f, if $f \in \mathbb{C}$), or possibly zero functions if $f \notin \mathbb{C}$. Related results were also given by Hellerstein, Pillaipakkamnatt, Raghavan, and Wilkins (1996).

We extend the results of Hegedüs (1995) in two ways. First, we extend both the lower and upper bounds to the case of general label spaces $\mathcal{Y}$. Second, we replace the factor $\log(|\mathbb{C}|)$ in the upper bound by the *Littlestone dimension* L. These results are also of independent interest, though the proofs merely represent a straightforward extension of the original proofs of Hegedüs (1995). Formally, we establish the following result. Its proof is presented in Section G.

Theorem 7 *For any concept class $\mathbb{C}$, $\log_{|\mathcal{Y}|}(|\mathbb{C}|) \leq \mathrm{QC}_{\mathrm{MQ}}(\mathbb{C}) \leq \mathrm{XTD}(\mathbb{C})\mathrm{L}$. Moreover, even if $|\mathcal{Y}| = \infty$, as long as every $x \in \mathcal{X}$ has $\{h(x) : h \in \mathbb{C}\}$ finite, if $|\mathbb{C}| = \infty$ then $\mathrm{QC}_{\mathrm{MQ}}(\mathbb{C}) = \infty$.*

Proof of Theorem 6 The theorem follows immediately from Theorem 7, in combination with a result of Hanneke and Yang (2015) establishing $\mathrm{XTD}(\mathbb{C}) \leq \mathfrak{s}$.[3] ■

4. The Eluder Dimension, Threshold Dimension, and Cardinality of the Class

Another well-studied combinatorial dimension is the *eluder dimension* (Russo and Van Roy, 2013; Osband and Van Roy, 2014; Foster, Rakhlin, Simchi-Levi, and Xu, 2021). In the literature on contextual bandits and reinforcement learning, the following definition is referred to as the *policy eluder dimension*, to distinguish it from the eluder dimension for real-valued value functions, called

3. Though their result was stated for binary classification, we note that their proof remains valid for any label space $\mathcal{Y}$.

the *value function eluder dimension* (Foster, Rakhlin, Simchi-Levi, and Xu, 2021); we discuss the latter in Section 5.

Definition 8 (Russo and Van Roy, 2013; Foster, Rakhlin, Simchi-Levi, and Xu, 2021) *For any concept class $\mathbb{C}$ and any concept h, the* eluder dimension *of $\mathbb{C}$ centered at h, denoted by $\mathfrak{e}_h = \mathfrak{e}_h(\mathbb{C})$,*[4] *is defined as the largest $n \in \mathbb{N} \cup \{0\}$ such that $\exists x_1, \ldots, x_n \in \mathcal{X}$ satisfying*

$$\forall i \in \{1, \ldots, n\}, \exists h_i \in \mathbb{C} \text{ with } h_i(x_i) \neq h(x_i) \text{ and } \forall j < i, h_i(x_j) = h(x_j).$$

Such a sequence $\{(x_1, h(x_1)), \ldots, (x_n, h(x_n))\}$ is called an eluder sequence *centered at h. If no largest such n exists, then define $\mathfrak{e}_h = \infty$.*

One useful interpretation of this is that the eluder dimension centered at some $h \in \mathbb{C}$ is the length of the longest sequence x_i such that, for each point x_i in the sequence, even knowing all $h(x_j)$ labels for all $j \leq i-1$, there is still uncertainty about the label $h(x_i)$. Equivalently (and revealingly, for our purposes below), for $h \in \mathbb{C}$, $\mathfrak{e}_h$ is the maximum length n of a sequence $x_1, \ldots, x_n$ such that

$$\forall i \leq n, x_i \in \mathrm{DIS}\left(\mathbb{C}_{\{(x_j, h(x_j)) : j < i\}}\right).$$

Also define the eluder dimension of $\mathbb{C}$, denoted $\mathfrak{e} = \mathfrak{e}(\mathbb{C})$, as $\mathfrak{e} = \sup_{h \in \mathbb{C}} \mathfrak{e}_h$. Equivalently, $\mathfrak{e}$ is the maximum length n of a sequence $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \mathcal{Y}$ such that $\forall i \leq n, x_i \in \mathrm{DIS}\left(\mathbb{C}_{(x_j, y_j) : j < i\}}\right)$, or else $\mathfrak{e} = \infty$ if there is no maximum such value. As in Definition 8, we refer to any such sequence $(x_1, y_1), \ldots, (x_n, y_n)$ as an *eluder sequence*.

Also, we say an infinite sequence $(x_1, y_1), (x_2, y_2), \ldots$ in $\mathcal{X} \times \mathcal{Y}$ is an *infinite eluder sequence* if $\forall n \in \mathbb{N}$, the prefix $S_{n-1} = \{(x_1, y_1), \ldots, (x_{n-1}, y_{n-1})\}$ satisfies $x_n \in \mathrm{DIS}(\mathbb{C}_{S_{n-1}})$. Note that the existence of an infinite eluder sequence is, on its surface, stronger than having $\mathfrak{e} = \infty$, since the latter merely implies the existence of arbitrarily large eluder sequences, rather than a single infinite eluder sequence. Nonetheless, our Theorem 11 below will imply that $\mathfrak{e} = \infty$ *if and only if* there exists an infinite eluder sequence.

The eluder dimension was introduced by Russo and Van Roy (2013) and has been used extensively in the literature on contextual bandits and reinforcement learning and other sequential interactive learning settings (Russo and Van Roy, 2013; Osband and Van Roy, 2014; Wen and Van Roy, 2017; Ayoub, Jia, Szepesvari, Wang, and Yang, 2020; Wang, Salakhutdinov, and Yang, 2020; Foster, Rakhlin, Simchi-Levi, and Xu, 2021; Agarwal, Jin, and Zhang, 2023; Sekhari, Sridharan, Sun, and Wu, 2023; Zhu and Nowak, 2022) to upper bound the regret of certain learning algorithms, and some of its combinatorial properties have been studied by Li, Kamath, Foster, and Srebro (2022). In particular, it follows immediately from its definition that $\mathfrak{e} \geq \max\{\mathfrak{s}, \mathrm{L}\}$: that is, any star set is an eluder sequence, and similarly any branch in a shattered Littlestone tree is also an eluder sequence (Li, Kamath, Foster, and Srebro, 2022).

Many works involving the eluder dimension have focused on a variant for real-valued functions (*value functions* or *Q-functions*, in the context of contextual bandits and reinforcement learning)

4. To be fully parallel to the star number, we should refer to this as the *extended eluder dimension*, and denote it by $\bar{\mathfrak{e}}_h$, since this definition does not enforce that $\{(x_1, h(x_1)), \ldots, (x_n, h(x_n))\}$ is realizable by $\mathbb{C}$. This distinction was important for the star number, to give precise equivalences, such as in Theorems 3 and 19. However, since we do not explore analogues of Theorems 19 or 23 for the eluder dimension, this distinction will not be important for our purposes in this section, so for brevity we simply refer to this as the *eluder dimension*.

which we discuss in Section 5 below. The version for discrete functions defined above (introduced by Foster, Rakhlin, Simchi-Levi, and Xu, 2021) is referred to as the *policy eluder dimension* (where, in contextual bandits and reinforcement learning, $\mathcal{X}$ is thought of as a *state space* and $\mathcal{Y}$ as an *action space*, and a *policy* is a function $\mathcal{X} \to \mathcal{Y}$; see e.g., Foster, Rakhlin, Simchi-Levi, and Xu, 2021). In particular, Foster, Rakhlin, Simchi-Levi, and Xu (2021) provide a kind of lower bound for the optimal regret in the adversarial contextual bandits problem in the realizable case with a value gap assumption, based on the policy eluder dimension.[5]

4.1. The Eluder Dimension is the Threshold Dimension of Version Spaces and Disagreements

This section continues the theme (from Section 2.1) of revealing the fact that combinatorial parameters which characterize learning turn out to coincide precisely with natural and well-known dimensions of convexity spaces. In the case of the eluder dimension, interestingly we find that it precisely coincides with the *threshold dimension* of version spaces and of regions of disagreement thereof. The threshold dimension itself is a well-known and important quantity in set theory. It quantifies the length of the longest *chain* in a collection of sets (where a chain is a sequence $D_1 \subsetneq D_2 \subsetneq \cdots$). Specifically, the threshold dimension is defined as follows.

Definition 9 (Shelah, 1978) *For any non-empty set $\mathcal{Z}$ and any $\mathcal{D} \subseteq 2^{\mathcal{Z}}$, the* threshold dimension, *denoted by $\mathbb{T}(\mathcal{D})$, is defined as the largest $n \in \mathbb{N} \cup \{0\}$ such that $\exists D_0, D_1, \ldots, D_n \in \mathcal{D}$ with $D_0 \subsetneq D_1 \subsetneq \cdots \subsetneq D_n$. Equivalently, $\exists x_1, \ldots, x_n \in \mathcal{Z}$ for which, $\forall t \in \{0, 1, \ldots, n\}$, $\exists D_t \in \mathcal{D}$ such that $D_t \cap \{x_1, \ldots, x_n\} = \{x_1, \ldots, x_t\}$. Such a sequence $\{x_1, \ldots, x_t\}$ is said to be a* threshold set *for $\mathcal{D}$. If no largest such n exists, define $\mathbb{T}(\mathcal{D}) = \infty$.*

The threshold dimension was introduced in the context of model theory by Shelah (1978). It has recently entered the learning theory literature, playing significant roles in a number of works. In most cases, this is due to its relation to the *Littlestone dimension* L, where it is known that, for $\mathcal{Y} = \{0, 1\}$, $\mathbb{T}(\mathbb{C}) = \Omega(\log(\mathrm{L}))$ and $\mathrm{L} = \Omega(\log(\mathbb{T}(\mathbb{C})))$ (Shelah, 1978; Hodges, 1997), where here $\mathbb{T}(\mathbb{C})$ is defined by equating each h with its set $\{x : h(x) = 1\}$ (i.e., $\mathbb{T}(\mathbb{C}) := \mathbb{T}(\{\{x : h(x) = 1\} : h \in \mathbb{C}\})$). As discussed above, the Littlestone dimension is the combinatorial dimension characterizing the optimal number of mistakes in realizable *online* classification (Littlestone, 1988) and the optimal regret bound in agnostic online classification (Ben-David, Pál, and Shalev-Shwartz, 2009; Daniely, Sabato, Ben-David, and Shalev-Shwartz, 2015; Hanneke, Moran, Raman, Subedi, and Tewari, 2023a). The Littlestone dimension is also known to characterize approximate *differentially private* learnability (Alon, Livni, Malliaris, and Moran, 2019; Bun, Livni, and Moran, 2020), and indeed, the proofs establishing the necessity of finite Littlestone dimension are directly based on the threshold dimension (Alon, Livni, Malliaris, and Moran, 2019). The relation to the threshold dimension also played an important role in establishing certain *closure properties* for classes of finite Littlestone dimension (Alon, Beimel, Moran, and Stemmer, 2020). Moreover, in recent work on online learning with only an *ERM* oracle (i.e., where there only access to the concept class is via an optimization algorithm which returns a concept correct on a given realizable data set), Dagan, Daskalakis, Assos, Attias, and Fishelson (2023) directly use the threshold dimension in their

5. The actual statement of this result is somewhat nuanced (see Theorem 2.11 of Foster, Rakhlin, Simchi-Levi, and Xu, 2021). It essentially says that, for any policy class $\mathbb{C}$ and any $h^\star \in \mathbb{C}$, there exists a set of value functions inducing $\mathbb{C}$ as the corresponding set of optimal policies, such that for any algorithm achieving regret of order $\frac{T}{\mathfrak{e}_{h^\star}}$, there exists a sequence on which $h^\star$ is the optimal policy in $\mathbb{C}$ for which the regret is at least of order $\mathfrak{e}_{h^\star}$.

analysis of the number of mistakes made by a particular online learning algorithm. As we discuss in detail below, the threshold dimension (and star number) of the class $\mathbb{C}$ were also used in recent work of Li, Kamath, Foster, and Srebro (2022) to provide upper and lower bounds on the eluder dimension. Additionally, an infinite variant of threshold sets (of the type in Definition 9) also played a pivotal role in characterizations of binary games satisfying a *minimax* theorem by Hanneke, Livni, and Moran (2021); Holzman (2023).

We now formally relate the eluder dimension and threshold dimension of version spaces and regions of disagreement thereof, observing that the eluder dimension is precisely the threshold dimensions of such sets. A (simple) proof of this observation is included in Section J.

Theorem 10 *For any $\mathbb{C}$ and any concept h, $\mathfrak{e}_h = \mathbb{T}(\mathcal{V}_h(\mathbb{C}))$. Moreover, $\forall h \in \mathbb{C}$, $\mathfrak{e}_h = \mathbb{T}(\mathcal{D}_h(\mathbb{C}))$.*

Unlike the analogous results for the star number (Theorem 3, with numerous implications in Sections A, B, C, E.1), at this time it is not clear what implications the result in Theorem 10 might have in learning-theoretic contexts. But it is nonetheless interesting to note such a precise relation between such natural and well-studied combinatorial dimensions.

4.2. The Eluder Dimension is No Smaller than Log Cardinality of the Concept Class

Our next elementary observation about the eluder dimension relates this dimension to the cardinality of the concept class. Specifically, we will argue that the eluder dimension is upper and lower bounded in terms of the cardinality $|\mathbb{C}|$ of the concept class. An upper bound of $|\mathbb{C}|-1$ is known (Li, Kamath, Foster, and Srebro, 2022) (and can sometimes be sharp). Moreover, it follows immediately from the relation $\mathfrak{e} \geq \max\{\mathfrak{s}, \mathrm{L}\}$ and Theorem 6 that $\mathfrak{e} \geq \sqrt{\log_{|\mathcal{Y}|}(|\mathbb{C}|)}$, which is particularly revealing, as it indicates the eluder dimension of infinite classes is *always* infinite. The following result establishes a sharper lower bound of $\mathfrak{e} \geq \log_{|\mathcal{Y}|}(|\mathbb{C}|)$, and moreover finds that any infinite $\mathbb{C}$ admits an infinite eluder *sequence* (when $|\mathcal{Y}| < \infty$). Formally, we have the following result. Its proof is presented in Section K.

Theorem 11 *Suppose $|\mathcal{Y}| < \infty$. For any concept class $\mathbb{C}$,*

$$\log_{|\mathcal{Y}|}(|\mathbb{C}|) \leq \mathfrak{e} \leq |\mathbb{C}| - 1.$$

*Moreover, if $|\mathbb{C}| = \infty$, then there exists an infinite eluder sequence for $\mathbb{C}$. The latter also holds when $|\mathcal{Y}| = \infty$ if every $x \in \mathcal{X}$ has $\{h(x) : h \in \mathbb{C}\}$ finite (*not *necessarily of uniformly bounded size).*

The proof of the $\log_{|\mathcal{Y}|}(|\mathbb{C}|)$ lower bound, presented in Section K, is in fact quite simple. We can construct an eluder sequence of length $\log_{|\mathcal{Y}|}(|\mathbb{C}|)$ inductively: take any $x_1 \in \mathrm{DIS}(\mathbb{C})$, and by the pigeonhole principle there must exist some $y_1 \in \mathcal{Y}$ with $|\mathbb{C}_{\{(x_1,y_1)\}}| \geq \frac{1}{|\mathcal{Y}|}|\mathbb{C}|$; likewise, take any $x_2 \in \mathrm{DIS}(\mathbb{C}_{\{(x_1,y_1)\}})$, and choose $y_2 \in \mathcal{Y}$ with $|\mathbb{C}_{\{(x_1,y_1),(x_2,y_2)\}}| \geq \frac{1}{|\mathcal{Y}|}|\mathbb{C}_{\{(x_1,y_1)\}}|$; we may continue this construction for at least $\log_{|\mathcal{Y}|}(|\mathbb{C}|)$ rounds before the first time n for which $\mathrm{DIS}(\mathbb{C}_{\{(x_i,y_i)\}_{i\leq n}}) = \emptyset$. By construction, each $x_i \in \mathrm{DIS}(\mathbb{C}_{\{(x_j,y_j)\}_{j<i}})$, so that $(x_1, y_1), \ldots, (x_n, y_n)$ is indeed an eluder sequence. While this proof is quite straightforward, this simple observation appears to have heretofore gone unnoticed in the literature.

We discuss sharpness of this result, and relations to prior literature, in Section D, along with additional remarks concerning the eluder dimension.

5. Extension to Real-Valued Functions

The results above on the eluder dimension concern the case of discrete classification (equivalently, *policies*, in a contextual bandit or reinforcement learning context). However, there are also many works based on a variant of the eluder dimension for *real-valued* functions. Specifically, in this section, we consider the case of $\mathcal{Y} = [0,1]$, and we consider a *scale-sensitive* eluder dimension, originally defined by Russo and Van Roy (2013) (the following is a slight refinement of the definition, due to Foster, Rakhlin, Simchi-Levi, and Xu, 2021).

Definition 12 *For $\mathcal{Y} = [0,1]$, for any concept class $\mathbb{C}$, for any $\varepsilon > 0$ and $n \in \mathbb{N}$, we say a sequence $(x_1, y_1), \ldots, (x_n, y_n)$ in $\mathcal{X} \times [0,1]$ is an ε-eluder sequence for $\mathbb{C}$ if*

$$\forall i \in \{1, \ldots, n\}, \quad \exists h_i \in \mathbb{C} \quad \textit{with} \quad |h_i(x_i) - y_i| > \varepsilon \quad \textit{and} \quad \sum_{j<i} (h_i(x_j) - y_j)^2 \leq \varepsilon^2.$$

For any concept h, an ε-eluder sequence is said to be centered *at h if each $y_i = h(x_i)$. The ε-eluder dimension centered at h, denoted by $\mathfrak{e}_h(\varepsilon) = \mathfrak{e}_h(\varepsilon, \mathbb{C})$, is defined as the largest $n \in \mathbb{N} \cup \{0\}$ such that $\exists \varepsilon' \geq \varepsilon$ for which there exists an ε'-eluder sequence (for $\mathbb{C}$) centered at h. If no largest such n exists, then define $\mathfrak{e}_h(\varepsilon) = \infty$. Also define the ε-eluder dimension of $\mathbb{C}$, denoted by $\mathfrak{e}(\varepsilon)$, as*

$$\mathfrak{e}(\varepsilon) = \sup_{h \in \mathbb{C}} \mathfrak{e}_h(\varepsilon).$$

Comparing with Definition 8, the main distinction is that we allow the functions h_i to merely *approximate* the values of h on the prefix $x_1, \ldots, x_{i-1}$, rather than being strictly equal to h at these points. Similarly to the discrete case, we say an infinite sequence $(x_1, y_1), (x_2, y_2), \ldots$ in $\mathcal{X} \times [0,1]$ is an *infinite ε-eluder sequence* for $\mathbb{C}$ if every finite prefix $(x_1, y_1), \ldots, (x_n, y_n)$ is an ε-eluder sequence.

The ε-eluder dimension has been used and studied extensively in the literatures on reinforcement learning and contextual bandits and other sequential interactive learning settings (Russo and Van Roy, 2013; Osband and Van Roy, 2014; Wen and Van Roy, 2017; Ayoub, Jia, Szepesvari, Wang, and Yang, 2020; Wang, Salakhutdinov, and Yang, 2020; Foster, Rakhlin, Simchi-Levi, and Xu, 2021; Agarwal, Jin, and Zhang, 2023; Sekhari, Sridharan, Sun, and Wu, 2023; Zhu and Nowak, 2022).

Li, Kamath, Foster, and Srebro (2022) have investigated the expressiveness of the eluder dimension, aiming to identify which types of functions classes it would be finite for, and found that its usefulness extends beyond classical analyses of generalized linear function classes for which it had previously been shown to be finite. As discussed above, they found that the Littlesone and star number together determine whether the eluder dimension is finite in the case of discrete classification (which we have shown, in Theorem 11, is also determined by finiteness of the even-simpler quantity $|\mathbb{C}|$). However, they left open the question of whether there is a familiar combinatorial notion which determines which classes have finite ε-eluder dimension. We answer this question, finding an answer analogous to Theorem 11 from the discrete case, but with the L_∞ covering numbers of $\mathbb{C}$ in place of the cardinality $|\mathbb{C}|$. Qualitatively, we show that the ε-eluder dimension is finite if and only if the L_∞ covering numbers of the class are finite. We also give quantitative upper and lower bounds in terms of the L_∞ covering numbers, analogous to the quantitative relation to $|\mathbb{C}|$ in the discrete case (Theorem 11). Formally, we recall the following standard definition.

Definition 13 *For $\mathcal{Y} = [0,1]$, for any $\mathbb{C}$ and any $\varepsilon > 0$, define the L_∞-covering number of $\mathbb{C}$, denote $\mathcal{N}(\varepsilon, \mathbb{C}, \mathrm{L}_\infty)$, as the smallest $n \in \mathbb{N}$ such that, $\exists h_1, \ldots, h_n$ (not necessarily in $\mathbb{C}$) such that,*

$$\forall h \in \mathbb{C}, \min_{1 \le i \le n} \sup_{x \in \mathcal{X}} |h_i(x) - h(x)| \le \varepsilon.$$

If no such finite n exists, define $\mathcal{N}(\varepsilon, \mathbb{C}, \mathrm{L}_\infty) = \infty$.

The following theorem shows that the scale-sensitive eluder dimension is finite if and only if the corresponding L_∞-covering numbers of the class are finite. Its proof is presented in Section M.

Theorem 14 *For $\mathcal{Y} = [0,1]$, for any concept class $\mathbb{C}$, for any $\varepsilon > 0$, and $\delta \in (0, 1/2)$,*

$$\begin{aligned} \mathcal{N}(\varepsilon, \mathbb{C}, \mathrm{L}_\infty) = \infty &\implies \mathfrak{e}(\varepsilon) = \infty \\ \text{and } \mathcal{N}(\varepsilon\delta, \mathbb{C}, \mathrm{L}_\infty) < \infty &\implies \mathfrak{e}(\varepsilon) < \infty. \end{aligned}$$

Moreover, if $\mathcal{N}(\varepsilon, \mathbb{C}, \mathrm{L}_\infty) = \infty$, then there exists an infinite ε-eluder sequence. Additionally, we have the following quantitative relations:[6]

$$\left\lfloor \frac{2\ln(\mathcal{N}(3\varepsilon, \mathbb{C}, \mathrm{L}_\infty))}{\ln\left(\frac{4}{\varepsilon^2}\ln(\mathcal{N}(3\varepsilon, \mathbb{C}, \mathrm{L}_\infty))\right)} \right\rfloor \le \mathfrak{e}(\varepsilon) \le \left\lceil \frac{1}{(1-2\delta)^2} \right\rceil \mathcal{N}(\varepsilon\delta, \mathbb{C}, \mathrm{L}_\infty).$$

The result is particularly interesting since many common function classes are known to have infinite L_∞-covering numbers (see Anthony and Bartlett, 1999).

Summary of Additional Results in the Appendices

In addition to the proofs of all results presented above, we present a number of additional results in the appendices. Section A presents a discussion of a fascinating connection to the subject of *abstract convexity theory*, which places these results in context as connected to a number of recent findings on the importance to learning theory of various natural dimensions of the space of version spaces. It additionally presents relations of the (centered) star numbers to the well-known *dual* VC dimension, and further establishes relations between the star number of version spaces and their regions of disagreement to the star number of the concept class, along with a result establishing that the star number is nearly *self-dual*. Additionally, Section B provides a simple exact characterization of the VC dimension of the unique minimal *intersection-closed* concept class $\bar{\mathbb{C}}$ containing a given concept class $\mathbb{C}$: namely, $s_{\mathbf{1}}(\mathbb{C})$, the star number centered at the constant-1 function. In light of the results of Section A, this has further implications relating $\mathrm{VC}(\bar{\mathbb{C}})$ to the dual VC dimension of $\mathbb{C}$. Moreover, further reflection on this embedding result presented in Section C reveals a new general compression scheme of size $\mathfrak{s}_{\min} := \inf_{h \in \mathcal{Y}^{\mathcal{X}}} \mathfrak{s}_h$, the *minimum* star number (in contrast to the more well-known compression scheme of size $\mathfrak{s} = \sup_{h \in \mathbb{C}} \mathfrak{s}_h$ from Hanneke and Yang, 2015). As discussed in Section C.1, $\mathfrak{s}_{\min}$ is often significantly smaller than $\mathfrak{s}$, and moreover, this general compression scheme unifies a number of existing compression schemes in the literature. Section D presents a number of remarks about the eluder dimension, the sharpness of the upper and lower bounds established in Theorem 11, and the relation of Theorem 11 to existing results in the literature.

6. For simplicity, we interpret $\frac{0}{\ln(0)} = 0$ to handle the case $\mathcal{N}(3\varepsilon, \mathbb{C}, \mathrm{L}_\infty) = 1$.

Additionally, Section D.1 presents a precise relation between the eluder dimension of version spaces and their disagreement regions to the eluder dimension of the concept class. Section E presents further implications of the results of Sections 2.1 and A. Specifically, it presents a new proof of a result of Hanneke (2016) bounding the probability in the region of disagreement of a version space induced by i.i.d. samples, via classic generalization bounds for empirical risk minimization (largely enabled by the relation between the star number and the VC dimension of disagreement regions, established in Theorem 3). Finally, Section E.2 presents a new relation between the star number and a complexity measure introduced by El-Yaniv and Wiener (2010, 2012) for the analysis of the perfect selective classification, implied by Proposition 4 and Theorem 3.

Acknowledgments

I would like to sincerely thank Shay Moran for countless helpful discussions, comments, and insightful questions, which significantly influenced the direction of this work in ways too numerous to list. I also thank Surbhi Goel, for a discussion about intersection-closed concept classes, which inspired me to revisit the question of characterizing the minimal VC dimension of embeddings into intersection-closed classes, ultimately leading to Theorem 19.

References

A. Agarwal, Y. Jin, and T. Zhang. VOQL: Towards optimal regret in model-free RL with nonlinear function approximation. In *Proceedings of* 36^{th} *Conference on Learning Theory*, 2023.

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

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

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

N. Alon, M. Bun, R. Livni, M. Malliaris, and S. Moran. Private and online learnability are equivalent. *Journal of the ACM*, 69(4):1–34, 2022.

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

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

M. Anthony and P. L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

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

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

A. Ayoub, Z. Jia, C. Szepesvari, M. Wang, and L. Yang. Model-based reinforcement learning with value-targeted regression. In *Proceedings of the* 37^{th} *International Conference on Machine Learning*, 2020.

M.-F. Balcan, A. Beygelzimer, and J. Langford. Agnostic active learning. In *Proceedings of the* 23^{rd} *International Conference on Machine Learning*, 2006.

M.-F. Balcan, S. Hanneke, R. Pukdee, and D. Sharma. Reliable learning in challenging environments. In *Advances in Neural Information Processing Systems 36*, 2024.

S. Ben-David. 2 notes on classes with Vapnik-Chervonenkis dimension 1. *arXiv:1507.05307*, 2015.

S. Ben-David and N. Eiron. Self-directed learning and its relation to the VC-dimension and to teacher-directed learning. *Machine Learning*, 33:87–104, 1998.

S. Ben-David, E. Kushilevitz, and Y. Mansour. Online learning versus offline learning. *Machine Learning*, 29(1):45–63, 1997.

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

A. Blum, S. Hanneke, J. Qian, and H. Shao. Robust learning under clean-label attack. In *Proceedings of the* 34^{th} *Annual Conference on Learning Theory*, 2021.

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

O. Bousquet, S. Hanneke, S. Moran, and N. Zhivotovskiy. Proper learning, Helly number, and an optimal SVM bound. *arXiv:2005.11818*, 2020b.

O. Bousquet, S. Hanneke, S. Moran, R. van Handel, and A. Yehudayoff. A theory of universal learning. In *Proceedings of the* 53^{rd} *Annual ACM Symposium on Theory of Computing*, 2021.

N. Brukhim, D. Carmon, I. Dinur, S. Moran, and A. Yehudayoff. A characterization of multiclass learnability. In *Proceedings of the* 63^{rd} *Annual IEEE Symposium on Foundations of Computer Science*, 2022.

M. Bun, R. Livni, and S. Moran. An equivalence between private classification and online prediction. In *Proceedings of the* 61^{st} *Annual IEEE Symposium on Foundations of Computer Science*, 2020.

