Proceedings of Machine Learning Research vol 237:1–15, 2024 35th International Conference on Algorithmic Learning Theory

# Multiclass Learnability Does Not Imply Sample Compression

**Chirag Pabbaraju** CPABBARA@CS.STANFORD.EDU
*Stanford University*

**Editors:** Claire Vernade and Daniel Hsu

## Abstract

A hypothesis class admits a sample compression scheme, if for every sample labeled by a hypothesis from the class, it is possible to retain only a small subsample, using which the labels on the entire sample can be inferred. The size of the compression scheme is an upper bound on the size of the subsample produced. Every learnable binary hypothesis class (which must necessarily have finite VC dimension) admits a sample compression scheme of size only a finite function of its VC dimension, independent of the sample size. For multiclass hypothesis classes, the analog of VC dimension is the DS dimension. We show that the analogous statement pertaining to sample compression is not true for multiclass hypothesis classes: every learnable multiclass hypothesis class, which must necessarily have finite DS dimension, does not admit a sample compression scheme of size only a finite function of its DS dimension.

**Keywords:** Sample Compression, Multiclass PAC Learning.

## 1. Introduction

Sample compression is a widely studied paradigm in learning theory. At a high level, the main question that sample compression asks is: given a labeled training dataset, is it possible to get by working only with a small fraction of the dataset? A valid *sample compression scheme* gets rid of all the uninformative points in the dataset, and *compresses* the sample to a much smaller subsample, such that there exists an algorithm that can *reconstruct* all the labels on the original sample correctly just from the compressed sample. A classical example of sample compression is exhibited by *support vector machines* for the task of classifying linearly separable data. Here, the compressor may only send the support vectors in the data to the reconstructor. The reconstructor goes on to build a hyperplane that maximally separates the support vectors with the largest possible margin; this, in turn, also recovers correct labels on the non-support-vector points. In the language of learning theory, if such compression-reconstruction is possible for every sample *realizable* by a hypothesis class $\overline{\mathcal{H}}$, we say that the class $\overline{\mathcal{H}}$ admits a sample compression scheme. In this case, the *size* of the compression scheme is the size $k(m)$ that a sample of size $m$ gets compressed down to.

In fact, sample compression is intrinsically tied up with the *learnability* of binary hypothesis classes (where the label space is $\{0, 1\}$). Formally, Littlestone and Warmuth (1986) showed that every binary class $\overline{\mathcal{H}}$ that admits a sample compression scheme of size $k(m)$ also defines a PAC (Probably-Approximately-Correct) (Valiant, 1984) learning algorithm for the class having sample complexity $O(k(m))$. Thus, compression implies learnability in the case of binary classes. In their work, Littlestone and Warmuth (1986) also asked the converse: does learnability imply compression? Since the PAC learnability of a binary class $\overline{\mathcal{H}}$ is completely characterized by the finiteness of its VC dimension $\mathrm{VC}(\overline{\mathcal{H}})$ (Vapnik and Chervonenkis, 1974, 2015; Blumer et al., 1989), this question is equivalent to asking: does every binary class $\overline{\mathcal{H}}$ having finite VC dimension $\mathrm{VC}(\overline{\mathcal{H}})$ admit a sample compression scheme of size only a finite function of $\mathrm{VC}(\overline{\mathcal{H}})$?

A long line of insightful works on this question culminated with Moran and Yehudayoff (2016) answering it in the affirmative. For any binary class $\overline{\mathcal{H}}$ having VC dimension $d_{\mathrm{VC}}$, Moran and Yehudayoff (2016) constructed a sample compression scheme of size $2^{O(d_{\mathrm{VC}})}$. Prior to their work, existing sample compression schemes had a dependence either on the sample size $m$ (e.g., compression of size $O(d_{\mathrm{VC}} \cdot \log(m))$ via boosting due to Freund (1995); Freund and Schapire (1997)), or on the size of $\overline{\mathcal{H}}$ (e.g., compression of size $O(2^{d_{\mathrm{VC}}} \cdot \log\log|\overline{\mathcal{H}}|)$ due to Moran et al. (2017)).[1] The work of Moran and Yehudayoff (2016) gets rid of both these dependencies and obtains a compression scheme of size only a function of the VC dimension, thus establishing the equivalence of learnability and sample compression for binary hypothesis classes. It is worth mentioning that constructing sample compression schemes of size even sub-exponential in $d_{\mathrm{VC}}$ is still open, and has been a longstanding famous problem in learning theory (Warmuth, 2003)!

For essentially the same reasons that compression implies learnability in the binary case, compression also implies learnability in the *multiclass* case (where the label space is not just $\{0, 1\}$ but much larger, possibly infinite too), as was observed by David et al. (2016). Here too, we can ask the counterpart of Littlestone and Warmuth (1986)'s question: does multiclass learnability imply sample compression? In fact, the notion of what learnability means in the multiclass setting was only fully established in a recent seminal work by Brukhim et al. (2022), who equated PAC learnability of a class with finiteness of its DS dimension, which was first introduced in the work of Daniely and Shalev-Shwartz (2014). Concretely, while finiteness of the DS dimension was known to be necessary for learnability, Brukhim et al. (2022) also constructed an algorithm that successfully learns classes having finite DS dimension. Thus, the natural question to ask is: does every multiclass hypothesis class $\overline{\mathcal{H}}$ having finite DS dimension $\mathrm{DS}(\overline{\mathcal{H}})$ admit a sample compression scheme of size only a finite function of $\mathrm{DS}(\overline{\mathcal{H}})$?

Interestingly, the route that Brukhim et al. (2022) take to construct a learning algorithm for hypothesis classes having finite DS dimension *is* via sample compression. Concretely, for any sample of size $m$ realizable by a hypothesis class of DS dimension $d_{\mathrm{DS}}$, they construct a sample compression scheme of size $\tilde{O}(d_{\mathrm{DS}}^{1.5} \cdot \mathrm{polylog}(m))$. This $\mathrm{polylog}(m)$ dependence on the size of the compression scheme is indeed reminiscent of the analogous dependence in the boosting-based compression scheme of Freund (1995); Freund and Schapire (1997) for binary classes. Given that this dependence was ultimately removed in the work of Moran and Yehudayoff (2016), could it also be altogether gotten rid of in the multiclass case?

We answer this question in the negative, and show that a dependence on the sample size $m$ is indeed necessary in the compression size for any valid sample compression scheme in the multiclass setting. Our main result is the following:

**Theorem 1 (Multiclass Learnability $\not\Rightarrow$ Compression)** *There exists a hypothesis class $\overline{\mathcal{H}}$ mapping a domain $\mathcal{X}$ to $\mathcal{Y} = \{0, 1, 2, \dots\}$ that satisfies:*

*(1)* $d_{\mathrm{DS}}(\overline{\mathcal{H}}) = 1$.

