A Fine-grained Characterization of PAC Learnability

Marco Bressan MARCO.BRESSAN@UNIMI.IT
Università degli Studi di Milano, Italy

Nataly Brukhim NBRUKHIM@PRINCETON.EDU
Institute for Advanced Study, USA

Nicolò Cesa-Bianchi NICOLO.CESA-BIANCHI@UNIMI.IT
Università degli Studi di Milano, Italy
Politecnico di Milano, Italy

Emmanuel Esposito EMMANUEL@EMMANUELESPOSITO.IT
Università degli Studi di Milano, Italy

Yishay Mansour MANSOUR.YISHAY@GMAIL.COM
Tel Aviv University, Israel
Google Research

Shay Moran SMORAN@TECHNION.AC.IL
Technion, Israel
Google Research

Maximilian Thiessen MAXIMILIAN.THIESSEN@TUWIEN.AC.AT
TU Wien, Austria

Editors: Nika Haghtalab and Ankur Moitra

Abstract

In the multiclass PAC setting, even when full learnability is unattainable, meaningful information can often be extracted to guide predictions. However, classical learning theory has mainly focused on the dichotomy "learnable vs. non-learnable", leaving notions of partial learnability largely unexplored. Indeed, even for a non-learnable class, a learner may still achieve partial success—for example, by making reliable predictions whenever the true label belongs to a fixed subset of the label space, even if it fails otherwise. Similarly, the rigid nature of PAC learnability makes it impossible to distinguish between classes where one can achieve favorable trade-offs between, say, false-positive and false-negative rates, and classes where such trade-offs are fundamentally unattainable. In a nutshell, standard PAC learnability precludes a fine-grained exploration of learnability.

To overcome this limitation, we develop a fine-grained theory of PAC learnability. For any hypothesis class $\mathcal{H}$, given a loss function (which quantifies the penalty for predicting $\widehat{y}$ instead of the true label y) and a target loss threshold z, our theory determines whether it is possible to achieve a loss of at most z. In contrast, classical PAC learning considers only the special case of the zero-one loss and $z = 0$, corresponding to a near perfect classification guarantee. We give a complete characterization of all attainable guarantees, captured by a *finite family* of combinatorial dimensions, which we term the *J-cube dimensions* of $\mathcal{H}$. These dimensions are defined for every subset J of at least two labels. This extends the fundamental theorem of realizable PAC learning based on the VC dimension. In fact, our results hold in a more general multi-objective setting where we fully characterize the Pareto frontier of guarantees attainable for the class $\mathcal{H}$.

Keywords: Multiclass learning, partial learnability, multi-objective learning, Pareto frontier

1. Introduction

A central goal in statistical learning theory is understanding the learnability of a hypothesis class $\mathcal{H}$. The classic way to formalize the problem is through the notion of *PAC learnability* introduced by Valiant (1984). PAC learnability is well understood; as of today, we have several explicit characterizations of it in terms of the combinatorial properties of the class $\mathcal{H}$. For binary classification, a seminal result establishes that PAC learnability is fully characterized by the Vapnik-Chervonenkis (VC) dimension (Vapnik and Chervonenkis, 1974; Blumer et al., 1989). For the multiclass setting, a long line of work initiated by Natarajan and Tadepalli (1988); Natarajan (1989) shows that PAC learnability is determined by the Natarajan dimension of $\mathcal{H}$ when the number of labels is finite (Ben-David, Cesa-Bianchi, Haussler, and Long, 1995), and by the Daniely-Shalev-Shwartz (DS) dimension of $\mathcal{H}$ in full generality (Brukhim et al., 2022).

One limitation of PAC learnability is its *coarseness*. Consider a label space $\mathcal{Y}$ consisting of, say, 10 labels, and two hypothesis classes $\mathcal{H}$ and $\mathcal{H}'$ over $\mathcal{Y}$. Assume that neither $\mathcal{H}$ nor $\mathcal{H}'$ is PAC learnable. However, the extent of this unlearnability differs between the two classes. While $\mathcal{H}$ remains unlearnable in the PAC sense, it still allows for label separability: for almost every pair of distinct labels—except for $\{1, 2\}$—there exists a learner that can, with high probability, eliminate one incorrect label. In contrast, $\mathcal{H}'$ exhibits a stronger form of unlearnability, as no pair of labels is separable. That is, for every pair $\{y', y''\}$, no learner can reliably rule out even a single incorrect label. It is reasonable to deem $\mathcal{H}$ "much more learnable" than $\mathcal{H}'$; and yet in standard PAC terms both $\mathcal{H}$ and $\mathcal{H}'$ are just not learnable, without further distinction.

A similar phenomenon happens if one measures the learner's performance through multiple losses. Suppose for instance that for each label $i = 1, 2, \dots, 10$ we measure the frequency z_i of the event that the true label is i and the learner's prediction is wrong. Even if on both $\mathcal{H}$ and $\mathcal{H}'$ no learner can simultaneously drive all z_i to 0, it could be the case that on $\mathcal{H}$ one could drive z_i to 0 for any subset of nine labels, while on $\mathcal{H}'$ this is possible for just one label. Again, $\mathcal{H}$ would arguably be much more learnable than $\mathcal{H}'$, yet again from a standard PAC perspective the two classes are non-learnable in just the same way. In other words, the extent to which a class can be *partially* learned makes a difference, even when that class is not PAC learnable. It is therefore natural to ask in which sense and to what extent a class $\mathcal{H}$ can be partially learned. Unfortunately, the existing literature does not capture or characterize partial learnability in the sense suggested by our examples above. A related notion of partial learnability is the framework of *list PAC learning* (Brukhim et al., 2022; Charikar and Pabbaraju, 2023), where the goal is to provide a short list of labels containing the correct one. Similarly, classical approaches to multiclass learning have often relied on reducing the task into a sequence of multiple binary problems, such as one-versus-all or all-pairs comparisons. Indeed, our framework captures all of these notions, as well as a more general multi-objective setting.

Our formalization of partial learnability relies on $(\boldsymbol{w}, \boldsymbol{z})$-*learnability*, a notion recently introduced by Bressan et al. (2025) to study cost-sensitive boosting in a multi-objective learning setting. Here $\boldsymbol{w} = (w_1, w_2, \dots)$ is a vector of losses and $\boldsymbol{z} = (z_1, z_2, \dots)$ a vector of guarantees (recall the z_i's from above). A class $\mathcal{H}$ is $(\boldsymbol{w}, \boldsymbol{z})$-learnable if one can learn hypotheses whose loss according to w_i can be made at most $z_i + \epsilon$ for arbitrarily small $\epsilon > 0$ and for all $i = 1, 2, \dots$ at once. This is a strict generalization of PAC learnability, which we recover when $\boldsymbol{w}$ is the single standard 0-1 loss and $\boldsymbol{z} = (0)$. Our main result is a complete characterization of the $(\boldsymbol{w}, \boldsymbol{z})$-learnability of any given class $\mathcal{H}$. This characterization is given by a simple collection of combinatorial dimensions of

$\mathcal{H}$, one for every subset J of labels, which we term *J-cube dimension* of $\mathcal{H}$, and which is a natural extension of the VC dimension. Our characterization says that $\mathcal{H}$ is $(\boldsymbol{w}, \boldsymbol{z})$-learnable if and only if $\mathcal{H}$ has a finite J-cube dimension for every subset J with the following property: when the distribution of the true labels is supported on J, one cannot achieve $\boldsymbol{w}$-loss at most $\boldsymbol{z}$ by just 'tossing a die' (that is, by predicting a label according to some distribution, without even looking at the test example). From the point of view of multi-objective learning, where each objective is encoded by one of the losses in $\boldsymbol{w}$, our result can be seen as a characterization of the Pareto frontier of the attainable guarantees $\boldsymbol{z}$ in terms of the class $\mathcal{H}$ (we give also some pictorial examples below). Besides this characterization, we give upper and lower bounds on the sample complexity of a $(\boldsymbol{w}, \boldsymbol{z})$-learner for $\mathcal{H}$, based again on the J-cube dimension. Moreover, for the natural choice of $\boldsymbol{w}$ described above, where w_i tracks the mispredictions when the true label is i, we derive a concise algebraic characterization of $(\boldsymbol{w}, \boldsymbol{z})$-learnability based on the values of $\sqrt{z_i} + \sqrt{z_j}$ for all pairs of labels i, j, and we prove sample complexity bounds with a low (near-quadratic) dependence on the number of labels.

From a technical standpoint, our work employs a mix of different techniques and proof strategies, including one-inclusion-graph orientations, shifting arguments, and conversions from list learners to standard label learners. The most technical part is the construction of a $(\boldsymbol{w}, \boldsymbol{z})$-learner given only the fact that a class $\mathcal{H}$ has finite J-cube dimension for certain label subsets J. It turns out that the connection between the two notions is given by the following *J-elimination* problem, of independent interest: given a test point x, and a subset of labels J guaranteed to contain the correct label of x, identify a label in J that is *not* the one of x. Clearly J-elimination generalizes standard learning, since it boils down to identifying the correct label when $|J| = 2$. We show that finite J-cube dimension implies the existence of a *PAC J-eliminator*: an algorithm that, given a sufficiently large training sample, outputs a J-elimination rule that with good probability performs well over the data distribution. With PAC J-eliminators in place, we can then predict a label by first ruling out as many incorrect labels as possible and then predicting a label from the residual list.

Note that our characterization of $(\boldsymbol{w}, \boldsymbol{z})$-learnability does not follow from the results of Bressan et al. (2025), which focus instead on the notion $(\boldsymbol{w}, \boldsymbol{z})$-boostability. Although learnability and boostability both relies on the game-theoretic notion of J-dice attainability (Definition 3), the tools used to prove these two results are sharply different, in much the same way the standard theory of boosting differs from classic PAC learning theory.

Organization. The rest of the paper is organized as follows. We state our main result—the characterization of fine-grained PAC learnability (Theorem 4) and bounds on its sample complexity in terms of the J-cube dimension (Theorem 5)—in Section 1.1. There we also discuss the special case of population-driven losses leading to an improved sample complexity bound (Theorem 7). In Section 2, we discuss PAC elimination, state the achieved sample complexity bounds (Theorem 10), and give an overview of our algorithm used for PAC J-elimination. Relying on these PAC J-eliminators, we describe our main algorithm for fine-grained learning in Section 3, which proves the sample complexity upper bound of Theorem 5. In Section 4 we prove a lemma enabling the improved sample complexity for the population-driven loss. We provide further context of our results in Section 5 and end with a discussion of related work in Section 6.

1.1. Main results

In this work we consider hypothesis classes $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ for an arbitrary domain $\mathcal{X}$ and a label space $\mathcal{Y} = [k]$ with $k \geq 2$. We define a *cost function*, or simply *cost*, to be any function $w : \mathcal{Y}^2 \to [0, 1]$

that satisfies $w(i,i) = 0$ for all $i \in \mathcal{Y}$. The value $w(i,j)$ should be thought of as the penalty incurred by predicting i when the true label is j. A multi-objective cost function is a vector $\boldsymbol{w} = (w_1, \dots, w_r)$ where w_i is a cost function for every $i = 1, \dots, r$. Thus, $\boldsymbol{w}(\widehat{y}, y) \in [0,1]^r$ for two labels $\widehat{y}, y \in \mathcal{Y}$. We say a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ is $\mathcal{H}$-realizable if there exists $c \in \mathcal{H}$ such that $\mathbb{P}_{(x,y)\sim\mathcal{D}}[c(x) = y] = 1$. We are interested in the following $(\boldsymbol{w}, \boldsymbol{z})$-learning guarantee with respect to any multi-objective loss within the PAC framework—see, e.g., Bressan et al. (2025).

Definition 1 ($(\boldsymbol{w}, \boldsymbol{z})$-learner) *Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, let $\boldsymbol{w} : \mathcal{Y}^2 \to [0,1]^r$ be a multi-objective cost, and let $\boldsymbol{z} \in [0,1]^r$. An algorithm $\mathcal{A}$ is a $(\boldsymbol{w}, \boldsymbol{z})$-learner for $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ if there is a function $m_0 : (0,1)^2 \to \mathbb{N}$ such that for every $\mathcal{H}$-realizable distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ and every $\epsilon, \delta \in (0,1)$ what follows holds. If S is a sample of $m_0(\epsilon, \delta)$ examples drawn i.i.d. from $\mathcal{D}$, then $\mathcal{A}(S)$ returns a predictor $h : \mathcal{X} \to \mathcal{Y}$ such that with probability at least $1 - \delta$:*

$$\forall\, i = 1, \dots, r, \qquad L_{\mathcal{D}}^{w_i}(h) \triangleq \mathbb{E}_{(x,y)\sim\mathcal{D}}\Big[w_i(h(x), y)\Big] \le z_i + \epsilon \,.$$

If that is the case then we say $\mathcal{H}$ is $(\boldsymbol{w}, \boldsymbol{z})$-learnable, and we define the sample complexity $m_{\mathcal{H}}^{\boldsymbol{w},\boldsymbol{z}}$ of $(\boldsymbol{w}, \boldsymbol{z})$-learning $\mathcal{H}$ to be the optimal $m_0(\epsilon, \delta)$ achievable by any learning algorithm.

For ease of notation we may write $L_{\mathcal{D}}^{\boldsymbol{w}}(h) \triangleq \mathbb{E}_{(x,y)\sim\mathcal{D}}\big[\boldsymbol{w}(h(x), y)\big]$ and, thus, the expected loss inequality in the definition above can be rewritten as $L_{\mathcal{D}}^{\boldsymbol{w}}(h) \le \boldsymbol{z} + \epsilon\mathbf{1}$ where $\mathbf{1}$ is the r-dimensional all-one vector and the inequality is taken coordinate-wise. We also allow for the predictor h to be a randomized function, in which case the expectation is taken over the randomization of h as well. We give a characterization of fine-grained PAC-learnability via the J-cube dimension, defined next.

Definition 2 (J-cube dimension) *Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ and $J \subseteq \mathcal{Y}$. A subset $X \subseteq \mathcal{X}$ is* J-shattered *by $\mathcal{H}$ if $J^X \subseteq \mathcal{H}_{|X}$. The* J-cube dimension *of $\mathcal{H}$, denoted $d_J(\mathcal{H})$, is the largest cardinality of a set J-shattered by $\mathcal{H}$. If there are arbitrarily large such sets then $d_J(\mathcal{H}) = \infty$.*

Observe that a J-shattered set for $J = \{0,1\}$ is simply a boolean cube, and that in this case the J-cube dimension is equivalent to the VC dimension. For arbitrary sets $J \subseteq \mathcal{Y}$, a J-shattered set is a natural extension of the boolean cube.

Before stating the main result, we recall a key definition from Bressan et al. (2025) concerning the *trivial* guarantees that can be attained by *any* hypothesis class. We say that a guarantee $(\boldsymbol{w}, \boldsymbol{z})$ is trivial with respect to a subset of labels $J \subseteq \mathcal{Y}$ if there exists a $(\boldsymbol{w}, \boldsymbol{z})$-learner whose output is always a hypothesis that can be simulated by a (biased) random die, as formally defined next.

Definition 3 (J-dice attainability) *Let $\mathcal{Y} = [k]$, let $\boldsymbol{w} = (w_1, \dots, w_r)$ be a multi-objective loss, let $\boldsymbol{z} \in [0,1]^r$, and let $J \subseteq \mathcal{Y}$. Then $(\boldsymbol{w}, \boldsymbol{z})$ is* J-dice-attainable *if:*

$$\forall \boldsymbol{q} \in \Delta_J,\ \exists \boldsymbol{p} \in \Delta_{\mathcal{Y}},\ \forall i = 1, \dots, r, \qquad w_i(\boldsymbol{p}, \boldsymbol{q}) \le z_i \,. \tag{1}$$

The J-dice-attainable region *of $\boldsymbol{w}$ is $D_J(\boldsymbol{w}) \triangleq \{\boldsymbol{z} \in [0,1]^r : (\boldsymbol{w}, \boldsymbol{z})$ is J-dice-attainable$\}$.*

We remark that each $D_J(\boldsymbol{w})$ is upward closed: if $\boldsymbol{z} \in D_J(\boldsymbol{w})$, then $\boldsymbol{z}' \in D_J(\boldsymbol{w})$ for all $\boldsymbol{z}' \ge \boldsymbol{z}$ coordinate-wise. Moreover, if $(\boldsymbol{w}, \boldsymbol{z})$ is J-dice-attainable, then $(\boldsymbol{w}, \boldsymbol{z})$ is also J'-dice-attainable for all $J' \subseteq J$ (i.e., $D_J(\boldsymbol{w}) \subseteq D_{J'}(\boldsymbol{w})$). In other words,

$$D_J(\boldsymbol{w}) = \bigcap_{J' \in \binom{J}{\ge 2}} D_{J'}(\boldsymbol{w}) \subseteq \bigcap_{J' \in \binom{J}{2}} D_{J'}(\boldsymbol{w}) \,.$$

The dice-attainable regions are visually illustrated in Figure 1 for a special case of $\boldsymbol{w}$, demonstrating these properties via a simplified example for 3 labels. We can now introduce our main result: the characterization of $(\boldsymbol{w}, \boldsymbol{z})$-learnability.