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

D. Cohn, L. Atlas, and R. Ladner. Improving generalization with active learning. *Machine Learning*, 15(2):201–221, 1994.

T. M. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, EC-14(3):326–334, 1965.

Y. Dagan, C. Daskalakis, A. Assos, I. Attias, and M. K. Fishelson. Online learning and solving infinite games with an ERM oracle. In *Proceedings of the* 36^{th} *Annual Conference on Learning Theory*, 2023.

V. Dalmau and P. Jeavons. Learnability of quantified formulas. *Theoretical Computer Science*, 306 (1–3):485–511, 2003.

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. *Journal of Machine Learning Research*, 16(12):2377–2404, 2015.

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

S. Dasgupta, D. Hsu, and C. Monteleoni. A general agnostic active learning algorithm. In *Advances in Neural Information Processing Systems* 20, 2007.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag New York, Inc., 1996.

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.

R. El-Yaniv and Y. Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(5):1605–1641, 2010.

R. El-Yaniv and Y. Wiener. Active learning via perfect selective classification. *Journal of Machine Learning Research*, 13(2):255–279, 2012.

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

D. J. Foster, A. Rakhlin, D. Simchi-Levi, and Y. Xu. Instance-dependent complexity of contextual bandits and reinforcement learning: A disagreement-based perspective. In *Proceedings of the* 34^{th} *Conference on Learning Theory*, 2021.

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

S. Hanneke. Teaching dimension and the complexity of active learning. In *Proceedings of the* 20^{th} *Conference on Learning Theory*, 2007a.

S. Hanneke. A bound on the label complexity of agnostic active learning. In *Proceedings of the* 24^{th} *International Conference on Machine Learning*, 2007b.

S. Hanneke. The complexity of interactive machine learning. Master's thesis, Machine Learning Department, Carnegie Mellon University, 2007c.

S. Hanneke. Adaptive rates of convergence in active learning. In *Proceedings of the* 22^{nd} *Conference on Learning Theory*, 2009a.

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

S. Hanneke. Rates of convergence in active learning. *The Annals of Statistics*, 39(1):333–361, 2011.

S. Hanneke. Activized learning: Transforming passive to active with improved label complexity. *Journal of Machine Learning Research*, 13(5):1469–1587, 2012.

S. Hanneke. Theory of disagreement-based active learning. *Foundations and Trends in Machine Learning*, 7(2–3):131–309, 2014.

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

S. Hanneke and A. Kontorovich. Stable sample compression schemes: New applications and an optimal SVM margin bound. In *Proceedings of the* 32^{nd} *International Conference on Algorithmic Learning Theory*, 2020.

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

S. Hanneke, A. Kontorovich, and M. Sadigurschi. Sample compression for real-valued learners. In *Proceedings of the* 30^{th} *International Conference on Algorithmic Learning Theory*, 2019.

S. Hanneke, R. Livni, and S. Moran. Online learning with simple predictors and a combinatorial characterization of minimax in 0/1 games. In *Proceedings of the* 34^{th} *Annual Conference on Learning Theory*, 2021.

S. Hanneke, S. Moran, V. Raman, U. Subedi, and A. Tewari. Multiclass online learning and uniform convergence. In *Proceedings of the* 36^{th} *Annual Conference on Learning Theory*, 2023a.

S. Hanneke, S. Moran, and J. Shafer. A trichotomy for transductive online learning. In *Advances in Neural Information Processing Systems 37*, 2023b.

D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100:78–150, 1992.

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

T. Hegedüs. Generalized teaching dimensions and the query complexity of learning. In *Proceedings of the 8th Conference on Computational Learning Theory*, 1995.

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

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

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

R. Holzman. The minimax property in infinite two-person win-lose games. *arXiv:2310.19314*, 2023.

C. Kuhlmann. On teaching and learning intersection-closed concept classes. In *Proceedings of the* 12^{th} *Conference on Learning Theory*, 1999.

G. Li, P. Kamath, D. J. Foster, and N. Srebro. Understanding the eluder dimension. In *Advances in Neural Information Processing Systems 35*, 2022.

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

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

T. Mitchell. Version spaces: A candidate elimination approach to rule learning. In *Proceedings of the* 5^{th} *International Joint Conference on Artificial Intelligence*, 1977.

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.

O. Montasser, S. Hanneke, and N. Srebro. Adversarially robust learning with unknown perturbation sets. In *Proceedings of the* 34^{th} *Annual Conference on Learning Theory*, 2021.

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.

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

A. B. Novikoff. On convergence proofs on perceptrons. In *Proceedings of the Symposium on the Mathematical Theory of Automata*, 1962.

I. Osband and B. Van Roy. Model-based reinforcement learning and the eluder dimension. In *Advances in Neural Information Processing Systems 27*, 2014.

D. Pálvölgyi and G. Tardos. Unlabeled compression schemes exceeding the VC-dimension. *Discrete Applied Mathematics*, 276:102–107, 2020.

R. L. Rivest and R. Sloan. Learning complicated concepts reliably and usefully. In *Proceedings of the First Annual Workshop on Computational Learning Theory*, 1988.

F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958.

J. Rubinstein and B. Rubinstein. Unlabelled sample compression schemes for intersection-closed classes and extremal classes. In *Advances in Neural Information Processing Systems 35*, 2022.

D. Russo and B. Van Roy. Eluder dimension and the sample complexity of optimistic exploration. In *Advances in Neural Information Processing Systems 26*, 2013.

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

A. Sekhari, K. Sridharan, W. Sun, and R. Wu. Selective sampling and imitation learning via online regression. In *Advances in Neural Information Processing Systems 36*, 2023.

S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

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

L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, November 1984.

M. L. J. van De Vel. *Theory of Convex Structures*. Elsevier, 1993.

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. *Theory of Probability and its Applications*, 16(2):264–280, 1971.

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

V. N. Vapnik and A. Y. Chervonenkis. On a class of perceptrons. *Automation and Remote Control*, 25(1), 1964a.

V. N. Vapnik and A. Y. Chervonenkis. On a class of algorithms of learning pattern recognition. *Automation and Remote Control*, 25(6), 1964b.

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

R. Wang, R. R. Salakhutdinov, and L. Yang. Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. In *Advances in Neural Information Processing Systems 33*, 2020.

M. K. Warmuth. Compressing to VC dimension many points. In *Proceedings of the* 16^{th} *Annual Conference on Learning Theory*, 2003.

Z. Wen and B. Van Roy. Efficient reinforcement learning in deterministic systems with value function generalization. *Mathematics of Operations Research*, 42(3):762–782, 2017.

Y. Wiener, S. Hanneke, and R. El-Yaniv. A compression technique for analyzing disagreement-based active learning. *Journal of Machine Learning Research*, 16(4):713–745, 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.

Y. Zhu and R. Nowak. Efficient active learning with abstention. In *Advances in Neural Information Processing Systems 35*, 2022.

Appendix A. Relating the Star Number, Dual VC Dimension, and Dual Star Number

The set $\mathcal{V}(\mathbb{C})$ is an instance of an abstract *convexity space*, a general subject which has been studied extensively in mathematics (see van De Vel, 1993). Formally, a convexity space is any set $\mathcal{V} \subseteq 2^{\mathcal{Z}}$ (for a set $\mathcal{Z}$) with $\{\mathcal{Z}, \emptyset\} \subseteq \mathcal{V}$ such that $\mathcal{V}$ is closed under intersections and monotone unions.[7]

Various dimensions for the convexity space $\mathcal{V}(\mathbb{C})$ have been found to play important roles in learning theory. In particular, when $\mathcal{Y} = \{0,1\}$, Bousquet, Hanneke, Moran, and Zhivotovskiy (2020a) have found that the *Helly* number of the convexity space $\mathcal{V}(\mathbb{C})$ characterizes the sample complexity of *proper* PAC learning (where the Helly number is a well-known quantity from abstract convexity theory; see van De Vel, 1993). Interestingly, the Helly number of $\mathcal{V}(\mathbb{C})$ is (typically) equivalent to a slightly (though consequentially) modified variant of the star number, which Bousquet, Hanneke, Moran, and Zhivotovskiy (2020a) term the *hollow star number*: namely, the maximum n such that there exists $S = \{(x_1, y_1), \ldots, (x_n, y_n)\} \in (\mathcal{X} \times \mathcal{Y})^n$ which is *not* realizable by $\mathbb{C}$ yet every i has a y_i' for which replacing (x_i, y_i) by (x_i, y_i') in S makes it realizable by $\mathbb{C}$. In other words, the only change in the definition compared to $\mathfrak{s}$ is that the "center" classification of the star should be *non-realizable* rather than realizable. Though only a small change to the definition, this turns out to have a significant effect on its value. For instance, for the concept class of linear classifiers on $\mathbb{R}^p$, the star number is infinite yet the hollow star number is $p+2$ (Bousquet, Hanneke, Moran, and Zhivotovskiy, 2020a). It is also worth noting that $\bar{\mathfrak{s}}$ is precisely the maximum of the star number and the hollow star number.

Moreover, when $\mathcal{Y} = \{0,1\}$, a (folklore) simple observation is that the well-known *dual VC dimension* (Assouad, 1983), denoted $\text{VC}^*(\mathbb{C})$, is (up to constants) equal the VC dimension of the *halfspaces* of $\mathcal{V}(\mathbb{C})$, where a halfspace in a convexity space $\mathcal{V}$ is a set V such that $\{V, V^c\} \subseteq \mathcal{V}$ where $V^c = \mathcal{Z} \setminus V$ denotes the complement (i.e., a halfspace is a convex set whose complement is also convex). In the case of $\mathcal{V}(\mathbb{C})$, the set of halfspaces is precisely given by the set

$$\text{HS}(\mathbb{C}) := \{\mathbb{C}_{\{(x,y)\}} : (x,y) \in \mathcal{X} \times \mathcal{Y}\} \cup \{\emptyset, \mathbb{C}\}$$

7. It is clear to see $\mathcal{V}(\mathbb{C})$ is a convexity space if $\mathcal{X}$ is finite, since intersections are achieved by concatenating the corresponding data sets. For infinite $\mathcal{X}$, it is strictly speaking not necessarily the case, since the finiteness of the data sets S in the definition of $\mathcal{V}(\mathbb{C})$ only ensure closure under finite intersections, and moreover there may be chains $V_1 \subset V_2 \subset \cdots$ whose limit is not in $\mathcal{V}(\mathbb{C})$ (e.g., this can occur for threshold classifiers $\mathbb{1}_{[t,\infty)}$ on $\mathbb{R}$, taking $\mathbb{C}_{\{(x_i,1)\}}$ for x_i a convergent increasing sequence). Such nuances are not important for our purposes in the present work, and indeed all of the results about $\mathcal{V}(\mathbb{C})$ in this work remain valid if we take the closure under arbitrary intersections and monotone unions. In contrast, these nuances were found to be quite consequential in the work of Bousquet, Hanneke, Moran, and Zhivotovskiy (2020a) characterizing the sample complexity of proper learning.

of version spaces $\mathbb{C}_{\{(x,y)\}}$ specified by a single labeled example $(x, y) \in \mathcal{X} \times \mathcal{Y}$ (plus the trivial halfspaces $\emptyset$ and $\mathbb{C}$). Indeed, the classical definition of $\mathrm{VC}^*(\mathbb{C})$ is in fact equivalent to the VC dimension of a further subset of $\mathrm{HS}(\mathbb{C})$, namely the sets $\mathbb{C}_{\{(x,1)\}}$ (or equivalently, the sets $\mathbb{C}_{\{(x,0)\}}$). In contrast, Theorem 3 characterizes the VC dimension of the entire convexity space $\mathcal{V}_{\mathbf{1}}(\mathbb{C})$, where $\mathbf{1}$ denotes the constant function $x \mapsto \mathbf{1}(x) := 1$. From this fact, we may immediately observe that $\bar{\mathfrak{s}}_{\mathbf{1}}$, the extended star number centered at the constant-1 function, satisfies $\bar{\mathfrak{s}} \geq \bar{\mathfrak{s}}_{\mathbf{1}} \geq \mathrm{VC}^*(\mathbb{C})$. The following corollary summarizes the above conclusions, and moreover extends this result to $\bar{\mathfrak{s}}_h$ for *any* concept h with only slight loss. It will have important implications in Sections B and C. Its proof is given in Section F.2.

Corollary 15 *In the case* $\mathcal{Y} = \{0, 1\}$, *let* $\mathrm{VC}^*(\mathbb{C})$ *denote the* dual VC dimension, *namely* $\mathrm{VC}^*(\mathbb{C}) = \mathrm{VC}(\{\mathbb{C}_{\{(x,1)\}} : x \in \mathcal{X}\})$. *It holds that*

- $\bar{\mathfrak{s}} \geq \mathrm{VC}(\mathrm{HS}(\mathbb{C})) \geq \mathrm{VC}^*(\mathbb{C})$.
- $\bar{\mathfrak{s}}_{\mathbf{1}} \geq \mathrm{VC}^*(\mathbb{C})$.
- *For any concept* h, $\bar{\mathfrak{s}}_h \geq \mathrm{VC}(\mathrm{HS}(\mathbb{C}))/2 \geq \mathrm{VC}^*(\mathbb{C})/2$.

It is also interesting to consider the star numbers of $\mathcal{V}_h(\mathbb{C})$ and $\mathcal{D}_h(\mathbb{C})$. In the case of $\mathcal{Y} = \{0, 1\}$, these are moreover related to the *dual star number* (defined analogously to the dual VC dimension). In particular, understanding these quantities will lead to concrete implications relevant to active learning and the analysis of empirical risk minimization in Section E.1. Let us overload the notation for the star number: for any non-empty set $\mathcal{Z}$, for any $D \subseteq \mathcal{Z}$ and non-empty set $\mathcal{D} \subseteq 2^{\mathcal{Z}}$, define $\mathfrak{s}_D(\mathcal{D})$, the star number of $\mathcal{D}$ centered at D, as $\mathfrak{s}_{\mathbb{1}_D}(\{\mathbb{1}_{D'} : D' \in \mathcal{D}\})$: that is, the star number of the corresponding set of indicator functions $\mathbb{1}_{D'} : \mathcal{Z} \to \{0, 1\}$ (where $\mathbb{1}_{D'}(x) = 1$ iff $x \in D'$). Similarly, define $\mathfrak{s}(\mathcal{D}) = \sup_{D \in \mathcal{D}} \mathfrak{s}_D(\mathcal{D})$.

We are interested in the relation between the star number $\mathfrak{s}_h$ of the class $\mathbb{C}$ and the star numbers of $\mathcal{V}_h(\mathbb{C})$ (the version spaces induced by h) and $\mathcal{D}_h(\mathbb{C})$ (the regions of disagreement of these version spaces). While, in general, the latter two quantities can be arbitrarily larger than $\mathfrak{s}_h$ (Proposition 17 below), it turns out the definitions all coincide in the case of certain centers: namely, the values $\mathfrak{s}_\emptyset(\mathcal{D}_h(\mathbb{C}))$ and $\mathfrak{s}_{\mathbb{C}}(\mathcal{V}_h(\mathbb{C}))$. Formally, we have the following result; its proof is presented in Section F.3. While the proof is again quite simple, this observation nonetheless has interesting implications regarding concentration inequalities for disagreement regions which we discuss in Section E.1 below.

Theorem 16 *For any* $\mathbb{C}$ *and concept* h, $\mathfrak{s}_{\mathbb{C}}(\mathcal{V}_h(\mathbb{C})) = \bar{\mathfrak{s}}_h(\mathbb{C})$. *Also,* $\forall h \in \mathbb{C}$, $\mathfrak{s}_\emptyset(\mathcal{D}_h(\mathbb{C})) = \mathfrak{s}_h(\mathbb{C})$.

In particular, it also follows from Theorems 3 and 16 that any $h \in \mathbb{C}$ satisfies

$$\mathfrak{s}_{\mathbb{C}}(\mathcal{V}_h(\mathbb{C})) = \mathrm{VC}(\mathcal{V}_h(\mathbb{C})) \quad \text{and} \quad \mathfrak{s}_\emptyset(\mathcal{D}_h(\mathbb{C})) = \mathrm{VC}(\mathcal{D}_h(\mathbb{C})).$$

One might wonder whether the equivalence in Theorem 16 extends beyond merely the $\mathbb{C}$- and $\emptyset$-centered star numbers for $\mathcal{V}_h(\mathbb{C})$ and $\mathcal{D}_h(\mathbb{C})$, respectively. The following proposition shows this is not the case, and indeed there can be an infinite gap. Its proof is presented in Section F.3.

Proposition 17 *For $\mathcal{Y} = \{0,1\}$, there exists $(\mathcal{X}, \mathbb{C})$ s.t. $\forall h \in \mathbb{C}$, $\mathfrak{s}_h(\mathbb{C}) = 2$ but $\mathfrak{s}_{\mathcal{X}}(\mathcal{D}_h(\mathbb{C})) = \infty$: namely, the class of homogeneous linear classifiers $\{x \mapsto \mathbb{1}[w^\top x \geq 0] : w \in \mathbb{R}^2, \|w\| = 1\}$ on $\mathcal{X} = \{x \in \mathbb{R}^2 : \|x\| = 1\}$, the unit circle in $\mathbb{R}^2$.*
Also, for $\mathcal{Y} = \mathcal{X} = \mathbb{N}$, there exists $\mathbb{C}$ and $h \in \mathbb{C}$ such that $\bar{\mathfrak{s}}(\mathbb{C}) = 2$ but $\mathfrak{s}_\emptyset(\mathcal{V}_h(\mathbb{C})) = \infty$: namely, the class $\{x \mapsto y : y \in \mathcal{Y}\} \cup \{x \mapsto x\}$ of all constant *functions $x \mapsto y$ plus the* identity *function $x \mapsto x$, where h is the identity function.*

It is also interesting to consider the *dual star number*, defined analogously to the dual VC dimension. Specifically, let $\mathbb{C}^* = \{h \mapsto f_x(h) := h(x) : x \in \mathcal{X}\}$, a space of *dual functions* $f_x : \mathbb{C} \to \mathcal{Y}$. Such dual functions have played a role in a number of recent works in learning theory, such as sample compression schemes (Moran and Yehudayoff, 2016), online learning (Hanneke, Livni, and Moran, 2021), and adversarially robust learning (Montasser, Hanneke, and Srebro, 2019). In particular, for $\mathcal{Y} = \{0,1\}$, the dual VC dimension $\mathrm{VC}^*(\mathbb{C})$ discussed above can equivalently be defined as $\mathrm{VC}^*(\mathbb{C}) = \mathrm{VC}(\mathbb{C}^*)$. For general label spaces $\mathcal{Y}$, we establish the following elementary relation for the *star number* of the dual class $\mathbb{C}^*$, revealing that the star number is *nearly self-dual*.

Let $\bar{\mathfrak{s}}_{\mathrm{const}}(\mathbb{C}) = \max_y \bar{\mathfrak{s}}_{x \mapsto y}(\mathbb{C})$ denote the maximum value of $\bar{\mathfrak{s}}_h(\mathbb{C})$ among all *constant* functions h (i.e., $\exists y \in \mathcal{Y}$ s.t. $h(x) = y$ for all $x \in \mathcal{X}$). Also define $\bar{\mathfrak{s}}_{\mathrm{const}}(\mathbb{C}^*)$ in this same way, but for the dual class $\mathbb{C}^*$. The proof of the following proposition is presented in Section F.3.

Proposition 18 *For any concept class $\mathbb{C}$, $\bar{\mathfrak{s}}(\mathbb{C}^*) \geq \bar{\mathfrak{s}}_{\mathrm{const}}(\mathbb{C}^*) = \bar{\mathfrak{s}}_{\mathrm{const}}(\mathbb{C}) \geq \frac{1}{|\mathcal{Y}|}\bar{\mathfrak{s}}(\mathbb{C})$ and moreover $\bar{\mathfrak{s}}(\mathbb{C}) \geq \bar{\mathfrak{s}}_{\mathrm{const}}(\mathbb{C}) = \bar{\mathfrak{s}}_{\mathrm{const}}(\mathbb{C}^*) \geq \frac{1}{|\mathcal{Y}|}\bar{\mathfrak{s}}(\mathbb{C}^*)$.*

Appendix B. The Minimal Dimension of Embedding into an Intersection-Closed Class

In this section, focusing on the case of $\mathcal{Y} = \{0,1\}$ (binary classification), we establish a new characterization of the minimum VC dimension of *embedding* any concept class $\mathbb{C}$ into an *intersection-closed* concept class. The minimal dimension turns out to be remarkably simple: namely $\mathfrak{s}_{\mathbf{1}}$, the star number centered at the constant 1 function.

Formally, a concept class $\mathbb{C}$ is said to be *intersection-closed* if, for every finite non-empty set $\mathbb{C}' \subseteq \mathbb{C}$, the concept

$$x \mapsto h_{\mathbb{C}'}(x) := \prod_{h \in \mathbb{C}'} h(x)$$

satisfies $h_{\mathbb{C}'} \in \mathbb{C}$.[8] A classic example of an intersection-closed concept class is *interval* classifiers on $\mathcal{X} = \mathbb{R}$: that is, $x \mapsto \mathbb{1}_{[a,b]}$, $a, b \in \mathbb{R}$. In $\mathbb{R}^n$, the natural generalization is the class of *axis-aligned rectangles* (Helmbold, Sloan, and Warmuth, 1990). Intersection-closed concept classes have been widely studied in learning theory, since they possess useful additional structure for specifying learning algorithms with improved sample complexity, and for yielding simple sample compression schemes (Helmbold, Sloan, and Warmuth, 1990; Haussler, Littlestone, and Warmuth, 1994; Floyd and Warmuth, 1995; Ben-David and Eiron, 1998; Kuhlmann, 1999; Dalmau and Jeavons, 2003; Auer and Ortner, 2007; Darnstädt, 2015; Hanneke, 2016; Blum, Hanneke, Qian, and Shao, 2021; Rubinstein and Rubinstein, 2022).

8. Some works distinguish between concept classes closed under finite intersections, as defined here, and concept classes closed under arbitrary intersections. This distinction will not be important for the results in this work: that is, all theorems and proofs will be valid for either definition of "intersection-closed".

Due to these favorable structures in intersection-closed classes, a natural question arises: for any given concept class $\mathbb{C}$, what is the minimum VC dimension of an intersection-closed class containing $\mathbb{C}$? The unique minimal intersection-closed concept class containing $\mathbb{C}$ can be expressed quite simply as (see e.g., Rubinstein and Rubinstein, 2022):

$$\bar{\mathbb{C}} := \left\{ h_{\mathbb{C}'} : \mathbb{C}' \subseteq \mathbb{C}, 1 \leq |\mathbb{C}'| < \infty \right\}.$$

It turns out that the VC dimension of $\bar{\mathbb{C}}$ is equally simple to state. Recall that $x \mapsto \mathbf{1}(x) = 1$ denotes the *constant* 1 function. The following theorem provides a simple characterization of $\mathrm{VC}(\bar{\mathbb{C}})$. An equally-simple proof of it is included in Section H.

Theorem 19 *For any concept class $\mathbb{C}$, it holds that $\mathrm{VC}(\bar{\mathbb{C}}) = \mathfrak{s}_{\mathbf{1}}(\mathbb{C})$.*

An immediate implication of Theorem 19, which we will discuss in great detail in Section C, is that any concept class admits a compression scheme of size $\mathfrak{s}_{\mathbf{1}}$. Indeed, Section C builds on this idea, ultimately leading to a compression scheme of size $\min_h \mathfrak{s}_h$. As another immediate corollary of Theorem 19, together with Corollary 15 and the fact that $\mathfrak{s}_{\mathbf{1}} \geq \bar{\mathfrak{s}}_{\mathbf{1}} - 1$, we can infer a relation between the intersection-closed embedding dimension and the *dual VC dimension* of $\mathbb{C}$:

$$\mathrm{VC}(\bar{\mathbb{C}}) \geq \mathrm{VC}^*(\mathbb{C}) - 1.$$

This moreover implies that any intersection-closed class $\mathbb{C}$ satisfies $\mathrm{VC}^*(\mathbb{C}) \leq \mathrm{VC}(\mathbb{C}) + 1$, which is an exponential improvement over the inequality $\mathrm{VC}^*(\mathbb{C}) \leq 2^{\mathrm{VC}(\mathbb{C})+1} - 1$ for general classes $\mathbb{C}$.

Remark 20 *We remark that Rubinstein and Rubinstein (2022) have also given a description of (a generalized variant of) $\mathrm{VC}(\bar{\mathbb{C}})$ in terms of properties of $\mathbb{C}$. However, we note that their description is substantially more-involved. Theorem 19 is noteworthy for the surprising simplicity of $\mathfrak{s}_{\mathbf{1}}$ as a characterization of $\mathrm{VC}(\bar{\mathbb{C}})$, which moreover clarifies its relation to other contexts where variants of the star number provide sharp characterizations.*

We also remark that Lemma 1 of Kuhlmann (1999) effectively states that, for any intersection-closed class $\mathbb{C}$, $\mathrm{VC}(\mathbb{C}) = \mathfrak{s}_{\mathbf{1}}(\mathbb{C})$. This implies that, for any concept class $\mathbb{C}$, $\mathrm{VC}(\bar{\mathbb{C}}) = \mathfrak{s}_{\mathbf{1}}(\bar{\mathbb{C}})$. In this light, the novelty in Theorem 19 is in observing that $\mathfrak{s}_{\mathbf{1}}(\bar{\mathbb{C}}) = \mathfrak{s}_{\mathbf{1}}(\mathbb{C})$: the $\mathbf{1}$-centered star number of the original *class $\mathbb{C}$. This itself is a rather immediate observation. However, the expression of $\mathrm{VC}(\bar{\mathbb{C}})$ in terms of a simple dimension for the* original *class $\mathbb{C}$ renders the result more useful. Moreover, as we will see in Section C, it provides strong insights which lead to more-powerful techniques going well beyond intersection-closed classes.*

Appendix C. A Sample Compression Scheme of Size Equal the Minimum Star Number

In this section, continuing to focus on the case $\mathcal{Y} = \{0, 1\}$ (binary classification), we discuss a new bound on the size of *sample compression schemes*, building from the insights of the previous section.

Sample compression schemes are a general family of learning algorithms, typically studied in the context of PAC learning, as they very easily yield generalization guarantees in that context. They are specified by a pair of functions (κ, ρ), called the *compression function* and *reconstruction function*, respectively. Given any data set S realizable by a concept class $\mathbb{C}$, $\kappa(S)$ returns a *subset* (or

subsequence) of S, and $\rho(\kappa(S))$ then evaluates to a *concept* which is required to be correct on the *entire* data set S (not just the subset $\kappa(S)$). Together, $S \mapsto \rho(\kappa(S))$ forms a PAC learning algorithm for $\mathbb{C}$ (i.e., *Probably Approximately Correct*; see Valiant, 1984; Vapnik and Chervonenkis, 1974 for background on PAC learning), with high probability error bounds scaling in the *size* of the compression scheme, meaning $|\kappa(S)|$. Compression schemes and their PAC learning guarantees were formally introduced in generality by Littlestone and Warmuth (1986) (though a number of specific compression schemes were well known in prior work; e.g., Vapnik and Chervonenkis, 1964a,b, 1974; Rosenblatt, 1958; Novikoff, 1962). By now, this subject has accumulated a substantial literature (e.g., Floyd and Warmuth, 1995; Helmbold, Sloan, and Warmuth, 1990; Devroye, Györfi, and Lugosi, 1996; Warmuth, 2003; Ben-David, 2015; Moran and Yehudayoff, 2016; Pálvölgyi and Tardos, 2020; Zhivotovskiy, 2017; Moran and Warmuth, 2016; Hanneke, Kontorovich, and Sadigurschi, 2019; Hanneke and Yang, 2015; Hanneke, 2016; Bousquet, Hanneke, Moran, and Zhivotovskiy, 2020a). Of particular interest in much of this literature is understanding the smallest possible *size* of compression schemes for a given concept class. In particular, Littlestone and Warmuth (1986) asked the question of whether every concept class $\mathbb{C}$ admits a sample compression scheme of size *bounded* as a function of the VC dimension $\mathrm{VC}(\mathbb{C})$, and later Floyd and Warmuth (1995) and Warmuth (2003) refined this question to a conjecture that every $\mathbb{C}$ admits a compression scheme of size $\mathrm{VC}(\mathbb{C})$ (or perhaps $O(\mathrm{VC}(\mathbb{C}))$), now known simply as the *sample compression conjecture*. More recently, Moran and Yehudayoff (2016) resolved the original *bounded* compression question of Littlestone and Warmuth (1986) *positively*, exhibiting a general compression scheme of size $\tilde{O}(\mathrm{VC}^*(\mathbb{C})\mathrm{VC}(\mathbb{C}))$, which is always at most $2^{O(\mathrm{VC}(\mathbb{C}))}$. However, the sharper question of Floyd and Warmuth (1995); Warmuth (2003) regarding whether compression schemes of size $O(\mathrm{VC}(\mathbb{C}))$ always exist remains open, and has been the subject of much work.