*(2) Any sample compression scheme for $\overline{\mathcal{H}}$ that compresses labeled samples of size $m$ to a subsample of size $k(m)$ must satisfy $k(m) = \Omega((\log(m))^{1-o(1)})$, where the $o(1)$ term goes to 0 as $m \to \infty$.*

1. For a detailed and exhaustive list of other prior compression schemes, we refer the reader to Section 1.2.2 in Moran and Yehudayoff (2016).

This result means that unlike the binary case, we cannot hope to obtain a sample compression scheme in the multiclass setting where the size of the scheme is a finite function of only the DS dimension of the hypothesis class. Instead, the size of any compression scheme must necessarily depend on the sample size. Note again that (1) above implies $\overline{\mathcal{H}}$ is learnable. Therefore, while compression implies learnability, learnability does not imply sample compression in the multiclass case, thus exhibiting a separation amongst the two paradigms in the binary and multiclass case. The rest of the paper is devoted to establishing Theorem 1 and discussing the result.

## 2. Preliminaries and Background

The input data domain is denoted as $\mathcal{X}$ and the label space as $\mathcal{Y}$. Concepts and hypotheses are interchangeably used to mean the same object. To prove our result, we will require dealing with *partial* concept classes ($\mathcal{X} \to \{\mathcal{Y} \cup \{\star\}\}$), where $\star$ is a special symbol, and hence we will denote the otherwise standard *total* concept classes ($\mathcal{X} \to \mathcal{Y}$) with symbols having a bar on top (e.g., $\overline{\mathcal{H}}$) and partial classes without a bar (e.g., $\mathcal{H}$). When we are thinking of partial classes and want the label space to additionally also include the special symbol $\star$, we will be explicit and use $\{\mathcal{Y} \cup \{\star\}\}$ — otherwise, $\mathcal{Y}$ should be assumed to not include $\star$. For a sequence $S \in \mathcal{X}^d$, we denote the restriction of a class $\mathcal{H}$ on $S$ (all the different ways in which members of $\mathcal{H}$ can label $S$) by $\mathcal{H}|_S$.

### 2.1. Partial Concept Classes and Disambiguation

Our lower bound heavily relies on the theory of partial concept classes introduced in the work of Alon et al. (2022). Concepts in a partial concept class are allowed to be undefined in certain regions of the input domain. These regions vary based on known structural assumptions on the data like margin-separatedness, data lying on a low-dimensional subspace, etc.

**Definition 2 (Partial Concept Classes (Alon et al., 2022))** *Given an input space $\mathcal{X}$, a label space $\mathcal{Y} = \{0, 1, 2, \dots\}$, and a special symbol $\star$, a partial concept class $\mathcal{H}$ maps $\mathcal{X}$ to $\mathcal{Y} \cup \{\star\}$ i.e., $\mathcal{H} \subseteq \{\mathcal{Y} \cup \{\star\}\}^{\mathcal{X}}$. For any $h \in \mathcal{H}$, if $h(x) = \star$, we say that $h$ is undefined at $x$. The support of a partial concept $h \in \mathcal{H}$ is defined as $\operatorname{supp}(h) = \{x \in \mathcal{X} : h(x) \neq \star\}$, and $\operatorname{supp}(\mathcal{H}) = \bigcup_{h \in \mathcal{H}} \operatorname{supp}(h)$. A labeled sequence $S = \{(x_1, y_1), \dots, (x_m, y_m)\}$ is realizable by $\mathcal{H}$ if there exists a partial concept $h \in \mathcal{H}$ such that $\forall i \in [m],\ h(x_i) \neq \star$ and $h(x_i) = y_i$.*

If every concept in the partial concept class has full support, the $\star$ symbol becomes irrelevant and we get the usual notion of a total concept class.

**Definition 3 (Total Concept Classes)** *A partial concept class $\overline{\mathcal{H}} \subseteq \{\mathcal{Y} \cup \{\star\}\}^{\mathcal{X}}$ that satisfies $\operatorname{supp}(\overline{h}) = \mathcal{X},\ \forall \overline{h} \in \overline{\mathcal{H}}$, is a total concept class.*

Total concept classes naturally define the notion of "disambiguation" of partial concept classes.

**Definition 4 (Disambiguation, Definition 9 in Alon et al. (2022))** *A total concept class $\overline{\mathcal{H}} \subseteq \mathcal{Y}^{\mathcal{X}}$ disambiguates a partial concept class $\mathcal{H}$ if for every finite labeled sequence $S = \{(x_1, y_1), \dots, (x_m, y_m)\}$ realizable by $\mathcal{H}$, there exists $\overline{h} \in \overline{\mathcal{H}}$ such that $\forall i \in [m],\ \overline{h}(x_i) = h(x_i)$.*

The hard-to-compress hypothesis class that realizes our lower bound in Theorem 1 will be a suitable disambiguation of a hard-to-compress partial concept class. Next, we define the relevant complexity parameter that completely captures learnability of a multiclass hypothesis class — the DS dimension.

### 2.2. DS Dimension

As mentioned above, while the VC dimension of a binary (total) concept class was long known to completely characterize its learnability, the corresponding problem of characterizing the learnability of a class on multiple classes was only resolved recently in the work of Brukhim et al. (2022). They showed that a combinatorial complexity parameter called the DS dimension (due to Daniely and Shalev-Shwartz (2014)) is the appropriate equivalent of the VC dimension in terms of characterizing learnability of multiclass concept classes. Since total classes are special cases of partial classes, we define the DS dimension more generally for multiclass *partial* concept classes below.

**Definition 5 (DS dimension (Daniely and Shalev-Shwartz, 2014))** *Let $\mathcal{H} \subseteq \{\mathcal{Y} \cup \{\star\}\}^{\mathcal{X}}$ be a partial concept class and let $S = \{x_1, \dots, x_d\} \in \mathcal{X}^d$ be an unlabeled sequence. For $i \in [d]$, we say that $f, g \in \mathcal{H}|_S$ are $i$-neighbours if $f(x_i) \neq g(x_i)$ and $f(x_j) = g(x_j)$, $\forall j \neq i$. We say that $\mathcal{H}$ DS-shatters $S$ if there exists $\mathcal{F} \subseteq \mathcal{H}, |\mathcal{F}| < \infty$ satisfying*

1. *$\forall f \in \mathcal{F}|_S,\ \forall i \in [d],\ f(x_i) \neq \star$.*
2. *$\forall f \in \mathcal{F}|_S,\ \forall i \in [d],\ f$ has at least one $i$-neighbor $g$ in $\mathcal{F}|_S$.*

*The DS dimension of $\mathcal{H}$, denoted as $d_{\mathrm{DS}} = d_{\mathrm{DS}}(\mathcal{H})$, is the largest integer $d$ such that $\mathcal{H}$ DS-shatters*[2] *some sequence $S$ of size $d$.*

### 2.3. Sample Compression Schemes

The way in which Brukhim et al. (2022) construct a learning algorithm for multiclass concept classes having finite DS dimension is through a sample compression scheme. This is sufficient because a successful sample compression scheme implies the existence of a learning algorithm (David et al., 2016). Here, we formally define sample compression schemes.