Theorem 4 (Characterization of fine-grained PAC learnability) *For every hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, every multi-objective loss $\boldsymbol{w} : \mathcal{Y}^2 \to [0,1]^r$, and every $\boldsymbol{z} \in [0,1]^r$:*

$$\mathcal{H} \textit{ is } (\boldsymbol{w}, \boldsymbol{z})\textit{-learnable} \iff \big(\forall J \subseteq \mathcal{Y} \,:\, J \in \mathcal{J}(\boldsymbol{w}, \boldsymbol{z}) \Rightarrow d_J(\mathcal{H}) < \infty\big).$$

where $\mathcal{J}(\boldsymbol{w}, \boldsymbol{z}) \triangleq \{J \subseteq \mathcal{Y} : \boldsymbol{z} \notin D_J(\boldsymbol{w})\}$. Equivalently,

$$\mathscr{L}_{\boldsymbol{w}}(\mathcal{H}) = \bigcap_{J \subseteq \mathcal{Y} \,:\, d_J(\mathcal{H}) = \infty} D_J(\boldsymbol{w}),$$

where $\mathscr{L}_{\boldsymbol{w}}(\mathcal{H}) = \{\boldsymbol{z} \in [0,1]^r : \mathcal{H} \text{ is } (\boldsymbol{w}, \boldsymbol{z})\text{-learnable}\}$.

Let us briefly elaborate on Theorem 4. Consider the first characterization. By definition of J-dice attainability, $J \in \mathcal{J}(\boldsymbol{w}, \boldsymbol{z})$ implies that no predictor that can be simulated by a (biased) random dice can achieve loss $\boldsymbol{z}$ under $\boldsymbol{w}$ when the distribution of the labels has support in J. Intuitively, then, any learner that achieves loss $\boldsymbol{z}$ under $\boldsymbol{w}$ (or comes arbitrarily close to it) must learn *something* about labels in J; that is, it must be able to discriminate the examples with labels in J to some nontrivial extent. Theorem 4 makes this connection formal by stating that being able to "discriminate" labels in J is precisely the condition $d_J(\mathcal{H}) < \infty$. In other words, in order to be $(\boldsymbol{w}, \boldsymbol{z})$-learnable, a class must ensure a finite J-cube dimension for every $J \in \mathcal{J}(\boldsymbol{w}, \boldsymbol{z})$, and this is in both a necessary and sufficient condition.

The second characterization gives a reversed perspective. Suppose $J \subseteq \mathcal{Y}$ has $d_J(\mathcal{H}) = \infty$. Then any learner for $\mathcal{H}$ cannot do better than guessing "at random" when the ground-truth label distribution is over J. That is, it cannot perform any better than a predictor that behaves as a die and, therefore, does not even look at the example at hand. Therefore, for the learner to be able to satisfy $\boldsymbol{z}$, it must be that $\boldsymbol{z} \in D_J(\boldsymbol{w})$. Again, Theorem 4 makes this formal by stating that the *learnable region* $\mathscr{L}_{\boldsymbol{w}}(\mathcal{H})$ of $\mathcal{H}$ w.r.t. $\boldsymbol{w}$ is precisely the intersection of all such $D_J(\boldsymbol{w})$.

The characterizations of Theorem 4 are only existential. The next result complements them by giving upper and lower bounds on the sample complexity of $(\boldsymbol{w}, \boldsymbol{z})$-learning $\mathcal{H}$.

Theorem 5 (Sample complexity of fine-grained PAC learning) *Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, let $\boldsymbol{w}$ a multi-objective loss, and let $\boldsymbol{z} \in \mathscr{L}_{\boldsymbol{w}}(\mathcal{H})$. The $(\boldsymbol{w}, \boldsymbol{z})$-learning sample complexity of $\mathcal{H}$ satisfies:*

$$m_{\mathcal{H}}^{\boldsymbol{w},\boldsymbol{z}}(\epsilon, \delta) = \mathcal{O}\left(\frac{N_{\boldsymbol{w},\boldsymbol{z}}\left(d_{\boldsymbol{w},\boldsymbol{z}} + \log \frac{N_{\boldsymbol{w},\boldsymbol{z}}}{\delta}\right)}{\epsilon} + \frac{k + \log \frac{1}{\delta}}{\epsilon^2}\right)$$

where $N_{\boldsymbol{w},\boldsymbol{z}} = |\mathcal{J}(\boldsymbol{w}, \boldsymbol{z})| < 2^k$ and $d_{\boldsymbol{w},\boldsymbol{z}} = \max_{J \in \mathcal{J}(\boldsymbol{w},\boldsymbol{z})} d_J(\mathcal{H})$. Moreover there exist $\boldsymbol{w}, \boldsymbol{z}$ such that if $\boldsymbol{z} \in \mathscr{L}_{\boldsymbol{w}}(\mathcal{H})$ then

$$m_{\mathcal{H}}^{\boldsymbol{w},\boldsymbol{z}}(\epsilon, \delta) = \Omega\left(\frac{d_{\boldsymbol{w},\boldsymbol{z}} + \log \frac{1}{\delta}}{\epsilon}\right).$$

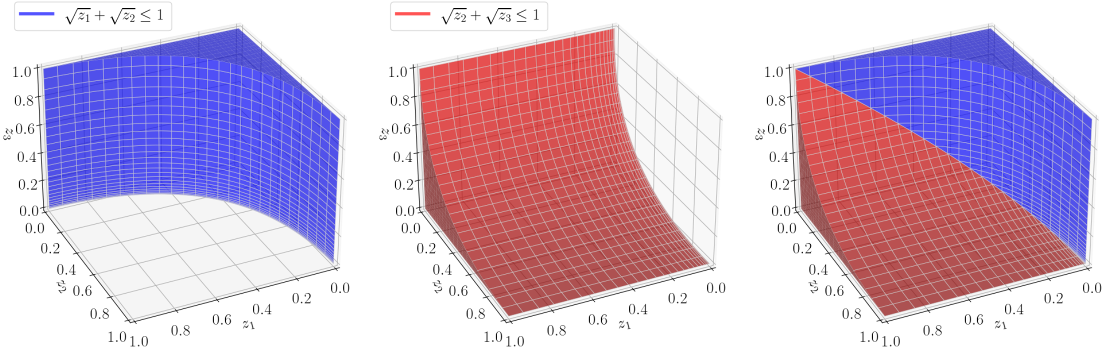

Figure 1: **Pareto frontiers.** The figures illustrate the Pareto frontiers as measured by the population-driven loss $\boldsymbol{w}^{\mathrm{p}}$ (see Equation (2)), for the case of 3 labels, shown in separate plots for clarity. Each $z_i \in [0,1]$ represents the error rate when the true label is i, and so a point $\boldsymbol{z} \in [0,1]^3$ tracks errors for each label separately. Each surface in the left and middle figures represents the Pareto frontier over pairs of labels $\{1,2\}$ and $\{2,3\}$, respectively, with the non-shaded area above each surface forming the corresponding dice-attainable regions $D_{\{1,2\}}(\boldsymbol{w}^{\mathrm{p}})$ and $D_{\{2,3\}}(\boldsymbol{w}^{\mathrm{p}})$. The right figure then shows the boundary of the intersection of the two dice-attainable regions, which by Theorem 6 corresponds exactly to the Pareto frontier of the learnable region $\mathscr{L}_{\boldsymbol{w}^{\mathrm{p}}}(\mathcal{H})$ for any hypothesis class $\mathcal{H}$ with $d_{\{1,2\}}(\mathcal{H}) = d_{\{2,3\}}(\mathcal{H}) = \infty$ and $d_{\{1,3\}}(\mathcal{H}) < \infty$. See Theorem 6 for the general result.

Population-driven loss. We focus on a natural multi-objective loss which we call *population-driven* loss. In a nutshell this is the standard 0-1 loss, but measured more finely by tracking the misprediction errors for each target label separately. Formally, for every $\ell \in \mathcal{Y}$ define the cost w_ℓ^{p} such that for all $i, j \in \mathcal{Y}$,

$$w_\ell^{\mathrm{p}}(i,j) \triangleq \mathbb{I}\{j = \ell \wedge i \neq \ell\} \ . \tag{2}$$

The population-driven loss is then simply the multi-objective loss $\boldsymbol{w}^{\mathrm{p}} = (w_1^{\mathrm{p}}, \dots, w_k^{\mathrm{p}})$. For every $i, j \in \mathcal{Y}$, the value of $\boldsymbol{w}^{\mathrm{p}}(i,j)$ is $\mathbf{0} + \mathbb{I}\{j \neq i\} \cdot \boldsymbol{e}_j$ — hence, as said, the j-th entry indicates whether the true label is j *and* the predicted label i is wrong.

Our main result is that whether a class $\mathcal{H}$ is $(\boldsymbol{w}^{\mathrm{p}}, \boldsymbol{z})$-learnable or not admits an elegant characterization determined by the J-cube dimension of certain *pairs* of labels. More precisely we prove the following analogue of Theorem 4:

Theorem 6 *For every hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ and every $\boldsymbol{z} \in [0,1]^k$,*

$$\mathcal{H} \textit{ is } (\boldsymbol{w}^{\mathrm{p}}, \boldsymbol{z})\textit{-learnable} \iff \left(\forall \{i,j\} \subseteq \mathcal{Y} \,:\, \sqrt{z_i} + \sqrt{z_j} < 1 \Rightarrow d_{\{i,j\}}(\mathcal{H}) < \infty\right).$$

Equivalently,

$$\mathscr{L}_{\boldsymbol{w}^{\mathrm{p}}}(\mathcal{H}) = \bigcap_{J = \{i,j\} \subseteq \mathcal{Y} \,:\, d_J(\mathcal{H}) = \infty} \left\{\boldsymbol{z} \in [0,1]^k \,:\, \sqrt{z_i} + \sqrt{z_j} \geq 1\right\}.$$

A simple example of the learnable region $\mathscr{L}_{\boldsymbol{w}^{\mathrm{p}}}(\mathcal{H})$ is depicted in Figure 1 for the case of 3 labels.

As a consequence of Theorem 6, when $\boldsymbol{z} \in \mathscr{L}_{\boldsymbol{w}^{\mathrm{p}}}(\mathcal{H})$ we obtain a $(\boldsymbol{w}^{\mathrm{p}}, \boldsymbol{z})$-learner for $\mathcal{H}$ whose sample complexity has a $\widetilde{\mathcal{O}}(k^2)$ dependence on k. This is in contrast to the general bounds of Theorem 5, where the same dependence might be exponential. Formally, we prove:

Theorem 7 *Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ and $\boldsymbol{z} \in \mathscr{L}_{\boldsymbol{w}^{\mathrm{p}}}(\mathcal{H})$. The $(\boldsymbol{w}^{\mathrm{p}}, \boldsymbol{z})$-learning complexity of $\mathcal{H}$ satisfies:*

$$m_{\mathcal{H}}^{\boldsymbol{w}^{\mathrm{p}},\boldsymbol{z}}(\epsilon,\delta) = \mathcal{O}\left(\frac{k^2\left(d_{\boldsymbol{z}} + \ln\frac{k}{\delta}\right)}{\epsilon} + \frac{k + \ln\frac{1}{\delta}}{\epsilon^2}\right)$$

where $d_{\boldsymbol{z}} = \max\left\{d_{\{i,j\}}(\mathcal{H}) \,:\, \sqrt{z_i} + \sqrt{z_j} < 1\right\}$.

2. PAC Elimination

This section describes *PAC J-elimination*, one of the key ingredients behind Theorem 4 and Theorem 5. In standard PAC learning, the goal is to learn a hypothesis that with good probability predicts the correct label of a given test point x. For classes that are not PAC learnable, however, this goal cannot be achieved. The idea is therefore to relax the goal, and try to at least *restrict* the set of candidate correct labels (akin to what is done in list learning). In fact, we consider an even simpler goal: eliminate at least one incorrect label.

Concretely, let $J \subseteq \mathcal{Y}$ with $|J| \geq 2$.[1] Given a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, we consider the task of learning a hypothesis $h : \mathcal{X} \to J$ that, given a test point x whose true label is y, predicts a label that with good probability is in $J \setminus \{y\}$. The performance of h can be formally captured by the following loss:

$$L_{\mathcal{D}}^{J}(h) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\ell_{(x,y)}^{J}(h)\right], \quad \text{where } \ell_{(x,y)}^{J}(h) = \mathbb{I}\{h(x) \notin J \vee h(x) = y\}\,.$$

As usual, when h is randomized, the expectation is understood over the randomness of h, too. Equipped with the loss above we can then introduce:

Definition 8 (PAC Elimination) *Let $J \subseteq \mathcal{Y}$ such that $|J| \geq 2$. We say that J is* PAC-eliminable *by a class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ if there exists a learning rule A and a sample complexity $m_{\mathcal{H}}^{J} : (0,1)^2 \to \mathbb{N}$ such that for all $\epsilon, \delta \in (0,1)$ and all $\mathcal{H}$-realizable distributions $\mathcal{D}$, when A receives an i.i.d. sample S from $\mathcal{D}$ of size $m \geq m_{\mathcal{H}}^{J}(\epsilon,\delta)$, the returned classifier $h = A(S)$ satisfies $L_{\mathcal{D}}^{J}(h) \leq \epsilon$ with probability at least $1-\delta$. Such a learning rule A is called a* PAC J-eliminator *with respect to $\mathcal{H}$ and the minimal sample complexity $m_{\mathcal{H}}^{J}$ is called the PAC J-elimination sample complexity of $\mathcal{H}$.*

Note that in the binary case, when $\mathcal{Y} = \{0,1\}$, a set J is PAC-eliminable with respect to $\mathcal{H}$ if and only if $\mathcal{H}$ is PAC-learnable (because, in the binary case, ruling out an incorrect label is equivalent to predicting the correct one). Thus PAC eliminability is a generalization of binary PAC learnability (and a relaxation of multiclass PAC learnability).

The main result of the present section is a characterization of PAC eliminability of a set J with respect to a class $\mathcal{H}$. This characterization is analogous to the characterization of PAC learnability based on the VC dimension of $\mathcal{H}$, and in fact boils down to it for $|J| = 2$. More precisely, we prove that J is PAC-eliminable with respect to $\mathcal{H}$ if and only if the J-cube dimension d_J of $\mathcal{H}$ is finite:

Theorem 9 (Characterization of PAC J-eliminability) *Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a hypothesis class, and let $J \subseteq \mathcal{Y}$ be such that $2 \leq |J| < \infty$. The following statements are equivalent:*

[1] We assume that J consists of at least two labels, as the elimination problem becomes infeasible if J contains only a single label. In such a case, the target function always predicts the label in J, leaving no incorrect label to eliminate. If J is infinite, it is PAC-eliminable with respect to every class $\mathcal{H}$ using a learner that, given an input sample S of size n, eliminates a random label from a subset $J_n \subseteq J$ of size $|J_n| = n$. This learner has error $1/n$.

1. *J is PAC-eliminable with respect to $\mathcal{H}$.*
2. *$d_J(\mathcal{H}) < \infty$.*

In fact, Theorem 9 is a corollary of the following characterization of the sample complexity of PAC J-eliminators in terms of $d_J(\mathcal{H})$:

Theorem 10 (PAC J-Elimination Sample Complexity) *Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a hypothesis class, and let $J \subseteq \mathcal{Y}$ be such that $2 \leq |J| < \infty$. The PAC J-elimination sample complexity $m^J_{\mathcal{H}}$ of $\mathcal{H}$ satisfies*

$$\Omega\left(\frac{d_J(\mathcal{H}) + \log(1/\delta)}{|J|\epsilon}\right) \leq m^J_{\mathcal{H}}(\epsilon, \delta) \leq \mathcal{O}\left(\frac{d_J(\mathcal{H}) + \log(1/\delta)}{\epsilon}\right).$$

Note that the upper and lower bounds of Theorem 10 exhibit a multiplicative gap of $\mathcal{O}(|J|)$; this gap is inevitable, meaning that either bound can become tight depending on the class (see below). However, qualitatively, the J-cube dimension does characterize learnability, in the sense that J is PAC-eliminable w.r.t. $\mathcal{H}$ if and only if $d_J(\mathcal{H}) < \infty$, as per Theorem 9. The rest of this section provides a walk-through of the techniques behind Theorem 9 and Theorem 10.

For what concerns the upper bounds, we are not aware of any empirical risk minimization (ERM) variant for PAC elimination. In fact, even basic notions like consistent hypotheses and proper learning do not have a clear definition. Thus, we adapt a common strategy based on the one-inclusion graph algorithm, which is broadly applicable (Haussler et al., 1994; Aden-Ali et al., 2023b; Brukhim et al., 2022). The general idea is the following. Instead of directly achieving a high probability guarantee, we start by seeking an error bound that just holds in expectation over the drawn sample. Through a standard symmetrization argument, that in-expectation error can be bounded in terms of the leave-one-out error on a random sample. In turn, the leave-one-out error can be determined by properties of the one-inclusion-graph.

Thus, we seek to adapt the classic one-inclusion-graph algorithm to the task of J-elimination. To this end we need to define, for every edge in the graph (which is actually a *hyperedge*, as the graph is actually a *hyper*graph), an orientation of that edge—that is, a vertex which is the label that will be predicted later. To obtain good leave-one-out bounds, we need that orientation to have low maximum indegree; that is, each vertex should be selected by as few edges as possible. (In contrast, standard PAC learning asks for a low *out*degree.) To find an orientation with low indegree, then, we define a new variant of graph density, which we call J-density, and apply a standard max-flow argument. Finally, we apply a shifting argument to simplify the hypothesis space. Interestingly, shifting for multi-class PAC learning requires to overcome several obstacles (see, e.g., the discussion by Brukhim et al. (2022)); but here, by virtue of the definitions of J-cube dimension and J-density, the shifting works smoothly, as in the VC case. In the end, we obtain an expected sample complexity of $\mathcal{O}\left(\frac{d_J(\mathcal{H})}{\epsilon}\right)$, analogous to the classic $\mathcal{O}\left(\frac{\mathrm{VC}(\mathcal{H})}{\epsilon}\right)$ bound for binary PAC classification (Haussler et al., 1994). Finally, the in-expectation guarantee can be turned into a predictor satisfying the required high-probability bound $\mathcal{O}\big(\frac{d_J(\mathcal{H})+\log(1/\delta)}{\epsilon}\big)$ relying on a technique by Aden-Ali et al. (2023a).