In this section, we present a new bound on the achievable size of compression schemes, quantified by the *minimum star number*. Moreover, the corresponding compression scheme is *unlabeled* and *stable* (defined below). While the bound we provide does not actually resolve the long-standing conjecture of Floyd and Warmuth (1995); Warmuth (2003), it does serve to unify a few different compression schemes from the literature, and greatly simplifies the verification of their compression size. For instance, it renders *completely trivial* the implication that classes with $\mathrm{VC}(\mathbb{C}) = 1$ admit a compression scheme of size 1, and unifies under a single theorem this fact and another well-known compression scheme of size $\mathrm{VC}(\mathbb{C})$ for intersection-closed classes (namely, the *Closure* algorithm). It also provides new results for compression schemes of size $\mathrm{VC}(\mathbb{C})$ for some families of concept classes for which such results were not previously known (e.g., all classes of $\mathrm{VC}(\mathbb{C}) = 2$ containing the subclass of *singletons*), and moreover provides the first known *stable* compression scheme for some classes. Formally, we begin with the following definition.

Definition 21 (Littlestone and Warmuth, 1986; Floyd and Warmuth, 1995) *Let $\mathbb{C}$ be any concept class. An* unlabeled *sample compression scheme is a pair of functions (κ, ρ), with $\rho : \mathcal{X}^* \to \mathcal{Y}^{\mathcal{X}}$ (called a* reconstruction function*) and with $\kappa : (\mathcal{X} \times \mathcal{Y})^* \to \mathcal{X}^*$ (called a* compression function*) such that $\forall n \in \mathbb{N} \cup \{0\}$ and $S = \{(x_1, y_1), \ldots, (x_n, y_n)\} \in (\mathcal{X} \times \mathcal{Y})^n$, $\kappa(S) \in \{x_1, \ldots, x_n\}^*$. (κ, ρ) is said to be* sample-consistent *for $\mathbb{C}$ if, for all data sets $S \in (\mathcal{X} \times \mathcal{Y})^*$ realizable by $\mathbb{C}$, $\rho(\kappa(S))(x) = y$ for every $(x, y) \in S$. The* size *of (κ, ρ) is defined as $\max_S |\kappa(S)|$ (where S ranges over data sets realizable by $\mathbb{C}$).*

A special type of sample compression scheme, known as *stable*, are of particular interest, as they are known to yield improved sample complexity guarantees when used as a learning algorithm

compared to general sample compression schemes (Bousquet, Hanneke, Moran, and Zhivotovskiy, 2020a). Formally, consider the following definition.

Definition 22 *For any concept class $\mathbb{C}$, a sample-consistent unlabeled sample compression scheme (κ, ρ) is said to be* stable *if $\forall n \in \mathbb{N} \cup \{0\}$, $\forall S = \{(x_1, y_1), \ldots, (x_n, y_n)\} \in (\mathcal{X} \times \mathcal{Y})^n$ realizable by $\mathbb{C}$, for any* subsequence *$S_\sigma \subseteq S$ with $\kappa(S) \subseteq \{x : \exists y, (x, y) \in S_\sigma\}$, it holds that $\rho(\kappa(S_\sigma)) = \rho(\kappa(S))$. That is, as long as S_σ contains the entire compression set $\kappa(S)$, the resulting classifier $\rho(\kappa(S_\sigma))$ will be the same as the classifier $\rho(\kappa(S))$ from the full original data set S.*

A fundamental result of Hanneke and Yang (2015) establishes that there always exists a (labeled, stable) sample compression scheme of size $\mathfrak{s} = \sup_{h \in \mathbb{C}} \mathfrak{s}_h$. Specifically, for any $S \in (\mathcal{X} \times \mathcal{Y})^*$ realizable by $\mathbb{C}$, there exists a subset S' of size at most $\mathfrak{s}$ such that $\mathbb{C}_{S'} = \mathbb{C}_S$. For this reason, this is sometimes referred to as a *version space compression scheme* (Wiener, Hanneke, and El-Yaniv, 2015; Hanneke and Yang, 2015). In particular, a compression function $\kappa(S)$ which returns this S' (in this case, allowing κ to include *labels* in its compression set) and $\rho(S')$ as returning any function in $\mathbb{C}_{S'}$, we immediately have a sample-consistent sample compression scheme (which can be made stable as long as there is a fixed preference order on $\mathbb{C}$ determining which element of $\mathbb{C}_{S'}$ the learner returns; see Bousquet, Hanneke, Moran, and Zhivotovskiy, 2020a).

However, since $\mathfrak{s}$ can often be quite large (already being infinite for simple classes, such as *one-dimensional intervals*), the above compression scheme is not ideal for most learning problems. However, it turns out there is a simple modification of this compression scheme which yields a *dramatic* reduction in the size. Indeed, we will propose a stable sample-consistent unlabeled compression scheme equal the *minimum* star number, defined as follows.

$$\mathfrak{s}_{\min} := \inf_{h \in \mathcal{Y}^{\mathcal{X}}} \mathfrak{s}_h.$$

It is immediate from its definition that $\mathrm{VC}(\mathbb{C}) \leq \mathfrak{s}_{\min} \leq \mathfrak{s}$, where the left inequality follows from the fact that any shattered set is a star set centered at every function. We will see in a number of examples presented in Section C.1 below that $\mathfrak{s}_{\min}$ is sometimes (though not always) closer to $\mathrm{VC}(\mathbb{C})$ than to $\mathfrak{s}$.

As alluded to, the compression scheme of this size $\mathfrak{s}_{\min}$ is actually based on a simple modification of the above version space compression scheme. For reasons we discuss below (see Remark 24), we refer to this general approach to sample compression as the *Generalized Closure Algorithm*. Specifically, consider the case that $\mathfrak{s}_{\min}$ is finite, and let $h_* = \mathrm{argmin}_h \mathfrak{s}_h$ (where h_* is not restricted to be in the class $\mathbb{C}$), breaking ties arbitrarily. For any $n \in \mathbb{N} \cup \{0\}$ and any $S = \{(x_1, y_1), \ldots, (x_n, y_n)\} \in (\mathcal{X} \times \mathcal{Y})^n$ realizable by $\mathbb{C}$, let us define $\kappa(S)$ as any minimal-size subset S' of $\{x : (x, h_*(x)) \in S\}$ such that

$$\{x : (x, h_*(x)) \in S\} \cap \mathrm{DIS}\big(\mathbb{C}_{(S', h_*(S'))}\big) = \emptyset.$$

In other words, all concepts $h \in \mathbb{C}$ with $h(S') = h_*(S')$ also agree with h_* on *every* $(x, y) \in S$ with $y = h_*(x)$. Let us then define a reconstruction function ρ, such that for such a set S' (i.e., any finite subset of $\mathcal{X}$ with $(S', h_*(S'))$ realizable by $\mathbb{C}$),

$$x \mapsto \rho(S', x) := \begin{cases} h(x), & \text{for any } h \in \mathbb{C}_{(S', h_*(S'))}, \text{ if } x \notin \mathrm{DIS}(\mathbb{C}_{(S', h_*(S'))}) \\ 1 - h_*(x), & \text{if } x \in \mathrm{DIS}(\mathbb{C}_{(S', h_*(S'))}) \end{cases}.$$

The size of the above compression scheme turns out to be bounded in terms of the minimum star number $\mathfrak{s}_{\min}$. Formally, we have the following theorem.

Theorem 23 *For any concept class $\mathbb{C}$ with $\mathfrak{s}_{\min} < \infty$, the (κ, ρ) defined by the Generalized Closure Algorithm is a stable sample-consistent unlabeled sample compression scheme of size at most $\mathfrak{s}_{\min}$.*

Remark 24 (The Generalized Closure Algorithm is the Closure algorithm) *We remark that the compression scheme (κ, ρ) defined by the Generalized Closure Algorithm has another natural interpretation as implementing the well-known* Closure algorithm *for intersection-closed concept classes, under a* re-definition *of the labels of each point $x \in \mathcal{X}$. Recall that Theorem 19 expresses the minimum VC dimension of an intersection-closed class $\bar{\mathbb{C}}$ containing $\mathbb{C}$ as precisely $\mathfrak{s}_{\mathbf{1}}$: the star number centered at the* all-1 *function. Intersection-closed concept classes $\mathbb{C}$ admit a well-known stable sample-consistent unlabeled compression scheme of size equal their VC dimension: namely, the* Closure *algorithm (Helmbold, Sloan, and Warmuth, 1990). For any data set S realizable by $\mathbb{C}$, the Closure algorithm returns a concept $h_S = \prod_{h \in \mathbb{C}_S} h$: that is, a concept which is 1 precisely on the examples for which* all *concepts in the version space $\mathbb{C}_S$ agree the label should be 1. Moreover, it turns out there always exists a subset $S' \subseteq S$ (called a* minimum spanning set*; Helmbold et al. (1990); Auer and Ortner (2007)) of size at most the VC dimension of $\mathbb{C}$, such that $h_{S'} = h_S$, and furthermore, such that every $(x, y) \in S'$ has $y = 1$. Thus, the set of x values such that $(x, 1) \in S'$ may be viewed as an unlabeled compression set, from which h_S can be reconstructed. It is also not hard to see this compression scheme is stable as well, from the definition of $h_{S'}$ (see Bousquet, Hanneke, Moran, and Zhivotovskiy, 2020a; Helmbold, Sloan, and Warmuth, 1990; Haussler, Littlestone, and Warmuth, 1994; Auer and Ortner, 2007).*

Now, to connect to the Generalized Closure Algorithm in Theorem 23, it is worth noting that there is nothing special about the all-1 *function in this context. We can choose any function h_* and simply* re-define *the labels, mapping $h_*(x)$ to 1 and $1 - h_*(x)$ to 0, for every $x \in \mathcal{X}$, so that h_** becomes *the all-1 function. Formally, this defines a concept class $\mathbb{C}^{h_*} = \{x \mapsto \mathbb{1}[h(x) = h_*(x)] : h \in \mathbb{C}\}$. We may then construct the minimal intersection-closed concept class $\bar{\mathbb{C}}$ for this $\mathbb{C}^{h_*}$, and by Theorem 19 its VC dimension will be the star number centered at the all-1 function, which, if we map the labels back to their original values, is precisely $\mathfrak{s}_{h_*}$. Thus, the Closure algorithm for this $\bar{\mathbb{C}}$ class is a stable sample-consistent unlabeled sample compression scheme of size $\mathfrak{s}_{h_*}$ (and can be easily converted to such a compression scheme for the original unmodified class $\mathbb{C}$ by reversing the re-mapping of the labels). We can then simply take h_* as the minimizer of $\mathfrak{s}_{h_*}$ to get the same guarantee as in Theorem 23. The expression of the sample compression scheme (κ, ρ) given above (the Generalized Closure Algorithm) is merely a direct statement of this compression scheme, since $\bar{\mathbb{C}}$ is itself based on taking intersections of concepts in $\mathbb{C}^{h_*}$. But these two definitions of compression schemes are in fact perfectly equivalent.*

C.1. Implications and Further Discussion of Theorem 23

Theorem 23 unifies a few different compression schemes from the literature, and moreover provides a much simpler way to calculate the achievable compression size for these classes. To start, we show this *trivially* implies that classes $\mathbb{C}$ with $\mathrm{VC}(\mathbb{C}) = 1$ admit compression schemes of size 1, by showing $\mathfrak{s}_{\min} = 1$ for such classes.

Example 1 *For any* $\mathbb{C}$ *with* $\mathrm{VC}(\mathbb{C}) = 1$*, it is quite easy to observe that* $\mathfrak{s}_{\min} = 1$*. Consider any* $h \in \mathbb{C}$*, and let* $h_* = 1 - h$*. For the sake of contradiction, suppose* $\{(x_1, h_*(x_1)), (x_2, h_*(x_2))\}$ *is a star set for* $\mathbb{C}$*. By definition, we have that* (x_1, x_2) *may be realizably labeled as* $(h_*(x_1), h_*(x_2))$*,* $(1-h_*(x_1), h_*(x_2))$*, and* $(h_*(x_1), 1-h_*(x_2))$*. But since* $h_* = 1-h$ *for some* $h \in \mathbb{C}$*, we also have that* $(h(x_1), h(x_2)) = (1 - h_*(x_1), 1 - h_*(x_2))$ *is a realizable labeling. Together, these labelings witness that* (x_1, x_2) *is shattered by* $\mathbb{C}$*, which contradicts* $\mathrm{VC}(\mathbb{C}) = 1$*. Therefore* $1 \geq \mathfrak{s}_{\min} \geq \mathrm{VC}(\mathbb{C}) = 1$*, hence we conclude that* $\mathfrak{s}_{\min} = 1$*.*

Compression schemes of size 1 *for such classes are already known to exist (Ben-David, 2015). Indeed, a careful examination reveals that the Generalized Closure Algorithm coincides precisely with the compression scheme of Ben-David (2015). Thus, we see that Theorem 23 effectively provides a unifying perspective, which expresses this compression scheme as a special case of a general principled approach to sample compression. It also provides by-far the simplest argument for the existence of such compression schemes for classes of VC dimension* 1*.*

Example 2 *Any* $\mathbb{C}$ *which is intersection-closed has* $\mathfrak{s}_{\min} = \mathrm{VC}(\mathbb{C})$*. This is clear from Theorem 19, from which we have* $\mathrm{VC}(\mathbb{C}) = \mathfrak{s}_{\mathbf{1}} \geq \mathfrak{s}_{\min} \geq \mathrm{VC}(\mathbb{C})$*.*

The next example is a new family of concept classes, for which compression schemes of size $\mathrm{VC}(\mathbb{C})$ were not previously known (nor were any *bounded*-size *stable* compression schemes).

Example 3 *For every class* $\mathbb{C}$ *with* $\mathrm{VC}(\mathbb{C}) = 2$ *such that* $\mathcal{X}$ *is a star set for* $\mathbb{C}$*, we have* $\mathfrak{s}_{\min} = 2$*. To see this, take* $h_0 \in \mathbb{C}$ *such that* $\mathcal{X}$ *is a star set centered at* h_0*, and let* $h_* = 1 - h_0$*, and we will argue* $\mathfrak{s}_{h_*} = 2$*, as follows. For every* 3 *distinct points* $S = \{x_1, x_2, x_3\}$*, since* S *is a star set centered at* h_0*, it cannot be the case that* S *is also a star set centered at* h_**, since otherwise the* 4 *classifications witnessing* S *being a star set centered at* h_0 *and the (disjoint)* 4 *classifications witnessing* S *being a star set centered at* h_**, would together comprise* 8 *distinct classifications of* S*, contradicting* $\mathrm{VC}(\mathbb{C}) = 2$*. Indeed, this shows* $\mathfrak{s}_{\min} = 2$ *also for any class* $\mathbb{C}$ *with* $\mathrm{VC}(\mathbb{C}) = 2$ *for which* $\exists h \in \mathbb{C}$ *such that every set of* 3 *distinct points* $\{x_1, x_2, x_3\}$ *is a star set centered at* h*.*

A natural question arises: How does $\mathfrak{s}_{\min}$ compare with the size $O(\mathrm{VC}(\mathbb{C}))$ from the sample compression conjecture of Warmuth (2003)? Is it always true that $\mathfrak{s}_{\min} = O(\mathrm{VC}(\mathbb{C}))$? This turns out not to be the case. Indeed, the following corollary is immediate from Corollary 15.

Corollary 25 *For any concept class* $\mathbb{C}$*,* $\mathfrak{s}_{\min} \geq \mathrm{VC}^*(\mathbb{C})/2 - 1$*.*

Since the dual VC dimension $\mathrm{VC}^*(\mathbb{C})$ can sometimes be as large as $2^{\mathrm{VC}(\mathbb{C})+1} - 1$ (e.g., dictator functions on $\{0,1\}^p$), we see that there can at least be exponential gaps between $\mathfrak{s}_{\min}$ and $\mathrm{VC}(\mathbb{C})$. Moreover, there exist classes, even with VC dimension 3, such that $\mathfrak{s}_{\min} = \infty$. We will provide such an example based on the following simple result.

Proposition 26 *For any finite* $\mathcal{Y}$*, for any concept class* $\mathbb{C}$*, let* $\mathfrak{s}_{\textit{all-const}}$ *denote the largest* n *such that there exists* $\{x_1, \ldots, x_n\} \in \mathcal{X}^n$ *which is a star set centered at* every *constant function* $x \mapsto h_y(x) := y$*,* $y \in \mathcal{Y}$*, or* $\mathfrak{s}_{\textit{all-const}} = \infty$ *if there is no largest such* n*. Then* $\mathfrak{s}_{\min} \geq \mathfrak{s}_{\textit{all-const}}/|\mathcal{Y}|$*.*

Proof Let $\{x_1, \ldots, x_n\}$ be a star set centered at every constant function h_y. For any concept h, there exists y such that $|\{x_i : h(x_i) = y\}| \geq n/|\mathcal{Y}|$. Since any subset of a star set centered at h_y

is also a star set centered at h_y, and since h agrees with h_y on the set $\{x_i : h(x_i) = y\}$, we may conclude that $\mathfrak{s}_h \geq |\{x_i : h(x_i) = y\}| \geq n/|\mathcal{Y}|$. ■

As a simple example of a class with $\mathrm{VC}(\mathbb{C}) = 3$ for which the above implies $\mathfrak{s}_{\min} = \infty$, consider the following.

Example 4 *Let $\mathcal{Y} = \{0, 1\}$, let $\mathcal{X}$ be an infinite set, and let $\mathbb{C} = \{\mathbb{1}_{\{t\}} : t \in \mathcal{X}\} \cup \{\mathbb{1}_{\mathcal{X}\setminus\{t\}} : t \in \mathcal{X}\}$, the class of* singletons *(which are 1 on exactly one point) and their complements (which are 0 on exactly one point). Then $\mathcal{X}$ is a star set centered at both constant functions $x \mapsto 0$ and $x \mapsto 1$, so that Proposition 26 implies $\mathfrak{s}_{\min} = \infty$. It is an easy exercise to verify that $\mathrm{VC}(\mathbb{C}) = 3$.*

Moreover, Proposition 26 reveals $\mathfrak{s}_{\min} = \infty$ even for many natural concept classes, such as *linear classifiers* in $\mathbb{R}^p$.

Example 5 *Let $p \geq 2$, $\mathcal{X} = \mathbb{R}^p$ and let $\mathbb{C}$ be the class of* linear classifiers *on $\mathbb{R}^p$: that is, $\mathbb{C} = \{x \mapsto \mathbb{1}[w^\top x + b \geq 0] : w \in \mathbb{R}^p, b \in \mathbb{R}\}$. Then $\mathrm{VC}(\mathbb{C}) = p + 1$ (Cover, 1965; Vapnik and Chervonenkis, 1974) but $\mathfrak{s}_{\min} = \infty$. To see that $\mathfrak{s}_{\min} = \infty$, consider any number n of points $x_1, \ldots, x_n$ positioned as the vertices of a convex polytope. These points are a star set centered at both constant functions $x \mapsto 0$ and $x \mapsto 1$, so that Proposition 26 implies $\mathfrak{s}_{\min} \geq n/2$. Since such points can be constructed for any $n \in \mathbb{N}$ (e.g., positioning them on a circle), we have that $\mathfrak{s}_{\min} = \infty$.*

Interestingly, for classes $\mathbb{C}$ of $\mathrm{VC}(\mathbb{C}) = 2$, the situation is less clear, and indeed we leave open the question of whether $\mathfrak{s}_{\min} = O(1)$ when $\mathrm{VC}(\mathbb{C}) = 2$. However, we can show at least a mild gap between $\mathfrak{s}_{\min}$ and $\mathrm{VC}(\mathbb{C})$ for such classes.

Example 6 *Pálvölgyi and Tardos (2020) construct an example of a concept class[9] $\mathbb{C}$ with $\mathrm{VC}(\mathbb{C}) = 2$ but for which there does not exist a sample-consistent unlabeled compression scheme of size 2. Specifically, $|\mathcal{X}| = 5$ and $\mathbb{C}$ consists of the "rotations" of the patterns 00101 and 00111 (we refer the reader to the original work for the details). Since the Generalized Closure Algorithm provides a sample-consistent unlabeled compression scheme of size $\mathfrak{s}_{\min}$, we may conclude that $\mathfrak{s}_{\min} \geq 3$.*

Another family of concept classes known to have favorable properties for sample compression is *extremal* classes (see Moran and Warmuth, 2016, for definitions). Rubinstein and Rubinstein (2022) recently showed that any intersection-closed concept class $\mathbb{C}$ can be embedded in an extremal class of VC dimension at most $11\mathrm{VC}(\mathbb{C})$. As discussed in Remark 24, $\mathfrak{s}_{\min}$ may be viewed as the dimension of embedding any $\mathbb{C}$ into a *generalized* intersection-closed class (where $h_*(x)$ functions as "1" would in a traditional intersection-closed class). Moreover, extremal classes remain extremal under such transformations (i.e., changing every $h(x)$ to $1 - h(x)$ for some x's). Therefore, together with the result of Rubinstein and Rubinstein (2022), we also have the following corollary.

Corollary 27 *For any $\mathbb{C}$ with $\mathfrak{s}_{\min} < \infty$, there is an extremal class $\tilde{\mathbb{C}} \supseteq \mathbb{C}$ with $\mathrm{VC}(\tilde{\mathbb{C}}) \leq 11\mathfrak{s}_{\min}$.*

9. This is sometimes known as "Warmuth's example," as Warmuth had previously studied this as an example of a maximal class which is not maximum.

Appendix D. Further Remarks About the Eluder Dimension

In this section, we provide a number of additional remarks concerning the eluder dimension, sharpness of Theorem 11, relations to to existing results, and a result concerning the eluder dimension of version spaces as their disagreement regions. We begin with a remark concerning the relation of Theorem 11 to results of Li, Kamath, Foster, and Srebro (2022).

Remark 28 *As mentioned previously, it follows immediately from its definition that* $\mathfrak{e} \geq \max\{\mathfrak{s}, \mathrm{L}\}$*, since every star set is an eluder sequence and every branch in a shattered Littlestone tree is an eluder sequence. Indeed, this also implies* $\mathfrak{e}_h \geq \bar{\mathfrak{s}}_h$ *for any* h*. Moreover, in the case of* $\mathcal{Y} = \{0,1\}$*, every* h *realizes* some *branch in any shattered Littlestone tree, so* $\mathfrak{e}_h \geq \mathrm{L}$ *in this case.*

In the case of $\mathcal{Y} = \{0,1\}$*, Li, Kamath, Foster, and Srebro (2022) have shown a complementary relation between the eluder dimension* $\mathfrak{e}_h$*, star number* $\mathfrak{s}_h$*, and threshold dimension* $\mathbb{T}_h := \mathbb{T}(\{\mathrm{DIS}(\{h', h\}) : h' \in \mathbb{C}\})$ *of the concept class (and therefore also the Littlestone dimension* L*). Specifically, they have shown that for* $h \in \mathbb{C}$*,* $\max\{\mathfrak{s}_h, \mathbb{T}_h\} \leq \mathfrak{e}_h \leq 4^{\max\{\mathfrak{s}_h, \mathbb{T}_h\}}$*. Since it is also known that* $\log_2(\mathrm{L}) \leq \mathbb{T}_h < 2^{\mathrm{L}+1}$ *(Shelah, 1978; Alon, Bun, Livni, Malliaris, and Moran, 2022), it follows that* $\max\{\mathfrak{s}_h, \mathrm{L}\} \leq \mathfrak{e}_h \leq 4^{\max\{\mathfrak{s}_h, 2^{\mathrm{L}+1}\}}$*. This alone still does not provide a relation between these quantities and the cardinality* $|\mathbb{C}|$*. However, by* combining *their result with* either *our Theorem 6* or *our Theorem 11, implications relating to the other can be reached. Specifically, combining the relation* $\mathfrak{e} \geq \max\{\bar{\mathfrak{s}}, \mathrm{L}\}$ *with our Theorem 6 immediately implies (for general label spaces* $\mathcal{Y}$*) a lower bound* $\mathfrak{e} \geq \sqrt{\log_{|\mathcal{Y}|}(|\mathbb{C}|)}$*, and hence that the eluder dimension is infinite for infinite concept classes (for* $|\mathcal{Y}| < \infty$*). However, Theorem 11 provides a sharper relation between* $\mathfrak{e}$ *and* $|\mathbb{C}|$ *compared to this simple application of Theorem 6, and additionally establishes the existence of an infinite eluder* sequence *for all infinite concept classes.*

Conversely, in the case $\mathcal{Y} = \{0,1\}$*, combining the relation* $\mathfrak{e} \geq \log_2(|\mathbb{C}|)$ *from our Theorem 11 with the result of Li, Kamath, Foster, and Srebro (2022), we can also derive a relation between* $\mathfrak{s}$*,* L*, and* $|\mathbb{C}|$*, analogous to Theorem 6. This would result in a relation of the form* $\max\{\mathfrak{s}, 2^{\mathrm{L}+1}\} \geq \log_4(\log_2(|\mathbb{C}|))$*, which would still provide the conclusion that there is no infinite class* $\mathbb{C}$ *with both finite Littlestone dimension* L *and finite star number* $\mathfrak{s}$*. However, the direct analysis in Theorem 6 still provides the sharper relation* $\mathfrak{s}\mathrm{L} \geq \log_2(|\mathbb{C}|)$ *and moreover extends the result to any finite label space* $\mathcal{Y}$*, with the relation* $\mathfrak{s}\mathrm{L} \geq \log_{|\mathcal{Y}|}(|\mathbb{C}|)$*.*[10]

Remark 29 *It is easy to observe the claim in Theorem 11 for infinite classes is not always true when* $|\mathcal{Y}| = \infty$*, if there exist points* x *with an infinite number of possible labels: for instance, consider the class of* constant *functions* $x \mapsto h_y(x) = y$ $(y \in \mathcal{Y} = \mathbb{N})$*, which is infinite yet has eluder dimension* 1*.*

We can also argue that either side of the inequalities in Theorem 11 can be sharp for some classes $\mathbb{C}$. Formally, we have the following result.

10. It is straightforward to observe that the result $\max\{\mathfrak{s}_h, \mathbb{T}_h\} \leq \mathfrak{e}_h \leq 4^{\max\{\mathfrak{s}_h, \mathbb{T}_h\}}$ $(h \in \mathbb{C})$ of Li, Kamath, Foster, and Srebro (2022) remains valid for general $\mathcal{Y}$ spaces, since all of $\mathfrak{s}_h$, $\mathbb{T}_h$, and $\mathfrak{e}_h$ are only concerned with the classes of *binary* loss-composed functions $\{x \mapsto \mathbb{1}[h'(x) \neq h(x)] : h' \in \mathbb{C}\}$, so that applying their result to this class (with center function the all-0 function) extends their result to any $\mathcal{Y}$ space. Thus, combining Theorem 11 with this generalization of the result of Li, Kamath, Foster, and Srebro (2022) generally implies $\max\{\mathfrak{s}, \mathbb{T}\} \geq \log_4(\log_{|\mathcal{Y}|}(|\mathbb{C}|))$. Moreover, for finite $\mathcal{Y}$, there remains a quantitative relation between L and $\mathbb{T}$ (a relation $\mathrm{L} \geq \lfloor \log_2(\log_{|\mathcal{Y}|}(\mathbb{T}/|\mathcal{Y}|)) \rfloor$ can be inferred from results of Hanneke, Moran, and Shafer, 2023b), so that in this case we may recover a result in a similar spirit to Theorem 6. However, Theorem 6 remains quantitatively much sharper, and applies even when we merely have that every x has $|\{h(x) : h \in \mathbb{C}\}| < \infty$.