**Definition 6 (Sample Compression, Definition 29 in Alon et al. (2022))** *A compression scheme $(\kappa, \rho)$ consists of a compression function $\kappa : (\mathcal{X} \times \mathcal{Y})^* \to (\mathcal{X} \times \mathcal{Y})^* \times \{0,1\}^*$ and a reconstruction function $\rho : (\mathcal{X} \times \mathcal{Y})^* \times \{0,1\}^* \to \mathcal{Y}^{\mathcal{X}}$. $\kappa$ and $\rho$ must satisfy the following property: for any sequence $S \in (\mathcal{X} \times \mathcal{Y})^*$, $\kappa(S) = (S', B)$ such that the elements in $S'$ necessarily also exist in $S$. The size $k(m)$ of the compression scheme for a given sample size $m$ is*

$$k(m) = \max_{S \in (\mathcal{X} \times \mathcal{Y})^m, (S', B) = \kappa(S)} \max(|S'|, |B|). \tag{1}$$

*The (unqualified) size $k$ of the compression scheme is the maximum size $k(m)$ over all $m$, or infinite if the size can be unbounded.*

*A compression scheme $(\kappa, \rho)$ is a sample compression scheme for a partial concept class $\mathcal{H} \in \{\mathcal{Y} \cup \{\star\}\}^{\mathcal{X}}$ if for all finite labeled sequences $S = \{(x_1, y_1), \dots, (x_m, y_m)\}$ realizable by $\mathcal{H}$, $\rho(\kappa(S))$ is correct on all of $S$ i.e., $\forall i \in [m],\ \rho(\kappa(S))(x_i) = y_i$.*

**Remark 7** *Observe that we only care about compressing sequences realizable by the class (no points in the sequence should be labeled with a $\star$), and that the reconstructor $\rho$ always outputs a total concept.*

2. When $\mathcal{Y} = \{0, 1\}$, DS-shattering is equivalent to the standard notion of VC-shattering i.e., realizability of all binary patterns.

### 2.4. Sample Compression Scheme of Moran and Yehudayoff (2016)

The $2^{O(d_{\mathrm{VC}})}$-sized compression scheme of Moran and Yehudayoff (2016) requires crucially using the *uniform convergence principle* (Vapnik and Chervonenkis, 2015) and also a bound on the *dual VC dimension* of binary classes having finite VC dimension. These ingredients are combined with a clever application of von Neumann's minimax theorem (v. Neumann, 1928) to yield their sample compression scheme. While they are able to use their compression scheme for binary classes in a blackbox manner to derive compression schemes for certain multiclass hypothesis classes having finite *graph dimension*, the graph dimension does not characterize multiclass learnability (more on this in Section 4.1). Instead, we discuss here why their techniques don't translate directly to the multiclass setting in light of the *DS dimension* being the more relevant dimension of interest, as shown by Brukhim et al. (2022).

Firstly, the principle of uniform convergence ceases to hold in the multiclass setting, and the sample complexity of different ERM (Empirical Risk Minimizer) learners can differ by an arbitrarily large factor when the the number of labels is infinite (Daniely et al., 2015). Moreover, the compression scheme of Moran and Yehudayoff (2016) crucially makes use of *proper* learners for binary classes, i.e., learning algorithms whose output hypotheses always belong to the class. On the other hand, there exist multiclass hypothesis classes that provably cannot be learned by any proper learner (Daniely and Shalev-Shwartz, 2014)! Additionally, for binary classes having VC dimension $d_{\mathrm{VC}}$, the dual VC dimension of the class is bounded above by $2^{d_{\mathrm{VC}}+1}$ (Assouad, 1983). However, in the multiclass setting, the corresponding dual DS dimension may not be bounded above by any finite function of the DS dimension (see Table 1 for an illustration)! In particular, we can have every concept in the class use its own set of labels, disjoint from any other concept's labels. If we do this, then it is easy to see that the class cannot DS-shatter any pair of points. However, the dual class can readily DS-shatter arbitrarily large sets.

| $\overline{\mathcal{H}}$ \ $\mathcal{X}$ | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ | $x_7$ | $x_8$ |
|---|---|---|---|---|---|---|---|---|
| $\overline{h}_1$ | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 |
| $\overline{h}_2$ | 3 | 3 | 4 | 4 | 3 | 3 | 4 | 4 |
| $\overline{h}_3$ | 5 | 6 | 5 | 6 | 5 | 6 | 5 | 6 |

Table 1: The DS dimension of $\overline{\mathcal{H}}$ is 1, since every hypothesis is using its own distinct set of labels, and hence no pair $(x_j, x_k)$ can be DS-shattered by $\overline{\mathcal{H}}$. However, notice how the dual class (wherein $\overline{h}_j$'s (rows) become the input domain, and $x_j$'s (columns) form the hypothesis class) can easily realize $i$-neighbors (as in Definition 5). In particular, the dual DS dimension above is 3, since the columns corresponding to $x_1, \dots, x_8$ DS-shatter the 3 rows $\{\overline{h}_1, \overline{h}_2, \overline{h}_3\}$. In fact, these columns are realizing the entire Cartesian product $\{1,2\} \times \{3,4\} \times \{5,6\}$ on $\{\overline{h}_1, \overline{h}_2, \overline{h}_3\}$, and are hence even Natarajan-shattering $\{\overline{h}_1, \overline{h}_2, \overline{h}_3\}$ (Natarajan, 1989). More generally, for any $n$, we can have $2^n$ columns $x_1, \dots, x_{2^n}$ and realize the Cartesian product $\{1,2\} \times \{3,4\} \times \cdots \times \{2n-1, 2n\}$ on $\{\overline{h}_1, \dots, \overline{h}_n\}$ — this class has DS dimension 1 but dual DS dimension $n$.

Due to these reasons, it seems apparent that the techniques from Moran and Yehudayoff (2016) don't port over to the multiclass setting, at least in a straightforward manner. In fact, our main result (Theorem 1), which we will now proceed towards proving, shows that this pursuit of constructing a bounded-DS dimension sample compression scheme in the multiclass setting is indeed fruitless.