For the lower bounds, by adapting no-free-lunch based arguments (Blumer et al., 1989) we prove that $\Omega\big(\frac{d_J(\mathcal{H})}{|J|\epsilon}\big)$ samples are necessary in general to guarantee an *expected* loss of ϵ. (Note that a loss of at most $1/|J|$ is easily achieved by selecting a label uniformly at random from J.) Interestingly, the $\mathcal{O}(1/|J|)$ gap with the upper bounds is inevitable, in the sense that there are classes where the sample complexity of PAC elimination is in $\Theta\big(\frac{d_J(\mathcal{H})}{\epsilon}\big)$, and classes where it is in $\Theta\big(\frac{d_J(\mathcal{H})}{|J|\epsilon}\big)$. Thus, in contrast to binary PAC learning (where the in-expectation sample complexity is $\Theta(\mathrm{VC}(\mathcal{H})/\epsilon)$, the J-cube dimension of a class does not fully characterize its sample complexity. In the high probability setting, we prove a lower bound of $\Omega\big(\frac{d_J(\mathcal{H})+\log(1/\delta)}{|J|\epsilon}\big)$.

3. Fine-grained PAC Learnability

This section presents another key ingredient behind Theorem 4 and Theorem 5: a construction that turns a family of PAC J-eliminators into a $(\boldsymbol{w}, \boldsymbol{z})$-learner, as well as the upper bounds that contruction yields. As a byproduct of these bounds, one gets the sufficient condition for $(\boldsymbol{w}, \boldsymbol{z})$-learnability in Theorem 4, and the sample complexity upper bounds of Theorem 5. The remaining direction of Theorem 4, and the lower bounds of Theorem 5, are proven in Appendix C.

Recall the family $\mathcal{J}(\boldsymbol{w}, \boldsymbol{z}) = \{J \subseteq \mathcal{Y} : \boldsymbol{z} \notin D_J(\boldsymbol{w})\}$. We show that, if for every $J \in \mathcal{J}(\boldsymbol{w}, \boldsymbol{z})$ there is a PAC J-eliminator with respect to $\mathcal{H}$, then we can construct a $(\boldsymbol{w}, \boldsymbol{z})$-learner for $\mathcal{H}$; and the sample complexity of such a learner depends on the sample complexity of the eliminators in this construction, as well as on their number. To this end let us give a more formal definition.

Definition 11 *Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, let $\boldsymbol{w} : \mathcal{Y} \times \mathcal{Y} \to [0,1]^r$ be a multi-objective loss, and let $\boldsymbol{z} \in [0,1]^r$. A family $\mathcal{R}$ of PAC eliminators w.r.t. $\mathcal{H}$ is* sufficient *for $(\boldsymbol{w}, \boldsymbol{z})$ if it contains a PAC J-eliminator for every $J \in \mathcal{J}(\boldsymbol{w}, \boldsymbol{z})$.*

Now we prove that, given a family of PAC eliminators that is sufficient for a certain pair $(\boldsymbol{w}, \boldsymbol{z})$, we can indeed build a $(\boldsymbol{w}, \boldsymbol{z})$-learner.

Theorem 12 (PAC eliminators to $(\boldsymbol{w}, \boldsymbol{z})$-learners) *Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, let $\boldsymbol{w} : \mathcal{Y} \times \mathcal{Y} \to [0,1]^r$ be a multi-objective loss, and let $\boldsymbol{z} \in [0,1]^r$. Moreover let $\mathcal{R}$ be a family of PAC eliminators w.r.t. $\mathcal{H}$ that is sufficient for $(\boldsymbol{w}, \boldsymbol{z})$. Then there exists a $(\boldsymbol{w}, \boldsymbol{z})$-learner for $\mathcal{H}$ with sample complexity*

$$\mathcal{O}\left(m_{\mathcal{R}}\left(\frac{\epsilon}{2|\mathcal{R}|}, \frac{\delta}{2|\mathcal{R}|}\right) + \frac{k + \ln\frac{1}{\delta}}{\epsilon^2}\right)$$

where $m_{\mathcal{R}}$ is any upper bound on the sample complexities of the eliminators in $\mathcal{R}$.

Proof The intuition is as follows. First, we train the PAC eliminators in $\mathcal{R}$. Then, given an unlabeled example x, we run those eliminators to rule out as many incorrect labels as possible. This gives us a residual list $J_x \subseteq \mathcal{Y}$ of candidate labels that with good probability contains the true label of x and satisfies $\boldsymbol{z} \in D_{J_x}(\boldsymbol{w})$. At this point, by definition of J_x-dice attainability, we can satisfy the loss bound $\boldsymbol{z}$ by predicting a label from a suitable distribution $\boldsymbol{p}$ over J_x. Computing $\boldsymbol{p}$ requires to estimate the label marginal $\boldsymbol{q}$ of the underlying data distribution, which we do in the training phase.

Let us turn to the formal proof. Fix a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ realizable by $\mathcal{H}$, and choose $\epsilon, \delta > 0$. We describe separately the training phase of the learner and the behavior of the output predictor.

The training phase. Draw a multiset S of $m_{\mathcal{A}}(\epsilon_0, \delta_0)$ examples i.i.d. from $\mathcal{D}$, where $\epsilon_0 = \frac{\epsilon}{2|\mathcal{R}|}$ and $\delta_0 = \frac{\delta}{2|\mathcal{R}|}$. Now consider any $\mathcal{A} \in \mathcal{R}$. Run $\mathcal{A}(S)$ to obtain a predictor $r_{\mathcal{A}} : \mathcal{X} \to \mathcal{Y}$. Let:

$$\mathcal{J}_{\mathcal{A}} = \{J \subseteq \mathcal{Y} : \mathcal{A} \text{ is a PAC } J\text{-eliminator w.r.t. } \mathcal{H}\} .$$

Thus $\mathcal{J}_\mathcal{A}$ is the set of all J eliminated by $\mathcal{A}$. Note that $|\mathcal{J}_\mathcal{A}| \geq 1$, and in fact we may have $|\mathcal{J}_\mathcal{A}| > 1$: a PAC J-eliminator is a PAC J'-eliminator for every $J' \supseteq J$, too, so $\mathcal{A}$ can be a PAC J-eliminator for multiple sets J. By definition of PAC elimination (Definition 8), with probability $1 - \frac{\delta}{2|\mathcal{R}|}$:

$$L_\mathcal{D}^J(r_\mathcal{A}) \leq \frac{\epsilon}{2|\mathcal{R}|} \quad \forall J \in \mathcal{J}_\mathcal{A} \,.$$

Let then $\mathcal{J} = \cup_{\mathcal{A} \in \mathcal{R}} \mathcal{J}_\mathcal{A}$, and for every $J \in \mathcal{J}$ let $r_J = r_\mathcal{A}$ for some $\mathcal{A}$ such that $J \in \mathcal{J}_\mathcal{A}$. By a union bound over $\mathcal{R}$, with probability at least $1 - \frac{\delta}{2}$ the set of predictors $\{r_J : J \in \mathcal{J}\}$ satisfies:

$$\mathbb{P}_{(x,y)\sim\mathcal{D}}\Big[\forall J \in \mathcal{J} \,:\, y \neq r_J(x) \in J\Big] \geq 1 - \frac{\epsilon}{2} \,. \tag{3}$$

Then, since $\|\boldsymbol{w}\|_\infty \leq 1$, we can charge $\frac{\delta}{2}$ to the probability of failure, as well as $\frac{\epsilon}{2}\,\mathbf{1}$ to the expected loss $L_\mathcal{D}^{\boldsymbol{w}}(h)$ of our final predictor h, and proceed assuming $\mathbb{P}_{(x,y)\sim\mathcal{D}}[\forall J \in \mathcal{J} : y \neq r_J(x) \in J] = 1$.

Next, let $\boldsymbol{q}$ be the marginal of $\mathcal{D}$ over $\mathcal{Y}$. The learner computes a distribution $\widehat{\boldsymbol{q}}$ over $\mathcal{Y}$ whose total variation distance $\|\boldsymbol{q} - \widehat{\boldsymbol{q}}\|_{\mathrm{TV}}$ from $\boldsymbol{q}$ is at most $\frac{\epsilon}{2}$ with probability at least $1 - \frac{\delta}{2}$. It is well known that this can be done by taking $m = \mathcal{O}\big(\frac{k+\ln(1/\delta)}{\epsilon^2}\big)$ independent labeled examples from $\mathcal{D}$, see for instance (Canonne, 2020, Theorem 1). Again, since $\|\boldsymbol{w}\|_\infty \leq 1$, we can then charge a further $\frac{\delta}{2}$ to the failure probability, a further $\frac{\epsilon}{2}\,\mathbf{1}$ to $L_\mathcal{D}^{\boldsymbol{w}}(h)$, and proceed assuming that $\boldsymbol{q} = \widehat{\boldsymbol{q}}$.

We can now describe the predictor.

The prediction phase. Predicting a label for a given $x \in \mathcal{X}$ consists in two steps: computing a list of labels $J_x \subseteq \mathcal{Y}$ through the eliminators, and choosing a (randomized) label from J_x.

Step 1. Compute:

$$J_x = \mathcal{Y} \setminus \{r_J(x) \,:\, J \in \mathcal{J}\} \,.$$

In words, J_x is the list of all labels not eliminated by some r_J. Now, by Equation (3), we can charge an $\frac{\epsilon}{2} \cdot \mathbf{1}$ to the loss and proceed assuming $y \neq r_J(x) \in J$ for all $J \in \mathcal{J}$. Note that this implies $y \in J_x$ by construction of J_x. Moreover, if $|J_x| = 1$ then one can incur loss $\mathbf{0}$ by predicting the only label in J_x, thus we can proceed assuming $|J_x| \geq 2$. This, in turn, implies $\boldsymbol{z} \in D_{J_x}(\boldsymbol{w})$. Suppose indeed by contradiction that $\boldsymbol{z} \notin D_{J_x}(\boldsymbol{w})$. Since $|J_x| \geq 2$, and since by assumption $\mathcal{R}$ is sufficient for $(\boldsymbol{w}, \boldsymbol{z})$, then $\mathcal{R}$ contains a J_x-eliminator. Now consider the output of that eliminator, $r_{J_x}(x)$. By assumption $r_{J_x}(x) \in J_x$, but the construction of J_x ensures $r_{J_x}(x) \notin J_x$, a contradiction. We conclude that $\boldsymbol{z} \in D_{J_x}(\boldsymbol{w})$, as claimed.

Step 2. We shall describe an h such that $L_\mathcal{D}^{\boldsymbol{w}}(h) \leq \boldsymbol{z}$. Let $\mathrm{supp}(J_x)$ be the set of all $J \subseteq \mathcal{Y}$ such that $\mathbb{P}_{(x,y)\sim\mathcal{D}}(J_x = J) > 0$. For every $J \in \mathrm{supp}(J_x)$ let $\mathcal{D}_J$ be the probability distribution obtained by conditioning $\mathcal{D}$ on $J_x = J$. Since we are assuming $y \in J_x$, the distribution $\mathcal{D}_J$ can be equivalently seen as obtained by conditioning $\mathcal{D}$ on $y \in J$. Now consider the marginal of $\mathcal{D}_J$ over $\mathcal{Y}$, denoted by $\boldsymbol{q}_J$. Note that the learner knows $\boldsymbol{q}$ (as we are assuming $\boldsymbol{q} = \widehat{\boldsymbol{q}}$, see above), hence it knows $\boldsymbol{q}_J$, too. Recall from above that $\boldsymbol{z} \in D_J(\boldsymbol{w})$. By definition of J-dice attainability, then, there exists a distribution $\boldsymbol{p}_J \in \Delta_\mathcal{Y}$ such that $\boldsymbol{w}(\boldsymbol{p}_J, \boldsymbol{q}_J) \leq \boldsymbol{z}$; in fact, $\boldsymbol{p}_J$ can be computed in polynomial time via linear programming. Let then $h_J : \mathcal{X} \to \mathcal{Y}$ be the randomized hypothesis that upon evaluation returns a label drawn independently from $\boldsymbol{p}_J$, regardless of the input example and of past evaluations. By construction, $L_{\mathcal{D}_J}^{\boldsymbol{w}}(h_J) = \boldsymbol{w}(\boldsymbol{p}_J, \boldsymbol{q}_J) \leq \boldsymbol{z}$.

The predictor thus proceeds as follows: after computing J_x, it returns h_J where $J = J_x$. The value of $L_\mathcal{D}^{\boldsymbol{w}}(h)$ then satisfies:

$$L_\mathcal{D}^{\boldsymbol{w}}(h) = \sum_{J \in \mathrm{supp}(J_x)} \mathbb{P}_{(x,y)\sim\mathcal{D}}(J_x = J) \cdot L_{\mathcal{D}_J}^{\boldsymbol{w}}(h_J) \leq \boldsymbol{z} \,.$$

Wrap-up. Recalling the assumptions above (the failure probability δ and the $\epsilon\,\mathbf{1}$ charged to the expected loss), we conclude that with probability $1-\delta$ the training phase produces a randomized predictor h such that $L_{\mathcal{D}}^{\boldsymbol{w}}(h) \le \boldsymbol{z} + \epsilon \cdot \mathbf{1}$. The algorithm described is therefore a $(\boldsymbol{w}, \boldsymbol{z})$-learner for $\mathcal{H}$, as desired. Summing the sample complexities completes the proof. ■

We additionally remark that we can remove the $1/\epsilon^2$ term in the case of learning with a single objective, as shown by the following result.

Corollary 13 *Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, let $w : \mathcal{Y} \times \mathcal{Y} \to [0,1]$ be a single-objective loss, and let $z \in [0,1]$. Moreover, let $\mathcal{R}$ be a family of PAC eliminators w.r.t. $\mathcal{H}$ that is sufficient for (w, z). Then there exists a (w, z)-learner for $\mathcal{H}$ with sample complexity*

$$\mathcal{O}\left(m_{\mathcal{R}}\left(\frac{\epsilon}{2|\mathcal{R}|}, \frac{\delta}{2|\mathcal{R}|}\right)\right)$$

where $m_{\mathcal{R}}$ is any upper bound on the sample complexities of the eliminators in $\mathcal{R}$.

Proof The statement follows from the proof of Theorem 12 with the only difference that we do not have to estimate the marginal distribution q on $\mathcal{Y}$. In particular, we proceed as before and compute a list J_x that with high probability is dice attainable. This means that for any marginal $\boldsymbol{q}$ on J_x there exists a distribution $\boldsymbol{p}$ such that $w(\boldsymbol{p}, \boldsymbol{q}) \le z$. By von Neumann's minimax theorem (von Neumann, 1928) we thus know that the minimax distribution $\boldsymbol{p}^*$ achieving $\min_{\boldsymbol{p}} \max_{\boldsymbol{q}} w(\boldsymbol{p}, \boldsymbol{q})$ also satisfies $w(\boldsymbol{p}^*, \boldsymbol{q}) \le z$ for all $\boldsymbol{q}$. ■

4. Population-Driven Cost

This section gives the key insights behind our results for the population-driven cost $\boldsymbol{w}^{\mathrm{p}}$, namely Theorem 6 and Theorem 7 (Section 1.1). Note that those results mirror closely their counterparts for general losses $\boldsymbol{w}$, that is, Theorem 4 and Theorem 5. The key difference is that, when $\boldsymbol{w} = \boldsymbol{w}^{\mathrm{p}}$, both the existential characterization of $(\boldsymbol{w}^{\mathrm{p}}, \boldsymbol{z})$-learnability and the corresponding sample complexity bounds have a dependence on $k = |\mathcal{Y}|$ of only $\widetilde{\mathcal{O}}(k^2)$, rather than $\widetilde{\mathcal{O}}(2^k)$. The purpose of this section is to illustrate how this improved dependence arises. The crux is the following result:

Lemma 14 *Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a hypothesis class and $\boldsymbol{z} \in [0,1]^k$. For every $J \subseteq \mathcal{Y}$ with $|J| \ge 2$ the following claims are equivalent:*

(1) $\boldsymbol{z} \in D_J(\boldsymbol{w}^{\mathrm{p}})$,

(2) $\boldsymbol{z} \in D_{\{i,j\}}(\boldsymbol{w}^{\mathrm{p}})$, *for all* $\{i,j\} \in \binom{J}{2}$,

(3) $\sqrt{z_i} + \sqrt{z_j} \ge 1$, *for all* $\{i,j\} \in \binom{J}{2}$.

To appreciate Lemma 14, recall the characterization of the learnable region $\mathscr{L}_{\boldsymbol{w}}(\mathcal{H})$ in terms of the dice-attainable regions $D_J(\boldsymbol{w})$ given by Theorem 4. For $\boldsymbol{w} = \boldsymbol{w}^{\mathrm{p}}$, Lemma 14 gives two key insights. First, every dice-attainable region $D_J(\boldsymbol{w}^{\mathrm{p}})$ can be "decomposed" into simpler elements—the dice-attainable regions of label pairs $D_{\{i,j\}}(\boldsymbol{w}^{\mathrm{p}})$. This is not true in general: one can devise simple examples of cost functions $\boldsymbol{w}$ for which this equivalence fails (see Section 5). Second, every $D_J(\boldsymbol{w}^{\mathrm{p}})$ in fact admits a simple algebraic characterization through quadratic inequalities: indeed, $D_J(\boldsymbol{w}^{\mathrm{p}})$ is just the set of all $\boldsymbol{z} \in [0,1]^k$ such that $\sqrt{z_i} + \sqrt{z_j} \ge 1$ for all distinct $i, j \in J$.