Theorem 30 *For any $n, k \in \mathbb{N} \setminus \{1\}$, for $\mathcal{Y} = \{0, \ldots, k-1\}$, there exists $\mathcal{X}$, and concept classes $\mathbb{C}$ and $\mathbb{C}'$ with $|\mathbb{C}| = |\mathbb{C}'| = n$ such that (supposing n is a power of k) $\mathfrak{e}(\mathbb{C}) = \log_k(|\mathbb{C}|)$ and $\mathfrak{e}(\mathbb{C}') = |\mathbb{C}'| - 1$.*

Specifically, for $\mathcal{X} = \{1, \ldots, n-1\}$, the class $\mathbb{C}'$ witnessing the above (as already found by Li, Kamath, Foster, and Srebro, 2022) is the set of *singletons*, $x \mapsto \mathbb{1}_{\{i\}}(x)$, $i \in \mathcal{X}$, along with the all-zero function $x \mapsto 0$. On the other hand, the class $\mathbb{C}$ witnessing the $\log_k(|\mathbb{C}|)$ eluder dimension is simply the set $\mathcal{Y}^{\{1,\ldots,\log_k(n)\}} \times \{0\}^{\{\log_k(n)+1,\ldots,n-1\}}$.

D.1. Eluder Dimension of Version Spaces and Disagreement Regions

Analogously to Theorem 16 and Proposition 17, we may also study the eluder dimension of the sets $\mathcal{V}_h(\mathbb{C})$ and $\mathcal{D}_h(\mathbb{C})$ themselves: that is, the eluder dimension of version spaces and their disagreement regions. Let us again overload the notation, this time for the eluder dimension: for any $D \subseteq \mathcal{X}$ and any non-empty set $\mathcal{D} \subseteq 2^{\mathcal{X}}$, define $\mathfrak{e}_D(\mathcal{D})$, the eluder dimension of $\mathcal{D}$ centered at D, as $\mathfrak{e}_{\mathbb{1}_D}(\{\mathbb{1}_{D'} : D' \in \mathcal{D}\})$: that is, the eluder dimension of the corresponding set of indicator functions $\mathbb{1}_{D'} : \mathcal{X} \to \{0,1\}$. Similarly, $\mathfrak{e}(\mathcal{D}) = \sup_{D \in \mathcal{D}} \mathfrak{e}_D(\mathcal{D})$. The following theorem is analogous to Theorem 16. Its proof is presented in Section L

Theorem 31 *For any concept class $\mathbb{C}$ and concept h, $\mathfrak{e}_{\mathbb{C}}(\mathcal{V}_h(\mathbb{C})) = \mathfrak{e}_h(\mathbb{C})$. Also $\forall h \in \mathbb{C}$, $\mathfrak{e}_{\emptyset}(\mathcal{D}_h(\mathbb{C})) = \mathfrak{e}_h(\mathbb{C})$.*

Similarly to Proposition 17, one might wonder whether there can be large gaps between $\mathfrak{e}(\mathcal{D}_h(\mathbb{C}))$ and $\mathfrak{e}(\mathbb{C})$. However, unlike the star number in Proposition 17, this is *not* the case for the eluder dimension. In light of Theorem 11, we may immediately note that finite $\mathfrak{e}(\mathbb{C})$ implies finite $\mathfrak{e}(\mathcal{D}_h(\mathbb{C}))$ and $\mathfrak{e}(\mathcal{V}_h(\mathbb{C}))$, since the latter two may be upper bounded by the number of distinct possible version spaces, which is clearly less than $2^{|\mathbb{C}|}$. While it seems likely that significant quantitative gaps are possible, we leave the exploration of this issue for future work.

Appendix E. Implications and Further Discussion of the Results

In this section, we present a number of implications of the above results. Specifically, we provide a new proof of a bound (originally due to Hanneke, 2016) on the probability in the region of disagreement of a version space, via classic VC-based generalization bounds for empirical risk minimization. We complement this with a lower bound, showing such bounds are sharp. We also establish a relation between the star number and a complexity measure proposed by El-Yaniv and Wiener (2010, 2012) in their analysis of the perfect selective classification problem.

E.1. A New Proof Bounding the Probability in the Region of Disagreement of a Version Space

As a direct implication of the results of Sections 2.1 and A (relating the star number and VC dimension of disagreement regions), we may state a new proof of a bound (originally proven by Hanneke, 2016) on the probability measure of the region of disagreement of a version space induced by an i.i.d. sample. Specifically, fix any concept class $\mathbb{C}$, any marginal distribution P_X over $\mathcal{X}$, and any *target concept* $h^\star \in \mathbb{C}$, and let P denote a distribution over $\mathcal{X} \times \mathcal{Y}$ such that $(X, Y) \sim P$ has

$X \sim P_X$ and $Y := h^\star(X)$.[11] Fix any $\delta \in (0,1)$ and let $S_n \sim P^n$ be an i.i.d. data set of size $n \in \mathbb{N}$ with $n \geq 2\mathfrak{s}_{h^\star}$.

Using arguments based on *sample compression schemes* (Littlestone and Warmuth, 1986; Floyd and Warmuth, 1995), an argument of Hanneke and Yang (2015) establishes that, with probability at least $1-\delta$,

$$P_X(\mathrm{DIS}(\mathbb{C}_{S_n})) = O\left(\frac{1}{n}\left(\mathfrak{s}_{h^\star}\log\left(\frac{n}{\mathfrak{s}_{h^\star}}\right) + \log\left(\frac{1}{\delta}\right)\right)\right). \tag{1}$$

Such a bound plays a crucial role in the analysis of so-called *disagreement-based* active learning methods (Cohn, Atlas, and Ladner, 1994; Hanneke and Yang, 2015; Hanneke, 2016), and has further implications for refined generalization bounds for supervised learning with empirical risk minimization (Hanneke, 2016). By using monotonicity properties of the set $\mathrm{DIS}(\mathbb{C}_{S_n})$ (i.e., $\mathrm{DIS}(\mathbb{C}_{S_n})$ is non-increasing with adding more data to S_n), an argument of Hanneke (2016) refines this bound, again using sample compression arguments: with probability at least $1-\delta$,

$$P_X(\mathrm{DIS}(\mathbb{C}_{S_n})) = O\left(\frac{1}{n}\left(\mathfrak{s}_{h^\star} + \log\left(\frac{1}{\delta}\right)\right)\right). \tag{2}$$

The leading constant factor in the bound of Hanneke (2016) is 21. This constant has since been refined to $2\ln(4)$ by Bousquet, Hanneke, Moran, and Zhivotovskiy (2020a,b) using a stronger generalization bound they prove for *stable* compression schemes (and by noting that the compression scheme of Hanneke and Yang, 2015; Hanneke, 2016, for $\mathrm{DIS}(\mathbb{C}_{S_n})$, is indeed stable).

One immediate implication of Theorem 3 is a new proof of (1), which, rather than being based on sample compression arguments, instead relies on the classic analysis of generalization bounds for empirical risk minimization in PAC learning (Vapnik and Chervonenkis, 1974; Blumer, Ehrenfeucht, Haussler, and Warmuth, 1989). Specifically, Blumer, Ehrenfeucht, Haussler, and Warmuth (1989) prove that, for any non-empty set $\mathcal{Z}$ and any set $\mathcal{D} \subseteq 2^{\mathcal{Z}}$, for any probability measure P over $\mathcal{Z}$, any $n \in \mathbb{N}$ with $n \geq \mathrm{VC}(\mathcal{D})$, and any $\delta \in (0,1)$, for $S_n \sim P^n$, with probability at least $1-\delta$,[12] every $D \in \mathcal{D}$ with $D \cap S_n = \emptyset$ satisfies

$$P(D) \leq \frac{2}{n}\left(\mathrm{VC}(\mathcal{D})\log_2\left(\frac{en}{\mathrm{VC}(\mathcal{D})}\right) + \log_2\left(\frac{2}{\delta}\right)\right). \tag{3}$$

In particular, for the distribution P defined based on marginal P_X and target concept $h^\star$, if we let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and $\mathcal{D} = \{D \times \mathcal{Y} : D \in \mathcal{D}_{h^\star}(\mathbb{C})\}$, then Theorem 3 implies $\mathrm{VC}(\mathcal{D}) = \mathfrak{s}_{h^\star}$ (noting that $\mathrm{VC}(\mathcal{D}) = \mathrm{VC}(\mathcal{D}_{h^\star}(\mathbb{C}))$ since the $\mathcal{Y}$ component of $D \times \mathcal{Y}$ is the same for all $D \times \mathcal{Y} \in \mathcal{D}$). Thus, for any $n \in \mathbb{N}$ with $n \geq \mathfrak{s}_{h^\star}$, and any $\delta \in (0,1)$, for $S_n \sim P^n$, since every $h \in \mathbb{C}_{S_n}$ have $\forall (x,y) \in S_n$, $h(x) = y$ (by definition of $\mathbb{C}_{S_n}$), we have $(\mathrm{DIS}(\mathbb{C}_{S_n}) \cap \mathcal{Y}) \cap S_n = \emptyset$, and therefore, with probability at least $1-\delta$, (3) implies

$$P_X(\mathrm{DIS}(\mathbb{C}_{S_n})) = P(\mathrm{DIS}(\mathbb{C}_{S_n}) \times \mathcal{Y}) \leq \frac{2}{n}\left(\mathfrak{s}_{h^\star}\log_2\left(\frac{en}{\mathfrak{s}_{h^\star}}\right) + \log_2\left(\frac{2}{\delta}\right)\right).$$

11. It is straightforward to generalize this analysis to any P with $\inf_{h\in\mathbb{C}} P((x,y) : h(x) \neq y) = 0$, in which case a slight modification of the argument below holds for $h^\star$ defined as any element of $\mathbb{C}$ with, say, $P((x,y) : h^\star(x) \neq y) < \delta/(2n)$. By a slightly more involved analysis, the result can in fact be stated with $h^\star$ defined as any measurable function (not necessarily in $\mathbb{C}$) with $P((x,y) : h^\star(x) \neq y) = 0$ and $\inf_{h\in\mathbb{C}} P_X(x : h(x) \neq h^\star(x)) = 0$ (which necessarily exists for any class with $\mathfrak{s} < \infty$; see Hanneke, 2012). For simplicity, we omit these details.

12. Here, and throughout this section, we do not discuss nuances arising from measurability considerations. Such issues have been thoroughly discussed in the literature, such as by Blumer, Ehrenfeucht, Haussler, and Warmuth (1989); van der Vaart and Wellner (1996); van Handel (2013); in particular, if $\mathcal{Z}$ is countable, or $\mathcal{D}$ satisfies certain topological conditions, then the stated events will indeed be measurable.

Thus, we have recovered the bound (1) of Hanneke and Yang (2015); Hanneke (2016) by a new proof: that is, via the classic PAC bound (3), rather than via generalization bounds for sample compression schemes. This is not to say that there is anything wrong with the latter; on the contrary, the proof of the compression scheme generalization bound is considerably simpler than the proof of VC-based generalization bounds. Nonetheless, it is often valuable to have multiple proofs of results, and understanding these different perspectives may lead to further insights in the future, such as in contexts where compression-based analysis has so-far not yielded sharp guarantees (such as, perhaps, in active learning with classification noise; see the discussion in Wiener, Hanneke, and El-Yaniv, 2015).

We can also use Theorem 3 to obtain the sharper bound (2) of Hanneke (2016), again via VC-based generalization bounds rather than compression schemes. We present two proofs of this, based on different results from Hanneke (2016). The first argument is based on the fact that $\mathrm{DIS}(\mathbb{C}_S)$ is *non-increasing* in S: that is, for any $S, T \in \bigcup_{n\geq 0}(\mathcal{X}\times\mathcal{Y})^n$, $\mathrm{DIS}(\mathbb{C}_{S\cup T}) \subseteq \mathrm{DIS}(\mathbb{C}_S)$. Hanneke (2016, Theorem 1) states a generalization bound for such monotone functions $S \mapsto \hat{D}(S) \in \mathcal{D}$, for any set $\mathcal{D}$, in terms of $\mathrm{VC}(\mathcal{D})$. Specifically, (for any P, δ, n, as described above, and $S_n \sim P^n$), with probability at least $1-\delta$,

$$P(\hat{D}(S_n)) \leq \frac{4}{n}\left(17\mathrm{VC}(\mathcal{D}) + 4\ln\left(\frac{4}{\delta}\right)\right).$$

In particular, taking P, P_X, and $h^\star$ as above, and $\mathcal{D} = \{D\times\mathcal{Y} : D \in \mathcal{D}_{h^\star}(\mathbb{C})\}$ as above, and recalling that Theorem 3 implies $\mathrm{VC}(\mathcal{D}) = \mathfrak{s}_{h^\star}$, Theorem 1 of Hanneke (2016) implies that with probability at least $1-\delta$,

$$P_X(\mathrm{DIS}(\mathbb{C}_{S_n})) = P(\mathrm{DIS}(\mathbb{C}_{S_n})\times\mathcal{Y}) \leq \frac{4}{n}\left(17\mathfrak{s}_{h^\star} + 4\ln\left(\frac{4}{\delta}\right)\right). \tag{4}$$

This recovers the form of the bound (2) proven by Hanneke (2016), again proven via VC generalization bounds rather than sample compression-based bounds.

We can also give yet another proof of this bound, with slightly sharpened numerical constant factors, using Theorem 16, in conjunction with a generalization bound of Hanneke (2016, Corollary 12) for general empirical risk minimizers, which refines the classic PAC bound (3). Specifically (continuing with any abstract space $\mathcal{D}$, and any P, δ, n, and for $S_n \sim P^n$), Theorem 11 of Hanneke (2016) and Lemma 44 of Hanneke and Yang (2015) together imply[13] that, with probability at least $1-\delta$, every $D \in \mathcal{D}$ with $D \cap S_n = \emptyset$ satisfies

$$P(D) \leq \frac{8}{n}\left(\mathrm{VC}(\mathcal{D})\ln\left(\frac{49e\mathfrak{s}_\emptyset(\mathcal{D})}{\mathrm{VC}(\mathcal{D})} + 37\right) + 8\ln\left(\frac{6}{\delta}\right)\right). \tag{5}$$

Returning to the definitions of P, P_X, and $h^\star$ from (2), for $n \in \mathbb{N}$ and $S_n \sim P^n$, with probability one every (x,y) in S_n has $h^\star(x) = y$, which further implies $\mathrm{DIS}(\mathbb{C}_{S_n}) \in \mathcal{D}_{h^\star}(\mathbb{C})$. Thus, applying

13. In the context of Theorem 11 of Hanneke (2016), we are interpreting the "target concept" as $\emptyset$ (equivalently, the *all-zero* function), so that the error rate of $\mathbb{1}_D$ is $P(D)$. Theorem 11 of Hanneke (2016) then gives a bound of this form, but with $\mathfrak{s}_\emptyset(\mathcal{D})$ replaced by a quantity $\max_{t\leq n}\hat{n}_t$, where $\hat{n}_t$ is the minimum size of a *version space compression set* for the first t data points (Hanneke, 2007a; El-Yaniv and Wiener, 2010; Wiener, Hanneke, and El-Yaniv, 2015): that is, the smallest size of a subset $\hat{S}$ of the first t data points S_t such that every $D \in \mathcal{D}$ with $D \cap \hat{S} = \emptyset$ also has $D \cap S_t = \emptyset$. Lemma 14 of Hanneke and Yang (2015) then implies that such a minimal-sized $\hat{S}$ is a star set centered at $\emptyset$, and hence has size at most $\mathfrak{s}_\emptyset(\mathcal{D})$.

(5) to the set $\mathcal{D} = \{D \times \mathcal{Y} : D \in \mathcal{D}_{h^\star}(\mathbb{C})\}$, and recalling that Theorem 3 implies $\mathrm{VC}(\mathcal{D}) = \mathfrak{s}_{h^\star}$, while Theorem 16 implies $\mathfrak{s}_\emptyset(\mathcal{D}) = \mathfrak{s}_{h^\star}$, for any $\delta \in (0,1)$, we have that with probability at least $1-\delta$,

$$P_X(\mathrm{DIS}(\mathbb{C}_{S_n})) = P(\mathrm{DIS}(\mathbb{C}_{S_n}) \times \mathcal{Y}) \leq \frac{8}{n}\left(\mathfrak{s}_{h^\star}\ln(49e+37) + 8\ln\left(\frac{6}{\delta}\right)\right).$$

This gives a numerical constant on the lead term (i.e., the $\mathfrak{s}_{h^\star}$ factor) at most 42, slightly sharper than the bound (4) above (which is 68), though still not as sharp as the factor 21 from the compression-based bound from Hanneke (2016), or the factor $2\ln(4)$ from the stable compression bound of Bousquet, Hanneke, Moran, and Zhivotovskiy (2020a,b). Therefore, again the contribution of this result (as with (4)) is merely in that it presents a new proof of the form of the bound.

We conclude this subsection by noting that the above bound is *sharp* up to constant factors, as implied by the following theorem; this result is new, though is related to known lower bound arguments from Hanneke (2016).

Theorem 32 *For any concept class $\mathbb{C}$ and any concept $h^\star$ with $\mathfrak{s}_{h^\star} \geq 2$, for any $\delta \in (0,1/2)$ and any $n \in \mathbb{N}$, there exists a marginal distribution P_X over $\mathcal{X}$ such that, letting P denote the joint distribution over $\mathcal{X} \times \mathcal{Y}$ such that $(X,Y) \sim P$ has $X \sim P_X$ and $Y := h^\star(X)$, P is realizable with respect to $\mathbb{C}$ and for $S_n \sim P^n$, with probability greater than δ,*

$$P_X(\mathrm{DIS}(\mathbb{C}_{S_n})) \geq \max\left\{\frac{\mathfrak{s}_{h^\star}-1}{4n}, \frac{1}{2n}\log_2\left(\frac{1}{\delta}\right)\right\} \wedge \frac{1}{2} = \Omega\left(\min\left\{\frac{1}{n}\left(\mathfrak{s}_{h^\star} + \log\left(\frac{1}{\delta}\right)\right), \frac{1}{2}\right\}\right). \tag{6}$$

Proof We establish the lower bound in four parts via a (simplified variant of a) standard argument from PAC learning (Vapnik and Chervonenkis, 1974; Ehrenfeucht, Haussler, Kearns, and Valiant, 1989). The first two parts establish a lower bound $\min\left\{\frac{1}{2n}\log_2\left(\frac{1}{\delta}\right), \frac{1}{2}\right\}$. For these, we will let P_X be supported on any star set $\{x_1, x_2\}$ centered at $h^\star$.

First consider the case $n \leq \log_2\left(\frac{1}{\delta}\right)$. In this case, set $P_X(\{x_1\}) = P_X(\{x_2\}) = \frac{1}{2}$, and let P be constructed as in the theorem (with marginal P_X on $\mathcal{X}$, and target concept $h^\star$), noting that Definition 2 implies P is realizable. Let $S_n = \{(X_1,Y_1),\ldots,(X_n,Y_n)\} \sim P^n$, and note that the probability that either every $i \leq n$ has $(X_i,Y_i) = (x_1, h^\star(x_1))$ or every $i \leq n$ has $(X_i,Y_i) = (x_2, h^\star(x_2))$ is precisely $2\left(\frac{1}{2}\right)^n \geq 2\delta$. On this event, we have $|\mathrm{DIS}(\mathbb{C}_{S_n}) \cap \{x_1,x_2\}| = 1$, so that $P_X(\mathrm{DIS}(\mathbb{C}_{S_n})) = \frac{1}{2}$.

Second, consider the case $n > \log_2\left(\frac{1}{\delta}\right)$. In this case, set $P_X(\{x_2\}) = \frac{1}{2n}\log_2\left(\frac{1}{\delta}\right)$ and $P_X(\{x_1\}) = 1 - P_X(\{x_2\})$. Again let P be as described in the theorem (with marginal P_X and target concept $h^\star$), noting that Definition 2 implies P is realizable. Let $S_n = \{(X_1,Y_1),\ldots,(X_n,Y_n)\} \sim P^n$. We are again interested in the event that every $i \leq n$ has $(X_i,Y_i) = (x_1, h^\star(x_1))$: namely,

$$P_X(\{x_1\})^n = \left(1 - \frac{1}{2n}\log_2\left(\frac{1}{\delta}\right)\right)^n.$$

The right hand side is an increasing function of n in the range $n > \log_2\left(\frac{1}{\delta}\right)$, and is therefore greater than $\left(1-\frac{1}{2}\right)^{\log_2(1/\delta)} = \delta$. On this event, we have $\mathrm{DIS}(\mathbb{C}_{S_n}) \cap \{x_1,x_2\} = \{x_2\}$, so that $P_X(\mathrm{DIS}(\mathbb{C}_{S_n})) = \frac{1}{2n}\log_2\left(\frac{1}{\delta}\right)$.

Next we prove a lower bound $\min\left\{\frac{\mathfrak{s}_{h^\star}-1}{4n}, \frac{1}{2}\right\}$. Again, there will be two cases. First, consider the case that $n \leq \frac{\mathfrak{s}_{h^\star}-1}{2}$. Let $\{x_1, \ldots, x_{2n}\}$ be any star set centered at $h^\star$ (which must exist since $2n \leq \mathfrak{s}_{h^\star}$). Define P_X as uniform on $\{x_1, \ldots, x_{2n}\}$, and note that for any $X_1, \ldots, X_n \in \{x_1, \ldots, x_{2n}\}$, for $S_n = \{(X_1, h^\star(X_1)), \ldots, (X_n, h^\star(X_n))\}$, $\mathrm{DIS}(\mathbb{C}_{S_n}) \cap \{x_1, \ldots, x_{2n}\} = \{x_1, \ldots, x_{2n}\} \setminus \{X_1, \ldots, X_n\}$, which has size at least n and has P_X probability mass at least $\frac{1}{2}$. Let P be as in the theorem statement (i.e., marginal P_X on $\mathcal{X}$, and target concept $h^\star$), and note that Definition 2 implies P is realizable. For $S_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\} \sim P^n$, since every $Y_i = h^\star(X_i)$, we have (always)

$$P_X(\mathrm{DIS}(\mathbb{C}_{S_n})) = P_X(\{x_1, \ldots, x_{2n}\} \setminus \{X_1, \ldots, X_n\}) \geq \frac{1}{2}.$$

To complete the proof, we consider the case that $n > \frac{\mathfrak{s}_{h^\star}-1}{2}$ and establish a lower bound $\frac{\mathfrak{s}_{h^\star}-1}{4n}$ holding with probability at least $\frac{1}{2} > \delta$. Let $x_1, \ldots, x_{\mathfrak{s}_{h^\star}}$ be a star set centered at $h^\star$, and define the marginal distribution P_X by $P_X(\{x_i\}) = \frac{1}{2n}$ for $i \in \{2, \ldots, \mathfrak{s}_{h^\star}\}$ and $P_X(\{x_1\}) = 1 - \frac{\mathfrak{s}_{h^\star}-1}{2n}$, noting that this is positive by the condition $n > \frac{\mathfrak{s}_{h^\star}-1}{2}$. Define P as in the theorem statement (i.e., marginal P_X on $\mathcal{X}$, and target concept $h^\star$), and note that Definition 2 implies P is realizable. Let $S_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\} \sim P^n$. Note that the number of $i \leq n$ with $X_i \in \{x_2, \ldots, x_{\mathfrak{s}_{h^\star}}\}$ is a $\mathrm{Binomial}\left(n, \frac{\mathfrak{s}_{h^\star}-1}{2n}\right)$ random variable, which therefore has $\left\lfloor \frac{\mathfrak{s}_{h^\star}-1}{2} \right\rfloor$ as a median value. Thus, with probability at least $\frac{1}{2}$, there are at most $\frac{\mathfrak{s}_{h^\star}-1}{2}$ values $i \leq n$ with $X_i \in \{x_2, \ldots, x_{\mathfrak{s}_{h^\star}}\}$. In particular, this also implies there are at most $\frac{\mathfrak{s}_{h^\star}-1}{2}$ values $i \in \{2, \ldots, \mathfrak{s}_{h^\star}\}$ with $x_i \notin \{X_1, \ldots, X_n\}$. Since $x_1, \ldots, x_{\mathfrak{s}_{h^\star}}$ is a star set centered at $h^\star$, we have

$$\mathrm{DIS}(\mathbb{C}_{S_n}) = \{x_1, \ldots, x_{\mathfrak{s}_{h^\star}}\} \setminus \{X_1, \ldots, X_n\}.$$

Therefore, on the above event of probability at least $\frac{1}{2} > \delta$, we have $P_X(\mathrm{DIS}(\mathbb{C}_{S_n})) \geq \frac{\mathfrak{s}_{h^\star}-1}{4n}$.

Altogether, we have established that, for all values of n, there exists a distribution P which is realizable with target concept $h^\star$ such that, for $S_n \sim P^n$, with probability greater than δ,

$$\begin{aligned} P_X(\mathrm{DIS}(\mathbb{C}_{S_n})) &\geq \max\left\{\min\left\{\frac{\mathfrak{s}_{h^\star}-1}{4n}, \frac{1}{2}\right\}, \min\left\{\frac{1}{2n}\log_2\left(\frac{1}{\delta}\right), \frac{1}{2}\right\}\right\} \\ &= \max\left\{\frac{\mathfrak{s}_{h^\star}-1}{4n}, \frac{1}{2n}\log_2\left(\frac{1}{\delta}\right)\right\} \wedge \frac{1}{2}. \end{aligned}$$

■

E.2. Additional Relations

The works of El-Yaniv and Wiener (2010, 2012) study the problem of *perfect selective classification* and the related setting of *active learning*. They adopt the approach of *disagreement-based* learning, as in other works on active learning and reliable prediction (e.g., Cohn, Atlas, and Ladner, 1994; Balcan, Beygelzimer, and Langford, 2006; Hanneke, 2007b; Dasgupta, Hsu, and Monteleoni, 2007; Rivest and Sloan, 1988; Balcan, Hanneke, Pukdee, and Sharma, 2024). Their analysis then boils down to bounding the probability $P_X(\mathrm{DIS}(\mathbb{C}_{S_n}))$ of the region of disagreement of the version space induced by a data set $S_n \sim P^n$, for realizable distributions P. However, rather than bounding this

in terms of the disagreement coefficient (as in prior works of Hanneke, 2007b, 2009b,a, 2011; Dasgupta, Hsu, and Monteleoni, 2007), they propose a novel analysis based on a quantity they introduce, expressed as the VC dimension of regions of disagreement of certain version spaces.[14] Specifically, they introduce a quantity $\gamma(k)$ which they call the *order-k characterizing set complexity*: namely,

$$\gamma(k) = \mathrm{VC}\Big(\Big\{\mathrm{DIS}(\mathbb{C}_S) : S \in (\mathcal{X} \times \mathcal{Y})^k\Big\}\Big),$$

the VC dimension of regions of disagreement of version spaces induced by data sets of size k. Their bound on $P_X(\mathrm{DIS}(\mathbb{C}_{S_n}))$ is then expressed in terms of $\gamma(k)$, for a particular choice of k. Specifically, for any $n \in \mathbb{N}$ and data set $S_n = \{(x_1, y_1), \ldots, (x_n, y_n)\} \in \mathcal{X} \times \mathcal{Y}$, they let $\hat{n}(S_n)$ denote the *version space compression set size*: namely, the size of the minimal subsequence $\hat{S} \subseteq S_n$ for which $\mathbb{C}_{\hat{S}} = \mathbb{C}_{S_n}$ (also known as the *empirical teaching dimension*, or minimal specifying set size, in the works of Hanneke (2007a,c, 2014); Hanneke and Yang (2015)). They then establish an upper bound on $P_X(\mathrm{DIS}(\mathbb{C}_{S_n}))$: for any realizable distribution P, for any $n \in \mathbb{N}$ and $S_n \sim P^n$, with probability at least $1 - \delta$,

$$P_X(\mathrm{DIS}(\mathbb{C}_{S_n})) = O\left(\frac{1}{n}\left(\gamma(\hat{n}(S_n))\log\left(\frac{n}{\gamma(\hat{n}(S_n))}\right) + \log\left(\frac{1}{\delta}\right)\right)\right). \tag{7}$$

This analysis was later refined by Wiener, Hanneke, and El-Yaniv (2015), who showed that in fact this same bound holds with $\hat{n}(S_n)$ in place of $\gamma(\hat{n}(S_n))$, and indeed Hanneke (2016); Hanneke and Kontorovich (2020) further refined this result by entirely removing the resulting $\log(n/\hat{n}(S_n))$ factor on the lead term. Nevertheless, it is interesting to consider the original bound (7) of El-Yaniv and Wiener (2010, 2012) as expressed in terms of $\gamma(\hat{n}(S_n))$, and attempt to relate it to other known quantities, and in particular, compare it to the optimal distribution-free bound on $P_X(\mathrm{DIS}(\mathbb{C}_{S_n}))$: that is, $\Theta\big(\frac{1}{n}\big(\mathfrak{s} + \log\big(\frac{1}{\delta}\big)\big) \wedge 1\big)$; indeed, if $h^\star$ is the *target concept* for a realizable distribution P (i.e., $P((x,y) : h^\star(x) \neq y) = 0$), then the results of Hanneke (2016); Hanneke and Yang (2015) and Section E imply an optimal $h^\star$-dependent bound $\Theta\big(\frac{1}{n}\big(\mathfrak{s}_{h^\star} + \log\big(\frac{1}{\delta}\big)\big) \wedge 1\big)$ (see also Section E.1).