## 3. Lower Bound via Disambiguation

The main ingredient we use to establish Theorem 1 is the following result from the work of Alon et al. (2022), which refutes the sample compression conjecture for binary *partial* concept classes. While Alon et al. (2022) stated their result only for $\mathcal{Y} = \{0, 1\}$, we can freely think of the label set in their construction to be $\mathcal{Y} = \{0, 1, 2, \dots\}$ instead, where we don't use the extra labels available at any point. Walking through their proof pointwise then already gives a lower bound for sample compression in (multiclass) partial concept classes. Moreover, as mentioned above, if we only ever use $\{0, 1\}$ labels, DS-shattering is equivalent to VC-shattering. With these considerations, we can state the result from Alon et al. (2022) in the following form (for completeness, we provide a proof in Appendix A):

**Lemma 8 (Theorem 7 in Alon et al. (2022))** *There exists a partial concept class $\mathcal{H} \subseteq \{\mathcal{Y} \cup \{\star\}\}^{\mathcal{X}}$ where $\mathcal{Y} = \{0, 1, 2, \dots\}$ that has the following properties:*

1. $\forall h \in \mathcal{H},\ \forall x \in \mathcal{X},\ h(x) \in \{0, 1, \star\}$,
2. $\mathcal{H} = \bigcup_{n=1}^{\infty} H_n$ *where each* $\mathcal{H}_n \subseteq \{\mathcal{Y} \cup \{\star\}\}^{\mathcal{X}}$,
3. $|\mathcal{H}_n|, |\text{supp}(\mathcal{H}_n)| < \infty$ *for every* $n$,
4. $\text{supp}(\mathcal{H}_n) \cap \text{supp}(\mathcal{H}_m) = \emptyset$ *for every* $n \neq m$,
5. $d_{\text{DS}}(\mathcal{H}_n) = 1$ *for every* $n$,
6. $d_{\text{DS}}(\mathcal{H}) = 1$,

*and additionally satisfies: any sample compression scheme for $\mathcal{H}$ must have size $\Omega((\log(m))^{1-o(1)})$, where $m$ is the size of an input sequence realizable by $\mathcal{H}$, and the $o(1)$ term goes to 0 as $m \to \infty$. In particular, there does not exist a sample compression scheme for $\mathcal{H}$ having a size that is a finite function solely of the DS dimension of $\mathcal{H}$.*

Now, we make a simple observation: as a consequence of the definition of disambiguation (Definition 4), sequences realizable by a disambiguating total class are necessarily a superset of the sequences realizable by the corresponding partial class. This leads to the following proposition, whose proof is immediate from the definitions of sample compression and disambiguation.

**Proposition 9 (Compression monotonic in disambiguation)** *Let $\mathcal{H} \subseteq \{\mathcal{Y} \cup \{\star\}\}^{\mathcal{X}}$ be a partial concept class and $\overline{\mathcal{H}} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a total concept class that disambiguates $\mathcal{H}$. Then, if there exists a sample compression scheme $(\kappa, \rho)$ for $\overline{\mathcal{H}}$ of size $k$, then $(\kappa, \rho)$ is also a sample compression scheme for $\mathcal{H}$.*

There is now a natural strategy to prove a sample compression lower bound for total concept classes: find a partial concept class of small DS dimension that is hard to compress to a small size. Then, construct a *disambiguation* of this class to a total concept class. Given the proposition above, the disambiguating class would be at least as hard to compress as the partial class. However, we would want the DS dimension of the disambiguating class to also be small, in order for the lower bound to be meaningful.

The first ingredient in the above strategy is already available to us — choose the partial class $\mathcal{H}$ given by Lemma 8 that has DS dimension 1. The crucial task that remains is constructing a disambiguation $\overline{\mathcal{H}}$ of $\mathcal{H}$ that also has small DS dimension. But given the power of conjuring new labels at will in the multiclass setting, this is not all too hard — we can construct a disambiguating $\overline{\mathcal{H}}$ that also has DS dimension 1 in a straightforward manner. We disambiguate each partial concept $h \in \mathcal{H}$ with a total concept $\overline{h}$ that assigns the $\star$'s in $h$ a unique label which is never again used by any other disambiguating concept. This preserves the DS dimension of the disambiguating class.

**Lemma 10 (Disambiguation with no DS blow-up)** *There exists a total concept class $\overline{\mathcal{H}} \subseteq \mathcal{Y}^{\mathcal{X}}$ where $\mathcal{Y} = \{0, 1, 2, \dots\}$ such that $\overline{\mathcal{H}}$ disambiguates $\mathcal{H}$ from Lemma 8 and satisfies $d_{\text{DS}}(\overline{\mathcal{H}}) = 1$.*

**Proof** Recall that $\mathcal{H} = \bigcup_{n=1}^{\infty} \mathcal{H}_n$, where each $|\mathcal{H}_n| < \infty$. This means that $\mathcal{H}$ is countably large. Let $h_1, h_2, h_3, \dots$ be an enumeration of all the concepts in $\mathcal{H}$. Then, for each $i \in \{1, 2, \dots\}$, define the total concept $\overline{h}_i : \mathcal{X} \to \mathcal{Y}$ as follows

$$\overline{h}_i(x) = \begin{cases} h_i(x) & \text{if } h_i(x) \neq \star, \\ i+1 & \text{otherwise.} \end{cases}$$

Consider the total concept class $\overline{\mathcal{H}} = \bigcup_{i=1}^{\infty} \overline{h}_i$. By construction, $\overline{\mathcal{H}}$ disambiguates $\mathcal{H}$. Further, any sequence DS-shattered by the partial class is certainly also DS-shattered by $\overline{\mathcal{H}}$, and hence $d_{\text{DS}}(\overline{\mathcal{H}}) \geq 1$. Now, let $S = \{x_1, \dots, x_d\}$ be any sequence that is DS-shattered by $\overline{\mathcal{H}}$. Then, according to Definition 5 above, let $\overline{\mathcal{F}} = \{\overline{f}_1, \dots \overline{f}_m\} \subseteq \overline{\mathcal{H}}$ be the finite subset of $\overline{\mathcal{H}}$ that realizes this shattering. Namely, if we think of all the distinct patterns $\overline{f}_1|_S, \overline{f}_2|_S, \dots, \overline{f}_m|_S$ that $\overline{\mathcal{F}}$ realizes on $S$, then every pattern $\overline{f}_i|_S$ has a neighbor $\overline{f}_j|_S$ in every direction $l \in [d]$ ($\overline{f}_i|_S$ and $\overline{f}_j|_S$ are the same everywhere but at index $l$). There can be two cases: either every string $\overline{f}_i|_S$ is such that $\overline{f}_i|_S \in \{0, 1\}^d$. But this would mean that the partial concepts $\{f_1, \dots, f_d\}$ themselves realize the DS-shattering of $S$, implying that $d \leq 1$. In the other case, there is some $\overline{f}_i$ which satisfies $\overline{f}_i(x_l) \in \{2, 3, \dots\}$ for some $l \in [d]$. But now observe that no other function in $\overline{\mathcal{H}}$ other than $\overline{f}_i$ attains the label $\overline{f}_i(x_l)$ on $x_l$ — this is merely an artefact of our construction of $\overline{\mathcal{H}}$. In particular, this means that $\overline{\mathcal{H}}$ cannot realize a neighbor for $\overline{f}_i|_S$ in any direction other than $l$, meaning also that $d \leq 1$. This completes the proof that $d_{\text{DS}}(\overline{\mathcal{H}}) \leq 1$. ■

**Remark 11** *Lemma 10, Proposition 9 and Lemma 8 together imply Theorem 1.*

## 4. Discussion

We showed that unlike the binary setting, compression and learnability are not equivalent in the multiclass learning setting. Namely, if we are allowed infinite labels, it is possible that a hypothesis class is learnable, but the size of a compressed sample must necessarily scale with the size of the original sample, and cannot be independent of it. Our result illustrates a separation between the paradigms of compression and learnability in the binary and multiclass settings. In the following, we discuss the relevance of our result in the context of a past result on multiclass compression by Moran and Yehudayoff (2016), and also discuss why the disambiguation technique from above does not work in order to prove lower bounds for sample compression when the disambiguating class is only allowed to use finitely many labels.