As an immediate corollary of Lemma 14, every $J \subseteq \mathcal{Y}$ with $|J| \geq 2$ satisfies:

$$D_J(\boldsymbol{w}^{\mathrm{p}}) = \bigcap_{\{i,j\} \in \binom{J}{2}} D_{\{i,j\}}(\boldsymbol{w}^{\mathrm{p}}) \ .$$

Together with Theorem 4, this proves Theorem 6. To conclude the proof of Theorem 7 too, we need to rely once more on Theorem 12 and further observe that, given some $\boldsymbol{z}$, we can always have a family $\mathcal{R}$ of PAC eliminators with $|\mathcal{R}| \leq k^2$ that is sufficient for $(\boldsymbol{w}^{\mathrm{p}}, \boldsymbol{z})$. This argument is made more formal in the full proof of Theorem 7, see Appendix D. The proof of Lemma 14 can be found in Appendix D, too, and employs some techniques and arguments from Bressan et al. (2025), including dualities between multi-objective and scalar costs and von Neumann's minimax theorem (von Neumann, 1928) applied to a label-prediction game over sets of labels J.

5. Discussion and Open Problems

We briefly discuss related learning settings, combinatorial dimensions, and open problems.

Related combinatorial dimensions. The J-cube dimension can be naturally related to other combinatorial parameters for binary and multiclass PAC learning. In the binary case ($J = \mathcal{Y}$ with $|\mathcal{Y}| = 2$) PAC J-eliminability and PAC learning are equivalent, as we already discussed. Moreover, in both cases the optimal sample complexity $\Theta\left(\frac{d+\log 1/\delta}{\epsilon}\right)$ is achieved by the one-inclusion graph algorithm, or our variant of it (Haussler et al., 1994; Aden-Ali et al., 2023a). In the multiclass case with a finite label space ($|\mathcal{Y}| < \infty$) the Natarajan dimension (Natarajan, 1989) characterizes learnability. Recall that a set $S \subseteq \mathcal{X}$ is N-shattered by a class $\mathcal{H}$ if there exists $f, f' \in \mathcal{Y}^S$ such that $f(s) \neq f'(s)$ for all $s \in S$, and for each $S' \subseteq S$ there exists an $h \in \mathcal{H}$ such that $h_{|S'} = f_{|S'}$ and $h_{|S \setminus S'} = f'_{|S \setminus S'}$. The Natarajan dimension $d_N(\mathcal{H})$ of $\mathcal{H}$ is the size of the largest set N-shattered by $\mathcal{H}$. It is easy to see that a set that is J-shattered with $|J| \geq 2$ is also N-shattered, implying $d_N(\mathcal{H}) \geq \max_J d_J(\mathcal{H})$. Hence, any class that is not J-eliminable for some J is also not PAC learnable. The other direction does not hold in general: N-shattered sets are typically not J-shattered for some J. However, an infinite Natarajan dimension implies an infinite J-cube dimension for some J with $|J| = 2$, see Proposition 41 in Appendix E.

Intuitively this means that if class if not learnable, then there exists at least a single pair of labels that are difficult to distinguish. Similar statements can be made about related parameters such as the graph dimension (Ben-David et al., 1995). We do not fully recover the sample complexity bound $\mathcal{O}\big(\frac{d_N(\mathcal{H}) \log(1/\epsilon) \log |\mathcal{Y}| + \log(1/\delta)}{\epsilon}\big)$ for finite $d_N(\mathcal{H})$ (Daniely et al., 2015). An inspection of Theorem 12 reveals that we get a dependence of $\mathcal{O}\big(2^{|\mathcal{Y}|}\big)$ for general multi-objective losses (see also the list of open problems at the end of this section). However, for unweighted classification loss this dependence drops to $|\mathcal{Y}|^2$ (see the discussion on list learning below). For multi-objective losses we also have an additional $1/\epsilon^2$ term, which however is not needed in the single-objective case (see Corollary 13) nearly matching the d_N-based bound. General multiclass learning with arbitrary (potentially infinite) $\mathcal{Y}$ is characterized by the DS-dimension $d_{\mathrm{DS}}(\mathcal{H})$ (Brukhim et al., 2022). Here we see again that $d_{\mathrm{DS}}(\mathcal{H}) \geq \max_J d_J(\mathcal{H})$. In particular, any J-shattered set S yields an $|S|$-pseudo-cube—see Brukhim et al. (2022) for the definition of pseudo-cube. In contrast to the Natarajan dimension, an infinite DS-dimension does not imply that there is some J with infinite J-dimension. Brukhim et al. (2022) show a class with infinite DS-dimension and Natarajan dimension equal to 1.

By the above discussion we thus have $d_J(\mathcal{H}) \leq d_N(\mathcal{H}) = 1$ for all J. PAC eliminability with infinite J is not characterized by the finiteness of the J-cube dimension. In particular, as mentioned in Section 2, PAC eliminability is always possible with error $1/\epsilon$ for infinite J. In terms of sample complexity, achieving a bound of $\mathcal{O}\big(\frac{d_{\mathrm{DS}}(\mathcal{H})+\log(1/\delta)}{\epsilon}\big)$ is open (Brukhim et al., 2022; Aden-Ali et al., 2023a; Hanneke et al., 2024). In particular, all known DS-based bounds contain additional $\log(1/\epsilon)$ terms. In contrast, while we do not have these additional terms, we have a potentially exponential dependence on $|\mathcal{Y}|$.

List learning. Theorem 4 provides a novel characterization of list learning. Recall that ℓ-list PAC learning (Brukhim et al., 2022; Charikar and Pabbaraju, 2023) is a variant of multiclass PAC learning where the learner can return a list of up to ℓ labels from $\mathcal{Y}$ with $\ell < |\mathcal{Y}|$. The prediction of such a list learner is correct if the list contains the correct label and is wrong otherwise.

Corollary 15 *A class $\mathcal{H}$ is ℓ-list PAC learnable if and only if $d_J(\mathcal{H}) < \infty$ for all $J \subseteq \mathcal{Y}$ with $|J| = \ell + 1$.*

Proof Denote by w the unweighted zero-one loss. Fix $\ell < |\mathcal{Y}|$. Brukhim et al. (2023) establishes that the existence of an ℓ-list learner for $\mathcal{H}$ is equivalent to the existence of a (w, z)-learner for $\mathcal{H}$ with $z \in \left(\frac{\ell-1}{\ell}, \frac{\ell}{\ell+1}\right]$. Let $J \subseteq \mathcal{Y}$. It holds that (w, z) is J-dice attainable if and only if $z \geq \frac{|J|-1}{|J|}$, see also Bressan et al. (2025). Thus, $z \in D_J(w)$ if and only if $|J| < \ell + 1$, i.e., $|J| \leq \ell$. Thus $\mathcal{J}(w, z) = \{J \subseteq \mathcal{Y} : z \notin D_J(w)\} = \{J \subseteq \mathcal{Y} : |J| > \ell\}$. Now, by Theorem 4 we know that the learnability of (w, z) is equivalent to $d_J(\mathcal{H}) < \infty$ for all $J \in \mathcal{J}(w, z)$. Finally note that $d_J(\mathcal{H}) \geq d_{J'}(\mathcal{H})$ for all $J \subseteq J'$. The claim follows. ■

Necessity to examine all $J \subseteq \mathcal{Y}$ for weighted losses. We remark that, although Theorem 6 shows that for the population-driven cost $\boldsymbol{w}^{\mathrm{P}}$ it suffices to examine only *pairs* of labels, that is not the case for general $\boldsymbol{w}$. It can be necessary to inspect all subsets $J \subseteq \mathcal{Y}$ and their J-cube dimensions to determine $(\boldsymbol{w}, \boldsymbol{z})$-learnability. Consider the following example.

Example 1 *Let $\mathcal{J} \subsetneq 2^{\mathcal{Y}}$ be an arbitrary subset-closed family (e.g., $\mathcal{J} = \binom{\mathcal{Y}}{2}$). For each maximal $J \in \mathcal{J}$, let S_J be an infinite set and let $\mathcal{X}$ be the disjoint union of all such S_J. Define $\mathcal{H}$ as the disjoint union of all J^{S_J} for each maximal J. Now let $J' \subseteq \mathcal{Y}$ such that $J' = J \cup \{y\}$ for a maximal $J \in \mathcal{J}$ and a $y \in \mathcal{Y} \setminus J$, and define $\mathcal{H}'$ as $\mathcal{H} \cup (J')^{S_J}$. From this definition we see that $d_J(\mathcal{H}) = d_J(\mathcal{H}') = \infty$ for all $J \in \mathcal{J}$. For J' we have $d_{J'}(\mathcal{H}') = \infty$ while $d_{J'}(\mathcal{H}) < \infty$. This means that for a $\boldsymbol{w}$ and $\boldsymbol{z}$ with $J' \in \mathcal{J}(\boldsymbol{w}, \boldsymbol{z})$ it holds that $\mathcal{H}$ has a $(\boldsymbol{w}, \boldsymbol{z})$-learner, while $\mathcal{H}'$ does not (see Theorem 4).*

Open problems. A main open problem is to remove the potentially exponential dependence on $|\mathcal{Y}| = k$ in the upper bound of Theorem 5; or showing that this is unavoidable in general. The same bound contains a $1/\epsilon^2$ term. This term is unavoidable in our approach—where we estimate the marginal distribution $\boldsymbol{q}$ in total variation distance—but it could perhaps be improved via a different technique. Note that when considering a single weighted loss function, this term is not required as shown by Corollary 13. Finally, it would be interesting to find characterizations beyond the J-cube dimensions that allow to fully recover known multiclass sample complexity bounds. For example, the multiplicative gap of $|J|$ in Theorem 10 is unavoidable using the J-cube dimension. A characterization removing this gap would be interesting.

6. Further Related Work

As mentioned in the introduction, our notion of partial learnability is directly connected to multi-objective learning with cost-sensitive multiclass loss functions, which is significantly different from other notions of partial learnability studied in the past. For example, the learnability of partial concept classes (Alon et al., 2022) or the partial learning of recursively enumerable languages (Gao et al., 2016).

Although cost-sensitive learning (Ling and Sheng, 2008) is a topic firmly rooted in machine learning, a rigorous analysis of the achievable guarantees is missing. The seminal work of Elkan (2001) characterizes the Bayes-optimal predictor for cost-sensitive prediction in the binary case and shows applications to decision tree learning. Zadrozny et al. (2003) study a very general cost-sensitive PAC learning model where the misclassification cost can also depend on the individual data point, and provide sample size bounds for the cost-proportionate rejection sampling algorithm. Zhou and Liu (2010) consider the cost-sensitive multiclass setting from a mostly empirical perspective. Scott (2012) studies the problem of surrogate loss consistency for cost-sensitive losses in binary classification and provides risk bounds. This is extended to a setting with label noise by Natarajan et al. (2018). Cost-sensitive learning is also key to the analysis of precision-recall trade-offs (Puthiya Parambath et al., 2014) and to the analysis of class-imbalanced classification (Xu et al., 2020). Multi-objective machine learning (MOL)—see, e.g., Jin (2007); Jin and Sendhoff (2008)—is also a popular topic that draws from the rich body of literature in multi-criteria optimization (Ehrgott, 2005). MOL is often studied in the context of multi-task learning (Lin et al., 2019; Sener and Koltun, 2018), also focusing on the limits of scalarization techniques (Hu et al., 2024; Súkeník and Lampert, 2024). However, none of these works provide general characterizations of multi-objective learnability in a PAC-style statistical framework. Mannor et al. (2014) explored connections between MOL and Blackwell approachability, whereas more recent studies have considered MOL within the framework of online convex optimization (Jiang et al., 2023). A different angle to MOL is provided by multicalibration (Hébert-Johnson et al., 2018), where one seeks multiclass predictors that are calibrated on specific subsets of the learning domain (Haghtalab et al., 2024). Online versions of multicalibration are connected to online MOL (Lee et al., 2022).

Acknowledgments

MB, NCB, and EE are partially supported by the MUR PRIN grant 2022EKNE5K (Learning in Markets and Society), the FAIR (Future Artificial Intelligence Research) project, funded by the NextGenerationEU program within the PNRR-PE-AI scheme, the EU Horizon CL4-2022-HUMAN-02 research and innovation action under grant agreement 101120237, project ELIAS (European Lighthouse of AI for Sustainability), and the One Health Action Hub, University Task Force for the resilience of territorial ecosystems, funded by Università degli Studi di Milano (PSR 2021-GSA-Linea 6).

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

YM has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 882396), by the Israel Science Foundation, the Yandex Initiative for Machine Learning at Tel Aviv University and a grant from the Tel Aviv University Center for AI and Data Science (TAD).

MT acknowledges support from a DOC fellowship of the Austrian academy of sciences (ÖAW) and travel support from ELISE (GA no 951847).

References

Ishaq Aden-Ali, Yeshwanth Cherapanamjeri, Abhishek Shetty, and Nikita Zhivotovskiy. Optimal PAC bounds without uniform convergence. In *Symposium on Foundations of Computer Science (FOCS)*, 2023a.

Ishaq Aden-Ali, Yeshwanth Cherapanamjeri, Abhishek Shetty, and Nikita Zhivotovskiy. The one-inclusion graph algorithm is not always optimal. In *Conference on Learning Theory (COLT)*, 2023b.

Noga Alon, Steve Hanneke, Ron Holzman, and Shay Moran. A theory of pac learnability of partial concept classes. In *Symposium on Foundations of Computer Science (FOCS)*, 2022.

Martin Anthony and Peter L Bartlett. *Neural network learning: Theoretical foundations*. Cambridge University Press, 1999.

Shai Ben-David, Nicolò Cesa-Bianchi, David Haussler, and Philip M Long. Characterizations of learnability for classes of $\{0, \ldots, n\}$-valued functions. *Journal of Computer and System Sciences*, 50(1):74–86, 1995.

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

Marco Bressan, Nataly Brukhim, Nicolò Cesa-Bianchi, Emmanuel Esposito, Yishay Mansour, Shay Moran, and Maximilian Thiessen. Of dice and games: A theory of generalized boosting. In *Conference on Learning Theory (COLT)*, 2025.

Nataly Brukhim, Daniel Carmon, Irit Dinur, Shay Moran, and Amir Yehudayoff. A characterization of multiclass learnability. In *Symposium on Foundations of Computer Science (FOCS)*, 2022.

Nataly Brukhim, Amit Daniely, Yishay Mansour, and Shay Moran. Multiclass boosting: simple and intuitive weak learning criteria. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Clément L Canonne. A short note on learning discrete distributions. *arXiv preprint arXiv:2002.11457*, 2020.

Moses Charikar and Chirag Pabbaraju. A characterization of list learnability. In *Symposium on Theory of Computing (STOC)*, 2023.

Amit Daniely and Shai Shalev-Shwartz. Optimal learners for multiclass problems. In *Conference on Learning Theory (COLT)*, 2014.

Amit Daniely, Sivan Sabato, Shai Ben-David, and Shai Shalev-Shwartz. Multiclass learnability and the ERM principle. *The Journal of Machine Learning Research*, 16:2377–2404, 2015.

G. B. Dantzig and D. R. Fulkerson. *12. On the Max-Flow Min-Cut Theorem of Networks*, pages 215–221. Princeton University Press, Princeton, 1956.

Matthias Ehrgott. *Multicriteria optimization*, volume 491. Springer Science & Business Media, 2005.

Charles Elkan. The foundations of cost-sensitive learning. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2001.

Ziyuan Gao, Frank Stephan, and Sandra Zilles. Partial learning of recursively enumerable languages. *Theoretical Computer Science*, 620:15–32, 2016.

Nika Haghtalab, Michael Jordan, and Eric Zhao. A unifying perspective on multi-calibration: Game dynamics for multi-objective learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2024.

Steve Hanneke, Shay Moran, and Qian Zhang. Improved sample complexity for multiclass PAC learning. In *Advances on Neural Information Processing Systems (NeurIPS)*, 2024.

David Haussler, Nick Littlestone, and Manfred K Warmuth. Predicting $\{0,1\}$-functions on randomly drawn points. In *Conference on Learning Theory (COLT)*, pages 280–296, 1988.

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

Ursula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning (ICML)*, 2018.

Yuzheng Hu, Ruicheng Xian, Qilong Wu, Qiuling Fan, Lang Yin, and Han Zhao. Revisiting scalarization in multi-task learning: A theoretical perspective. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2024.

Jiyan Jiang, Wenpeng Zhang, Shiji Zhou, Lihong Gu, Xiaodong Zeng, and Wenwu Zhu. Multi-objective online learning. In *International Conference on Learning Representations (ICLR)*, 2023.

Yaochu Jin. *Multi-objective machine learning*. Springer Science & Business Media, 2007.

Yaochu Jin and Bernhard Sendhoff. Pareto-based multiobjective machine learning: An overview and case studies. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(3):397–415, 2008.

Daniel Lee, Georgy Noarov, Mallesh Pai, and Aaron Roth. Online minimax multiobjective optimization: Multicalibeating and other applications. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:29051–29063, 2022.

Xi Lin, Hui-Ling Zhen, Zhenhua Li, Qing-Fu Zhang, and Sam Kwong. Pareto multi-task learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019.

Charles X Ling and Victor S Sheng. Cost-sensitive learning and the class imbalance problem. *Encyclopedia of Machine Learning*, 2011:231–235, 2008.

Shie Mannor, Vianney Perchet, and Gilles Stoltz. Approachability in unknown games: Online learning meets multi-objective optimization. In *Conference on Learning Theory (COLT)*, pages 339–355. PMLR, 2014.

Balas K Natarajan. On learning sets and functions. *Machine Learning*, 4(1):67–97, 1989.

Balas K. Natarajan and Prasad Tadepalli. Two new frameworks for learning. In *International Conference on Machine Learning (ICML)*, 1988.

Nagarajan Natarajan, Inderjit S Dhillon, Pradeep Ravikumar, and Ambuj Tewari. Cost-sensitive learning with noisy labels. *Journal of Machine Learning Research*, 18(155):1–33, 2018.

Shameem Puthiya Parambath, Nicolas Usunier, and Yves Grandvalet. Optimizing F-measures by cost-sensitive classification. *Advances in Neural Information Processing Systems (NeurIPS)*, 27, 2014.