As one immediate observation, by Proposition 4, we have $\gamma(k) = O(\mathfrak{s})$, so that the bound (7) implies a bound $O\big(\frac{1}{n}\big(\mathfrak{s}\log\big(\frac{n}{\mathfrak{s}}\big) + \log\big(\frac{1}{\delta}\big)\big) \wedge 1\big)$, which matches up to a log factor the optimal target-independent bound $\Theta\big(\frac{1}{n}\big(\mathfrak{s} + \log\big(\frac{1}{\delta}\big)\big) \wedge 1\big)$ (which follows from the upper bound of Hanneke, 2016 and the lower bound from Theorem 32 of Section E). Moreover, we may further refine the above bound to be target-dependent, noting that (by the same proof as in the original work of El-Yaniv and Wiener, 2010, 2012) it is possible to replace $\gamma(k)$ in (7) with

$$\gamma_{h^\star}(k) := \mathrm{VC}\Big(\Big\{\mathrm{DIS}(\mathbb{C}_S) : S \in \{(x, h^\star(x)) : x \in \mathcal{X}\}^k\Big\}\Big)$$

for realizable distributions P with target concept $h^\star \in \mathbb{C}$. We may then note that Theorem 3 implies $\gamma_{h^\star}(k) \leq \mathfrak{s}_{h^\star}$. Thus, this $h^\star$-dependent refinement of (7) implies an upper bound of $O\Big(\frac{1}{n}\Big(\mathfrak{s}_{h^\star}\log\Big(\frac{n}{\mathfrak{s}_{h^\star}}\Big) + \log\big(\frac{1}{\delta}\big)\Big) \wedge 1\Big)$, which matches up to a log factor the optimal target-dependent bound $\Theta\big(\frac{1}{n}\big(\mathfrak{s}_{h^\star} + \log\big(\frac{1}{\delta}\big)\big) \wedge 1\big)$ (see Section E).

In light of our Theorem 32, significant further refinements of the inequality $\gamma_{h^\star}(\hat{n}(S_n)) \leq \mathfrak{s}_{h^\star}$ are not generally possible. Indeed, examining the proof of Theorem 3, we may note that the proof of the lower bound $\mathrm{VC}(\mathcal{D}_h(\mathbb{C})) \geq \mathfrak{s}_h$ has a further implication that, for any finite $n \leq \mathfrak{s}_h$, the

14. Indeed, understanding the relation of this definition to the star number was a key inspiration for the present work.

inequality $\mathrm{VC}(\mathcal{D}_h(\mathbb{C})) \geq n$ is witnessed by data sets of size at most n, which therefore implies $\gamma_{h^\star}(k) \geq \min\{\mathfrak{s}_{h^\star}, k\}$. Moreover, $\forall n \in \mathbb{N}$, the value $\hat{\mathring{n}}_{h^\star}(n) := \max\{\hat{n}(S_n) : S_n \in \{(x, h^\star(x)) : x \in \mathcal{X}\}^n\}$ is known to satisfy $\hat{\mathring{n}}_{h^\star}(n) = \min\{\mathfrak{s}_{h^\star}, n\}$ (Hanneke and Yang, 2015). It follows that

$$\max\{\gamma_{h^\star}(\hat{n}_{h^\star}(S_n)) : S_n \in \{(x, h^\star(x)) : x \in \mathcal{X}\}^n\} \geq \min\{\mathfrak{s}_{h^\star}, n\}.$$

Appendix F. Proofs of Results from Section 2.1: The Star Number

F.1. Proof of Theorem 3 (Relating Star Number and VC Dimension)

This section presents the proofs of Theorem 3 and Proposition 4.

Proof of Theorem 3 We first argue that $\mathrm{VC}(\mathcal{V}_h(\mathbb{C})) \geq \bar{\mathfrak{s}}_h$ for any concept h. Specifically, consider any extended star set $x_1, \ldots, x_n$ (for $\mathbb{C}$) centered at h. Let $h_1, \ldots, h_n \in \mathbb{C}$ be as in the definition of $\bar{\mathfrak{s}}_h$: that is, $\forall i, j \in \{1, \ldots, n\}$, $h_i(x_j) = h(x_j)$ iff $i \neq j$. We will show that $\{h_1, \ldots, h_n\}$ is shattered by $\mathcal{V}_h(\mathbb{C})$. For any $I \subseteq \{1, \ldots, n\}$, letting $S = \{(x_j, h(x_j))\}_{j \in \{1,\ldots,n\} \setminus I}$, we claim that $\mathbb{C}_S \cap \{h_1, \ldots, h_n\} = \{h_i : i \in I\}$. To see this, note that each $i \in I$ has $h_i(x_j) = h(x_j)$ for every $j \neq i$, and therefore for every $j \in \{1, \ldots, n\} \setminus I$, so that $h_i \in \mathbb{C}_S$; moreover, every $i \in \{1, \ldots, n\} \setminus I$ has $(x_i, h(x_i)) \in S$, whereas $h_i(x_i) \neq h(x_i)$, so that $h_i \notin \mathbb{C}_S$. Thus, $\{h_1, \ldots, h_n\}$ is shattered by $\mathcal{V}_h(\mathbb{C})$, and hence $\mathrm{VC}(\mathcal{V}_h(\mathbb{C})) \geq n$. Since such an extended star set exists for any finite $n \leq \bar{\mathfrak{s}}_h$, we conclude that $\mathrm{VC}(\mathcal{V}_h(\mathbb{C})) \geq \bar{\mathfrak{s}}_h$. Moreover, since $\mathrm{VC}(\mathcal{V}(\mathbb{C})) \geq \mathrm{VC}(\mathcal{V}_h(\mathbb{C}))$ for any concept h, we further conclude that $\mathrm{VC}(\mathcal{V}(\mathbb{C})) \geq \sup_h \mathrm{VC}(\mathcal{V}_h(\mathbb{C})) \geq \sup_h \bar{\mathfrak{s}}_h = \bar{\mathfrak{s}}$.

Next we argue that every concept h has $\mathrm{VC}(\mathcal{V}_h(\mathbb{C})) \leq \bar{\mathfrak{s}}_h$. Let $\{h_1, \ldots, h_n\} \subseteq \mathbb{C}$ be any set shattered by $\mathcal{V}_h(\mathbb{C})$. In particular, from the definition of shattering, this implies that for every $i \in \{1, \ldots, n\}$, there exists a data set S_i consistent with h (i.e., a sequence of pairs $(x, h(x))$) such that $\mathbb{C}_{S_i} \cap \{h_1, \ldots, h_n\} = \{h_j : j \in \{1, \ldots, n\} \setminus \{i\}\}$. Since $h_i \notin \mathbb{C}_{S_i}$, there must exist at least one example $(x_i, h(x_i))$ in S_i with $h_i(x_i) \neq h(x_i)$. Moreover, since every $j \neq i$ has $h_j \in \mathbb{C}_{S_i}$, it must also be that $h_j(x_i) = h(x_i)$ for this example. We thus have a sequence $x_1, \ldots, x_n \in \mathcal{X}$ such that, $\forall i, j \in \{1, \ldots, n\}$, $h_i(x_j) = h(x_j)$ iff $i \neq j$. Thus, $x_1, \ldots, x_n$ is an extended star set (for $\mathbb{C}$) centered at h, so that $\bar{\mathfrak{s}}_h \geq n$. Since such a shattered set $\{h_1, \ldots, h_n\} \subseteq \mathbb{C}$ exists for any finite $n \leq \mathrm{VC}(\mathcal{V}_h(\mathbb{C}))$, we conclude that $\bar{\mathfrak{s}}_h \geq \mathrm{VC}(\mathcal{V}_h(\mathbb{C}))$.

Together with the fact that $\mathrm{VC}(\mathcal{V}_h(\mathbb{C})) \leq \bar{\mathfrak{s}}_h$ established above, we have that $\mathrm{VC}(\mathcal{V}_h(\mathbb{C})) = \bar{\mathfrak{s}}_h$. Moreover, if $h \in \mathbb{C}$, we can always take $h_0 = h$ to witness that any extended star set centered at h is also a star set centered at h, so that $\bar{\mathfrak{s}}_h = \mathfrak{s}_h$ in this case. Therefore, for $h \in \mathbb{C}$, we have $\mathrm{VC}(\mathcal{V}_h(\mathbb{C})) = \mathfrak{s}_h$.

We may argue that $\mathrm{VC}(\mathcal{V}(\mathbb{C})) \leq \bar{\mathfrak{s}}$ by a nearly-identical argument to the above. Consider any set $\{h_1, \ldots, h_n\} \subseteq \mathbb{C}$ shattered by $\mathcal{V}(\mathbb{C})$. In particular, this implies that $\forall i \in \{1, \ldots, n\}$, there exists a data set S_i such that $\mathbb{C}_{S_i} \cap \{h_1, \ldots, h_n\} = \{h_j : j \in \{1, \ldots, n\} \setminus \{i\}\}$. Since $h_i \notin \mathbb{C}_{S_i}$, there must exist at least one example (x_i, y_i) in S_i such that $h_i(x_i) \neq y_i$. Moreover, for each $j \neq i$, since $h_j \in \mathbb{C}_{S_i}$, it must be that $h_j(x_i) = y_i$. We have therefore constructed a sequence $(x_1, y_1), \ldots, (x_n, y_n)$ such that, $\forall i, j \in \{1, \ldots, n\}$, $h_i(x_j) = y_j$ iff $j \neq i$: that is, an extended star set centered at some h with $h(x_i) = y_i$ for all i (noting that the above property implies every x_i is necessarily distinct, so that such an h exists). Thus, we have that $\bar{\mathfrak{s}} \geq n$. Since such a shattered set $\{h_1, \ldots, h_n\} \subseteq \mathbb{C}$ exists for every finite $n \leq \mathrm{VC}(\mathcal{V}(\mathbb{C}))$, we conclude that $\bar{\mathfrak{s}} \geq \mathrm{VC}(\mathcal{V}(\mathbb{C}))$. Together with the above argument establishing that $\mathrm{VC}(\mathcal{V}(\mathbb{C})) \geq \bar{\mathfrak{s}}$, we conclude that $\mathrm{VC}(\mathcal{V}(\mathbb{C})) = \bar{\mathfrak{s}}$.

Similarly to the above, we can argue that $\mathrm{VC}(\mathcal{D}_h(\mathbb{C})) \geq \mathfrak{s}_h$ for any $h \in \mathbb{C}$. Let $\{x_1, \ldots, x_n\}$ be a star set (for $\mathbb{C}$) centered at h, and let $h_0, \ldots, h_n$ be as in the definition of $\mathfrak{s}_h$: that is, $\forall i \in \{0, \ldots, n\}$, $\forall j \in \{1, \ldots, n\}$, $h_i(x_j) = h(x_j)$ iff $i \neq j$. We will argue that $\{x_1, \ldots, x_n\}$ is shattered by $\mathcal{D}_h(\mathbb{C})$. For any $I \subseteq \{1, \ldots, n\}$, again letting $S = \{(x_j, h(x_j))\}_{j \in \{1,\ldots,n\} \setminus I}$, we claim that $\mathrm{DIS}(\mathbb{C}_S) \cap \{x_1, \ldots, x_n\} = \{x_i : i \in I\}$. To see this, note that the definition of the h_i functions guarantees that, for each $i \in I \cup \{0\}$ and $j \in \{1, \ldots, n\} \setminus I$, since $j \neq i$ we have $h_i(x_j) = h(x_j)$, so that $h_i \in \mathbb{C}_S$; thus, since each $i \in I$ has $h_i(x_i) \neq h_0(x_i)$, and we have argued that $\{h_i, h_0\} \subseteq \mathbb{C}_S$, we have that $\{x_i : i \in I\} \subseteq \mathrm{DIS}(\mathbb{C}_S)$. Also note that every $h' \in \mathbb{C}_S$ has $h'(x_j) = h(x_j)$ for all $j \in \{1, \ldots, n\} \setminus I$ by definition of $\mathbb{C}_S$, so that any such j has $x_j \notin \mathrm{DIS}(\mathbb{C}_S)$. Together, we have $\mathrm{DIS}(\mathbb{C}_S) \cap \{x_1, \ldots, x_n\} = \{x_i : i \in I\}$. Thus, $\{x_1, \ldots, x_n\}$ is shattered by $\mathcal{D}_h(\mathbb{C})$, so that $\mathrm{VC}(\mathcal{D}_h(\mathbb{C})) \geq \mathfrak{s}_h$. Since such a star set exists for any finite $n \leq \mathfrak{s}_h$, we conclude that $\mathrm{VC}(\mathcal{V}_h(\mathbb{C})) \geq \mathfrak{s}_h$.

We next argue that $\mathfrak{s}_h \geq \mathrm{VC}(\mathcal{D}_h(\mathbb{C}))$ for any $h \in \mathbb{C}$, as follows. Let $\{x_1, \ldots, x_n\} \subseteq \mathcal{X}$ be a set shattered by $\mathcal{D}_h(\mathbb{C})$. In particular, from the definition of shattering, for every $i \in \{1, \ldots, n\}$, there must exist a data set S_i consistent with h (i.e., a sequence of pairs $(x, h(x))$) such that $\mathrm{DIS}(\mathbb{C}_{S_i}) \cap \{x_1, \ldots, x_n\} = \{x_i\}$. Since $x_i \in \mathrm{DIS}(\mathbb{C}_{S_i})$, there must exist some $h_i \in \mathbb{C}_{S_i}$ with $h_i(x_i) \neq h(x_i)$. Moreover, since S_i is consistent with h, and $h \in \mathbb{C}$, we know that $h \in \mathbb{C}_{S_i}$; thus, since every $j \neq i$ have $x_j \notin \mathrm{DIS}(\mathbb{C}_{S_i})$, and $\{h_i, h\} \subseteq \mathbb{C}_{S_i}$, it must be that $h_i(x_j) = h(x_j)$. Defining $h_0 = h$, we have thus found a sequence $h_0, \ldots, h_n$ such that, $\forall i \in \{0, 1, \ldots, n\}$, $\forall j \in \{1, \ldots, n\}$, $h_i(x_j) = h(x_j)$ iff $j \neq i$: that is, $x_1, \ldots, x_n$ is a star set (for $\mathbb{C}$) centered at h. Therefore, we have that $\mathfrak{s}_h \geq n$. Since there exist sets $\{x_1, \ldots, x_n\}$ shattered by $\mathcal{D}_h(\mathbb{C})$ for every finite $n \leq \mathrm{VC}(\mathcal{D}_h(\mathbb{C}))$, we conclude that $\mathfrak{s}_h \geq \mathrm{VC}(\mathcal{D}_h(\mathbb{C}))$. Combining this with the above argument that $\mathrm{VC}(\mathcal{D}_h(\mathbb{C})) \geq \mathfrak{s}_h$, we conclude that $\mathrm{VC}(\mathcal{D}_h(\mathbb{C})) = \mathfrak{s}_h$ for any $h \in \mathbb{C}$. ■

Next we present the proof of Proposition 4.

Proof of Proposition 4 Suppose $x_1, \ldots, x_n \in \mathcal{X}$ are shattered by $\mathcal{D}(\mathbb{C})$, and for each $I \subseteq \{1, \ldots, n\}$ let $S_I \in (\mathcal{X} \times \mathcal{Y})^*$ be such that $\mathrm{DIS}(\mathbb{C}_{S_I}) \cap \{x_1, \ldots, x_n\} = \{x_i : i \in I\}$. We first argue that, without loss of generality, we may take these S_I sets to be labelings of subsets of $\{x_1, \ldots, x_n\}$. For each $I \subseteq \{1, \ldots, n\}$ with $I \neq \emptyset$, it must be that $\mathbb{C}_{S_I} \neq \emptyset$, and hence for each $j \in \{1, \ldots, n\} \setminus I$, there is a label $y_j \in \mathcal{Y}$ such that every $h \in \mathbb{C}_{S_I}$ has $h(x_j) = y_j$. Hence, denoting by $S'_I = \{(x_j, y_j) : j \in \{1, \ldots, n\} \setminus I\}$, we have $\mathbb{C}_{S'_I} \supseteq \mathbb{C}_{S_I}$, so that $\mathrm{DIS}(\mathbb{C}_{S'_I}) \supseteq \mathrm{DIS}(\mathbb{C}_{S_I})$, and yet clearly we still have $x_j \notin \mathrm{DIS}(\mathbb{C}_{S'_I})$ for each $j \in \{1, \ldots, n\} \setminus I$ (i.e., every $h \in \mathbb{C}_{S'_I}$ has $h(x_j) = y_j$, by definition). Thus, $\mathrm{DIS}(\mathbb{C}_{S'_I}) \cap \{x_1, \ldots, x_n\} = \{x_i : i \in I\}$. Moreover, for $I = \emptyset$, we can take any $h_0 \in \mathbb{C}$ and define $S'_I = \{(x_1, h_0(x_1)), \ldots, (x_n, h_0(x_n))\}$ so that again we have $\mathbb{C}_{S'_I} \neq \emptyset$ and $\mathrm{DIS}(\mathbb{C}_{S'_I}) \cap \{x_1, \ldots, x_n\} = \emptyset = \{x_i : i \in \mathcal{I}\}$. We therefore have that $\mathcal{D}' := \{\mathrm{DIS}(\mathbb{C}_{S'_I}) : I \subseteq \{1, \ldots, n\}\}$ shatters $\{x_1, \ldots, x_n\}$. In particular, since there are exactly 2^n sets in this collection $\mathcal{D}'$, this further implies that the $\mathbb{C}_{S'_I}$ version spaces are all *distinct*, and are non-empty by definition.

We aim to show that $n \leq 2\mathfrak{s} \log_2(e|\mathcal{Y}|)$. If $n \leq \mathfrak{s}$, this is trivially satisfied. For the remaining case, let $n > \mathfrak{s}$. By the above observation, we know $|\mathcal{D}'| = 2^n$. Moreover, recall that for any data set S'_I, since $\mathbb{C}_{S'_I} \neq \emptyset$, S'_I contains a subset S''_I of size at most $\mathfrak{s}$ for which $\mathbb{C}_{S''_I} = \mathbb{C}_{S'_I}$: that is, a version space compression set of size at most $\mathfrak{s}$ (see Theorem 13 of Hanneke and Yang, 2015; as remarked in footnote 3 of Section 3, though their result was stated for binary classification, their proof also applies to general $\mathcal{Y}$ spaces). Thus, the number of distinct sets in $\mathcal{D}'$ is at most the number

of possible realizable labeled data sets S'' of size at most $\mathfrak{s}$, which is at most $\binom{n}{\leq \mathfrak{s}}|\mathcal{Y}|^{\mathfrak{s}} \leq \left(\frac{en}{\mathfrak{s}}\right)^{\mathfrak{s}}|\mathcal{Y}|^{\mathfrak{s}}$. Together, we have that

$$2^n \leq \left(\frac{en}{\mathfrak{s}}\right)^{\mathfrak{s}}|\mathcal{Y}|^{\mathfrak{s}}.$$

Taking $\log_2$ of both sides yields

$$n \leq \mathfrak{s}\log_2\left(\frac{en}{\mathfrak{s}}\right) + \mathfrak{s}\log_2(|\mathcal{Y}|).$$

Together with Lemma 4.6 of Vidyasagar (2003), this implies

$$n < 2\mathfrak{s}\log_2(e) + 2\mathfrak{s}\log_2(|\mathcal{Y}|) = 2\mathfrak{s}\log_2(e|\mathcal{Y}|).$$

Since such a set $\{x_1, \ldots, x_n\}$ shattered by $\mathcal{D}(\mathbb{C})$ exists for every finite $n \leq \mathrm{VC}(\mathcal{D}(\mathbb{C}))$, we conclude that $\mathrm{VC}(\mathcal{D}(\mathbb{C})) \leq 2\mathfrak{s}\log_2(e|\mathcal{Y}|)$. ■

F.2. Proof of Corollary 15 (Relating Star Number and Dual VC Dimension)

The facts that $\bar{\mathfrak{s}} \geq \mathrm{VC}(\mathrm{HS}(\mathbb{C})) \geq \mathrm{VC}^*(\mathbb{C})$ and $\bar{\mathfrak{s}}_{\mathbf{1}} \geq \mathrm{VC}^*(\mathbb{C})$ follow immediately from Theorem 3, since $\bar{\mathfrak{s}} = \mathrm{VC}(\mathcal{V}(\mathbb{C}))$ and $\mathcal{V}(\mathbb{C}) \supseteq \mathrm{HS}(\mathbb{C})$, whereas $\bar{\mathfrak{s}}_{\mathbf{1}} = \mathrm{VC}(\mathcal{V}_{\mathbf{1}}(\mathbb{C}))$ and $\mathrm{VC}^*(\mathbb{C}) = \mathrm{VC}(\{\mathbb{C}_{\{(x,1)\}} : x \in \mathcal{X}\})$, while $\mathcal{V}_{\mathbf{1}}(\mathbb{C}) \supseteq \{\mathbb{C}_{\{(x,1)\}} : x \in \mathcal{X}\}$. It also follows immediately from $\mathrm{HS}(\mathbb{C}) \supseteq \{\mathbb{C}_{\{(x,1)\}} : x \in \mathcal{X}\}$ that $\mathrm{VC}(\mathrm{HS}(\mathbb{C})) \geq \mathrm{VC}^*(\mathbb{C})$.

It only remains to establish the claim that, for any concept h, $\bar{\mathfrak{s}}_h \geq \mathrm{VC}(\mathrm{HS}(\mathbb{C}))/2$. Consider any $\{h_1, \ldots, h_n\} \subseteq \mathbb{C}$ shattered by $\mathrm{HS}(\mathbb{C})$. Since $\emptyset$ and $\mathbb{C}$ are contained in $\mathrm{HS}(\mathbb{C})$ by definition, these may serve as the sets $D, D' \in \mathrm{HS}(\mathbb{C})$ with $D \cap \{h_1, \ldots, h_n\} = \emptyset$ and $D' \cap \{h_1, \ldots, h_n\} = \{h_1, \ldots, h_n\}$. However, for the remaining subsets, they must be witnessed by the non-trivial half-spaces: that is, sets of the form $\mathbb{C}_{\{(x,y)\}}$, $(x,y) \in \mathcal{X} \times \mathcal{Y}$. Thus, for every $H \in 2^{\{h_1,\ldots,h_n\}} \setminus \{\emptyset, \{h_1, \ldots, h_n\}\}$, there exists $(x,y) \in \mathcal{X} \times \mathcal{Y}$ such that $\mathbb{C}_{\{(x,y)\}} \cap \{h_1, \ldots, h_n\} = H$. Note that, for each such (x,y), there must be a "mirror" point (x',y') such that $\mathbb{C}_{\{(x',y')\}} \cap \{h_1, \ldots, h_n\} = \{h_1, \ldots, h_n\} \setminus H = \mathbb{C}_{\{(x,1-y)\}} \cap \{h_1, \ldots, h_n\}$, and therefore without loss of generality we may suppose $(x',y') = (x, 1-y)$. Thus, there exist $2^{n-1}-1$ points $x_1, \ldots, x_{2^{n-1}-1} \in \mathcal{X}$ such that $\{\mathbb{C}_{\{(x_i,y)\}} : i \leq 2^{n-1}-1, y \in \mathcal{Y}\} \cup \{\emptyset, \mathbb{C}\}$ shatters $\{h_1, \ldots, h_n\}$. Moreover, for each H and (x,y) as above, either $H = \mathbb{C}_{\{(x,h(x))\}} \cap \{h_1, \ldots, h_n\}$ or $\{h_1, \ldots, h_n\} \setminus H = \mathbb{C}_{\{(x,h(x))\}} \cap \{h_1, \ldots, h_n\}$. Since we also have $\mathbb{C} = \mathbb{C}_{\emptyset} \in \mathcal{V}_h(\mathbb{C})$, we see that

$$|\{V \cap \{h_1, \ldots, h_n\} : V \in \mathcal{V}_h(\mathbb{C})\}| \geq 2^{n-1}.$$

We may also complement the above inequality with the well-known *Sauer's lemma* (Sauer, 1972; Shelah, 1978; Vapnik and Chervonenkis, 1974), which states that if $n \geq \mathrm{VC}(\mathcal{V}_h(\mathbb{C}))$, then

$$|\{V \cap \{h_1, \ldots, h_n\} : V \in \mathcal{V}_h(\mathbb{C})\}| \leq \sum_{i=0}^{\mathrm{VC}(\mathcal{V}_h(\mathbb{C}))} \binom{n}{i}.$$

Since $\sum_{i=0}^{n}\binom{n}{i} = 2^n$ and each i satisfies $\binom{n}{i} = \binom{n}{i-1}$, we may observe that $\sum_{i=0}^{\lceil n/2 \rceil - 1}\binom{n}{i} < 2^{n-1}$, so that the above two inequalities for $|\{V \cap \{h_1, \ldots, h_n\} : V \in \mathcal{V}_h(\mathbb{C})\}|$ together imply $\mathrm{VC}(\mathcal{V}_h(\mathbb{C})) \geq \lceil n/2 \rceil$. Since Theorem 3 implies $\mathrm{VC}(\mathcal{V}_h(\mathbb{C})) = \bar{\mathfrak{s}}_h$, this completes the proof. ■

F.3. Proofs of the Relation Between the Star Number of Disagreement Regions and the Star Number of the Concept Class

This section presents the proofs of Theorem 16, Proposition 17, and Proposition 18. We begin with the proof of Theorem 16, establishing that $\mathfrak{s}_\emptyset(\mathcal{V}_h(\mathbb{C})) = \mathfrak{s}_\emptyset(\mathcal{D}_h(\mathbb{C})) = \mathfrak{s}_h(\mathbb{C})$.

Proof of Theorem 16 Fix any $\mathbb{C}$ and any concept h. First consider any (extended) star set $\{x_1, \ldots, x_n\}$ (for $\mathbb{C}$) centered at h. By definition, there exist $h_1, \ldots, h_n \in \mathbb{C}$ such that $\forall i, j \in \{1, \ldots, n\}$, $h_i(x_j) = h(x_j)$ iff $j \neq i$. Note that for any $i \in \{1, \ldots, n\}$, we have $\mathbb{C}_{\{(x_i, h(x_i)\}} \cap \{h_1, \ldots, h_n\} = \{h_j : j \in \{1, \ldots, n\} \setminus \{i\}\}$. Moreover, we trivially have $\mathbb{C} \cap \{h_1, \ldots, h_n\} = \{h_1, \ldots, h_n\}$. Since $\mathbb{C} \in \mathcal{V}_h(\mathbb{C})$ and each $i \in \{1, \ldots, n\}$ has $\mathbb{C}_{\{(x_i, h(x_i))\}} \in \mathcal{V}_h(\mathbb{C})$, this establishes that $\{h_1, \ldots, h_n\}$ is a star set (for $\mathcal{V}_h(\mathbb{C})$) centered at $\mathbb{C}$. Thus, $\mathfrak{s}_\mathbb{C}(\mathcal{V}_h(\mathbb{C})) \geq n$. Since such an (extended) star set $\{x_1, \ldots, x_n\}$ exists for every finite $n \leq \bar{\mathfrak{s}}_h$, we have that $\mathfrak{s}_\mathbb{C}(\mathcal{V}_h(\mathbb{C})) \geq \bar{\mathfrak{s}}_h$.