### 4.1. Upper Bound from Moran and Yehudayoff (2016) in Terms of Graph Dimension

As mentioned above, the seminal work of Moran and Yehudayoff (2016) answered the question "does learnability imply compression?" in the affirmative for binary hypothesis classes. Namely, for any binary hypothesis class of VC dimension $d_{\mathrm{VC}}$, Moran and Yehudayoff (2016) construct a sample compression scheme of size $2^{O(d_{\mathrm{VC}})}$. In fact, there is a nice reduction (outlined in Appendix B for completeness) from the multiclass setting to the binary setting that allows them to use their compression scheme as is, and obtain a sample compression scheme of size $2^{O(d_G)}$ for any (multiclass) hypothesis class, where $d_G$ is the *graph dimension* of the class. The graph dimension $d_G$ is defined as follows:

**Definition 12 (Graph dimension Natarajan (1989))** *Let $\overline{\mathcal{H}} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a hypothesis class and let $S \in \mathcal{X}^d$ be a sequence. We say that $\overline{\mathcal{H}}$ G-shatters $S$ if there exists an $\overline{f} \in \overline{\mathcal{H}}$ (which realizes the G-shattering), such that for every subsequence $T \subseteq S$, there exists an $\overline{h} \in \overline{\mathcal{H}}$ such that*

$$\forall x \in T,\ \overline{h}(x) = \overline{f}(x),\ \text{and}\ \forall x \in S \setminus T,\ \overline{h}(x) \neq \overline{f}(x).$$

*The size of the largest sequence that $\overline{\mathcal{H}}$ G-shatters[3] is called the graph dimension of $\overline{\mathcal{H}}$, denoted as $d_G(\overline{\mathcal{H}})$.*

However, even if it is the case that a hypothesis class always permits a sample compression scheme of size $2^{O(d_G)}$, the graph dimension $d_G$ need not necessarily be finite for a learnable hypothesis class when the label space is allowed to be infinitely large. In particular, Daniely and Shalev-Shwartz (2014) constructed a learnable hypothesis class $\overline{\mathcal{H}}$ that has $d_G(\overline{\mathcal{H}}) = \infty$. This is precisely why the compression scheme of Moran and Yehudayoff (2016) in terms of the graph dimension does not allow them to immediately conclude that "learnability implies compression" in the multiclass case. Furthermore, our lower bound does not contradict their upper bound. This is because the disambiguation we construct in Lemma 10 above only preserves the DS dimension[4]. On the other hand, the graph dimension of the disambiguating class can (and must) increase arbitrarily, so that the $2^{O(d_G)}$ bound is still a valid (but not meaningful) upper bound on the size of the compression scheme. Here is a simple example to see why this might be the case: fix a sequence $S = \{x_1, x_2, x_3\}$, and say that some $h$ in the partial class $\mathcal{H}$ realizes the pattern $(0,0,0)$ on this sequence. Say also that the set of distinct patterns that is realized on $S$ by the rest of the partial concepts in $\mathcal{H}$ is $(\star,0,0), (0,\star,0), (0,0,\star), (\star,\star,0), (\star,0,\star), (0,\star,\star), (\star,\star,\star)$. Observe that this sequence is not remotely DS/VC-shattered by the partial class. Now, let us think of the patterns that the disambiguating class realizes on this sequence. Since each total concept in the disambiguating class labels the $\star$'s in the partial concept it represents with a distinct number, the patterns on $S$ realized by $\overline{\mathcal{H}}$ would be something like:

$$\begin{aligned} (0,0,0) &\longrightarrow (0,0,0) & \qquad (\star,\star,0) &\longrightarrow (20,20,0)\\ (\star,0,0) &\longrightarrow (3,0,0) & (\star,0,\star) &\longrightarrow (39,0,39)\\ (0,\star,0) &\longrightarrow (0,7,0) & (0,\star,\star) &\longrightarrow (0,53,53)\\ (0,0,\star) &\longrightarrow (0,0,11) & (\star,\star,\star) &\longrightarrow (100,100,100). \end{aligned}$$

3. Note that just like DS-shattering, G-shattering is also equivalent to VC-shattering when $\mathcal{Y} = \{0,1\}$.
4. and also, the Natarajan dimension Natarajan (1989), which is also required to be finite for PAC learnability.

We can readily see that $S$ is G-shattered by $\overline{\mathcal{H}}$ ($\overline{h}$ realizes the G-shattering). This phenomenon must indeed be occurring at a larger scale — arbitrarily large sequences must be getting G-shattered by $\overline{\mathcal{H}}$ in the manner illustrated above, so as to ensure that the $2^{O(d_G)}$ upper bound does not contradict our lower bound from Theorem 1.

### 4.2. Upper/Lower Bounds in Terms of Natarajan Dimension for Finite Labels

In a sense, the DS dimension really only comes into picture when dealing with hypothesis classes that have an infinite label space. When the label space of the hypothesis class is finite, i.e., $|\mathcal{Y}| = c < \infty$, learnability of the class is completely characterized (e.g., Theorem 4 in Daniely et al. (2015)) by another folklore quantity called the *Natarajan dimension* $d_N$. The Natarajan dimension of a class must unconditionally be finite for learnability (irrespective of finitely many/infinite labels); however, its finiteness is sufficient for learnability only if the label space is finite. The Natarajan dimension is defined as follows:

**Definition 13 (Natarajan dimension (Natarajan, 1989))** *Let $\overline{\mathcal{H}} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a hypothesis class and let $S \in \mathcal{X}^d$ be a sequence. We say that $\overline{\mathcal{H}}$ N-shatters $S$ if there exist $\overline{f}_1, \overline{f}_2 \in \overline{\mathcal{H}}$ (which realize the $N$-shattering), such that $\forall x \in S,\ f_1(x) \neq f_2(x)$, and further, for every subsequence $T \subseteq S$, there exists an $\overline{h} \in \overline{\mathcal{H}}$ such that*

$$\forall x \in T,\ \overline{h}(x) = \overline{f}_1(x),\ \text{and}\ \forall x \in S \setminus T,\ \overline{h}(x) = \overline{f}_2(x).$$

*The size of the largest sequence that $\overline{\mathcal{H}}$ N-shatters[5] is called the Natarajan dimension of $\overline{\mathcal{H}}$, denoted as $d_N(\overline{\mathcal{H}})$.*

Observe that a sequence that is N-shattered by $\overline{\mathcal{H}}$ is also G-shattered by it, implying $d_N \leq d_G$. Additionally, due to a result by Ben-David et al. (1995), the graph dimension can also be upper-bounded in terms of the Natarajan dimension and the number of classes $c$, as

$$d_N \leq d_G \leq O(d_N \cdot \log(c)). \tag{2}$$

This relation, combined with the sample compression scheme due to Moran and Yehudayoff (2016) above, immediately implies a sample compression scheme of size $c^{O(d_N)}$ for any hypothesis class on $c$ classes having Natarajan dimension $d_N$. Since finiteness of the Natarajan dimension is a necessary condition for leanability, we conclude that compression and learnability are in fact equivalent in the multiclass setting *when the number of labels is finite*.