Clayton Scott. Calibrated asymmetric surrogate losses. *Electronic Journal of Statistics*, 6:958–992, 2012.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. *Advances in Neural Information Processing Systems (NeurIPS)*, 31, 2018.

Peter Súkeník and Christoph Lampert. Generalization in multi-objective machine learning. *Neural Computing and Applications*, pages 1–15, 2024.

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*. Nauka, Moscow, 1974.

John von Neumann. Zur Theorie der Gesellschaftsspiele. *Mathematische Annalen*, 100(1):295–320, 1928.

Ziyu Xu, Chen Dan, Justin Khim, and Pradeep Ravikumar. Class-weighted classification: Trade-offs and robust approaches. In *International Conference on Machine Learning (ICML)*, 2020.

Andrew Chi-Chin Yao. Probabilistic computations: Toward a unified measure of complexity. In *Symposium on Foundations of Computer Science (FOCS)*, 1977.

Bianca Zadrozny, John Langford, and Naoki Abe. Cost-sensitive learning by cost-proportionate example weighting. In *International Conference on Data Minin (ICDM)*, 2003.

Zhi-Hua Zhou and Xu-Ying Liu. On multi-class cost-sensitive learning. *Computational Intelligence*, 26(3):232–257, 2010.

Appendix A. Learning to Eliminate

We prove the following bounds on the PAC elimination sample complexity, which we restate for convenience.

Theorem 10 (PAC J-Elimination Sample Complexity) *Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a hypothesis class, and let $J \subseteq \mathcal{Y}$ be such that $2 \leq |J| < \infty$. The PAC J-elimination sample complexity $m^J_{\mathcal{H}}$ of $\mathcal{H}$ satisfies*

$$\Omega\left(\frac{d_J(\mathcal{H}) + \log(1/\delta)}{|J|\epsilon}\right) \leq m^J_{\mathcal{H}}(\epsilon, \delta) \leq \mathcal{O}\left(\frac{d_J(\mathcal{H}) + \log(1/\delta)}{\epsilon}\right).$$

We will prove Theorem 10 by first providing bounds that hold only in expectation.

Definition 16 (Elimination in Expectation) *Let $J \subseteq \mathcal{Y}$ such that $|J| \geq 2$. We say that J can be* eliminatable in expectation *by a class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ if there exists a learning rule A and a sample complexity $m^{\texttt{expt}}_{J,\mathcal{H}} : (0,1) \to \mathbb{N}$ such that for all $\epsilon \in (0,1)$, all realizable distributions $\mathcal{D}$, and all sample sizes $m \geq m^{\texttt{expt}}_{J,\mathcal{H}}(\epsilon)$ it holds that*

$$\mathop{\mathbb{E}}_{S \sim D^m}[L^J_{\mathcal{D}}(A(S))] \leq \epsilon\,.$$

The minimal such sample complexity $m^{\texttt{expt}}_{J,\mathcal{H}}(\epsilon)$ is called the J-elimination in-expectation sample complexity of $\mathcal{H}$.

Theorem 17 (In-Expectation Sample Complexity) *Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a hypothesis class, and let $J \subseteq \mathcal{Y}$ be such that $2 \leq |J| < \infty$. The J-elimination in-expectation sample complexity $m^{\texttt{expt}}_{J,\mathcal{H}}(\epsilon)$ satisfies*

$$\Omega\left(\frac{d_J(\mathcal{H})}{|J|\epsilon}\right) \leq m^{\texttt{expt}}_{J,\mathcal{H}}(\epsilon) \leq \mathcal{O}\left(\frac{d_J(\mathcal{H})}{\epsilon}\right).$$

Furthermore, the lower and upper bound is tight for specific families of hypothesis spaces.

We adapt standard no-free lunch type arguments (e.g., Blumer et al. (1989)) to prove the lower bound of Theorem 17 and transform it into the high probability PAC lower bound of Theorem 10. The details are in Appendix A.1. For the upper bound we adapt another common technique based on the leave-one-out error of the one-inclusion graph predictor (Haussler et al., 1994). Relying on the recent reverse online-to-batch technique from (Aden-Ali et al., 2023a), we then get the required high probability PAC upper bound. For full details see Appendix A.2. The tightness is shown in Appendix B.

A.1. Lower Bound

We prove the two bounds in Theorem 17 separately and start with the following no-free-lunch based lower bound.

Theorem 18 *Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a hypothesis class, and let $J \subseteq \mathcal{Y}$ be such that $2 \leq |J| < \infty$ and $d_J(\mathcal{H}) \geq 2$. Then, $m^{\texttt{expt}}_{J,\mathcal{H}}(\epsilon) = \Omega\left(\frac{d_J(\mathcal{H})}{|J|\epsilon}\right)$.*

Proof The proof follows by adapting the no-free-lunch theorem for the binary case (Vapnik and Chervonenkis, 1974; Blumer et al., 1989) to the setting of elimination. Let $d = d_J(\mathcal{H}) \geq 2$, let $X = \{x_1, \ldots, x_d\} \subseteq \mathcal{X}$ be J-shattered by $\mathcal{H}$ and let m be an arbitrary integer satisfying $m \geq \frac{d-1}{4}$. We will show that there exists a distribution leading to an error of at least $\Omega\left(\frac{d}{m|J|}\right)$ in expectation. Rearranging terms yields the claim.

Define the marginal distribution $\mathcal{D}_X^m$ to assign probability $1 - \frac{d-1}{4m}$ to x_1 and $\frac{1}{4m}$ to each x_i for $i > 1$. Let the random concept c be chosen uniformly at random from the $|J|^d$ concepts in $J^X \subseteq \mathcal{H}$.

As in the classic proof, if a learner A receives m examples then, with constant probability, its training set will omit a constant fraction of the examples in $X_{-1} = \{(x_i, c(x_i)) : i > 1\}$. In particular, let Z be the number of examples received by A that belong to X_{-1}. The expectation of Z is $\mathbb{E}[Z] = \frac{(d-1)}{4}$ and hence by a Chernoff bound we have

$$\mathbb{P}\left(Z \geq 2\mathbb{E}[Z]\right) \leq e^{-\frac{(d-1)}{12}} .$$

That is, with probability at least $1 - e^{-\frac{(d-1)}{12}} \geq 1 - e^{-\frac{1}{12}} > 0.07$, at least half (i.e., $\frac{d-1}{2}$) of the points from X_{-1} are omitted in the training sample. On this event, the test point $(x, c(x))$ is one of these omitted points with probability at least $\frac{d-1}{8m}$ (each point with probability $\frac{1}{4m}$). Now, as X is J-shattered and $y = c(x)$ is drawn uniformly at random from J, the probability of selecting y for elimination is at least $\frac{1}{|J|}$. Thus any algorithm wrongly eliminates y (and thus gets loss 1) with probability at least $\frac{1}{|J|}$. Note that as the random concept c is independent of the learner A, there exists a deterministic yet algorithm dependent concept c_A achieving the same loss; this follows by, for example, Yao's minimax principle (Yao, 1977). Thus $\mathcal{D}_X^m$ and c_A together form the $\mathcal{H}$-realizable distribution yielding the lower bound for any particular learner A. Combining these probabilities, the expected error satisfies:

$$\epsilon = \Omega\left(\frac{d}{m|J|}\right) .$$

■

By adapting standard PAC lower bounds (Anthony and Bartlett, 1999) to the elimination setting the lower bound in Theorem 10 follows.

Theorem 19 *Let $J \subseteq \mathcal{Y}$ with $|J| \geq 2$ and $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a hypothesis space with $d_J(\mathcal{H}) \geq 2$. The PAC J-elimination sample complexity of any learner satisfies $m_{\mathcal{H}}^J(\epsilon, \delta) = \Omega\left(\frac{d_J(\mathcal{H}) + \log(1/\delta)}{|J|\epsilon}\right)$.*

Proof The proof of Theorem 18 yields a lower bound of $m_{\mathcal{H}}^J(\epsilon, \delta) = \Omega\left(\frac{d_J(\mathcal{H})}{|J|\epsilon}\right)$ for $\delta < 0.07$. We get the missing $\Omega\left(\frac{\log(1/\delta)}{|J|\epsilon}\right)$ additive term by adapting standard PAC lower bounds (e.g., Anthony and Bartlett (1999)). Let $J = \{z_1, \ldots, z_k\}$. Let $\epsilon > 0$ and $\epsilon' = \epsilon k$. As $d_J(\mathcal{H}) \geq 2$ there exists $a, b \in \mathcal{X}$ and hypotheses $h_1, \ldots, h_k \in \mathcal{H}$ such that $h_1(a) = \cdots = h_k(a) = 1$ and $h_i(b) = z_i$ for $i \in [k]$. Let the ground truth c be chosen uniformly at random from the hypotheses $\{h_1, \ldots, h_k\}$. Consider the distribution $\mathcal{D}$ with $\mathcal{D}(a) = 1 - \epsilon'$ and $\mathcal{D}(b) = \epsilon$. Let $m \leq \frac{\log(1/\delta)}{2\epsilon'}$ and note that an iid. sample from $\mathcal{D}$ of size m contains no b with probability greater than δ; call such this sample S_a. Let A be a learning algorithm and $h = A(S_a)$ be the returned predictor returned by A on the sample S_a. With probability ϵ' the test point is b and h will err with probability at least $1/k$, as $c(b)$

is drawn uniformly at random from $[k]$. Thus, with probability at least δ (over the iid sample), the learner A will output a predictor with expected J-loss at least $\frac{\epsilon'}{k} = \epsilon$ and hence will fail the PAC elimination requirement. ■

A.2. Upper Bound

This direction is more challenging, as classical principles do not extend to the setting of learning to eliminate. Empirical Risk Minimization (ERM) is not well defined here, nor is the broader notion of proper learning, since the goal is to eliminate an incorrect label rather than predict the correct one. For example, let $y \in J$ be a label that is never predicted by any $h \in \mathcal{H}$. A trivial J-eliminator would always eliminate y and be always correct, however, no hypothesis in $\mathcal{H}$, in particular no consistent one, would ever pick this label and thus potentially get a large J-elimination loss. However, an adaptation of the One-Inclusion Graph algorithm provides a solution, which we explain next.

For readers familiar with the One-Inclusion Graph algorithm in the classical PAC setting, we highlight key differences and challenges in this adaptation. In the classical case, the performance of the One-Inclusion Predictor is tied to the out-degree of orientations in the graph. In contrast, the performance of our One-Inclusion Eliminator depends on the in-degree. Additionally, we develop J-specific variants of edge density and shifting to address the requirements of elimination tasks. Remarkably, unlike the classical setting, the simpler approach of greedily peeling a vertex of minimum degree, as used by Haussler et al. (1988); Daniely and Shalev-Shwartz (2014), does not seem to work in the general case. Instead, we employ a max-flow min-cut argument, similar to Haussler et al. (1994) (journal version of Haussler et al. (1988)), which applies to general hypergraphs (see Proposition 31).

A.2.1. Expected and Leave-One-Out Error

Let A be a learning rule and let $\mathcal{D}$ be a distribution. The expected error of A on a random sample $S \sim \mathcal{D}^n$ is

$$\epsilon_n(A; \mathcal{D}) = \mathop{\mathbb{E}}_{S \sim \mathcal{D}^n} L_D^J[A(S)]\,.$$

For every $n \in \mathbb{N}$, define

$$\epsilon_n(\mathcal{H}) = \inf_A \sup_{\mathcal{D}} \epsilon_n(A; \mathcal{D}),$$

where A ranges over all learning rules and $\mathcal{D}$ ranges over all $\mathcal{H}$-realizable distributions.

Let $S = \{(x_i, y_i)\}_{i=1}^n$ be a sequence of examples. The *leave-one-out* loss of A with respect to S is defined as:

$$\epsilon^{\texttt{LOO}}(A; S) = \mathop{\mathbb{E}}_{i \sim \mathcal{U}_{[n]}} [\ell^J_{(x_i,y_i)}(A(S_{-i}))],$$

where S_{-i} is the sample obtained by omitting the i-th example from S. Thus, the test point (x_i, y_i) is selected uniformly at random from S, and the learner observes the examples (x_j, y_j) for all $j \neq i$. The learner is then tasked with eliminating a label $\widehat{y}_i \in J$ such that $\widehat{y}_i \neq y_i$.

Definition 20 (Leave-One-Out Learning Rate) *For every $n \in \mathbb{N}$ define:*

$$\epsilon^{\texttt{LOO}}_{\mathcal{H}}(n) = \inf_A \sup_S \epsilon^{\texttt{LOO}}(A; S),$$

where A ranges over all learning rules and S ranges over all realizable sequences of size n.

It is well known that the leave-one-out error and the expected error are related through a symmetrization argument:

Proposition 21 *For every n, it holds that*

$$\epsilon_n(\mathcal{H}) \leq \epsilon_{\mathcal{H}}^{\texttt{LOO}}(n+1).$$

Proof Let A be a learning rule with a leave-one-out rate at most $\epsilon_{\mathcal{H}}^{\texttt{LOO}}(n+1)+\tau$, where $\tau > 0$ is arbitrarily small, and let $\mathcal{D}$ be an $\mathcal{H}$-realizable distribution. The expected loss of A with respect to $\mathcal{D}$ can be expressed as:

$$\begin{aligned}
\mathop{\mathbb{E}}_{S\sim\mathcal{D}^n}[L_{\mathcal{D}}^J(A(S))] &= \mathop{\mathbb{E}}_{S\sim\mathcal{D}^n,(x,y)\sim\mathcal{D}}\left[\ell_{(x,y)}^J(A(S))\right] \\
&= \mathop{\mathbb{E}}_{S'\sim\mathcal{D}^{n+1}}\left[\mathop{\mathbb{E}}_{i\sim\mathcal{U}_{[n+1]}}\left[\ell_{(x_i,y_i)}^J(A(S'_{-i}))\right]\right] && \text{(exchangeability)} \\
&= \mathop{\mathbb{E}}_{S'\sim\mathcal{D}^{n+1}}[\epsilon^{\texttt{LOO}}(A;S')] \\
&\leq \mathop{\mathbb{E}}_{S'\sim\mathcal{D}^{n+1}}[\epsilon_{\mathcal{H}}^{\texttt{LOO}}(n+1)+\tau] \\
&= \epsilon_{\mathcal{H}}^{\texttt{LOO}}(n+1)+\tau.
\end{aligned}$$

Since τ can be made arbitrarily small, the proof follows. ■

By Proposition 21, it is sufficient to prove that $\epsilon_{\mathcal{H}}^{\texttt{LOO}}(n) \leq \frac{d_J(\mathcal{H})}{n}$ to derive the upper bound in Theorem 17. We do this in the next section.

A.2.2. THE ONE-INCLUSION J-ELIMINATOR

In this section, we show that $\epsilon_{\mathcal{H}}^{\texttt{LOO}}(n) \leq \frac{d_J(\mathcal{H})}{n}$ using a variant of the One-Inclusion Graph Predictor from the classical PAC setting. We term this variant the One-Inclusion Graph Eliminator.

Definition 22 (One-Inclusion Hypergraph) *Let $\underline{S} = \{x_1,\dots,x_n\}$ be a multiset of n unlabeled examples (the order does not matter, but repetitions are counted). The one-inclusion hypergraph $G(\underline{S}) = (V,E)$ is an n-regular hypergraph defined as follows:*

- *The vertex set $V = \mathcal{H}|_{\underline{S}}$ consists of all $\mathcal{H}$-realizable hypotheses restricted to $\underline{S}$.*
- *Each $h \in \mathcal{H}$ is incident to exactly n hyperedges e_i, for $i = 1,\dots,n$, such that:*

  $$e_i = \{h' \in V : h'(x_j) = h(x_j) \textit{ for all } j \neq i\}.$$

  The index i is called the direction *of the edge e_i, so each vertex is incident to exactly one edge in every direction.*

We emphasize that singleton edges (self-loops) are allowed: e_i is a singleton edge if no $h' \neq h$ agrees with h on x_j for all $j \neq i$. Additionally, parallel edges are permitted: if e_i and e_j contain the same vertices as sets for $i \neq j$, they are treated as distinct edges corresponding to different directions. We have the following simple observation.

Proposition 23 *Let e_i, e'_j be two distinct hyperedges of a one-inclusion hypergraph. Then e_i and e'_j intersect in at most one vertex.*

Proof Let e_i and e'_j be distinct edges with $e_i \cap e'_j \neq \emptyset$. First assume $i \neq j$ and let $h_a, h_b \in e_i \cap e'_j$. As $h_a, h_b \in e_i$ they agree on all x_ℓ with $\ell \neq i$. As $h_a, h_b \in e_j$ they also agree on x_i. Hence $h_a = h_b$.

Now assume $i = j$. Without loss of generality there exists an $h_b \in e'_j \setminus e_i$. Let $h_a \in e_i \cap e'_j$. Note that h_b agrees with h_a on all $x \neq x_i = x_j$ and thus $h_b \in e_i$, leading to a contradiction. ■

Hence the only possible parallel edges are singleton edges. We next introduce the one-inclusion elimination rule. Let $S = \{(x_i, c(x_i))\}_{i=1}^n$ be an $\mathcal{H}$-realizable sequence of examples, and let $\underline{S} = \{x_1, \ldots, x_n\}$ be the induced multiset of unlabeled examples. Note that each possible input S_{-i} corresponds to an edge in the hypergraph $G(\underline{S}) = (V, E)$, namely, the edge in direction i that is incident to the vertex $c|_{\underline{S}}$. Thus, in the language of hypergraphs, a J-elimination rule corresponds to a map:

$$A : E \to J,$$

which assigns to each edge (test point) the label from J that A eliminates. Moreover, the error of A on S is given by:

$$\epsilon^{\texttt{LOO}}(A; S) = \frac{|\{i : A(e_i) = c(x_i)\}|}{n}, \tag{4}$$

where $e_1, \ldots, e_n$ are the n hyperedges incident to $c|_{\underline{S}}$. This motivates the following definition:

Definition 24 *Let $\mathcal{H}$ be a concept class, $\underline{S} = \{x_1, \ldots, x_n\}$ be a multiset, and $G = G(\underline{S}) = (V, E)$ be the corresponding one-inclusion hypergraph. For a mapping $A : E \to J$ and a vertex $c \in V$, define:*

$$\texttt{deg}_{\texttt{in}}(c; A) = |\{i : A(e_i) = c(x_i)\}|,$$

where $e_1, \ldots, e_n$ are the n hyperedges incident to c. Also, define the maximal in-degree:

$$\texttt{deg}_{\max}(A; G) = \max_{c \in V} \texttt{deg}_{\texttt{in}}(c; A).$$

We are now ready to introduce the one-inclusion elimination rule:

Algorithm: One-Inclusion J-Eliminator
Input: $\mathcal{H}$-realizable training set $S = \{(x_j, y_j)\}_{j \neq i}$, test point x_i.