Moreover, in the case $h \in \mathbb{C}$, for any $i \in \{1, \ldots, n\}$, letting $S_i = \{(x_j, h(x_j)) : j \in \{1, \ldots, n\} \setminus \{i\}\}$, by definition of h_i we have $\{h, h_i\} \subseteq \mathbb{C}_{S_i}$, and $h_i(x_i) \neq h(x_i)$, so that $x_i \in \mathrm{DIS}(\mathbb{C}_{S_i})$. Furthermore, since S_i contains $(x_j, h(x_j))$ for every $j \neq i$, we have that every $h' \in \mathbb{C}_{S_i}$ has $h'(x_j) = h(x_j)$ for every $j \neq i$, so that $x_j \notin \mathrm{DIS}(\mathbb{C}_{S_i})$. Altogether, we have that every $i \in \{1, \ldots, n\}$ satisfies $\mathrm{DIS}(\mathbb{C}_{S_i}) \cap \{x_1, \ldots, x_n\} = \{x_i\}$. Additionally, letting $S_0 = \{(x_1, h(x_1)), \ldots, (x_n, h(x_n))\}$, every $h' \in \mathbb{C}_{S_0}$ has $h'(x_j) = h(x_j)$ for all $j \in \{1, \ldots, n\}$, and therefore $\mathrm{DIS}(\mathbb{C}_{S_0}) \cap \{x_1, \ldots, x_n\} = \emptyset$. Since $\mathbb{C}_{S_i} \in \mathcal{V}_h(\mathbb{C})$ for every $i \in \{0, \ldots, n\}$, we have thus established that $\{x_1, \ldots, x_n\}$ is a star set (for $\mathcal{D}_h(\mathbb{C})$) centered at $\emptyset$. Since such a star set $\{x_1, \ldots, x_n\}$ (for $\mathbb{C}$) centered at h exists for every finite $n \leq \mathfrak{s}_h$, we conclude that $\mathfrak{s}_\emptyset(\mathcal{D}_h(\mathbb{C})) \geq \mathfrak{s}_h$. It remains to complement these with the opposite inequalities.

Consider again any concept h, and consider any star set $\{h_1, \ldots, h_n\} \subseteq \mathbb{C}$ (for $\mathcal{V}_h(\mathbb{C})$) centered at $\mathbb{C}$. By definition, there exist data sets $S_1, \ldots, S_n$ consistent with h (i.e., sequences of pairs $(x, h(x))$) such that every $i \in \{1, \ldots, n\}$ has $\mathbb{C}_{S_i} \cap \{h_1, \ldots, h_n\} = \{h_j : j \in \{1, \ldots, n\} \setminus \{i\}\}$. In particular, there must exist at least one point $(x_i, h(x_i))$ in S_i such that $h_i(x_i) \neq h(x_i)$. Moreover, since every $j \neq i$ has $h_j \in \mathbb{C}_{S_i}$, we also have that $h_j(x_i) = h(x_i)$. We have thus found a sequence $\{x_1, \ldots, x_n\}$ such that $\forall i, j \in \{1, \ldots, n\}$, $h_i(x_j) = h(x_j)$ iff $i \neq j$. We have therefore established that $\{x_1, \ldots, x_n\}$ is an extended star set (for $\mathbb{C}$) centered at h, so that $\bar{\mathfrak{s}}_h \geq n$. Since there exists such a star set $\{h_1, \ldots, h_n\}$ (for $\mathcal{V}_h(\mathbb{C})$) centered at $\mathbb{C}$ for every finite $n \leq \mathfrak{s}_\mathbb{C}(\mathcal{V}_h(\mathbb{C}))$, we conclude that $\bar{\mathfrak{s}}_h \geq \mathfrak{s}_\mathbb{C}(\mathcal{V}_h(\mathbb{C}))$. Together with the fact that $\mathfrak{s}_\mathbb{C}(\mathcal{V}_h(\mathbb{C})) \geq \bar{\mathfrak{s}}_h$ (established above), we further conclude that $\bar{\mathfrak{s}}_\mathbb{C}(\mathcal{V}_h(\mathbb{C})) = \mathfrak{s}_h$.

Next, consider the case $h \in \mathbb{C}$, and consider any star set $\{x_1, \ldots, x_n\}$ (for $\mathcal{D}_h(\mathbb{C})$) centered at $\emptyset$. By definition, there exist data sets $S_1, \ldots, S_n$ consistent with h (i.e., sequences of pairs $(x, h(x))$), such that $\forall i \in \{1, \ldots, n\}$, $\mathrm{DIS}(\mathbb{C}_{S_i}) \cap \{x_1, \ldots, x_n\} = \{x_i\}$. In particular, this implies there exists $h_i \in \mathbb{C}_{S_i}$ with $h_i(x_i) \neq h(x_i)$. Moreover, since every $j \neq i$ has $x_j \notin \mathrm{DIS}(\mathbb{C}_{S_i})$, and since $h \in \mathbb{C}_{S_i}$ (S_i being a sequence of $(x, h(x))$ pairs), it must be that every $h' \in \mathbb{C}_{S_i}$ has $h'(x_j) = h(x_j)$, and therefore in particular, $h_i(x_j) = h(x_j)$. Letting $h_0 = h$, we have thus found a sequence $h_0, h_1, \ldots, h_n$ such that, $\forall i \in \{0, \ldots, n\}$, $\forall j \in \{1, \ldots, n\}$, $h_i(x_j) = h(x_j)$ iff $j \neq i$: that is, $\{x_1, \ldots, x_n\}$ is a star set (for $\mathbb{C}$) centered at h. It follows that $\mathfrak{s}_h \geq n$. Since there exists such a star set $\{x_1, \ldots, x_n\}$ (for $\mathcal{D}_h(\mathbb{C})$) centered at $\emptyset$ for every finite $n \leq \mathfrak{s}_\emptyset(\mathcal{D}_h(\mathbb{C}))$, we conclude that $\mathfrak{s}_h \geq \mathfrak{s}_\emptyset(\mathcal{D}_h(\mathbb{C}))$. Together with the fact that $\mathfrak{s}_\emptyset(\mathcal{D}_h(\mathbb{C})) \geq \mathfrak{s}_h$ (established above), we further conclude that $\mathfrak{s}_\emptyset(\mathcal{D}_h(\mathbb{C})) = \mathfrak{s}_h$, which completes the proof. ■

Next we present the proof of Proposition 17, establishing that in general there can be large gaps between $\mathfrak{s}(\mathbb{C})$ and the value $\mathfrak{s}(\mathcal{D}_h(\mathbb{C}))$.

Proof of Proposition 17 Consider the concept class $\mathbb{C}$ of homogeneous linear classifiers $\mathbb{C} = \{x \mapsto h_w(x) := \mathbb{1}[w^\top x \geq 0] : w \in \mathbb{R}^2, \|w\| = 1\}$ on $\mathcal{X} = \{x \in \mathbb{R}^2 : \|x\| = 1\}$ the unit circle in $\mathbb{R}^2$. Let $h = h_w \in \mathbb{C}$. Since any two distinct non-antipodal points x_1, x_2 are *shattered* by $\mathbb{C}$ (and are therefore a star set centered at *every* concept), it is clear that $\mathfrak{s}_h \geq 2$. To see that $\mathfrak{s}_h < 3$, consider any three points $x_1, x_2, x_3 \in \mathcal{X}$ (and without loss of generality, suppose they are distinct, or else clearly they cannot be a star set). If two of the points, x_a and x_b are antipodal and $h(x_a) = h(x_b) = 1$, then *any* $h' \in \mathbb{C}$ with $h' \neq h$ has either $h'(x_a) = 0$ or $h'(x_b) = 0$, so in particular, there does not exist $h' \in \mathbb{C}$ with $h'(x_c) \neq h(x_c)$ (for the third point x_c) while $h'(x_a) = h(x_a)$ and $h'(x_b) = h(x_b)$, so that x_1, x_2, x_3 are not a star set centered at h. On the other hand, suppose no such x_a, x_b exist. Consider the line (through the origin) corresponding to the *decision boundary* of h (i.e., the points $x \in \mathbb{R}^2$ with $w^\top x = 0$). Note that we can view any $h' \in \mathbb{C}$ with $h' \neq h$ as a *rotation* of h. However, among x_1, x_2, x_3, there is a point x_i such that rotating the line corresponding to h clockwise will intersect x_i first, and a point x_j such that rotating the line counterclockwise will intersect x_j first (in either case, if there is a "tie", it must be that there were two antipodal points, in which case we choose the point whose $h(x)$ value was 0). Thus, for any $h' \in \mathbb{C}$ with $h'(x_k) \neq h(x_k)$ for the third point x_k (the remaining element in $\{x_1, x_2, x_3\} \setminus \{x_i, x_j\}$), the minimal-angle rotation transforming h to h' must cross one of x_i or x_j, so that either $h'(x_i) \neq h(x_i)$ or $h'(x_j) \neq h(x_j)$. Therefore, x_1, x_2, x_3 is not a star set.

Next consider $\mathfrak{s}_{\mathcal{X}}(\mathcal{D}_h(\mathbb{C}))$. For this, note that $\mathbb{C} \in \mathcal{V}_h(\mathbb{C})$, and $\mathrm{DIS}(\mathbb{C}) = \mathcal{X}$. On the other hand, for any $x \in \mathcal{X}$, $\mathrm{DIS}(\mathbb{C}_{\{(x,h(x))\}}) = \mathcal{X} \setminus \{x\}$. Thus, effectively *all of* $\mathcal{X}$ is a star set (for $\mathcal{D}_h(\mathbb{C})$) centered at $\mathcal{X}$ (and in particular, any finite subset $\mathcal{X}' \subset \mathcal{X}$ is a finite star set centered at $\mathcal{X}$), so that $\mathfrak{s}_{\mathcal{X}}(\mathcal{D}_h(\mathbb{C})) = \infty$.

Next we turn to the claim about $\mathcal{V}_h(\mathbb{C})$. Let $\mathcal{Y} = \mathcal{X} = \mathbb{N}$, define $x \mapsto h(x) := x$ (the identity function), and for each $y \in \mathbb{N}$ define $x \mapsto h_y(x) = y$ (the constant functions), and let $\mathbb{C} = \{h_y : y \in \mathcal{Y}\} \cup \{h\}$. Any distinct $x, x' \in \mathcal{X}$ are a star set centered at h, as witnessed by $h_{x'}$ and h_x, so $\bar{\mathfrak{s}}(\mathbb{C}) \geq \mathfrak{s}_h(\mathbb{C}) \geq 2$. On the other hand, for any $x, x', x'' \in \mathcal{X}$, if we suppose (for the sake of contradiction) that they are a star set centered at some h_0, then since the 3 functions witnessing this must all be distinct functions in $\mathbb{C}$, it must be that at least two of them are constant functions h_y, $h_{y'}$, which (by definition of a star set) means h_y and $h_{y'}$ must *agree* with h_0 on at least *one* of x, x', x'', which implies $y = y'$: a contradiction (since they are distinct *constant* functions). Thus, $\bar{\mathfrak{s}}(\mathbb{C}) \leq 2$. On the other hand, for every $y \in \mathcal{Y}$, $\{h, h_y\} = \mathbb{C}_{\{(y,y)\}} \in \mathcal{V}_h(\mathbb{C})$, so that $\mathbb{C}_{\{(y,y)\}} \cap \{h_{y'} : y' \in \mathcal{Y}\} = \{h_y\}$: that is, $\{h_y : y \in \mathcal{Y}\}$ is an infinite star set for $\mathcal{V}_h(\mathbb{C})$ centered at $\emptyset$. Therefore, $\mathfrak{s}_\emptyset(\mathcal{V}_h(\mathbb{C})) = \infty$. ■

We conclude this section by presenting the proof of Proposition 18, establishing that the star number is *nearly self-dual*.

Proof of Proposition 18 Noting that $\mathbb{C}^{**} = \mathbb{C}$, the second claimed sequence of inequalities follows immediately from the first claimed sequence of inequalities, so we need only establish the first claim. The inequality $\bar{\mathfrak{s}}(\mathbb{C}^*) \geq \bar{\mathfrak{s}}_{\mathrm{const}}(\mathbb{C}^*)$ follows immediately from the definition of $\bar{\mathfrak{s}}(\mathbb{C}^*)$: that is, $\bar{\mathfrak{s}}(\mathbb{C}^*) = \sup_f \bar{\mathfrak{s}}_f(\mathbb{C}^*) \geq \sup_{y \in \mathcal{Y}} \bar{\mathfrak{s}}_{h \mapsto y}(\mathbb{C}^*)$.

Next we show that $\bar{\mathfrak{s}}_{\mathrm{const}}(\mathbb{C}^*) = \bar{\mathfrak{s}}_{\mathrm{const}}(\mathbb{C})$. Let $\{x_1, \ldots, x_n\}$ be an extended star set for $\mathbb{C}$ centered at a constant function $x \mapsto y_0$ for some $y_0 \in \mathcal{Y}$: that is, $\exists h_1, \ldots, h_n \in \mathbb{C}$ such that $\forall i, j \in$

$\{1,\dots,n\}$, $h_i(x_j) = y_0$ iff $i \neq j$. By definition, this can be stated equivalently as $f_{x_j}(h_i) = y_0$ iff $i \neq j$, so that $f_{x_1},\dots,f_{x_n} \in \mathbb{C}^*$ witness the fact that $h_1,\dots,h_n$ are an extended star set for $\mathbb{C}^*$ centered at the constant function $h \mapsto y_0$. Since such an extended star set $x_1,\dots,x_n$ centered at some constant function exists for every finite $n \leq \bar{\mathfrak{s}}_{\text{const}}(\mathbb{C})$, we have $\bar{\mathfrak{s}}_{\text{const}}(\mathbb{C}^*) \geq \bar{\mathfrak{s}}_{\text{const}}(\mathbb{C})$. Moreover, applying this inequality to the concept class $\mathbb{C}^*$, we have $\bar{\mathfrak{s}}_{\text{const}}(\mathbb{C}) = \bar{\mathfrak{s}}_{\text{const}}(\mathbb{C}^{**}) \geq \bar{\mathfrak{s}}_{\text{const}}(\mathbb{C}^*)$ as well, so that $\bar{\mathfrak{s}}_{\text{const}}(\mathbb{C}^*) = \bar{\mathfrak{s}}_{\text{const}}(\mathbb{C})$.

Finally, we argue that $\bar{\mathfrak{s}}_{\text{const}}(\mathbb{C}) \geq \frac{1}{|\mathcal{Y}|}\bar{\mathfrak{s}}(\mathbb{C})$. Let $x_1,\dots,x_n$ be an extended star set for $\mathbb{C}$ centered at some function $h : \mathcal{X} \to \mathcal{Y}$: that is, $\exists h_1,\dots,h_n \in \mathbb{C}$ such that $\forall i,j \in \{1,\dots,n\}$, $h_i(x_j) = h(x_j)$ iff $i \neq j$. By the pigeonhole principle, there exists $y_0 \in \mathcal{Y}$ such that $|\{x_j : h(x_j) = y_0, j \in \{1,\dots,n\}\}| \geq \frac{n}{|\mathcal{Y}|}$. By definition, the functions h_i with $h(x_i) = y_0$ witness the fact that $\{x_j : h(x_j) = y_0, j \in \{1,\dots,n\}\}$ is an extended star set for $\mathbb{C}$ centered at the constant function $x \mapsto y_0$, so that $\bar{\mathfrak{s}}_{\text{const}}(\mathbb{C}) \geq |\{x_j : h(x_j) = y_0, j \in \{1,\dots,n\}\}| \geq \frac{n}{|\mathcal{Y}|}$. Since such an extended star set for $\mathbb{C}$ exists for every finite $n \leq \bar{\mathfrak{s}}(\mathbb{C})$, we have that $\bar{\mathfrak{s}}_{\text{const}}(\mathbb{C}) \geq \frac{1}{|\mathcal{Y}|}\bar{\mathfrak{s}}(\mathbb{C})$. This completes the proof. ■

Appendix G. Proof of Theorem 7 (Exact Learning with Membership Queries)

The proof follows closely the proofs of Hegedüs (1995), which established analogous results except with $\log(|\mathbb{C}|)$ in place of L, and only for the case of $|\mathcal{Y}| = 2$. Both the generalization to any finite $\mathcal{Y}$, and replacing $\log(|\mathbb{C}|)$ by L, require only minor adjustments to the proof. Most notably, whereas the algorithm of Hegedüs (1995) was essentially based on the *Halving* algorithm (an online learning algorithm guaranteeing mistake bound $\log(|\mathbb{C}|)$), we will instead substitute the *Standard Optimal Algorithm*: SOA (defined below).

The Lower Bound: We begin with the lower bound, establishing that if $|\mathbb{C}| = \infty$ then $\text{QC}_{\text{MQ}}(\mathbb{C}) = \infty$, and if $|\mathbb{C}| < \infty$, then $\text{QC}_{\text{MQ}}(\mathbb{C}) \geq \log_{|\mathcal{Y}|}(|\mathbb{C}|)$. The key observation is that, since the learning algorithm is deterministic, the setting is equivalent to one in which an adversary may respond to the learner's queries x_t with *any* label $y_t \in \mathcal{Y}$, as long as the entire sequence $\{(x_1,y_1),\dots,(x_Q,y_Q)\}$ in the end is realizable by $\mathbb{C}$. Then, for the algorithm to guarantee success, it must guarantee that even for such adversarial responses, it will always end up with $|\mathbb{C}_{\{(x_1,y_1),\dots,(x_Q,y_Q)\}}| = 1$ (otherwise, for whichever concept the learner would return, an adversary can always choose the other as $h^\star$, so that in this case the learner would fail to return $h^\star$ despite all queries being answered according to $h^\star(x_i)$ as required). Based on this equivalent formulation, we have that any $\mathbb{C}$ with $|\mathbb{C}| = 1$ has $\text{QC}_{\text{MQ}}(\mathbb{C}) = 0$, whereas any $\mathbb{C}$ with $|\mathbb{C}| \geq 2$ satisfies

$$\text{QC}_{\text{MQ}}(\mathbb{C}) = 1 + \min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \text{QC}_{\text{MQ}}(\mathbb{C}_{\{(x,y)\}}). \tag{8}$$

Consider first the case of $|\mathbb{C}| < \infty$ and $|\mathcal{Y}| < \infty$. We proceed by induction. If $|\mathbb{C}| = 1$, then $\text{QC}_{\text{MQ}}(\mathbb{C}) = 0 = \log_{|\mathcal{Y}|}(|\mathbb{C}|)$. This will serve as our base case. Now take as an inductive hypothesis that, for some $\mathbb{C}$ with $|\mathbb{C}| \geq 2$, for any $\mathbb{C}' \subsetneq \mathbb{C}$, it holds that $\text{QC}_{\text{MQ}}(\mathbb{C}') \geq \log_{|\mathcal{Y}|}(|\mathbb{C}'|)$. Let $\hat{x} \in \mathcal{X}$ be such that $\max_{y\in\mathcal{Y}} \text{QC}_{\text{MQ}}(\mathbb{C}_{\{(\hat{x},y)\}}) = \min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \text{QC}_{\text{MQ}}(\mathbb{C}_{\{(x,y)\}})$. In particular, note that such an $\hat{x}$ must have that $\mathbb{C}_{\{(\hat{x},y)\}} \subsetneq \mathbb{C}$ for every $y \in \mathcal{Y}$, or else (8) would imply $\text{QC}_{\text{MQ}}(\mathbb{C}) = \infty \geq \log_{|\mathcal{Y}|}(|\mathbb{C}|)$. Moreover, for such an $\hat{x}$, since $\mathbb{C} = \bigcup_{y\in\mathcal{Y}} \mathbb{C}_{\{(\hat{x},y)\}}$, by the pigeonhole principle

there must exist at least one $\hat{y} \in \mathcal{Y}$ with $|\mathbb{C}_{\{(\hat{x},\hat{y})\}}| \geq \frac{1}{|\mathcal{Y}|}|\mathbb{C}|$. Therefore,

$$\mathrm{QC}_{\mathrm{MQ}}(\mathbb{C}) = 1 + \max_{y \in \mathcal{Y}} \mathrm{QC}_{\mathrm{MQ}}(\mathbb{C}_{\{(\hat{x},y)\}}) \geq 1 + \mathrm{QC}_{\mathrm{MQ}}(\mathbb{C}_{\{(\hat{x},\hat{y})\}})$$
$$\geq 1 + \log_{|\mathcal{Y}|}(|\mathbb{C}_{\{(\hat{x},\hat{y})\}}|) \geq 1 + \log_{|\mathcal{Y}|}\left(\frac{1}{|\mathcal{Y}|}|\mathbb{C}|\right) = \log_{|\mathcal{Y}|}(|\mathbb{C}|),$$

where the second inequality is by the inductive hypothesis. It follows that every finite $\mathbb{C}$ satisfies $\mathrm{QC}_{\mathrm{MQ}}(\mathbb{C}) \geq \log_{|\mathcal{Y}|}(|\mathbb{C}|)$ by the principle of induction.

We address the case of $|\mathbb{C}| = \infty$ similarly. Here we only suppose that every $x \in \mathcal{X}$ has $|\{h(x) : h \in \mathbb{C}\}| < \infty$. Consider the execution of some learning algorithm, guaranteed to make at most $Q < \infty$ queries. Let x_1 be the algorithm's first query point. By the pigeonhole principle, there must exist at least one $y_1 \in \{h(x_1) : h \in \mathbb{C}\}$ with $|\mathbb{C}_{\{(x_1,y_1)\}}| = \infty$. Supposing the adversary replies with y_1 to this query, let x_2 be its next query point. Similarly, by the pigeonhole principle, there exists at least one $y_2 \in \{h(x_2) : h \in \mathbb{C}_{\{(x_1,y_1)\}}\}$ such that $|\mathbb{C}_{\{(x_1,y_1),(x_2,y_2)\}}| = \infty$. Let the adversary reply with y_2 to this second query, and let x_3 be the algorithm's third query point. This continues inductively, so that after any number $q \leq Q$ of queries, we still have $|\mathbb{C}_{\{(x_i,y_i)\}_{i\leq q}}| = \infty$. The algorithm will terminate after Q queries, and we will still have $|\mathbb{C}_{\{(x_i,y_i)\}_{i\leq Q}}| = \infty > 1$. Thus, such a learning algorithm fails to be correct for this learning problem. Since this is true of *any* learning algorithm guaranteeing any finite number $Q < \infty$ of queries, we conclude that $\mathrm{QC}_{\mathrm{MQ}}(\mathbb{C}) = \infty$.[15]

The Standard Optimal Algorithm: Before giving the proof of the upper bounds, we first formally define a useful (for the purpose of the upper bound, and in general) online learning algorithm known as the *Standard Optimal Algorithm*, or SOA. The SOA was proposed by Littlestone (1988) for the special case $\mathcal{Y} = \{0, 1\}$, and was extended to handle any label space $\mathcal{Y}$ by Daniely, Sabato, Ben-David, and Shalev-Shwartz (2015). It implements a function $\mathrm{SOA} : (\mathcal{X} \times \mathcal{Y})^* \times \mathcal{X} \to \mathcal{Y}$, defined as follows: for any $n \in \mathbb{N} \cup \{0\}$ and any data set $S = \{(x_1, y_1), \ldots, (x_n, y_n)\} \in (\mathcal{X} \times \mathcal{Y})^n$ realizable by $\mathbb{C}$, for any $x \in \mathcal{X}$,

$$\mathrm{SOA}(S, x) = \operatorname*{argmax}_{y \in \mathcal{Y}} \mathrm{L}(\mathbb{C}_{S \cup \{(x,y)\}}),$$

where we interpret $\mathrm{L}(\emptyset) = -1$ for convenience. Littlestone (1988); Daniely, Sabato, Ben-David, and Shalev-Shwartz (2015) make the elementary observation that, for any concept class $\mathbb{C}'$ with $\mathrm{L}(\mathbb{C}') < \infty$, for any $x \in \mathcal{X}$, there is *at most one* $y \in \mathcal{Y}$ with $\mathrm{L}(\mathbb{C}'_{\{(x,y)\}}) = \mathrm{L}(\mathbb{C}')$: otherwise we could make x a root node, with two edges labeled by the y, y' which witness a violation of this property, and upon each of these edges we could hang a subtree of depth $\mathrm{L}(\mathbb{C}')$ shattered by $\mathbb{C}_{\{(x,y)\}}$ and $\mathbb{C}_{\{(x,y')\}}$ respectively, thus overall creating a shattered tree of depth $\mathrm{L}(\mathbb{C}')+1$, contradicting the definition of $\mathrm{L}(\mathbb{C}')$. Based on this fact, Littlestone (1988); Daniely, Sabato, Ben-David, and Shalev-Shwartz (2015) immediately conclude that, for any sequence $\{(x_1, y_1), \ldots, (x_n, y_n)\} \in (\mathcal{X} \times \mathcal{Y})^n$ realizable by $\mathbb{C}$,

$$\sum_{t=1}^{n} \mathbb{1}[\mathrm{SOA}(x_1, y_1, \ldots, x_{t-1}, y_{t-1}, x_t) \neq y_t] \leq \mathrm{L}, \tag{9}$$

15. We have in fact proven a stronger result: namely, that for every deterministic MQ-algorithm, there exist responses $y_1, y_2, \ldots$ to its queries $x_1, x_2, \ldots$, for which all $q < \infty$ have $\{(x_i, y_i)\}_{i\leq q}$ realizable by $\mathbb{C}$, such that there is no finite number of queries Q for which the algorithm has "finished" learning. Thus, the negative result for MQ-learning also extends to the *universal* quantification of learnability studied by Bousquet, Hanneke, Moran, van Handel, and Yehudayoff (2021).

since each "mistake" in the sequence reduces the Littlestone dimension:

$$\mathrm{L}(\mathbb{C}_{\{(x_i,y_i)\}_{i\leq t}}) \leq \mathrm{L}(\mathbb{C}_{\{(x_i,y_i)\}_{i<t}}) - 1.$$

This *mistake bound* guarantee in (9) (a seminal result in the theory of online learning) will be the only property of SOA salient to our proof of Theorem 7.

The Upper Bound: To complete the proof, we prove the claimed upper bound, establishing that $\mathrm{QC}_{\mathrm{MQ}}(\mathbb{C}) \leq \mathrm{XTD}(\mathbb{C})\mathrm{L}$. This part of the proof applies to any label space $\mathcal{Y}$. A key ingredient in this proof is the notion of a *minimal specifying set*. Specifically, for any concept g (not necessarily in $\mathbb{C}$), a minimal specifying set for g is a set $S_g \subseteq \mathcal{X}$ of minimal cardinality such that $|\{h \in \mathbb{C} : h(S) = g(S)\}| \leq 1$. It follows from the definition of $\mathrm{XTD}(\mathbb{C})$ that, for any g, any minimal specifying set S_g for g has $|S_g| \leq \mathrm{XTD}(\mathbb{C})$.

We are now ready to show $\mathrm{QC}_{\mathrm{MQ}}(\mathbb{C}) \leq \mathrm{XTD}(\mathbb{C})\mathrm{L}$. Consider the following learning algorithm for the problem of Exact Learning with Membership Queries for concept class $\mathbb{C}$. This algorithm is identical to an algorithm of Hegedüs (1995) known as MEMB-HALVING-1, except that we substitute the SOA in place of the *Halving* algorithm (which is another online learning algorithm also proposed by Littlestone, 1988).