As for lower bounds, a straightforward lower bound can be obtained by a counting argument, similar to (Floyd and Warmuth, 1995, Theorem 14). Concretely, let $\overline{\mathcal{H}} \subseteq \mathcal{Y}^{\mathcal{X}}$ such that $|\mathcal{Y}| = c$ and $|\mathcal{X}| = m$. For any sample compression scheme of size $k$, the number of distinct "compression sets" possible are at most $\sum_{i=0}^{k} \binom{m}{i} \cdot c^i \cdot 2^{\Theta(k)}$ (choose $i$ distinct elements from $\mathcal{X}$, label it in one of at most $c^i$ possible ways, and append an additional bit string of size at most $\Theta(k)$ to it), which is at most $\left(\frac{cme}{k}\right)^{\Theta(k)}$. If we now think of compressing the entire domain of each hypothesis in the class, each of these compression sets should point to a distinct hypothesis in the class, and hence there should at least be one compression set for every hypothesis in the class. Consequently, if the size of $\overline{\mathcal{H}}$ were to be large, while keeping its Natarajan dimension bounded, we would get a lower

5. Again, N-shattering is equivalent to VC-shattering when $\mathcal{Y} = \{0, 1\}$.

bound on $k$. As (Haussler and Long, 1995, Theorem 4) show, for any $c, d_N \leq m$, one can construct $\overline{\mathcal{H}} \subseteq \mathcal{Y}^{\mathcal{X}}$ having Natarajan dimension $d_N$, $|\mathcal{Y}| = c$ and $|\mathcal{X}| = m$ such that

$$|\overline{\mathcal{H}}| = \sum_{i=0}^{d_N} \binom{m}{i} (c-1)^i.$$

If we set $m = d_N$, we get $|\overline{\mathcal{H}}| = c^{d_N}$. Recalling the upper bound on the number of compression sets from before, we can conclude that if $k < C \cdot d_N$ for some absolute constant $C$, the number of possible compression sets will be smaller than $|\overline{\mathcal{H}}|$. In summary, the size $k$ of any valid sample compression must satisfy

$$\Omega(d_N) \leq k \leq c^{O(d_N)}.$$

Note that setting $c = 2$ (which also makes $d_N = d_{\mathrm{VC}}$) recovers the longstanding unsettled exponential gap between upper and lower bounds in the size of compression schemes in the binary case. For larger but constant $c$, we are morally faced with the same unsettled exponential gap, where the Natarajan dimension replaces the VC dimension. Perhaps unsettling is the regime where we think of $d_N$ as constant. In this case, note that the lower bound is $\Omega(1)$ and does not even depend on the number of classes $c$, whereas the upper bound is $\mathrm{poly}(c)$. It seems plausible that the lower bound on the compression size should grow with the number of classes.

**Open Problem 1** *Let $k$ be the (unqualified) size of any valid sample compression scheme for a hypothesis class $\overline{\mathcal{H}} \subseteq \mathcal{Y}^{\mathcal{X}}$ having Natarajan dimension $d_N = \Theta(1)$, where $|\mathcal{Y}| = c < \infty$. Is $k = \omega(1)$ with respect to $c$? Is $k = \Omega(\mathrm{polylog}(c))$? Is $k = \Omega(c^\delta)$ for some $\delta > 0$?*

### 4.3. Disambiguation Using Only Finitely Many Labels

In the disambiguation that we constructed above, we crucially used the power of *infinite* labels available to us. In fact, using infinite labels is necessary for this proof technique. If instead, we only considered disambiguations that label $\star$'s with one of $c < \infty$ labels, we cannot hope to preserve learnability of the disambiguating class. For $c = 2$, this is immediate from Theorem 1 in Alon et al. (2022), which says that any binary (total) concept class disambiguating the partial class from Lemma 10 *must* have infinite VC dimension. However, even for $c > 2$ but finite, this approach will not work. This is crucially because any total class $\overline{\mathcal{H}}$ disambiguating the partial class $\mathcal{H}$ from Lemma 8 also disambiguates each of the $\mathcal{H}_n$s individually for increasing $n$. We can then instantiate Lemma 16 in Section A.1 and the contrapositive of the *multiclass* version of the Sauer-Shelah-Perles lemma (Haussler and Long, 1995) to conclude that $d_N(\overline{\mathcal{H}}) = \infty$. Thus, this approach with finite-label disambiguators will only let us derive a lower bound on the compression size for what has become an *unlearnable* class, a not-so-interesting result. In contrast, and perhaps intriguingly, disambiguating with *infinite* labels allows us to retain the learnability of the disambiguating class, while also inheriting the lower bound on the compression size from the underlying partial class.

## Acknowledgments

I am deeply grateful to Moses Charikar for his discussions and valuable feedback on the manuscript. This work is supported by his and Gregory Valiant's Simons Investigator Awards, and NSF awards 1704417 and 1813049.

## References


Noga Alon, Steve Hanneke, Ron Holzman, and Shay Moran. A theory of PAC learnability of partial concept classes. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 658–671. IEEE, 2022.

Patrick Assouad. Densité et dimension. In *Annales de l'Institut Fourier*, volume 33, pages 233–282, 1983.

Kaspars Balodis, Shalev Ben-David, Mika Göös, Siddhartha Jain, and Robin Kothari. Unambiguous DNFs and Alon-Saks-Seymour. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 116–124. IEEE, 2022.

S Ben-David, N Cesabianchi, D Haussler, and PM Long. Characterizations of learnability for classes of $\{0,..., n\}$-valued functions. *Journal of Computer and System Sciences*, 50(1):74–86, 1995.

Shalev Ben-David, Pooya Hatami, and Avishay Tal. Low-sensitivity functions from unambiguous certificates. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, volume 67, page 28. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2017.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM (JACM)*, 36(4):929–965, 1989.

Nicolas Bousquet, Aurélie Lagoutte, and Stéphan Thomassé. Clique versus independent set. *European Journal of Combinatorics*, 40:73–92, 2014.

Nataly Brukhim, Daniel Carmon, Irit Dinur, Shay Moran, and Amir Yehudayoff. A characterization of multiclass learnability. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 943–955. IEEE, 2022.

Amit Daniely and Shai Shalev-Shwartz. Optimal learners for multiclass problems. In *Conference on Learning Theory*, pages 287–316. PMLR, 2014.