Steps:

1. Compute the one-inclusion hypergraph $G(\underline{S}) = (V, E)$, where $\underline{S} = \{x_1, \ldots, x_n\}$ is the multiset corresponding to the unlabeled training examples and the test point.

2. Compute $A : E \to J$ that minimizes the maximal in-degree $\texttt{deg}_{\max}(A; G(\underline{S}))$.

3. Predict the eliminated label:

    - Identify the hyperedge $e_i = \{c \in V : c(x_j) = y_j \text{ for all } j \neq i\}$.
    - Set the predicted label $\widehat{y_i} = A(e_i)$.

Equation (4) implies the following bound on the leave-one-out error of the one-inclusion elimination rule:

Proposition 25 *Let $A^\star$ denote the one-inclusion elimination rule, and let $S = \{(x_i, y_i)\}_{i=1}^n$ be an $\mathcal{H}$-realizable sequence. Then,*

$$\epsilon^{\texttt{LOO}}(A^\star; S) \;\le\; \frac{\min_A \texttt{deg}_{\max}(A; G(\underline{S}))}{n},$$

where A ranges over all mappings $A : E \to J$, and E is the edge set of $G(\underline{S})$.

A.2.3. Bounding the Minimum Max-Degree

By Proposition 25, the upper bound follows from the following result:

Theorem 26 *For every concept class $\mathcal{H}$ and every one-inclusion hypergraph $G(\underline{S})$ corresponding to $\mathcal{H}$,*

$$\min_A \texttt{deg}_{\max}(A; G(\underline{S})) \;\le\; d_J(\mathcal{H}),$$

where A ranges over all mappings $A : E \to J$.

Toward this end, we follow the classical approach by Haussler et al. (1994) of relating $\min_A \texttt{deg}_{\max}(A; G(\underline{S}))$ with an appropriate notion of edge density, which we define next. First, we introduce the following notion of induced hypergraphs.

Definition 27 (Induced Hypergraph) *Let $G = (V, E)$ be a hypergraph and let $U \subseteq V$. The induced hypergraph on U is the hypergraph $G_U = (U, E_U)$ with $E_U = \{e \cap U : e \in E, e \cap U \neq \emptyset\}$.*

Let $G(\underline{S}) = (V, E)$ be the one-inclusion hypergraph of $\mathcal{H}$ with respect to $\underline{S}$, and let $U \subseteq V$. The induced hypergraph G_U is also a one-inclusion hypergraph, corresponding to the subclass $\mathcal{H}' \subseteq \mathcal{H}$ of concepts satisfying $c|_{\underline{S}} \in U$.

Definition 28 (J-Density) *Let $G = G(\underline{S}) = (V, E)$ be a one-inclusion hypergraph, and let $e_i \in E$ be an edge in direction i. We say that e_i is a J-*edge *if $J \subseteq \{h(x_i) : h \in e_i\}$. That is, e_i is a J-edge if, for every $j \in J$, there exists a vertex $h \in e_i$ such that $h(x_i) = j$.*

Let $E_J \subseteq E$ denote the set of J-edges in E. The J-density of G is defined as:

$$\texttt{dens}_J(G) = \frac{|E_J|}{|V|}.$$

The maximal J-density of $G(\underline{S})$ is:

$$\texttt{maxdens}_J(G(\underline{S})) = \max_{\emptyset \subsetneq U \subseteq V} \texttt{dens}_J(G_U),$$

where G_U is the hypergraph G induces on U.

Theorem 26 follows by combining the next two inequalities:

Lemma 29 *For every $\mathcal{H}$ and $\underline{S}$ as above,*

$$\min_A \texttt{deg}_{\max}(A; G(\underline{S})) \;\le\; \texttt{maxdens}_J(G(\underline{S})),$$

where A ranges over all mappings $A : E \to J$.

Lemma 30 *For every $\mathcal{H}$ and $\underline{S}$ as above,*

$$\texttt{maxdens}_J(G(\underline{S})) \;\le\; d_J(\mathcal{H}).$$

A.2.4. BOUNDING THE MAX DEGREE BY J-DENSITY

Proof of Lemma 29. The proof hinges on the following general result for hypergraphs. Let $G = (V, E)$ be a hypergraph. An *orientation* (or choice function) is a mapping $f : E \to V$ such that $f(e) \in e$ for all $e \in E$. A *randomized orientation* assigns to each $e \in E$ a distribution f_e over its vertices, serving as a fractional relaxation similar to those in other combinatorial problems. The in-degree of a vertex $v \in V$ is defined as:

$$\mathtt{deg}_{\mathtt{in}}(v; f) = |\{e \in E : f(e) = v\}|.$$

For randomized orientations, the in-degree of v is given by:

$$\mathtt{deg}_{\mathtt{in}}(v; f) = \sum_{e \in E} f_e(v),$$

where $f_e(v)$ is the probability assigned to $v \in e$ by the distribution f_e. The maximal in-degree of f is:

$$\mathtt{deg}_{\max}(f; G) = \max_{v \in V} \mathtt{deg}_{\mathtt{in}}(v; f).$$

The *maximal density* of G is defined by:

$$\mathtt{maxdens}(G) = \max_{\emptyset \subsetneq U \subseteq V} \frac{|E_U|}{|U|},$$

where $E_U = \{e \cap U : e \in E, e \cap U \neq \emptyset\}$ is the set of hyperedges in the U-induced subgraph.

Proposition 31 *Let $G = (V, E)$ be a hypergraph. Then:*

$$\min_f \mathtt{deg}_{\max}(f; G) = \mathtt{maxdens}(G),$$

where the minimum is over all (possibly randomized) orientations. Furthermore, there exists a deterministic orientation f such that:

$$\mathtt{deg}_{\max}(f; G) = \lceil \mathtt{maxdens}(G) \rceil.$$

Proof We adapt the proof of Haussler et al. (1988) to our case. We construct a digraph $\overrightarrow{G} = (V', E')$ based on G. Let $V' = \{s, t\} \dot\cup V \dot\cup E$ where s, t are, respectively, the distinct source and target vertex. We call E the first layer and V the second layer of $\overrightarrow{G}$. The edge set is given by $E' = E_s \cup E_t \cup E_G$ where

$$E_s = \{(s, e) \mid e \in E\},$$
$$E_G = \{(e, v) \mid e \in E, v \in e\}, \text{ and}$$
$$E_t = \{(v, t) \mid v \in V\}.$$

Each edge in E_s has capacity 1, the edges in E_G capacity ∞, and the edges in E_t capacity $\mathtt{maxdens}(G)$. Further consider the slightly modified flow network $\overrightarrow{G_I}$ with only capacities in E_t replaced by $\lceil \mathtt{maxdens}(G) \rceil$.

We start with the following claim and later relate the in-degree to the max flow in $\overrightarrow{G}$.

Claim 32 *The max flow in $\overrightarrow{G}$ is equal to $|E|$ and there exists an integral flow in $\overrightarrow{G_I}$ equal to $|E|$.*

Note that any flow in $\overrightarrow{G}$ is at most $|E|$ as there are $|E|$ edges from the source to the first layer in $\overrightarrow{G}$ each with capacity one. It remains to show that the max flow is at least $|E|$. The max flow equals the size of the min cut. Take any min-cut C. Clearly the total capacity of the min cut is finite and hence no edges in E_G are cut. Consider $S = E_s \setminus C$ the set of source edges that are not cut by C. Denote by W the set of vertices in the second layer of $\overrightarrow{G}$ incident to any of edge in S. As the edges in E_G are not cut, for each vertex $w \in W$ the cut C has to contain the edge $(w, t) \in E_t$ with capacity $\texttt{maxdens}(G)$. Let us bound $|S|$. Each edge $(s, e) \in S$ corresponds to the hyperedge e in the subhypergraph of G induced by W. This means that $|S|$ is bounded by $D'|W|$ where $D' \leq \texttt{maxdens}(G)$ is the density of this subgraph. Hence for the number of cut edges in E_s is at least $|E_s \setminus S| = |E| - D'|W|$. Overall the total capacity of C is

$$|W| \cdot \texttt{maxdens}(G) + |E_s \setminus S| \geq |W| \cdot \texttt{maxdens}(G) + |E| - D'|W| \geq |E| \,.$$

Finally, note that any flow for $\overrightarrow{G}$ is a flow for $\overrightarrow{G_I}$. As all capacities in $\overrightarrow{G_I}$ are integral, there also exists an integral flow for $\overrightarrow{G_I}$ of size $|E|$ (Dantzig and Fulkerson, 1956). This proves the claim.

Any max flow $f^\star$ in $\overrightarrow{G}$ distributes one unit of flow per edge in E_s on the vertices in V (the second layer). This yields an (possibly randomized) orientation $f : E \to V$ of G such that $\texttt{deg}_{\max}(f; G) \leq \texttt{maxdens}(G)$. In particular, let $f_e(v) = f^\star((e, v))$ for each $e \in E$ and $v \in e$. Any flow entering each vertex v in the second layer has to continue through the single edge with capacity $\texttt{maxdens}(G)$ to the target vertex and thus $\texttt{deg}_{\max}(f; G) = \sum_e f_e(v) \leq \texttt{maxdens}(G)$, where the sum is going over edges e incident to v. This shows the upper bound

$$\min_{f'} \texttt{deg}_{\max}(f'; G) \leq \texttt{deg}_{\max}(f; G) \leq \texttt{maxdens}(G)\,.$$

Analogously, we can define a deterministic orientation f_I with in-degree at most $\lceil \texttt{maxdens}(G) \rceil$ based on the integral max flow.

It remains to show that the optimal (possibly randomized) orientation f satisfies $\texttt{deg}_{\max}(f; G) \geq \texttt{maxdens}(G)$ (and the corresponding bound for the deterministic case). Assume $\texttt{deg}_{\max}(f; G) < \texttt{maxdens}(G)$. Any orientation f corresponds to a flow $\overrightarrow{f}$ of $\overrightarrow{G}$ given by full saturation on the unit-capacity source edges E_s and $\overrightarrow{f}((e, v)) = f_e(v)$ for the E_G edges (the flow on the target edges E_t is given by the flow coming into each vertex of the second layer). This means that any orientation f corresponds to a flow $\overrightarrow{f}$ with the max in degree of f being equal to the max flow on any of the target edges. Hence no target edge has full saturation with flow strictly smaller than $\texttt{maxdens}(G)$ (by assumption on f). We will show that this is not possible. Let H be an induced subgraph of G with density $\texttt{maxdens}(G)$. Each edge of H is responsible for one unit of flow in $\overrightarrow{f}$. This means that the total amount of flow arriving at the vertices in the second layer corresponding to the vertices $V(H)$ is at least $\texttt{maxdens}(G)|V(H)|$. Hence in any flow at least one of the target edges has to have flow at least $\texttt{maxdens}(G)$. This shows that any orientation f $\texttt{deg}_{\max}(f; G) < \texttt{maxdens}(G)$ is not possible, as it does not correspond to a valid flow $\overrightarrow{f}$. The statement for the deterministic flow follows analogously. ■

With Proposition 31 in hand, we now prove Lemma 29. Let $\mathcal{H}$ and $\underline{S}$ be as in Lemma 29, and let $G = G(\underline{S}) = (V, E)$ be the one-inclusion hypergraph associated with $\underline{S}$. We define a mapping $A : E \to J$ as follows:

1. **Non-J-edges:** for every edge $e_i \in E$ such that $J \setminus \{h(x_i) \mid h \in e_i\} \neq \emptyset$, set $A(e_i) = j$ for some $j \in J \setminus \{h(x_i) \mid h \in e_i\}$. Notice that, with respect to these edges, the in-degrees of all vertices in V is 0.

2. J**-edges:** define a new hypergraph $G' = (V, E')$ with the same vertex set V and edge set E' obtained by:
$$E' = \big\{e_i \setminus \{v : v(x_i) \notin J\} : e_i \in E_J\big\},$$
where E_J is the set of J-edges in E. Notice that each edge $e'_i \in E'$ is a subset of a unique edge $e_i \in E_J$, as any pair of distinct edges in E_J intersects in at most one vertex and contains exactly $|J| > 1$ vertices (see Proposition 23).

   Using Proposition 31, there exists an orientation $f : E' \to V$ such that:
$$\mathtt{deg}_{\max}(f; G') \;\leq\; \mathtt{maxdens}(G').$$
   We use this orientation f to define A on the edges in E_J: for each $e_i \in E_J$, map $A(e_i)$ to $j \in J$ such that $c(x_i) = j$, where $c = f(e'_i)$ and $e'_i \subseteq e_i$ is the corresponding edge in G'.

Finally, this construction yields a mapping $A : E \to J$ such that:
$$\mathtt{deg}_{\max}(A; G(\underline{S})) \;\leq\; \mathtt{maxdens}(G') \;\leq\; \mathtt{maxdens}_J(G(\underline{S})).$$

The second inequality, $\mathtt{maxdens}(G') \;\leq\; \mathtt{maxdens}_J(G(\underline{S}))$, follows because G' is obtained from G by replacing each hyperedge in E_J with a subedge of it, which cannot increase the density. This completes the proof.

A.2.5. Bounding J-Density by J-cube Dimension

Proof of Lemma 30. Here we use a shifting argument, adapted to J-edges, to transform the one-inclusion hypergraph $G(\underline{S})$ into a J*-monotone* one-inclusion hypergraph $G'(\underline{S})$ (the notion of J-monotonicity will be defined shortly) such that:
$$\mathtt{maxdens}_J(G(\underline{S})) \;\leq\; \mathtt{maxdens}_J(G'(\underline{S})) \;\leq\; d_J(\mathcal{H}).$$

Throughout this proof, we fix a linear order on $J = \{z_1, \ldots, z_k\}$:
$$z_1 < z_2 < \cdots < z_k.$$

The basic operation for transforming G into G' is defined as follows:

Definition 33 (x_i-shifting) *Let $x_i \in \underline{S}$. An i-shifting operation on a one-inclusion hypergraph $G(\underline{S}) = (V, E)$ takes* every *edge $e_i \in E$ in direction i and replaces it with an edge e'_i, defined as follows:*

Let $v_1, \ldots, v_t \in e_i$ be all vertices in e_i such that $v_j(x_i) \in J$. Replace these vertices with vertices $v'_1, \ldots, v'_t$ such that $v'_j(x_i) = z_j$, and otherwise v'_j agrees with v_j on $\underline{S} \setminus \{x_i\}$. Thus, e_i is replaced by:
$$e_i \leftarrow \Big(e_i \setminus \{v_1, \ldots, v_t\}\Big) \cup \{v'_1, \ldots, v'_t\}.$$

Repeatedly applying i-shifting eventually transforms $G(\underline{S})$ into a J-monotone hypergraph:

Definition 34 (J-monotonicity) *Let $x_i \in \underline{S}$. We say that $G(\underline{S}) = (V, E)$ is J-monotone in direction i if for every $v \in V$ such that $v(x_i) \in J$ and for every $z \in J$ with $z \leq v(x_i)$, there exists $v' \in V$ such that $v'(x_i) = z$ and $v'(x_j) = v(x_j)$ for all $j \neq i$. We say that $G(\underline{S}) = (V, E)$ is J-monotone if it is J-monotone in every direction i.*

Notice that applying an i-shifting operation produces a graph that is J-monotone in direction i. We can iterate this process to get a J-monotone graph.[2]

Lemma 35 *Repeatedly applying an i-shifting operation for all i to G results in a graph G' that is J-monotone.*

Proof Denote $\texttt{id}_J : z_j \mapsto j$ for $z_j \in J$ and $\texttt{id}_J : y \mapsto 0$ for $y \in Y \setminus J$. Let $s : v \mapsto \sum_{x_i \in \underline{S}} \texttt{id}_J(v(x_i))$ and $s(G) = \sum_v s(v)$. Note that any shifting operation (for an arbitrary direction) decreases $s(G)$ by at least one; for each edge e that was shifted to e' we have $\sum_{v' \in e'} s(v') < \sum_{v \in e} s(v)$. Hence as the initial $s(G)$ is finite after a finite number of shifting operations the resulting graph G' cannot be changed anymore by shifting (in arbitrary directions). Thus G' is J-monotone. ■

Thus, shifting G in every direction i transforms it into a J-monotone graph. The following lemma establishes key properties of the resulting graph:

Lemma 36 *Let $G'(\underline{S})$ denote the result of applying an i-shifting operation on $G(\underline{S})$. Then:*

1. $\texttt{maxdens}_J(G(\underline{S})) \leq \texttt{maxdens}_J(G'(\underline{S}))$,
2. *The J-cube dimension of $G'(\underline{S})$ is at most the J-cube dimension of $G(\underline{S})$.*

Proof For the first claim take any set U of vertices in $G(\underline{S})$ with J-edges $E_J(U)$. Note that shifting does not affect J-edges. Hence the density on U in $G'(\underline{S})$ is at least $|E_J(U)|/|U|$ and the claim follows.