Algorithm: MEMB-SOA-1
0. Initialize $S = \{\}$
1. While $|\mathbb{C}_S| \geq 2$
2. Let $g(\cdot) = \mathrm{SOA}(S, \cdot)$
3. Let S_g be a minimal specifying set for g
4. Query each $x \in S_g$ (in any order) to observe $h^\star(x)$
5. Let $S \leftarrow S \cup \{(x, h^\star(x)) : x \in S_g\}$
6. Return the sole remaining element $\hat{h} \in \mathbb{C}_S$

We claim that for any $h^\star \in \mathbb{C}$, MEMB-SOA-1 makes a total number of queries at most $\mathrm{XTD}(\mathbb{C})\mathrm{L}$ before terminating. Moreover, since it only terminates when $|\mathbb{C}_S| < 2$, and S is a data set consistent with $h^\star$, we have $h^\star \in \mathbb{C}_S$, so that it always returns $\hat{h} = h^\star$ (as required for correctness). Since each execution of steps 2-5 only queries the elements of a minimal specifying set, the total number of queries per execution of these steps is at most $\mathrm{XTD}(\mathbb{C})$. It therefore suffices to show that the algorithm executes steps 2-5 at most L number of times. In particular, we will argue that after each execution of steps 2-5, $\mathrm{L}(\mathbb{C}_S)$ is reduced by at least 1. To see this, note that on each execution of step 5, we append to S a data set $\{(x, h^\star(x)) : x \in S_g\}$, where $g = \mathrm{SOA}(S)$ (for the data set S before S_g is appended). In particular, if at this time we have $g(x) = h^\star(x)$ for every $x \in S_g$, then by definition of the minimal specifying set we have $|\mathbb{C}_{S\cup S_g}| \leq 1$. For us to enter steps 2-5 we must have $|\mathbb{C}_S| \geq 2$ (before appending S_g), which implies $\mathrm{L}(\mathbb{C}_S) \geq 1$, whereas if $|\mathbb{C}_{S\cup S_g}| \leq 1$ we have $\mathrm{L}(\mathbb{C}_{S\cup S_g}) \leq 0$ (and indeed, it is equal 0 since we will in fact have $\mathbb{C}_{S\cup S_g} = \{h^\star\}$). On the other hand, if there exists $x \in S_g$ such that $g(x) \neq h^\star(x)$, then by definition of $\mathrm{SOA}(S)$ and the aforementioned property that any $y \neq \mathrm{SOA}(S, x)$ has $\mathrm{L}(\mathbb{C}_{S\cup\{(x,y)\}}) < \mathrm{L}(\mathbb{C}_S)$, we conclude that $\mathrm{L}(\mathbb{C}_{S\cup S_g}) \leq \mathrm{L}(\mathbb{C}_S) - 1$. Since we initialize $S = \emptyset$, and therefore $\mathrm{L}(\mathbb{C}_S) = \mathrm{L}(\mathbb{C})$ at the start, and since we always retain $h^\star \in \mathbb{C}_S$ so that $\mathrm{L}(\mathbb{C}_S) \geq 0$, it follows that the algorithm executes steps 2-5 at most $\mathrm{L}(\mathbb{C})$ number of times. This completes the proof. ■

Appendix H. Proof of Theorem 19 (on the VC Dimension of Embedding in an Intersection-Closed Class)

Consider any star set $\{x_1, \ldots, x_n\}$ (for $\mathbb{C}$) centered at $\mathbf{1}$. By definition, there exist $h_0, h_1, \ldots, h_n \in \mathbb{C}$ such that, for every $i, j \in \{1, \ldots, n\}$, $h_0(x_j) = 1$ and $h_i(x_j) = \mathbb{1}[j \neq i]$: that is, h_0 classifies all n points as 1, whereas for every x_i, the concept h_i classifies all $n-1$ *other* points 1 while $h_i(x_i) = 0$. We will argue that $\{x_1, \ldots, x_n\}$ is shattered by $\bar{\mathbb{C}}$. For any $I \subseteq \{1, \ldots, n\}$, let $h_I = \prod_{j \in \{0,\ldots,n\} \setminus I} h_j \in \bar{\mathbb{C}}$. For each $i \in I$, since every $j \in \{0, \ldots, n\} \setminus I$ has $h_j(x_i) = 1$, we have $h_I(x_i) = 1$. For each $i \in \{1, \ldots, n\} \setminus I$ (if the set is non-empty), since $h_i(x_i) = 0$, we also have $h_I(x_i) = 0$. We therefore have that $\{h_I : I \subseteq \{1, \ldots, n\}\}$ shatters $\{x_1, \ldots, x_n\}$, so that $\bar{\mathbb{C}}$ does as well. Since such a star set $\{x_1, \ldots, x_n\}$ exists for every finite $n \leq \mathfrak{s}_{\mathbf{1}}$, we conclude that $\mathrm{VC}(\bar{\mathbb{C}}) \geq \mathfrak{s}_{\mathbf{1}}$.

Next we complement this with the opposite inequality. Let $\{x_1, \ldots, x_n\}$ be any set shattered by $\bar{\mathbb{C}}$. In particular, since all 2^n classifications are realizable by $\bar{\mathbb{C}}$ and every $h \in \bar{\mathbb{C}}$ can be expressed as a function $h_{\mathbb{C}'}$ for some finite non-empty set $\mathbb{C}' \subseteq \mathbb{C}$, there must exist finite non-empty sets $\mathbb{C}'_0, \mathbb{C}'_1, \ldots, \mathbb{C}'_n \subseteq \mathbb{C}$ such that $\forall i \in \{0, \ldots, n\}$, $\forall j \in \{1, \ldots, n\}$, $h_{\mathbb{C}'_i}(x_j) = 1$ iff $i \neq j$. Letting h_0 be any concept in $\mathbb{C}'_0$, every $j \in \{1, \ldots, n\}$ has $h_0(x_j) \geq h_{\mathbb{C}'_0}(x_j) = 1$, and therefore h_0 classifies all n points as 1. Moreover, for each $i \in \{1, \ldots, n\}$, since $h_{\mathbb{C}'_i}(x_i) = 0$, there must exist some $h_i \in \mathbb{C}'_i$ with $h_i(x_i) = 0$. Furthermore, every $j \in \{1, \ldots, n\} \setminus \{i\}$ then has $h_i(x_j) \geq h_{\mathbb{C}'_i}(x_j) = 1$, so that $h_i(x_j) = 1$. We have thus found a sequence $h_0, \ldots, h_n \in \mathbb{C}$ such that $\forall i \in \{0, \ldots, n\}$, $\forall j \in \{1, \ldots, n\}$, $h_i(x_j) = 1$ iff $i \neq j$: that is, we have established that $\{x_1, \ldots, x_n\}$ is a star set (for $\mathbb{C}$) centered at $\mathbf{1}$, so that $\mathfrak{s}_{\mathbf{1}} \geq n$. Since there exists such a set $\{x_1, \ldots, x_n\}$ shattered by $\bar{\mathbb{C}}$ for every finite $n \leq \mathrm{VC}(\bar{\mathbb{C}})$, we conclude that $\mathfrak{s}_{\mathbf{1}} \geq \mathrm{VC}(\bar{\mathbb{C}})$. Combining these two inequalities yields that $\mathrm{VC}(\bar{\mathbb{C}}) = \mathfrak{s}_{\mathbf{1}}$. ■

Appendix I. Proof of Theorem 23 (A Compression Scheme of Size Minimum Star Number)

Consider any $n \in \mathbb{N} \cup \{0\}$ and any data set $S = \{(x_1, y_1), \ldots, (x_n, y_n)\} \in (\mathcal{X} \times \mathcal{Y})^n$ realizable by $\mathbb{C}$. In particular, the latter implies $\mathbb{C}_S \neq \emptyset$, so that $\exists h_S \in \mathbb{C}$ with $h_S(x_i) = y_i$ for every $i \leq n$. By definition of the compression set $S' = \kappa(S)$, we have that $y_i = h_*(x_i)$ for every $x_i \in S'$, so that $h_S \in \mathbb{C}_{(S', h_*(S'))}$ as well. In particular, this means that for every $x_i \notin \mathrm{DIS}(\mathbb{C}_{(S', h_*(S'))})$, $\rho(S', x_i) = h_S(x_i) = y_i$. Moreover, by definition of S', every x_i with $y_i = h_*(x_i)$ satisfies $x_i \notin \mathrm{DIS}(\mathbb{C}_{(S', h_*(S'))})$. Therefore, for every $x_i \in \mathrm{DIS}(\mathbb{C}_{(S', h_*(S'))})$, we have $y_i = 1 - h_*(x_i)$, so that $\rho(S', x_i) = y_i$ in this case as well. Altogether, we have established that (κ, ρ) is sample-consistent.

Next we bound the *size* of the compression set. Let $S_* = \{x : (x, h_*(x)) \in S\}$. By definition, $S' = \kappa(S)$ is a minimal subset of S_* for which $S_* \cap \mathrm{DIS}(\mathbb{C}_{(S', h_*(S'))}) = \emptyset$. Since the agreed-upon label of any $x \in S_*$ by every $h \in \mathbb{C}_{(S', h_*(S'))}$ must be $h_*(x)$ (since again, $h_S \in \mathbb{C}_{(S', h_*(S'))}$, and $h_S(x_i) = y_i = h_*(x_i)$ for every $x_i \in S_*$), this means S' may be described *equivalently* as a minimal $S' \subseteq S_*$ such that $\mathbb{C}_{(S', h_*(S'))} = \mathbb{C}_{(S_*, h_*(S_*))}$. Such a set S' is known as a minimal *version space compression set* for $(S_*, h_*(S_*))$ (also known as a minimal *empirical teaching set*) (Hanneke, 2007a, 2014; Wiener, Hanneke, and El-Yaniv, 2015; El-Yaniv and Wiener, 2010, 2012; Hanneke and Yang, 2015; Hanneke, 2016). It was shown by Hanneke and Yang (2015, Lemma 14) that a minimal version space compression set for a realizable data set $(S_*, h_*(S_*))$ is necessarily a *star set* centered at h_*. Indeed, this is clear from *minimality* of S', since being a star set centered at h_* can be

expressed concisely by the condition that, $\forall x \in S'$, $x \in \mathrm{DIS}(\mathbb{C}_{(S'\setminus\{x\},h_*(S'\setminus\{x\}))})$; if this condition were not satisfied by S', then some $x \in S'$ has $x \notin \mathrm{DIS}(\mathbb{C}_{(S'\setminus\{x\},h_*(S'\setminus\{x\}))})$, which (since $h_S \in \mathbb{C}_{(S'\setminus\{x\},h_*(S'\setminus\{x\}))}$) implies the label agreed-upon for x by all $h \in \mathbb{C}_{(S'\setminus\{x\},h_*(S'\setminus\{x\}))}$ must be $h_S(x) = h_*(x)$ (since $x \in S_*$), and therefore $\mathbb{C}_{(S'\setminus\{x\},h_*(S'\setminus\{x\}))} = \mathbb{C}_{(S',h_*(S'))} = \mathbb{C}_{(S_*,h_*(S_*))}$ contradicting minimality of $|S'|$ among subsets of S_* satisfying $\mathbb{C}_{(S',h_*(S'))} = \mathbb{C}_{(S_*,h_*(S_*))}$. Thus, since S' is a star set centered at h_*, we have $|S'| \leq \mathfrak{s}_{h_*} = \mathfrak{s}_{\min}$.

It remains only to argue that (κ, ρ) is also *stable*. Consider any subsequence $S_\sigma \subseteq S$ such that the set $S' = \kappa(S)$ satisfies $S' \subseteq \{x_i : (x_i, y_i) \in S_\sigma\}$. In this case, $\kappa(S_\sigma)$ is some set $S'' \subseteq S_* \cap \{x_i : (x_i, y_i) \in S_\sigma\}$ such that

$$\{x : (x, h_*(x)) \in S_\sigma\} \cap \mathrm{DIS}(\mathbb{C}_{(S'',h_*(S''))}) = \emptyset.$$

Since we still have $h_S \in \mathbb{C}_{(S'',h_*(S''))}$ (since $S'' \subseteq S_*$), it must be that for every x_i with $(x_i, h_*(x_i)) \in S_\sigma$, we have that every $h \in \mathbb{C}_{(S'',h_*(S''))}$ satisfies $h(x_i) = h_S(x_i) = h_*(x_i)$. In particular, since $S' \subseteq \{x : (x, h_*(x)) \in S_\sigma\}$, we have that every $h \in \mathbb{C}_{(S'',h_*(S''))}$ satisfies $h(x) = h_*(x)$ for all $x \in S'$. We therefore have that $\mathbb{C}_{(S'',h_*(S''))} = \mathbb{C}_{(S''\cup S',h_*(S''\cup S'))} = \mathbb{C}_{(S_*,h_*(S_*))} = \mathbb{C}_{(S',h_*(S'))}$. It then follows immediately from the definition that $\rho(\kappa(S_\sigma)) = \rho(S'') = \rho(S') = \rho(\kappa(S))$. Hence (κ, ρ) is *stable*. This completes the proof. ■

Appendix J. Proof: The Eluder Dimension is the Threshold Dimension of Version Spaces and Disagreements

Fix any concept class $\mathbb{C}$ and concept h. Consider an eluder sequence $\{(x_1, h(x_1)), \ldots, (x_n, h(x_n))\}$ (for $\mathbb{C}$) centered at h, and let $h_1, \ldots, h_n$ be as in the definition: that is, $\forall i \in \{1, \ldots, n\}$, $h_i(x_i) \neq h(x_i)$ and $\forall j < i$, $h_i(x_j) = h(x_j)$. We will argue that $\{h_n, \ldots, h_1\}$ witness a *threshold* set for $\mathcal{V}_h(\mathbb{C})$. Specifically, for each $i \in \{0, 1, \ldots, n\}$, define $S_i = \{(x_j, h(x_j))\}_{j \leq i}$. Since each h_j has $h_j(x_j) \neq h(x_j)$, but every $i > j$ has $h_i(x_j) = h(x_j)$, we have that, for each $i \in \{0, \ldots, n\}$, $\mathbb{C}_{S_i} \cap \{h_1, \ldots, h_n\} = \{h_{i+1}, \ldots, h_n\}$. In particular, we have $\mathbb{C}_{S_0} \cap \{h_1, \ldots, h_n\} = \mathbb{C} \cap \{h_1, \ldots, h_n\} = \{h_1, \ldots, h_n\}$, and $\mathbb{C}_{S_n} \cap \{h_1, \ldots, h_n\} = \{\}$, and every $i \in \{1, \ldots, n-1\}$ has $\mathbb{C}_{S_i} \cap \{h_1, \ldots, h_n\}$ equal to the set of the last $n - i$ concepts: $\{h_{i+1}, \ldots, h_n\}$. Thus, the sequence $\{h_n, \ldots, h_1\}$ is a threshold set for $\mathcal{V}_h(\mathbb{C})$, so that $\mathbb{T}(\mathcal{V}_h(\mathbb{C})) \geq n$. Since such an eluder sequence (for $\mathbb{C}$) centered at h exists for every finite $n \leq \mathfrak{e}_h$, we conclude that $\mathbb{T}(\mathcal{V}_h(\mathbb{C})) \geq \mathfrak{e}_h$.

In the other direction, consider any sequence $\{h_n, \ldots, h_1\}$ in $\mathbb{C}$ which is a threshold set for $\mathcal{V}_h(\mathbb{C})$. By definition, for each $i \in \{0, \ldots, n\}$, there exists a data set S_i consistent with h (i.e., a finite sequence of pairs $(x, h(x))$) such that $\mathbb{C}_{S_i} \cap \{h_1, \ldots, h_n\} = \{h_{i+1}, \ldots, h_n\}$. In particular, for $i \in \{1, \ldots, n\}$, there must exist at least one $(x_i, h(x_i))$ in S_i with $h_i(x_i) \neq h(x_i)$. Moreover, for every $i, j \in \{1, \ldots, n\}$ with $j < i$, since $h_i \in \{h_{j+1}, \ldots, h_n\}$, we have that $h_i \in \mathbb{C}_{S_j}$, so that $h_i(x_j) = h(x_j)$. We have thus found a sequence $x_1, \ldots, x_n$ such that, $\forall i \in \{1, \ldots, n\}$, $h_i(x_i) \neq h(x_i)$ and $\forall j < i$, $h_i(x_j) = h(x_j)$: that is, $\{(x_1, h(x_1)), \ldots, (x_n, h(x_n))\}$ is an eluder sequence (for $\mathbb{C}$) centered at h. We therefore have $\mathfrak{e}_h \geq n$. Since there exists such a threshold set $\{h_n, \ldots, h_1\}$ for $\mathcal{V}_h(\mathbb{C})$ for every finite $n \leq \mathbb{T}(\mathcal{V}_h(\mathbb{C}))$, we conclude that $\mathfrak{e}_h \geq \mathbb{T}(\mathcal{V}_h(\mathbb{C}))$. Together with the fact $\mathbb{T}(\mathcal{V}_h(\mathbb{C})) \geq \mathfrak{e}_h$ established above, we have that $\mathbb{T}(\mathcal{V}_h(\mathbb{C})) = \mathfrak{e}_h$.

Next we turn to disagreement sets. In this case, we fix any $h \in \mathbb{C}$. As above, consider an eluder sequence $\{(x_1, h(x_1)), \ldots, (x_n, h(x_n))\}$ (for $\mathbb{C}$) centered at h, and let $h_1, \ldots, h_n$ be as in the definition: that is, $\forall i \in \{1, \ldots, n\}$, $h_i(x_i) \neq h(x_i)$ and $\forall j < i$, $h_i(x_j) = h(x_j)$. We will argue that $\{x_n, \ldots, x_1\}$ is a threshold set for $\mathcal{D}_h(\mathbb{C})$. For each $i \in \{0, \ldots, n\}$, let $S_i = \{(x_j, h(x_j))\}_{j \leq i}$.

By definition, every $h' \in \mathbb{C}_{S_i}$ has $h'(x_j) = h(x_j)$ for all $j \leq i$, so that $\text{DIS}(\mathbb{C}_{S_i}) \cap \{x_j : j \leq i\} = \emptyset$. Moreover, for every $j \in \{i+1, \ldots, n\}$, since $h_j(x_{j'}) = h(x_{j'})$ for all $j' < j$, we have $h_j \in \mathbb{C}_{S_i}$. Since we also always have $h \in \mathbb{C}_{S_i}$ (since $h \in \mathbb{C}$ and S_i is consistent with h), and $h_j(x_j) \neq h(x_j)$, we conclude that every $j \in \{i+1, \ldots, n\}$ satisfies $x_j \in \text{DIS}(\mathbb{C}_{S_i})$. Altogether, we have $\text{DIS}(\mathbb{C}_{S_i}) \cap \{x_1, \ldots, x_n\} = \{x_{i+1}, \ldots, x_n\}$. Thus, $\{x_n, \ldots, x_1\}$ is a threshold set for $\mathcal{D}_h(\mathbb{C})$, so that $\mathbb{T}(\mathcal{D}_h(\mathbb{C})) \geq n$. Since such an eluder sequence (for $\mathbb{C}$) centered at h exists for every finite $n \leq \mathfrak{e}_h$, we conclude that $\mathbb{T}(\mathcal{D}_h(\mathbb{C})) \geq \mathfrak{e}_h$.

Turning to the other direction, let $\{x_n, \ldots, x_1\}$ be a threshold set for $\mathcal{D}_h(\mathbb{C})$. By definition, there exist data sets $S_0, \ldots, S_n$ consistent with h (i.e., finite sequences of $(x, h(x))$ pairs) such that $\forall i \in \{0, \ldots, n\}$, $\text{DIS}(\mathbb{C}_{S_i}) \cap \{x_1, \ldots, x_n\} = \{x_{i+1}, \ldots, x_n\}$. We will argue that $\{(x_1, h(x_1)), \ldots, (x_n, h(x_n))\}$ is an eluder sequence (for $\mathbb{C}$) centered at h. For each $i \in \{1, \ldots, n\}$, since $x_i \in \text{DIS}(\mathbb{C}_{S_{i-1}})$, there must exist at least one $h_i \in \mathbb{C}_{S_{i-1}}$ with $h_i(x_i) \neq h(x_i)$. Moreover, since $\text{DIS}(\mathbb{C}_{S_{i-1}}) \cap \{x_j : j \leq i-1\} = \emptyset$, and $h \in \mathbb{C}_{S_{i-1}}$ (since $h \in \mathbb{C}$ and S_{i-1} is consistent with h), it must be that every $h' \in \mathbb{C}_{S_{i-1}}$ has $h'(x_j) = h(x_j)$ for every $j \leq i-1$. In particular, this means $h_i(x_j) = h(x_j)$ for all $j < i$. Altogether, we have found a sequence $h_1, \ldots, h_n \in \mathbb{C}$ such that, $\forall i \in \{1, \ldots, n\}$, $h_i(x_i) \neq h(x_i)$ and $\forall j < i$, $h_i(x_j) = h(x_j)$: that is, $\{x_1, \ldots, x_n\}$ is an eluder sequence (for $\mathbb{C}$) centered at h. We therefore have $\mathfrak{e}_h \geq n$. Since there exists such a threshold set $\{x_n, \ldots, x_1\}$ for $\mathcal{D}_h(\mathbb{C})$ for every finite $n \leq \mathbb{T}(\mathcal{D}_h(\mathbb{C}))$, we conclude that $\mathfrak{e}_h \geq \mathbb{T}(\mathcal{D}_h(\mathbb{C}))$. Together with the fact that $\mathbb{T}(\mathcal{D}_h(\mathbb{C})) \geq \mathfrak{e}_h$ established above, we conclude that $\mathbb{T}(\mathcal{D}_h(\mathbb{C})) = \mathfrak{e}_h$. ■

Appendix K. Proof of Theorem 11 (Eluder Dimension Not Smaller Than Log Cardinality)

The fact that $\mathfrak{e} \leq |\mathbb{C}| - 1$ follows immediately from the definition, since each h_i, $i \in \{1, \ldots, \mathfrak{e}\}$ must be distinct, and distinct from the center $h \in \mathbb{C}$. The remainder of the proof focuses on the lower bound on $\mathfrak{e}$.

For simplicity, we separate the case $|\mathbb{C}| < \infty$ from $|\mathbb{C}| = \infty$. We begin with the case $|\mathbb{C}| < \infty$. Let $N = \lceil \log_{|\mathcal{Y}|}(|\mathbb{C}|) \rceil$. We construct an eluder sequence $(x_1, y_1), \ldots, (x_N, y_N)$ inductively, to satisfy that, for any $n \leq \log_{|\mathcal{Y}|}(|\mathbb{C}|)$, the prefix $S_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ satisfies $|\mathbb{C}_{S_n}| \geq |\mathcal{Y}|^{-n}|\mathbb{C}|$. As a base case, $S_0 = \{\}$ is trivially an eluder sequence of length 0 satisfying $|\mathbb{C}_{S_0}| = |\mathbb{C}|$. This also trivially completes the proof for the case $|\mathbb{C}| = 1$; for the remainder, we suppose $|\mathbb{C}| \geq 2$. For $n \leq N$, suppose (for the purpose of induction) we have constructed an eluder sequence $S_{n-1} = \{(x_1, y_1) \ldots, (x_{n-1}, y_{n-1})\}$ such that $|\mathbb{C}|_{S_{n-1}} \geq |\mathcal{Y}|^{1-n}|\mathbb{C}|$. Since $|\mathcal{Y}|^{1-n}|\mathbb{C}| > 1$ (due to $n \leq N = \lceil \log_{|\mathcal{Y}|}(|\mathbb{C}|) \rceil$), we have that $\text{DIS}(\mathbb{C}_{S_{n-1}}) \neq \emptyset$. Let x_n be any element of $\text{DIS}(\mathbb{C}_{S_{n-1}})$. Since $\mathbb{C}_{S_{n-1}} = \bigcup_{y \in \mathcal{Y}} \mathbb{C}_{S_{n-1} \cup \{(x_n, y)\}}$, the pigeonhole principle implies there exists $y_n \in \mathcal{Y}$ with $|\mathbb{C}_{S_{n-1} \cup \{(x_n, y_n)\}}| \geq \frac{1}{|\mathcal{Y}|}|\mathbb{C}_{S_{n-1}}| \geq |\mathcal{Y}|^{-n}|\mathbb{C}|$. Defining $S_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, and recalling that $x_n \in \text{DIS}(\mathbb{C}_{S_{n-1}})$ by definition, this remains an eluder sequence, and satisfies $|\mathbb{C}_{S_n}| \geq |\mathcal{Y}|^{-n}|\mathbb{C}|$ by the choice of y_n. The existence of the eluder sequence $(x_1, y_1), \ldots, (x_N, y_N)$ follows by the principle of induction.

The case $|\mathbb{C}| = \infty$ is constructed similarly. For this, we only require that for every $x \in \mathcal{X}$, the set of possible labels $\mathcal{Y}_x := \{h(x) : h \in \mathbb{C}\}$ is finite (not necessarily of uniformly bounded size). We construct an infinite eluder sequence $(x_1, y_1), (x_2, y_2), \ldots$ inductively, to satisfy that, for any $n \in \mathbb{N} \cup \{0\}$, the prefix $S_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ satisfies $|\mathbb{C}_{S_n}| = \infty$. Again, the base case $S_0 = \{\}$ is trivially an eluder sequence of length 0 satisfying $|\mathbb{C}_{S_0}| = |\mathbb{C}| =$

∞. For $n \in \mathbb{N}$, suppose (for the purpose of induction) we have constructed an eluder sequence $S_{n-1} = \{(x_1, y_1) \ldots, (x_{n-1}, y_{n-1})\}$ such that $|\mathbb{C}|_{S_{n-1}} = \infty$. In particular, this implies we have $\mathrm{DIS}(\mathbb{C}_{S_{n-1}}) \neq \emptyset$. Let x_n be any element of $\mathrm{DIS}(\mathbb{C}_{S_{n-1}})$. Since $\mathbb{C}_{S_{n-1}} = \bigcup_{y \in \mathcal{Y}_{x_n}} \mathbb{C}_{S_{n-1} \cup \{(x_n, y)\}}$, and we have assumed $\mathcal{Y}_{x_n}$ is finite, the pigeonhole principle implies there exists $y_n \in \mathcal{Y}_{x_n}$ with $|\mathbb{C}_{S_{n-1} \cup \{(x_n, y_n)\}}| = \infty$. Defining $S_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, and recalling that $x_n \in \mathrm{DIS}(\mathbb{C}_{S_{n-1}})$ by definition, this sequence S_n remains an eluder sequence, and satisfies $|\mathbb{C}_{S_n}| = \infty$ by the choice of y_n. The existence of the infinite eluder sequence $(x_1, y_1), (x_2, y_2), \ldots$ follows by the principle of induction. ■

Appendix L. Proof of Theorem 31 (The Eluder Dimension of Version Spaces and Disagreement Regions)

One can show $\mathfrak{e}_{\mathbb{C}}(\mathcal{V}_h(\mathbb{C})) \geq \mathfrak{e}_h(\mathbb{C})$ by a simple direct argument constructing an eluder sequence for $\mathcal{V}_h(\mathbb{C})$ centered at $\mathbb{C}$ based on any eluder sequence for $\mathbb{C}$ centered at h. However, we can argue this even more simply via Theorem 10 as follows. consider any concept h and any threshold set $h_1, \ldots, h_n \in \mathbb{C}$ for $\mathcal{V}_h(\mathbb{C})$: that is, $\exists V_0, \ldots, V_n \in \mathcal{V}_h(\mathbb{C})$ such that, $\forall t \in \{0, 1, \ldots, n\}$, $V_t \cap \{h_1, \ldots, h_n\} = \{h_1, \ldots, h_t\}$. It follows immediately from this that $\{(h_1, 1), \ldots, (h_n, 1)\}$ is an eluder sequence for $\{\mathbb{1}_V : V \in \mathcal{V}_h(\mathbb{C})\}$ centered at $\mathbf{1}$, so that $\mathfrak{e}_{\mathbb{C}}(\mathcal{V}_h(\mathbb{C})) \geq n$. Since such a threshold set exists for every finite $n \leq \mathbb{T}(\mathcal{V}_h(\mathbb{C}))$, we have that $\mathfrak{e}_{\mathbb{C}}(\mathcal{V}_h(\mathbb{C})) \geq \mathbb{T}(\mathcal{V}_h(\mathbb{C}))$. Moreover, Theorem 10 implies $\mathbb{T}(\mathcal{V}_h(\mathbb{C})) = \mathfrak{e}_h(\mathbb{C})$, and therefore we also have $\mathfrak{e}_{\mathbb{C}}(\mathcal{V}_h(\mathbb{C})) \geq \mathfrak{e}_h(\mathbb{C})$.