Amit Daniely, Sivan Sabato, Shai Ben-David, and Shai Shalev-Shwartz. Multiclass learnability and the ERM principle. *J. Mach. Learn. Res.*, 16(1):2377–2404, 2015.

Ofir David, Shay Moran, and Amir Yehudayoff. On statistical learning via the lens of compression. *arXiv preprint arXiv:1610.03592*, 2016.

Sally Floyd and Manfred Warmuth. Sample compression, learnability, and the Vapnik-Chervonenkis dimension. *Machine learning*, 21:269–304, 1995.

Yoav Freund. Boosting a weak learning algorithm by majority. *Information and computation*, 121 (2):256–285, 1995.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

Mika Göös. Lower bounds for clique vs. independent set. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 1066–1076. IEEE, 2015.

David Haussler and Philip M Long. A generalization of Sauer's lemma. *Journal of Combinatorial Theory, Series A*, 71(2):219–240, 1995.

Jeff Kahn. *Recent results on some not-so-recent hypergraph matching and covering problems*. DIMACS, Center for Discrete Mathematics and Theoretical Computer Science, 1991.

Nick Littlestone and Manfred Warmuth. Relating data compression and learnability. 1986.

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

Shay Moran, Amir Shpilka, Avi Wigderson, and Amir Yehudayoff. Teaching and compressing for low VC-dimension. In *A Journey Through Discrete Mathematics: A Tribute to Jiří Matoušek*, pages 633–656. Springer, 2017.

Balas K Natarajan. On learning sets and functions. *Machine Learning*, 4:67–97, 1989.

J v. Neumann. Zur theorie der gesellschaftsspiele. *Mathematische annalen*, 100(1):295–320, 1928.

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

Vladimir Vapnik and Alexey Chervonenkis. Theory of pattern recognition, 1974.

Vladimir N Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity*, pages 11–30. Springer, 2015.

Manfred K Warmuth. Compressing to VC dimension many points. In *COLT*, volume 3, pages 743–744. Springer, 2003.

## Appendix A. Proof of Lemma 8

In this section, we essentially rewrite the proof by Alon et al. (2022), which is a beautiful reduction from a recent breakthrough result by Balodis et al. (2022), which in turn builds upon the works of Ben-David et al. (2017); Göös (2015) to show a nearly tight bound for the Alon-Saks-Seymour problem (Kahn, 1991; Bousquet et al., 2014).
Alon et al. (2022) translate the result by Balodis et al. (2022) into the construction of a partial concept class that is hard to disambiguate with a small total class, and also consequently hard to compress.

### A.1. A Partial Class That is Hard to Disambiguate

We recall some concepts from graph theory. Given a graph $G = (V, E)$, the chromatic number $\chi(G)$ of $G$ is the minimum number of colors required, such that each vertex can be assigned a color in a way that no two vertices connected by an edge have the same color assigned to them. The biclique partition number $\mathrm{bp}(G)$ is the minimum number of complete bipartite graphs required to successfully partition the edge set $E$ of $G$. Each complete bipartite graph in the decomposition consists of all the vertices in $V$ (some possibly isolated, but the rest forming a complete bipartite graph) and a subset of the edges in $E$. Balodis et al. (2022) proved the following result relating these two quantities, in response to a problem originally posed by Alon, Saks and Seymour (Kahn, 1991):

**Theorem 14 (Corollary 3 in Balodis et al. (2022))** *For every $n$, there exists a finite simple graph $G = (V, E)$ with $\mathrm{bp}(G) = n$ such that*

$$\chi(G) \geq n^{(\log(n))^{1-o(1)}},$$

*where the $o(1)$ term goes to 0 as $n \to \infty$.*

Now, we describe the clever reduction by Alon et al. (2022), who leverage the lower bound result above to construct a partial concept class that is hard to disambiguate with a small total concept class. Given $n$, let $G = (V, E)$ be the graph promised by Theorem 14, and let $B_i = (L_i, R_i, E_i)$ be $n$ complete bipartite graphs (identified with numbers in $[n]$) that witness the partitioning of the edge set of $G$ such that $\mathrm{bp}(G) = n$, and the edge sets $E_i$ are pairwise disjoint. Let $\mathcal{Y} = \{0, 1, 2, \dots\}$, and define the multiclass partial concept class $\mathcal{H}_n \subseteq \{\mathcal{Y} \cup \{\star\}\}^{[n]}$ as follows: for each vertex $v \in V$, $\mathcal{H}_n$ contains a partial concept $h_v$ such that for each $i \in [n]$,

$$h_v(i) = \begin{cases} 0 & \text{if } v \in L_i, \\ 1 & \text{if } v \in R_i, \\ \star & \text{otherwise.} \end{cases} \tag{3}$$

Since $\{0, 1\}$ are the only non-$\star$ labels in $\mathcal{H}_n$, DS-shattering reduces to VC-shattering. We have the following two lemmas:

**Lemma 15 (Lemma 31 in Alon et al. (2022))** $\mathrm{DS}(\mathcal{H}_n) = 1$.

**Proof** Since there exists at least one edge in $G$, the hypotheses corresponding to the endpoints of this edge shatter a set of size 1, and hence $\mathrm{DS}(\mathcal{H}_n) \geq 1$. We will show that for any $i \neq j$, $\mathcal{H}_n$ cannot simultaneously realize the patterns $(0, 0)$ and $(1, 1)$ on $(i, j)$, implying that $\mathrm{DS}(\mathcal{H}_n) < 2$. Towards a contradiction, assume that some $h_u$ satisfies $h_u(i) = 0$ and $h_u(j) = 0$. From (3) above, this means that $u \in L_i$ and $u \in L_j$. Now, assume also that some $h_v$ satisfies $h_v(i) = 1$ and $h_v(j) = 1$. This means that $v \in R_i$ and $v \in R_j$. But since the bipartite components $B_i$ are complete, this means that the edge $(u, v)$ exists in both $B_i$ and $B_j$, contradicting the disjointedness of the edge sets $E_i$ and $E_j$. ■

**Lemma 16 (Lemma 32 in Alon et al. (2022))** *Let $\overline{\mathcal{H}}_n \in \mathcal{Y}^{[n]}$ be a total concept class that disambiguates $\mathcal{H}_n$. Then, $\overline{\mathcal{H}}_n$ defines a coloring of $G$ using $|\overline{\mathcal{H}}_n|$ colors. Therefore, from Theorem 14 above,*

$$|\overline{\mathcal{H}}_n| \geq n^{(\log(n))^{1-o(1)}}.$$

**Proof** Let $\overline{h}_v \in \overline{\mathcal{H}}_n$ be the total concept that disambiguates $h_v \in \mathcal{H}_n$ i.e., every sequence realizable by $h_v$ is also realizable by $\overline{h}_v$. Then, we identify the concept $\overline{h}_v$ with a unique color $Id(\overline{h}_v)$, and assign the vertex $v$ this color. This defines a candidate coloring of the vertices of $G$. It remains to argue that no two endpoints of any edge in $G$ are assigned the same color. Indeed, let $(u, v)$ be en edge in $G$. Then, this edge necessarily exists in exactly one of the bipartite components $B_i = (L_i, R_i, E_i)$, meaning that either $h_u(i) = 0, h_v(i) = 1$ or $h_u(i) = 1, h_v(i) = 0$. Whatever be the case, $h_u$ and $h_v$ necessarily disagree on $i$, and therefore so do their disambiguators $\overline{h}_u$ and $\overline{h}_v$, implying $Id(\overline{h}_u) \neq Id(\overline{h}_v)$. ■