For the second claim take any set $R = \{r_1, \dots, r_d\}$ that is J-shattered by the vertex set of $G'(\underline{S})$. Let $f : R \to J$ with $f(x_i) = z_k$. Note that as R is J-shattered there is a vertex v'_k in G' that predicts as f on R. By J-monotonicity in direction i there exists for all $\ell \in [k-1]$ a vertex v'_ℓ in G' such that $v'_\ell(x_i) = z_\ell$ and $v'_\ell(x_j) = v'_k(x_j)$ for $j \neq i$. By the definition of shifting the set of vertices $\{v'_1, \dots, v'_k\}$ can only be in $G'(\underline{S})$ if there is a set of vertices $\{v_1, \dots, v_k\} = \{v'_1, \dots, v'_k\}$ in $G(\underline{S})$. As this holds for all f, this shows that R is also J-shattered by the vertex set of $G(\underline{S})$. Hence the J-cube dimension of $G'(\underline{S})$ is at most the one of $G(\underline{S})$. ■

Thus, we only need the following lemma to prove Lemma 30.

Lemma 37 *Let $G = (V, E)$ be a J-monotone graph with J-cube dimension d. It holds that $\texttt{maxdens}_J(G) \leq d$.*

[2] Actually it holds that we only have to perform the shifting operation once per direction. This follows by an argument similar to showing that if the columns of a matrix are sorted and then the rows are sorted, the columns remain sorted in the resulting matrix.

Proof Note that every $v \in V$ has at most d positions i such that $v(x_i) = z_k$ (recall that z_k is the largest label in J). Indeed, if this were not the case, then by J-monotonicity the J-cube dimension of G would exceed d. Thus, we can associate each J-edge $e_i \in E_J$ with the vertex $v \in e$ such that $v(x_i) = z_k$. By the above, every $v \in V$ is associated with at most d edges, implying $|E_J| \leq d \cdot |V|$. Applying the same argument to every $U \subseteq V$, we infer that $\texttt{maxdens}_J(G) \leq d$. ■

A.2.6. From Leave-One-Out to PAC

Relying on Aden-Ali et al. (2023a) we can turn the above leave-one-out guarantee into a PAC guarantee with an additive $\log(1/\delta)$ term.

Theorem 38 *Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a hypothesis class, $\mathcal{D}$ a realizable distribution on $\mathcal{X} \times \mathcal{Y}$, and let $J \subseteq \mathcal{Y}$ be such that $2 \leq |J| < \infty$. There is a predictor $\widehat{f}_{\texttt{MAJ}}(S)$ trained on an i.i.d. sample $S \sim \mathcal{D}^m$ of size $m \geq m^J_{\mathcal{H}}(\epsilon, \delta)$ with*

$$m^J_{\mathcal{H}}(\epsilon, \delta) = \mathcal{O}\left(\frac{d_J(\mathcal{H}) + \log(1/\delta)}{\epsilon}\right)$$

such that with probability at least $1 - \delta$ the predictor $\widehat{f}_{\texttt{MAJ}}(S)$ achieves expected J-loss of

$$L^J_{\mathcal{D}}(\widehat{f}_{\texttt{MAJ}}(S)) \leq \epsilon\,.$$

Proof We will rely on the following theorem by Aden-Ali et al. (2023a), which we repeat here for completeness. We will denote by $S_{\leq t}$ the t-prefix of a sample S, i.e., the first t labeled data points.

Theorem 39 (Aden-Ali et al. 2023a) *Fix a hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$, and a learning algorithm A satisfying both*

1. *there exists an upper bound M_n/n on the leave-one-out error loss of A on any realizable sample S of size n: $\epsilon^{\texttt{LOO}}_{\ell}(A; S) \leq \frac{M_n}{n}$.*
2. *the returned predictor does not depend on the order of the training sample, i.e., for any sample S and permuted sample S' it holds that $A(S) = A(S')$.*

For any realizable distribution $\mathcal{D}$ and $\delta \in (0, 1)$ given an iid. sample $S \sim P^m$ we have

$$\frac{4}{3n} \sum_{t=n/4}^{n-1} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left[\ell(\,[A(S_{\leq t})](x), y\,)\right] \leq \mathcal{O}\left(\frac{M_n + \log(1/\delta)}{n}\right)\,,$$

with probability at least $1 - \delta$ over the randomness of S.

We can verify that our one-inclusion hypergraph based learner $A_{\texttt{OIG}}$ satisfies the requirements of the theorem with $M_n \leq d_J(\mathcal{H})$ and thus we get the bound

$$\frac{4}{3n} \sum_{t=n/4}^{n-1} L^J_{\mathcal{D}}(A_{\texttt{OIG}}(S_{\leq t})) \leq \mathcal{O}\left(\frac{d_J(\mathcal{H}) + \log(1/\delta)}{n}\right)\,,$$

plugging in the J-loss for ℓ. Finally, we can again follow Aden-Ali et al. (2023a) by constructing a majority vote $\widehat{f}_{\mathtt{MAJ}}$ over the $3/4n$ predictors $A(S_{\le t})$ for $t = n/4, \dots, n-1$ and achieve the bound:

$$\begin{aligned} L^J_{\mathcal{D}}(\widehat{f}_{\mathtt{MAJ}}) &= \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell^J_{(x,y)}(\widehat{f}_{\mathtt{MAJ}})\}] \\ &\le \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[2\left(\frac{4}{3n}\sum_{t=n/4}^{n-1}\ell^J_{(x,y)}(A_{\mathtt{OIG}}(S_{\le t}))\right)\right] \\ &= 2\frac{4}{3n}\sum_{t=n/4}^{n-1} L^J_{\mathcal{D}}(A_{\mathtt{OIG}}(S_{\le t})) \le \mathcal{O}\left(\frac{d_J(\mathcal{H}) + \log(1/\delta)}{n}\right). \end{aligned}$$

Here the first inequality holds, as the majority vote only has loss one if half of the one-inclusion graph based predictors have loss one. Rearranging ϵ and n gives the required bound. ■

Appendix B. Tightness of Bounds

B.1. Tightness of In-Expectation Upper Bound

We will show that the J-density upper bound by $d_J(\mathcal{H})$ in Theorem 26 is asymptotically tight with arbitrary $d = d_J(\mathcal{H})$ and $k = |J|$, for a specific family of hypothesis classes $\mathcal{H}_{d,k}$. This then implies that the $\mathcal{O}(d/\epsilon)$ upper bound in Theorem 17 is tight for specific hypothesis classes.

Let $d, k \ge 2$ and $n \in \mathbb{N}$ large enough to be determined later. Let $X = \{x_1, \dots, x_n\}$, $J = \{z_1, \dots, z_k\}$, and $V = \mathcal{H}_{d,k}$ consisting of all hypotheses on X with at most d times the label z_k. That is

$$V = \{v \in J^X : |\{i : v(x_i) = z_k\}| \le d\}.$$

Note that

$$|V| = \sum_{i=0}^{d}\binom{n}{i}(k-1)^{n-i}.$$

In the one-inclusion graph each $v \in V$ is incident to the n hyperedges $e_i = \{v' \in V : v'(x_j) = v(x_j) \text{ for all } j \ne i\}$ for $i \in [n]$. Recall that a J-edge is an edge e_i that for all $j \in [k]$ contains a vertex v with $v(x_i) = z_j$. Let $v \in V$ and consider two cases. Let V_d be the subset of vertices with d times the label z_k and $V_{<d} = V \setminus V_d$. Note that for any $v \in V_{<d}$, all n edges incident to v are J-edges (as all required adjacent v' have at $\le d$ times the label z_k and are thus in V). Each $v \in V_d$ has d incident J-edges; for a coordinate j with $v(x_j) \ne z_k$ the corresponding edge e_j requires a vertex with $d+1$ coordinates having the label z_k, which is not in V. As we sum up all incident J-edges of all vertices we count each J-edge k times and thus get

$$|E_J| = \frac{n|V_{<d}| + d|V_d|}{k} = \frac{(n-d)|V_{<d}| + d|V|}{k}.$$

Our goal is to show that $\frac{|E_J|}{|V|} \ge \Omega(d)$ for large enough n. To ease notation denote $t_i = \binom{n}{i}(k-1)^{n-i}$ such that $|V| = \sum_{i=0}^{d} t_i$ Note that

$$\frac{|E_J|}{|V|} = \frac{(n-d)}{k}\cdot\frac{|V_{<d}| + d|V|}{|V|} \ge \frac{(n-d)|V_{<d}|}{k|V|} \ge \frac{(n-d)t_{d-1}}{k|V|}. \tag{5}$$

We will show it holds that $2t_d \geq |V|$. In fact, let us compare t_i and t_{i-1} for for each $i \in [d]$. Indeed, it holds that

$$t_i = \binom{n}{i}(k-1)^{n-i} = \frac{n-i+1}{i}\binom{n}{i-1}(k-1)^{n-i}\frac{(k-1)}{(k-1)} = \frac{n-i+1}{(k-1)i}t_{i-1}\,.$$

Thus for

$$n \geq 2(k-1)d + d - 1 \geq 2(k-1)i + i - 1 \tag{6}$$

it holds that $t_i \geq 2t_{i-1}$. This implies

$$2t_d = t_d + t_d \geq t_d + 2t_{d-1} \geq \cdots \geq t_d + \cdots + t_0 = |V|\,.$$

This gives a lower bound to the r.h.s. of Equation (5) as

$$\frac{(n-d)t_{d-1}}{k|V|} \geq \frac{(n-d)t_{d-1}}{2t_d k} = \frac{(n-d)(k-1)d}{2k(n-d+1)} \geq \frac{d}{8}\,.$$

By Proposition 31 we thus get that the maximum in-degree for any orientation/algorithm is $\Omega(d)$, which by Equation (4) gives a lower bound on the leave-one-out error of any algorithm A:

$$\epsilon^{\mathtt{LOO}}(A; X) \geq \Omega(d/n)\,.$$

In fact, for any $X' \subseteq X$ with $|X'| = m$ satisfying Equation (6) as n, it also holds

$$\epsilon^{\mathtt{LOO}}(A; X') \geq \Omega(d/m)\,.$$

Now, let $\mathcal{D}_X$ be the uniform distribution over X. Fix a sample size m satisfying Equation (6) as n. Inspecting the proof of Proposition 21, we see that for any algorithm A and for any sample size m we have for the expected error

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}[L^J_{\mathcal{D}}(A(S))] = \mathop{\mathbb{E}}_{S'\sim\mathcal{D}^{m+1}}[\epsilon^{\mathtt{LOO}}(A; S')]\,.$$

Furthermore, as m is fixed and we can choose $n = |X|$ large enough, with probability, say, at least $1/2$ the sample S' has no duplicate entries. In this case,

$$\mathop{\mathbb{E}}_{S'\sim\mathcal{D}^{m+1}}[\epsilon^{\mathtt{LOO}}(A; S')] \geq 1/2 \mathop{\mathbb{E}}_{S'\sim\mathcal{D}^{m+1}}[\epsilon^{\mathtt{LOO}}(A; S') \mid S \text{ has no duplicates}]\,.$$

Note that for any such fixed S' it holds that $\epsilon^{\mathtt{LOO}}(A; S') \geq \Omega(d/m)$ and thus by monotonicity of expectations

$$\mathop{\mathbb{E}}_{S'\sim\mathcal{D}^{m+1}}[\epsilon^{\mathtt{LOO}}(A; S') \mid S \text{ has no duplicates}] \geq \Omega(d/m)\,.$$

Overall this means that there are cases where the expected error of any algorithm satisfies $\Omega(d/m)$, which shows that the upper bound in Theorem 17 is tight.

B.2. Tightness of In-Expectation Lower Bound

We will show that for a specific family of hypothesis classes the J-density is $\mathcal{O}(d_J(\mathcal{H})/|J|)$. This then implies that the $\Omega\left(\frac{d_J(\mathcal{H})}{|J|\epsilon}\right)$ lower bound in Theorem 17 is tight for specific hypothesis classes. Indeed, let $|J| = k$, let $d = d_J(\mathcal{H})$ and let X be any maximum J-shatterable set. This means that $V = \mathcal{H} = J^X$ and thus $|V| = k^d$. Also all edges are J-edges and hence $|E_J| = \frac{d|V|}{k}$. Thus, $|E_J|/|V| = d/k$. As the density can only decrease in this case for any subgraph, this shows that the maximum density is d/k. Thus, as we bounded the maximum density we get an in-expectation error bound of $\epsilon \leq \mathcal{O}(d/|J|)$ following the results in Appendix A.2.

Appendix C. Fine-grained PAC Learnability

Proof of Theorem 4. We prove the two directions separately.

$\Longleftarrow$ *direction.* Suppose $d_J(\mathcal{H}) < \infty$ for all $J \in \mathcal{J}(\boldsymbol{w}, \boldsymbol{z})$. Then, by Definition 8 and Theorem 10, for every $J \in \mathcal{J}(\boldsymbol{w}, \boldsymbol{z})$ there exists a PAC J-eliminator w.r.t. $\mathcal{H}$. This yields a family $\mathcal{R}$ of PAC eliminators that is sufficient for $(\boldsymbol{w}, \boldsymbol{z})$—see Definition 11—and $\mathcal{H}$ is $(\boldsymbol{w}, \boldsymbol{z})$-learnable by Theorem 12.

$\Longrightarrow$ *direction.* We prove the contrapositive. Suppose that there exists a $J \in \mathcal{J}(\boldsymbol{w}, \boldsymbol{z})$ such that $d_J(\mathcal{H}) = \infty$. Since $\boldsymbol{z} \notin D_J(\boldsymbol{w})$ by definition of J, we know that

$$\exists \boldsymbol{q} \in \Delta_J,\ \forall \boldsymbol{p} \in \Delta_{\mathcal{Y}},\ \exists i \in [r], \quad w_i(\boldsymbol{p}, \boldsymbol{q}) > z_i \ .$$

Let $\boldsymbol{q}^* \in \Delta_J$ be one such distribution. Now let $\mathcal{A}$ be any (deterministic) learning algorithm for $\mathcal{H}$ and let $m \geq 1$ be an arbitrary integer. Given that $d_J(\mathcal{H}) = \infty$, there exist arbitrarily large subsets $X' \subset \mathcal{X}$ such that the restriction of $\mathcal{H}$ to X' includes $J^{X'}$. Then, we can consider a sufficiently large integer $n > m$ (to be specified later), let $X = \{x_1, \dots, x_n\} \subset \mathcal{X}$ be one such subset of $\mathcal{X}$ of size n, and let $\mathcal{F} = J^X$. Define the randomized concept $f \in \mathcal{F}$ so that $f(x) \sim \boldsymbol{q}^*$ independently for every $x \in X$. Denote by $\mathcal{D}_X$ the uniform distribution over X and consider an i.i.d. sample $S \sim \mathcal{D}_X^m$ of size m. The learner $\mathcal{A}$, given S as input (together with the respective labels given by f), returns a (deterministic) hypothesis $\mathcal{A}(S) = h_S \in \mathcal{Y}^{\mathcal{X}}$. It follows that, by drawing $x \sim \mathcal{D}_X$, the expected cost $w_i(h_S(x), f(x))$ for any $i \in [r]$ satisfies

$$\begin{aligned}
\mathbb{E}_f \mathbb{E}_S \mathbb{E}_x w_i(h_S(x), f(x)) &\geq \mathbb{E}_f \mathbb{E}_S \Big[\mathbb{P}_{x \sim \mathcal{D}_X}[x \notin S] \cdot \mathbb{E}_{x \sim \mathcal{D}_{X \setminus S}}[w_i(h_S(x), f(x))]\Big] \\
&\geq \Big(1 - \frac{m}{n}\Big) \mathbb{E}_f \mathbb{E}_S \mathbb{E}_{x \sim \mathcal{D}_{X \setminus S}} w_i(h_S(x), f(x)) \\
&= \Big(1 - \frac{m}{n}\Big) \mathbb{E}_f \mathbb{E}_S \mathbb{E}_{x \sim \mathcal{D}_{X \setminus S}} w_i(h_S(x), \boldsymbol{q}^*) \\
&\qquad\qquad \text{(independence of } h_S(x) \text{ and } f(x) \text{ for } x \sim \mathcal{D}_{X \setminus S}) \\
&= \Big(1 - \frac{m}{n}\Big) w_i(\boldsymbol{p}_{\mathcal{A}}, \boldsymbol{q}^*) \ ,
\end{aligned}$$

where $\mathcal{D}_{X \setminus S}$ is the distribution $\mathcal{D}_X$ conditioned on $X \setminus S$ for any given $S \subseteq X$, and $\boldsymbol{p}_{\mathcal{A}} \in \Delta_{\mathcal{Y}}$ is such that $p_{\mathcal{A}}(y) = \mathbb{P}_{f, S \sim \mathcal{D}_X^m, x \sim \mathcal{D}_{X \setminus S}}[h_S(x) = y]$ for all $y \in \mathcal{Y}$. Recall that $\|\boldsymbol{w}\|_\infty = \max_{i \in [r]} \|w_i\|_\infty$ and define $\rho = \min_{\boldsymbol{p} \in \Delta_{\mathcal{Y}}} \max_{i \in [r]} \big(w_i(\boldsymbol{p}, \boldsymbol{q}^*) - z_i\big)$, which satisfies $\rho > 0$ by definition of $\boldsymbol{q}^*$. Also let $j \in \arg\max_{i \in [r]} \big(w_i(\boldsymbol{p}_{\mathcal{A}}, \boldsymbol{q}^*) - z_i\big)$. Taking $\epsilon \leq \rho/2$ and $n \geq \lfloor \frac{2m\|\boldsymbol{w}\|_\infty}{\rho} \rfloor + 1$, we can show for such j that

$$\begin{aligned}
\mathbb{E}_f \mathbb{E}_S \mathbb{E}_x w_j(h_S(x), f(x)) &\geq \Big(1 - \frac{m}{n}\Big) w_j(\boldsymbol{p}_{\mathcal{A}}, \boldsymbol{q}^*) \\
&\geq \Big(1 - \frac{\rho - \epsilon}{\|\boldsymbol{w}\|_\infty}\Big) w_j(\boldsymbol{p}_{\mathcal{A}}, \boldsymbol{q}^*) \\
&\geq w_j(\boldsymbol{p}_{\mathcal{A}}, \boldsymbol{q}^*) - \rho + \epsilon \\
&\geq w_j(\boldsymbol{p}_{\mathcal{A}}, \boldsymbol{q}^*) - \max_{i \in [r]} \big(w_i(\boldsymbol{p}_{\mathcal{A}}, \boldsymbol{q}^*) - z_i\big) + \epsilon \\
&= z_j + \epsilon \ ,
\end{aligned}$$

where the last step follows by definition of j.