For the other direction, consider any eluder sequence $(h_1, 1), \ldots, (h_n, 1) \in \mathbb{C} \times \{1\}$ for $\{\mathbb{1}_V : V \in \mathcal{V}_h(\mathbb{C})\}$ centered at $\mathbf{1}$. By definition, there exist $V_1, \ldots, V_n \in \mathcal{V}_h(\mathbb{C})$ such that $\forall i \in \{1, \ldots, n\}$, $V_i \cap \{h_1, \ldots, h_i\} = \{h_1, \ldots, h_{i-1}\}$. By definition of $\mathcal{V}_h(\mathbb{C})$, for each $i \in \{1, \ldots, n\}$, there exists a finite sequence S_i in $\mathcal{X}$ such that $V_i = \mathbb{C}_{(S_i, h(S_i))}$. In particular, since $h_i \in \mathbb{C}$ and $\mathbb{C}_{(S_i, h(S_i))} \cap \{h_1, \ldots, h_i\} = V_i \cap \{h_1, \ldots, h_i\} = \{h_1, \ldots, h_{i-1}\}$, there exists at least one point $x_i \in S_i$ such that $h_i(x_i) \neq h(x_i)$, and since $\{h_1, \ldots, h_{i-1}\} \subseteq \mathbb{C}_{(S_i, h(S_i))} \subseteq \mathbb{C}_{\{(x_i, h(x_i))\}}$, every $j < i$ has $h_j(x_i) = h(x_i)$. Thus, $(x_n, h(x_n)), \ldots, (x_1, h(x_1))$ is an eluder sequence for $\mathbb{C}$ centered at h: that is, denoting by $x'_i = x_{n-i+1}$ and $h'_i = h_{n-i+1}$, we have $\forall i \in \{1, \ldots, n\}$, $h'_i(x'_i) = h_{n-i+1}(x_{n-i+1}) \neq h(x_{n-i+1}) = h(x'_i)$ and $\forall j < i$, since $n - j + 1 > n - i + 1$, $h'_i(x'_j) = h_{n-i+1}(x_{n-j+1}) = h(x_{n-j+1}) = h(x'_j)$. Thus, $\mathfrak{e}_h(\mathbb{C}) \geq n$. Since such an eluder sequence $(h_1, 1), \ldots, (h_n, 1) \in \mathbb{C} \times \{1\}$ for $\{\mathbb{1}_V : V \in \mathcal{V}_h(\mathbb{C})\}$ centered at $\mathbf{1}$ exists for every finite $n \leq \mathfrak{e}_{\mathbb{C}}(\mathcal{V}_h(\mathbb{C}))$, we have that $\mathfrak{e}_h(\mathbb{C}) \geq \mathfrak{e}_{\mathbb{C}}(\mathcal{V}_h(\mathbb{C}))$. Combining these two parts, we conclude that $\mathfrak{e}_{\mathbb{C}}(\mathcal{V}_h(\mathbb{C})) = \mathfrak{e}_h(\mathbb{C})$.

Turning to $\mathcal{D}_h(\mathbb{C})$, consider any concept $h \in \mathbb{C}$ and any *threshold set* $x_1, \ldots, x_n \in \mathcal{X}$ for $\mathcal{D}_h(\mathbb{C})$: that is, $\exists D_0, \ldots, D_n \in \mathcal{D}_h(\mathbb{C})$ such that, $\forall t \in \{0, 1, \ldots, n\}$, $D_t \cap \{x_1, \ldots, x_n\} = \{x_1, \ldots, x_t\}$. It immediately follows from this that $\{(x_n, 0), (x_{n-1}, 0), \ldots, (x_1, 0)\}$ is an eluder sequence for $\{\mathbb{1}_D : D \in \mathcal{D}_h(\mathbb{C})\}$ centered at $\mathbf{0}$, so that $\mathfrak{e}_{\emptyset}(\mathcal{D}_h(\mathbb{C})) \geq n$. Since such a threshold set exists for every finite $n \leq \mathbb{T}(\mathcal{D}_h(\mathbb{C}))$, we have $\mathfrak{e}_{\emptyset}(\mathcal{D}_h(\mathbb{C})) \geq \mathbb{T}(\mathcal{D}_h(\mathbb{C}))$. Since Theorem 10 implies $\mathbb{T}(\mathcal{D}_h(\mathbb{C})) = \mathfrak{e}_h(\mathbb{C})$, we have that $\mathfrak{e}_{\emptyset}(\mathcal{D}_h(\mathbb{C})) \geq \mathfrak{e}_h(\mathbb{C})$.

In the other direction, consider any eluder sequence $(x_1, 0), \ldots, (x_n, 0)$ for $\{\mathbb{1}_D : D \in \mathcal{D}_h(\mathbb{C})\}$ centered at $\mathbf{0}$. By definition, there exist $D_1, \ldots, D_n \in \mathcal{D}_h(\mathbb{C})$ such that $\forall i \in \{1, \ldots, n\}$, $D_i \cap \{x_1, \ldots, x_i\} = \{x_i\}$. By definition of $\mathcal{D}_h(\mathbb{C})$, for each $i \in \{1, \ldots, n\}$, there exists a finite sequence S_i in $\mathcal{X}$ such that $D_i = \mathrm{DIS}(\mathbb{C}_{(S_i, h(S_i))})$. In particular, since each $i \in \{1, \ldots, n\}$ has $\mathrm{DIS}(\mathbb{C}_{(S_i, h(S_i))}) \cap \{x_1, \ldots, x_i\} = D_i \cap \{x_1, \ldots, x_i\} = \{x_i\}$, there exists $h_i \in \mathbb{C}_{(S_i, h(S_i))}$ with $h_i(x_i) \neq h(x_i)$, and moreover, since $h \in \mathbb{C}_{(S_i, h(S_i))}$ and $\mathrm{DIS}(\mathbb{C}_{(S_i, h(S_i))}) \cap \{x_1, \ldots, x_{i-1}\} = \emptyset$, it

must be that $\forall j < i$, $h_i(x_j) = h(x_j)$. Thus, $h_1, \ldots, h_n$ witness that $(x_1, h(x_1)), \ldots, (x_n, h(x_n))$ is an eluder sequence for $\mathbb{C}$ centered at h, so that $\mathfrak{e}_h(\mathbb{C}) \geq n$. Since such an eluder sequence $(x_1, 0), \ldots, (x_n, 0)$ for $\{\mathbb{1}_D : D \in \mathcal{D}_h(\mathbb{C})\}$ centered at $\mathbf{0}$ exists for every finite $n \leq \mathfrak{e}_\emptyset(\mathcal{D}_h(\mathbb{C}))$, we have that $\mathfrak{e}_h(\mathbb{C}) \geq \mathfrak{e}_\emptyset(\mathcal{D}_h(\mathbb{C}))$. Combining these two parts, we conclude that $\mathfrak{e}_\emptyset(\mathcal{D}_h(\mathbb{C})) = \mathfrak{e}_h(\mathbb{C})$. ■

Appendix M. Proofs of Results on the Scale-Sensitive Eluder Dimension

This section presents the proof of Theorem 14. The proof is similar in spirit to the simple proof of Theorem 11 (which relates $\mathfrak{e}(\mathbb{C})$ to $|\mathbb{C}|$), but is made significantly more complicated by the fact that the $h_i(x_j)$ values need only approximate the values y_j for $j < i$, rather than matching them exactly.

Infinite Eluder Sequence: We begin with the necessity direction: that is, $\mathcal{N}(\varepsilon, \mathbb{C}, \mathrm{L}_\infty) = \infty \implies \mathfrak{e}(\varepsilon) = \infty$. Toward this end, suppose $\mathbb{C}$ and $\varepsilon > 0$ are such that $\mathcal{N}(\varepsilon, \mathbb{C}, \mathrm{L}_\infty) = \infty$. We construct an infinite ε-eluder sequence by induction. Let $t \in \mathbb{N}$, and for the purpose of induction (with $t = 1$ as a base case where this holds trivially) suppose we have already constructed an ε-eluder sequence $(x_1, y_1), \ldots, (x_{t-1}, y_{t-1})$ in $\mathcal{X} \times [0, 1]$ with the further property that $\forall \varepsilon' > 0$, the set

$$\mathbb{C}_{t-1,\varepsilon'} := \left\{ h \in \mathbb{C} : \max_{1 \leq i \leq t-1} |h(x_i) - y_i| < \varepsilon' \right\}$$

has

$$\mathcal{N}(\varepsilon, \mathbb{C}_{t-1,\varepsilon'}, \mathrm{L}_\infty) = \infty. \tag{10}$$

Let $\varepsilon_{t-1} = \frac{\varepsilon}{\sqrt{t-1}}$.

Claim 1: There exists $x_t \in \mathcal{X}$ such that $\exists h_1, h_2 \in \mathbb{C}_{t-1,\varepsilon_{t-1}}$ with $|h_1(x_t) - h_2(x_t)| > 2\varepsilon$.

Such a choice of x_t must exist, since otherwise every $x \in \mathcal{X}$ has

$$\sup_{h \in \mathbb{C}_{t-1,\varepsilon_{t-1}}} h(x) - \inf_{h \in \mathbb{C}_{t-1,\varepsilon_{t-1}}} h(x) \leq 2\varepsilon,$$

in which case the function

$$x \mapsto \bar{h}(x) = \frac{1}{2} \left(\sup_{h \in \mathbb{C}_{t-1,\varepsilon_{t-1}}} h(x) + \inf_{h \in \mathbb{C}_{t-1,\varepsilon_{t-1}}} h(x) \right),$$

which takes the midpoint between the sup and inf values for each x, would satisfy

$$\sup_{h \in \mathbb{C}_{t-1,\varepsilon_{t-1}}} \left| \bar{h}(x) - h(x) \right| \leq \varepsilon,$$

meaning $\mathcal{N}(\varepsilon, \mathbb{C}_{t-1,\varepsilon_{t-1}}, \mathrm{L}_\infty) = 1$, contradicting (10) for $\varepsilon' = \varepsilon_{t-1}$. Thus, we have established Claim 1. We will take this choice of x_t to extend the sequence $x_1, \ldots, x_{t-1}$.

It remains to specify the value y_t to extend the inductive hypothesis. First note that, due to Claim 1, for *any* choice of $y_t \in [0, 1]$, the sequence $(x_1, y_1), \ldots, (x_t, y_t)$ will be an ε-eluder sequence. To see this, recall that $(x_1, y_1), \ldots, (x_{t-1}, y_{t-1})$ is an ε-eluder sequence (by the inductive hypothesis). Moreover, by the definition of x_1, h_1, and h_2 in Claim 1, both h_1 and h_2 are in $\mathbb{C}_{t-1,\varepsilon_{t-1}}$, which implies both $r \in \{1, 2\}$ satisfy

$$\sum_{i<t} (h_r(x_i) - y_i)^2 \leq (t-1) \max_{i<t} |h_r(x_i) - y_i|^2 < (t-1)\varepsilon_{t-1}^2 = \varepsilon^2.$$

Finally, since $|h_1(x_t) - h_2(x_t)| > 2\varepsilon$, by the triangle inequality, for any $y_t \in [0,1]$, at least one $b \in \{1,2\}$ satisfies $|h_b(x_t) - y_t| > \varepsilon$. Thus, we have established that *any* choice of y_t suffices to establish that $(x_1, y_1), \ldots, (x_t, y_t)$ is an ε-eluder sequence. All that remains is to specify a value y_t which also extends the additional property (10).

Claim 2: For each $n \in \mathbb{N}$, there exists $\tilde{y}^{(n)} \in \left\{\frac{k}{2n} : k \in \{1, \ldots, 2n-1\}\right\}$ such that

$$\mathcal{N}\left(\varepsilon, \left\{h \in \mathbb{C}_{t-1,1/n} : \left|h(x_t) - \tilde{y}^{(n)}\right| < \frac{1}{n}\right\}, \mathrm{L}_\infty\right) = \infty. \tag{11}$$

Claim 2 essentially follows from the Pigeonhole principle. To see this, note that if Claim 2 were not true, then since the intervals $\left(\frac{k}{2n} - \frac{1}{n}, \frac{k}{2n} + \frac{1}{n}\right)$, $k \in \{1, \ldots, 2n-1\}$, cover the interval $[0,1]$, the sets $\left\{h \in \mathbb{C}_{t-1,1/n} : \left|h(x_t) - \frac{k}{2n}\right| < \frac{1}{n}\right\}$, $k \in \{1, \ldots, 2n-1\}$ cover the set $\mathbb{C}_{t-1,1/n}$, so that if each of these $2n-1$ sets had a finite ε-cover (under L_∞), we could construct a finite ε-cover of $\mathbb{C}_{t-1,1/n}$ (under L_∞) by a union of these $2n-1$ finite covers. Thus, such a $\tilde{y}^{(n)}$ satisfying (11) must exist. Thus, we have established Claim 2.

Since $[0,1]$ is compact, there exists an increasing sequence n_j in $\mathbb{N}$ and a value $y_t \in [0,1]$ such that $\lim_{j\to\infty} \tilde{y}^{(n_j)} = y_t$. This will be our choice of y_t to extend the ε-eluder sequence.

It remains to argue that this choice of (x_t, y_t) indeed extends the inductive hypothesis. For any $\varepsilon' > 0$, let

$$\mathbb{C}_{t,\varepsilon'} = \left\{h \in \mathbb{C} : \max_{1 \le i \le t} |h(x_i) - y_i| < \varepsilon'\right\}.$$

By the definition of n_j and y_t, $\exists j \in \mathbb{N}$ such that $\frac{1}{n_j} < \frac{\varepsilon'}{2}$ and $\left|\tilde{y}^{(n_j)} - y_t\right| < \frac{\varepsilon'}{2}$. In particular, any h with $\left|h(x_t) - \tilde{y}^{(n_j)}\right| < \frac{1}{n_j}$ has $\left|h(x_t) - \tilde{y}^{(n_j)}\right| < \frac{\varepsilon'}{2}$, which therefore also satisfies $|h(x_t) - y_t| < \varepsilon'$ by the triangle inequality. Together with monotonicity of $\mathbb{C}_{t-1,\varepsilon'}$ in ε', this implies

$$\begin{aligned}
\mathbb{C}_{t,\varepsilon'} &= \left\{h \in \mathbb{C}_{t-1,\varepsilon'} : |h(x_t) - y_t| < \varepsilon'\right\} \\
&\supseteq \left\{h \in \mathbb{C}_{t-1,1/n_j} : |h(x_t) - y_t| < \varepsilon'\right\} \\
&\supseteq \left\{h \in \mathbb{C}_{t-1,1/n_j} : \left|h(x_t) - \tilde{y}^{(n_j)}\right| < \frac{1}{n_j}\right\}.
\end{aligned}$$

Since (11) implies this last set has infinite ε-covering number (under L_∞), we conclude that we have $\mathcal{N}(\varepsilon, \mathbb{C}_{t,\varepsilon'}, \mathrm{L}_\infty) = \infty$ as well. Thus, we have extended the property (10). By the principle of induction, this completes the proof of the existence of an infinite ε-eluder sequence. This further implies $\mathfrak{e}(\varepsilon) = \infty$.

Quantitative Lower Bound: Next we turn to establishing the quantitative lower bound in the case $\mathcal{N}(3\varepsilon, \mathbb{C}, \mathrm{L}_\infty) < \infty$: we will show $\mathfrak{e}(\varepsilon) \ge n$, where

$$n = \left\lfloor \frac{2\ln(\mathcal{N}(3\varepsilon, \mathbb{C}, \mathrm{L}_\infty))}{\ln\left(\frac{4\ln(\mathcal{N}(3\varepsilon, \mathbb{C}, \mathrm{L}_\infty))}{\varepsilon^2}\right)} \right\rfloor.$$

Let $G \subseteq \mathbb{C}$ be a maximal subset of $\mathbb{C}$ such that for every distinct $g, g' \in G$, $\sup_{x\in\mathcal{X}} |g(x) - g'(x)| > 3\varepsilon$: that is, a maximal 3ε-packing of $\mathbb{C}$. As is well-known, G is also a 3ε-cover of $\mathbb{C}$ under L_∞: $\sup_{h\in\mathbb{C}} \min_{g\in G} \sup_{x\in\mathcal{X}} |g(x) - h(x)| \le 3\varepsilon$ (otherwise, the $h \in \mathbb{C}$ violating this condition could be

added to G while preserving the 3ε-packing property, contradicting maximality of G). Therefore, $|G| \geq \mathcal{N}(3\varepsilon, \mathbb{C}, \mathrm{L}_\infty)$.

If $n = 0$, we trivially have $\mathfrak{e}(\varepsilon) \geq n$. To focus on the remaining case, let us suppose $n \geq 1$. In particular (in light of footnote 6) this also means we have $|G| \geq 2$ and hence also $\varepsilon \in (0, 1/3)$.

We construct an ε-eluder sequence of length n in two steps: first, we construct a sequence (x_t, y_t) which is an ε-eluder sequence (with slightly larger than ε gaps, and a stronger constraint on approximation of past values), but for which the y_t values aren't necessarily realized by some $h \in \mathbb{C}$, and second, we round these y_t values to be equal to $h(x_t)$ for some particular $h \in \mathbb{C}$ (since $\mathfrak{e}(\varepsilon) = \sup_{h \in \mathbb{C}} \mathfrak{e}_h(\varepsilon)$ requires that the eluder sequence be centered at some $h \in \mathbb{C}$). We construct the sequence (x_t, y_t) inductively. For the purpose of induction, suppose for some $t \in \{1, \ldots, n\}$ (with a base case of $t = 1$, for which this holds trivially) we have already constructed a sequence $(x_1, y_1), \ldots, (x_{t-1}, y_{t-1})$ and $g_1, \ldots, g_{t-1} \in G$ such that $\forall i \leq t-1$, $|g_i(x_i) - y_i| > \frac{3}{2}\varepsilon$ and $\forall j < i$, $|g_i(x_j) - y_j| \leq \frac{\varepsilon}{2\sqrt{n}}$, and moreover such that the set

$$G_{t-1} := \left\{ g \in G : \max_{i<t} |g(x_i) - y_i| \leq \frac{\varepsilon}{2\sqrt{n}} \right\}$$

satisfies

$$|G_{t-1}| \geq \left(\frac{\varepsilon}{2\sqrt{n}}\right)^{t-1} |G|. \tag{12}$$

In particular, note that our choice of n guarantees

$$(t-1)\ln\left(\frac{2\sqrt{n}}{\varepsilon}\right) < \frac{2\ln(|G|)}{\ln\left(\frac{4\ln(|G|)}{\varepsilon^2}\right)} \frac{1}{2} \ln\left(\frac{1}{\varepsilon^2} \frac{8\ln(|G|)}{\ln\left(\frac{4\ln(|G|)}{\varepsilon^2}\right)}\right) < \ln(|G|),$$

so that $\left(\frac{\varepsilon}{2\sqrt{n}}\right)^{t-1} > \frac{1}{|G|}$, and hence (12) implies $|G_{t-1}| > 1$. Since $G_{t-1} \subseteq G$, there exist $g', g'' \in G_{t-1}$ for which $\exists x_t \in \mathcal{X}$ with $|g'(x_t) - g''(x_t)| > 3\varepsilon$. We will choose the next point x_t in the ε-eluder sequence as any point satisfying this.

Next we define a corresponding y_t value. By the pigeonhole principle, there exists $y_t \in [0, 1]$ such that the set

$$G_t := \left\{ g \in G_{t-1} : |g(x_t) - y_t| \leq \frac{\varepsilon}{2\sqrt{n}} \right\}$$

satisfies

$$|G_t| \geq \frac{\varepsilon}{2\sqrt{n}} |G_{t-1}| \geq \left(\frac{\varepsilon}{2\sqrt{n}}\right)^{t} |G|. \tag{13}$$

To see this, consider covering the interval $[0, 1]$ by $\left\lceil \frac{\sqrt{n}}{\varepsilon} \right\rceil \leq \frac{2\sqrt{n}}{\varepsilon}$ closed intervals of width $\frac{\varepsilon}{\sqrt{n}}$; the pigeonhole principle implies at least one of these intervals contains at least an $\frac{\varepsilon}{2\sqrt{n}}$-fraction of the values $g(x_1)$, $g \in G_{t-1}$, in which case the y_t value at the midpoint of such an interval satisfies the claimed inequality. We will choose any such y_t as the label of x_t to extend the eluder sequence. In particular, by definition, this extends the property (12) to t, so that it only remains to define g_t satisfying the other claimed properties.

By our choice of x_t, the triangle inequality implies there must exist at least one $g_t \in \{g', g''\}$ with $|g_t(x_t) - y_t| > \frac{3}{2}\varepsilon$; we will define g_t as such a function, to extend the sequence $g_1, \ldots, g_{t-1}$

by one. In particular, since $g_t \in G_{t-1}$, this further satisfies that $\forall j < t$, $|g_t(x_j) - y_j| \leq \frac{\varepsilon}{2\sqrt{n}}$. Thus, we have defined a sequence $(x_1, y_1), \ldots, (x_t, y_t) \in \mathcal{X} \times [0,1]$ and $g_1, \ldots, g_t \in G$ such that $\forall i \leq t$, $|g_i(x_i) - y_i| > \frac{3}{2}\varepsilon$, and $\forall j < i$, $|g_i(x_j) - y_j| \leq \frac{\varepsilon}{2\sqrt{n}}$, and moreover the set G_t defined above satisfies (13): that is, we have extended the inductive hypothesis from $t-1$ to t.

By the principle of induction, it follows that there exists sequences $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times [0,1]$ and $g_1, \ldots, g_n \in G$ with $\forall i \leq n$, $|g_i(x_i) - y_i| > \frac{3}{2}\varepsilon$ and $\forall j < i$, $|g_i(x_j) - y_j| \leq \frac{\varepsilon}{2\sqrt{n}}$, and such that the set

$$G_n := \left\{ g \in G : \max_{i \leq n} |g(x_i) - y_i| \leq \frac{\varepsilon}{2\sqrt{n}} \right\}$$

satisfies

$$|G_n| \geq \left(\frac{\varepsilon}{2\sqrt{n}}\right)^n |G|. \tag{14}$$

Now we turn to the second step in the construction: identifying a suitable *center* function $g_0 \in G$. Similarly to above, we note that

$$n \ln\left(\frac{2\sqrt{n}}{\varepsilon}\right) \leq \frac{2\ln(|G|)}{\ln\left(\frac{4\ln(|G|)}{\varepsilon^2}\right)} \frac{1}{2} \ln\left(\frac{8}{\varepsilon^2} \frac{\ln(|G|)}{\ln\left(\frac{4\ln(|G|)}{\varepsilon^2}\right)}\right) < \ln(|G|).$$

Therefore, (14) implies $|G_n| \geq 1$. Let g_0 be any function in G_n. By the triangle inequality, we have that $\forall i \in \{1, \ldots, n\}$, $|g_i(x_i) - g_0(x_i)| \geq |g_i(x_i) - y_i| - |g_0(x_i) - y_i| > \frac{3}{2}\varepsilon - \frac{\varepsilon}{2\sqrt{n}} \geq \varepsilon$, and $\forall j < i$, $|g_i(x_j) - g_0(x_j)| \leq |g_i(x_j) - y_j| + |g_0(x_j) - y_j| \leq \frac{\varepsilon}{\sqrt{n}}$. This moreover implies that $\forall i \in \{1, \ldots, n\}$,

$$\sum_{j<i} (g_i(x_j) - g_0(x_j))^2 \leq (i-1)\frac{\varepsilon^2}{n} < \varepsilon^2.$$

We have thus identified an ε-eluder sequence (for $\mathbb{C}$) centered at $g_0 \in G \subseteq \mathbb{C}$ of length n, which completes the proof of the claimed lower bound on $\mathfrak{e}(\varepsilon)$.

Upper Bound: To complete the proof of Theorem 14, we turn to establishing the sufficiency direction: that is, $\mathcal{N}(\varepsilon\delta, \mathbb{C}, \mathrm{L}_\infty) < \infty \implies \mathfrak{e}(\varepsilon) < \infty$, for any $\delta \in (0, 1/2)$, and moreover, a quantitative bound

$$\mathfrak{e}(\varepsilon) \leq \left\lceil \frac{1}{(1-2\delta)^2} \right\rceil \mathcal{N}(\varepsilon\delta, \mathbb{C}, \mathrm{L}_\infty).$$

Since the right hand side is non-increasing in ε, it will suffice to show that any ε-eluder sequence has size at most equal the right hand side: that is, upon establishing this for *every* $\varepsilon > 0$, we may in particular apply it, for any $\varepsilon > 0$, to the $\varepsilon' \geq \varepsilon$ for which there is an ε'-eluder sequence of length $\mathfrak{e}(\varepsilon)$, concluding $\mathfrak{e}(\varepsilon) \leq \left\lceil \frac{1}{(1-2\delta)^2} \right\rceil \mathcal{N}(\varepsilon'\delta, \mathbb{C}, \mathrm{L}_\infty) \leq \left\lceil \frac{1}{(1-2\delta)^2} \right\rceil \mathcal{N}\varepsilon\delta, \mathbb{C}, \mathrm{L}_\infty)$, to complete the proof.

Let $\{f_1, \ldots, f_N\}$ be a set of $N = \mathcal{N}(\varepsilon\delta, \mathbb{C}, \mathrm{L}_\infty)$ functions $f_i : \mathcal{X} \to [0,1]$ such that $\forall h \in \mathbb{C}$, $\min_{1 \leq i \leq N} \sup_{x \in \mathcal{X}} |f_i(x) - h(x)| \leq \varepsilon\delta$. Such a set must exist by the definition of $\mathcal{N}(\varepsilon\delta, \mathbb{C}, \mathrm{L}_\infty)$. For each $i \in \{1, \ldots, N\}$, let $F_i = \{h \in \mathbb{C} : \sup_{x \in \mathcal{X}} |f_i(x) - h(x)| \leq \varepsilon\delta\}$. Note that $\bigcup_{i=1}^N F_i = \mathbb{C}$.

Now consider any ε-eluder sequence $(x_1, y_1), \ldots, (x_n, y_n)$ for $\mathbb{C}$, and let $h_1, \ldots, h_n \in \mathbb{C}$ be functions in $\mathbb{C}$ which witness this fact: that is,

$$\forall i \in \{1, \ldots, n\}, |h_i(x_i) - y_i| > \varepsilon \quad \text{and} \quad \sum_{j<i} (h_i(x_j) - y_j)^2 \leq \varepsilon^2. \tag{15}$$

Claim 3: For every $t \in \{1, \ldots, N\}$, $|\{h_1, \ldots, h_n\} \cap F_t| \leq \left\lceil \frac{1}{(1-2\delta)^2} \right\rceil$.

We prove this claim by contradiction. Let $T = \left\lceil \frac{1}{(1-2\delta)^2} \right\rceil$, and suppose there exists $t \leq N$ for which there exist $i_1 < i_2 < \cdots < i_{T+1}$ with $h_{i_1}, \ldots, h_{i_{T+1}} \in F_t$. For each $k \in \{1, \ldots, T\}$, by definition of h_{i_k}, we have $|h_{i_k}(x_{i_k}) - y_{i_k}| > \varepsilon$. By definition of F_t, we have $|f_t(x_{i_k}) - h_{i_k}(x_{i_k})| \leq \varepsilon\delta$ and $|f_t(x_{i_k}) - h_{i_{T+1}}(x_{i_k})| \leq \varepsilon\delta$. By the triangle inequality, this implies $|h_{i_{T+1}}(x_{i_k}) - h_{i_k}(x_{i_k})| \leq 2\varepsilon\delta$. By another application of the triangle inequality, we have that $|h_{i_{T+1}}(x_{i_k}) - y_{i_k}| \geq |h_{i_k}(x_{i_k}) - y_{i_k}| - |h_{i_{T+1}}(x_{i_k}) - h_{i_k}(x_{i_k})| > \varepsilon - 2\varepsilon\delta \geq (1-2\delta)\varepsilon$. Since this holds for every $k \leq T$, we have that

$$\sum_{j < i_{T+1}} \left(h_{i_{T+1}}(x_j) - y_j\right)^2 \geq \sum_{k \leq T} \left(h_{i_{T+1}}(x_{i_k}) - y_{i_k}\right)^2 > T(1-2\delta)^2\varepsilon^2 \geq \varepsilon^2.$$

This contradicts (15), which therefore completes the proof of Claim 3.

The claimed upper bound on the length n of any ε-eluder sequence follows immediately from Claim 3: that is, if each F_t contains at most T of the functions h_i witnessing the eluder sequence, and there are n such functions h_i total, then it must be that $n \leq TN = \left\lceil \frac{1}{(1-2\delta)^2} \right\rceil \mathcal{N}(\varepsilon\delta, \mathbb{C}, \mathrm{L}_\infty)$. ■