Now, for each $n$, we can instantiate $\mathcal{H}_n$ as defined above, each having its own separate domain, and extend the domain of every $\mathcal{H}_n$ to the union of the domains as follows: every $h \in \mathcal{H}_n$ labels the domain of any other $\mathcal{H}_m$ entirely with $\star$. For the domain-extended $\mathcal{H}_n$'s thus defined, by construction, we have that $\text{supp}(\mathcal{H}_n) \cap \text{supp}(\mathcal{H}_m) = \emptyset$ for all $n \neq m$. Furthermore, since each $\mathcal{H}_n$ has a support of size $n$ and is based on a finite simple graph, we already have $|\mathcal{H}_n|, |\text{supp}(\mathcal{H}_n)| < \infty$. Let $\mathcal{H} = \bigcup_{n=1}^{\infty} \mathcal{H}_n$. Since any shattered sequence would have to entirely lie in the support of a single $\mathcal{H}_n$, by Lemma 15 above, we also have that $\text{DS}(\mathcal{H}) = 1$. This justifies all the points describing $\mathcal{H}$ in Theorem 1.

### A.2. Compression Implies Disambiguation

The following lemma shows that sample compression schemes imply disambiguations of bounded size for partial concept classes.

**Lemma 17 (Proposition 14 in Alon et al. (2022))** *For $\mathcal{Y} = \{0, 1, 2, \dots\}$, let $\mathcal{H} \in \{\mathcal{Y} \cup \{\star\}\}^{\mathcal{X}}$ be such that*

1. $\forall h \in \mathcal{H},\ \forall x \in \mathcal{X},\ h(x) \in \{0, 1, \star\}$,
2. $|\text{supp}(\mathcal{H})| \leq n$.

*Then, if $(\kappa, \rho)$ is a sample compression scheme for $\mathcal{H}$ of (unqualified) size $k$, then there exists a disambiguation of $\mathcal{H}$ of size at most $n^{\Theta(k)}$.*

**Proof** By definition of a compression scheme, $\kappa$ must be able to compress the support of every $h \in \mathcal{H}$ to a short labeled sequence of size at most $k$, such that the output of $\rho$ on this short sequence (together with some appropriate bit string of size at most $\Theta(k)$) correctly labels the entire support of $h$. Thus, if we iterate over all possible realizable sequences and bit strings of size at most $\Theta(k)$, the reconstruction by $\rho$ on all of these necessarily disambiguates every single partial concept in $\mathcal{H}$. Since $|\text{supp}(\mathcal{H})| \leq n$, the total number of configurations that we need to apply $\rho$ to is at most $\sum_{i=0}^{k} \binom{n}{i} \cdot 2^i \cdot 2^{\Theta(k)}$ (choose $i$ distinct elements from $\text{supp}(\mathcal{H})$, label it in one of $2^i$ possible ways, and append a bit string of size at most $\Theta(k)$ to it), which is at most $n^{\Theta(k)}$ as required. ■

### A.3. Putting Things Together

Say there exists a compression scheme $(\kappa, \rho)$ for $\mathcal{H}$ defined in Section A.1 above that compresses labeled sequences of size $m$ to size $k(m)$. Then, observe that $(\kappa, \rho)$ defines a compression scheme of (unqualified) size $k = k(m)$ for $\mathcal{H}_m$ (for any sequence realizable by $\mathcal{H}_m$, which must be of size at most $m$, elongate the sequence (if required) to have size exactly $m$ with duplicate elements, and $(\kappa, \rho)$ now correctly compresses-reconstructs it). From Lemma 17 above, this implies a disambiguation of $\mathcal{H}_m$ of size at most $m^{\Theta(k(m))}$. But then, Lemma 16 necessitates that

$$m^{\Theta(k(m))} \geq m^{(\log(m))^{1-o(1)}}$$

which gives us that $k(m) = \Omega((\log(m))^{1-o(1)})$ as required.

## Appendix B. Sample Compression Scheme in Terms of Graph Dimension

We elaborate on the reduction from sample compression schemes for binary hypothesis classes to those for multiclass hypothesis classes from Section 4.1 from Moran and Yehudayoff (2016). Given a hypothesis class $\overline{\mathcal{H}} : \mathcal{X} \to \mathcal{Y}$ having graph dimension $d_G$, construct the *binary* hypothesis class $\overline{\mathcal{H}}' : (\mathcal{X} \times \mathcal{Y}) \to \{0, 1\}$, defined as follows: $\overline{\mathcal{H}}' = \{\overline{h}' : \overline{h} \in \overline{\mathcal{H}}\}$ where $\overline{h}'$ is defined as follows:

$$\overline{h}'(x, y) = \begin{cases} 1 & \text{if } \overline{h}(x) = y, \\ 0 & \text{otherwise.} \end{cases}$$

We can then see that $d_{\mathrm{VC}}(\overline{\mathcal{H}}') = d_G(\overline{\mathcal{H}})$. Now, given a sample $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ realizable by $\overline{\mathcal{H}}$, the compressor constructs the sequence $S' = \{((x_1, y_1), 1), \ldots, ((x_m, y_m), 1)\}$ that is realizable by $\overline{\mathcal{H}}'$. Theorem 1.4 in Moran and Yehudayoff (2016) then implies that there exists a sample compression scheme of size $2^{O(d_G)}$ for $\overline{\mathcal{H}}'$. Namely, let $\mathrm{ERM}_{\overline{\mathcal{H}}'}(S')$ be any hypothesis in $\overline{\mathcal{H}}'$ entirely consistent with $S'$. Then, there exist subsequences $S'_1, \ldots, S'_t$ of $S'$ (where $t = O(2^{d_G})$) each of size $O(d_G)$ such that the majority vote of $\mathrm{ERM}_{\overline{\mathcal{H}}'}(S'_1), \ldots, \mathrm{ERM}_{\overline{\mathcal{H}}'}(S'_t)$ is 1 on every $(x_i, y_i)$ pair in $S'$. This equivalently means that the majority vote of $\mathrm{ERM}_{\overline{\mathcal{H}}}(S_1), \ldots, \mathrm{ERM}_{\overline{\mathcal{H}}}(S_t)$ is the correct label $y_i$ for every $x_i$ in $S$. Thus, the compressor compresses $S$ to $S_1, \ldots, S_t$ (along with a bit string of size $2^{O(d_G)}$ specifying splits), and the reconstructor invokes ERM on each $S_i$ and takes the majority vote to obtain correct predictions on all of $S$.