If we now define $u_i : \mathcal{Y}^2 \to [-1,1]$ such that $u_i(y,y') = w_i(y,y') - z_i$ for every $y,y' \in \mathcal{Y}$ and every $i \in [r]$, and also let $u : \mathcal{Y}^2 \to [-1,1]$ be defined so that $u(y,y') = \max_{i\in[r]} u_i(y,y')$ for $y,y' \in \mathcal{Y}$, we have that

$$\mathbb{E}_f \mathbb{E}_S \mathbb{E}_x u(h_S(x), f(x)) \geq \mathbb{E}_f \mathbb{E}_S \mathbb{E}_x u_j(h_S(x), f(x)) \geq \epsilon \ .$$

Since this holds for any deterministic learning algorithm $\mathcal{A}$ for $\mathcal{H}$ and any positive integer m, by Yao's minimax principle (Yao, 1977) we can show that

$$\epsilon \leq \min_{\mathcal{A}} \mathbb{E}_f \mathbb{E}_{S\sim\mathcal{D}_X^m} \mathbb{E}_{x\sim\mathcal{D}_X} u(h_S(x), f(x)) \leq \max_{\mathcal{D}\in\Delta_{\mathcal{X}}, f\in\mathcal{H}} \mathbb{E}_{\mathcal{A}'} \mathbb{E}_{S\sim\mathcal{D}^m} \mathbb{E}_{x\sim\mathcal{D}} u(h_S(x), f(x))$$

for any possibly randomized learning algorithm $\mathcal{A}'$ for $\mathcal{H}$ and any sample size m. In other words, this shows by the probabilistic method and by definition of u that, for any $\mathcal{A}'$ and any m, there exist a distribution $\mathcal{D}$ over $\mathcal{X}$ and a function $f \in \mathcal{H}$ such that $L^{\boldsymbol{w}}_{\mathcal{D}\otimes f}(h_S) \not\leq \boldsymbol{z} + \epsilon\mathbf{1}$ for sufficiently small values of $\epsilon > 0$, with constant probability bounded away from zero w.r.t. the random draw of S and the internal randomness of $\mathcal{A}'$. Therefore, $\mathcal{H}$ is not $(\boldsymbol{w}, \boldsymbol{z})$-learnable. ■

Proof of Theorem 5. The proof of the upper bound is the same as the proof of $(\boldsymbol{w}, \boldsymbol{z})$-learnability of $\mathcal{H}$ in Theorem 4, with the only difference of using the explicit sample complexity bounds of Theorem 12 together with $|\mathcal{R}| \leq |\mathcal{J}(\boldsymbol{w}, \boldsymbol{z})|$. For the lower bound let w be the unweighted 0-1 loss and let $z^* = 0$. If $z^* \notin \mathscr{L}_{w^*}(\mathcal{H})$ then (w^*, z^*) is not learnable and no sample size is sufficient. Now assume $z^* \in \mathscr{L}_{w^*}(\mathcal{H})$. Let $J^* = \arg\max_{J\in\mathcal{J}(w^*,z^*)} d_J(\mathcal{H})$ and let S be a J^*-shattered set of size $d_{J^*}(\mathcal{H})$. Observe that S is also (Natarajan) N-shattered (see Section 5). As w^* is the unweighted 0-1 loss and $z^* = 0$, standard lower bounds for multi-class learning with labels restricted to J^* apply (Ben-David et al., 1995; Daniely et al., 2015), implying a lower bound of $\Omega\left(\frac{d_{J^*}(\mathcal{H}) + \log(1/\delta)}{\epsilon}\right)$. ■

Appendix D. Population-driven Loss

D.1. Proof of Lemma 14

We prove the equivalence between item (1) and item (3); both are clearly equivalent to item (2) by using $J = \{i,j\}$.

From (1) to (3). Assume $\boldsymbol{z} \in D_J(\boldsymbol{w}^{\mathrm{p}})$. This implies $\boldsymbol{z} \in D_{\{i,j\}}(\boldsymbol{w}^{\mathrm{p}})$ for all $i,j \in J$ with $i \neq j$. Then Bressan et al. (2025, Theorem 17) implies $\sqrt{z_i} + \sqrt{z_j} \geq 1$ again for all $i,j \in J$ with $i \neq j$.

From (3) to (1). The proof makes use of Lemma 40 below. Before starting we recall some necessary definitions and notation from Bressan et al. (2025). For a subset $J \subseteq \mathcal{Y}$ of labels, Δ_J is the family of all distributions supported on J. Given a cost $w : \mathcal{Y}^2 \to [0,1]$, the *value of the game* of w restricted to J is:

$$\mathrm{V}_J(w) \triangleq \max_{\boldsymbol{q}\in\Delta_J} \min_{\boldsymbol{p}\in\Delta_{\mathcal{Y}}} w(\boldsymbol{p}, \boldsymbol{q}) = \min_{\boldsymbol{p}\in\Delta_{\mathcal{Y}}} \max_{\boldsymbol{q}\in\Delta_J} w(\boldsymbol{p}, \boldsymbol{q}) \tag{7}$$

where the equality follows by von Neumann's minimax theorem (von Neumann, 1928). The *value of the game* of w is $\mathrm{V}(w) = \mathrm{V}_{\mathcal{Y}}(w)$. Given a multi-objective cost $\boldsymbol{w} : \mathcal{Y}^2 \to [0,1]^r$ and a distribution

$\boldsymbol{\alpha} \in \Delta_{[r]}$, the *scalarization* of $\boldsymbol{w}$ given by $\boldsymbol{\alpha}$, denoted by $\boldsymbol{\alpha} \cdot \boldsymbol{w}$, is the cost $w : \mathcal{Y}^2 \to [0,1]$ given by $(\boldsymbol{\alpha} \cdot w)(i,j) = \boldsymbol{\alpha} \cdot \boldsymbol{w}(i,j)$ for every $i,j \in \mathcal{Y}$.

Assume now $\sqrt{z_i} + \sqrt{z_j} \geq 1$ for all distinct $i,j \in J$. By Bressan et al. (2025, Theorem 17), $(z_i, z_j) \in D_{\{i,j\}}(\boldsymbol{w}^{\mathrm{p}})$ for all $i,j \in J$ with $i \neq j$. We want to show that

$$\boldsymbol{\alpha} \cdot \boldsymbol{z} \geq \mathrm{V}_J(\boldsymbol{\alpha} \cdot \boldsymbol{w}^{\mathrm{p}}) \qquad \text{for all } \boldsymbol{\alpha} \in \Delta_{\mathcal{Y}} \tag{8}$$

which—together with Bressan et al. (2025, Theorem 18)—implies $\boldsymbol{z} \in D_J(\boldsymbol{w}^{\mathrm{p}})$. Now note that

$$\mathrm{V}_J(\boldsymbol{\alpha} \cdot \boldsymbol{w}^{\mathrm{p}}) = \max_{\boldsymbol{q} \in \Delta_J} \min_{\boldsymbol{p} \in \Delta_{\mathcal{Y}}} \sum_{i \in J} \alpha_i \cdot q_i (1 - p_i) = \max_{\boldsymbol{q} \in \Delta_J} \min_{\boldsymbol{p} \in \Delta_J} \sum_{i \in J} \alpha_i \cdot q_i (1 - p_i)$$

because the minimizing player can only do worse by choosing $\boldsymbol{p}$ assigning non-zero mass outside of J. Hence, Equation (8) is equivalent to

$$\boldsymbol{\alpha} \cdot \boldsymbol{z} \geq \mathrm{V}'_J(\boldsymbol{\alpha} \cdot \boldsymbol{w}^{\mathrm{p}}) \qquad \text{for all } \boldsymbol{\alpha} \in [0,1]^{|J|} \tag{9}$$

where

$$\mathrm{V}'_J(w') = \min_{\boldsymbol{p} \in \Delta_J} \max_{\boldsymbol{q} \in \Delta_J} w'(\boldsymbol{p}, \boldsymbol{q})$$

for any cost function $w' : J^2 \to \mathbb{R}$. Define $I_{\boldsymbol{\alpha}} \triangleq \{i \in J : \alpha_i > 0\}$. Now, if $|I_{\boldsymbol{\alpha}}| \leq 1$, then Equation (9) trivially holds as $\mathrm{V}'_J(\boldsymbol{\alpha} \cdot \boldsymbol{w}^{\mathrm{p}}) = 0$. Hence, we may assume $|I_{\boldsymbol{\alpha}}| \geq 2$. Using Lemma 40 on V'_J with label set J, we have that Equation (9) is implied by

$$\boldsymbol{\alpha} \cdot \boldsymbol{z} \geq \max_{I \subseteq I_{\boldsymbol{\alpha}}} \frac{|I| - 1}{\sum_{i \in I} \frac{1}{\alpha_i}} .$$

Choose $I \subseteq I_{\boldsymbol{\alpha}}$ with $|I| \geq 2$. We have:

$$\boldsymbol{\alpha} \cdot \boldsymbol{z} \geq \sum_{i \in I} \alpha_i \, z_i = \frac{1}{|I| - 1} \sum_{\substack{i,j \in I \\ i<j}} (\alpha_i \, z_i + \alpha_j \, z_j) \geq \frac{1}{|I| - 1} \sum_{\substack{i,j \in I \\ i<j}} \frac{1}{\frac{1}{\alpha_i} + \frac{1}{\alpha_j}} \tag{10}$$

where the last step is due to

$$\alpha_i \, z_i + \alpha_j \, z_j \geq \frac{\alpha_i \, \alpha_j}{\alpha_i + \alpha_j} = \frac{1}{\frac{1}{\alpha_i} + \frac{1}{\alpha_j}}$$

which holds for all $\alpha_i, \alpha_j \geq 0$ and all $z_i, z_j \geq 0$ such that $\sqrt{z_i} + \sqrt{z_j} \geq 1$. Let $x_i = \frac{1}{\alpha_i}$ for all i. We show that:

$$\frac{1}{|I| - 1} \sum_{\substack{i,j \in I \\ i<j}} \frac{1}{x_i + x_j} \geq \frac{|I| - 1}{\sum_{i \in I} x_i} .$$

To see this, note that for a given value of $\sum_{i \in I} x_i$ the left-hand side is minimized when $x_i = x_j$ for all $i,j \in I$, in which case we have indeed:

$$\frac{1}{|I| - 1} \sum_{\substack{i,j \in I \\ i<j}} \frac{1}{x_i + x_j} = \frac{1}{|I| - 1} \sum_{\substack{i,j \in I \\ i<j}} \frac{1}{\frac{2}{|I|} \sum_{i \in I} x_i} = \frac{|I|^2}{4} \frac{1}{\sum_{i \in I} x_i} \geq \frac{|I| - 1}{\sum_{i \in I} x_i} .$$

Substituting x_i and chaining with Equation (10) proves Equation (9).

Lemma 40 *Let $\mathcal{Y} = [k]$ for $k \geq 2$, let $\boldsymbol{\alpha} \in [0,1]^k$ such that $I_{\boldsymbol{\alpha}} = \{i \in \mathcal{Y} : \alpha_i > 0\}$ has size $|I_{\boldsymbol{\alpha}}| \geq 2$. Then:*

$$\mathrm{V}(\boldsymbol{\alpha} \cdot \boldsymbol{w}^{\mathrm{p}}) = \max_{I \subseteq I_{\boldsymbol{\alpha}}} \frac{|I| - 1}{\sum_{i \in I} \frac{1}{\alpha_i}} \; . \tag{11}$$

Proof Let $w = \boldsymbol{\alpha} \cdot \boldsymbol{w}^{\mathrm{p}}$. By definition of w and of $\mathrm{V}(w)$, and by the minimax theorem,

$$\mathrm{V}(w) = \min_{\boldsymbol{p} \in \Delta_{\mathcal{Y}}} \max_{\boldsymbol{q} \in \Delta_{\mathcal{Y}}} \sum_{i \in \mathcal{Y}} q_i \, \alpha_i (1 - p_i) = \max_{\boldsymbol{q} \in \Delta_{\mathcal{Y}}} \min_{\boldsymbol{p} \in \Delta_{\mathcal{Y}}} \sum_{i \in \mathcal{Y}} q_i \, \alpha_i (1 - p_i) \; . \tag{12}$$

The terms with $\alpha_i = 0$ do not contribute. Hence we may restrict $\boldsymbol{q} \in \Delta_{I_{\boldsymbol{\alpha}}}$ and $i \in I_{\boldsymbol{\alpha}}$. This yields:

$$\mathrm{V}(w) = \min_{\boldsymbol{p} \in \Delta_{\mathcal{Y}}} \max_{\boldsymbol{q} \in \Delta_{I_{\boldsymbol{\alpha}}}} \sum_{i \in I_{\boldsymbol{\alpha}}} q_i \, \alpha_i (1 - p_i) = \max_{\boldsymbol{q} \in \Delta_{I_{\boldsymbol{\alpha}}}} \min_{\boldsymbol{p} \in \Delta_{\mathcal{Y}}} \sum_{i \in I_{\boldsymbol{\alpha}}} q_i \, \alpha_i (1 - p_i) \; . \tag{13}$$

We now prove Equation (11) by giving bounds in both directions.

First direction. We show that:

$$\mathrm{V}(w) \geq \max_{I \subseteq I_{\boldsymbol{\alpha}}} \frac{|I| - 1}{\sum_{i \in I} \frac{1}{\alpha_i}} \; . \tag{14}$$

Fix any $I \subseteq I_{\boldsymbol{\alpha}}$, and define $\boldsymbol{q} \in \Delta_I$ by:

$$q_i = \frac{\frac{1}{\alpha_i}}{\sum_{j \in I} \frac{1}{\alpha_j}} \qquad \forall i \in I \; .$$

Then for every $\boldsymbol{p} \in \Delta_{\mathcal{Y}}$:

$$\sum_{i \in I} q_i \cdot \alpha_i (1 - p_i) = \frac{\sum_{i \in I} (1 - p_i)}{\sum_{j \in I} \frac{1}{\alpha_j}} = \frac{|I| - \sum_{i \in I} p_i}{\sum_{j \in I} \frac{1}{\alpha_j}} \geq \frac{|I| - 1}{\sum_{j \in I} \frac{1}{\alpha_j}} \; .$$

Maximizing over $I \subseteq I_{\boldsymbol{\alpha}}$ proves Equation (14).

Second direction. We show that:

$$\mathrm{V}(w) \leq \max_{I \subseteq I_{\boldsymbol{\alpha}}} \frac{|I| - 1}{\sum_{i \in I} \frac{1}{\alpha_i}} \; . \tag{15}$$

To this end we define a distribution $\boldsymbol{\beta}$ for the predicting player such that $w^{\mathrm{p}}(\boldsymbol{p}, \boldsymbol{q})$ is bounded by the right-hand side of Equation (15) for every $\boldsymbol{q} \in \Delta_{\mathcal{Y}}$.

Without loss of generality assume $\alpha_1 \geq \cdots \geq \alpha_k$, and let:

$$V_i^{\boldsymbol{\alpha}} = \begin{cases} \frac{i-1}{\sum_{j=1}^{i} \frac{1}{\alpha_j}} & \text{if } i \in I_{\boldsymbol{\alpha}}, \\ 0 & \text{otherwise.} \end{cases}$$

Define the *cutoff index of* $\boldsymbol{\alpha}$ as $\ell_{\boldsymbol{\alpha}} = \max\{i \in [k] : \alpha_i \geq V_i^{\boldsymbol{\alpha}}\}$. Then define the distribution $\boldsymbol{\beta}$:

$$\beta_i = \begin{cases} 1 - \frac{1}{\alpha_i} V_{\ell_{\boldsymbol{\alpha}}}^{\boldsymbol{\alpha}} & \text{if } i \leq \ell_{\boldsymbol{\alpha}}, \\ 0 & \text{otherwise.} \end{cases}$$

Note that $\boldsymbol{\beta}$ is a valid distribution. Indeed, on the one hand, the ordering of $\boldsymbol{\alpha}$ together with the definition of $\ell_{\boldsymbol{\alpha}}$ implies $\beta_i \geq 0$ for all $i \in \mathcal{Y}$. On the other hand,

$$\sum_{i=1}^{k} \beta_i = \sum_{i=1}^{\ell_{\boldsymbol{\alpha}}} \left(1 - \frac{\frac{\ell_{\boldsymbol{\alpha}}-1}{\alpha_i}}{\sum_{j=1}^{\ell_{\boldsymbol{\alpha}}} \frac{1}{\alpha_j}} \right) = \ell_{\boldsymbol{\alpha}} - (\ell_{\boldsymbol{\alpha}} - 1) = 1 \,.$$

Now we prove the following fact used below: for every $i = 2, \ldots, |I_{\boldsymbol{\alpha}}|$, we have $\alpha_i \geq V_i^{\boldsymbol{\alpha}}$ if and only if $V_i^{\boldsymbol{\alpha}} \geq V_{i-1}^{\boldsymbol{\alpha}}$. To this end for every $i \in I_{\boldsymbol{\alpha}}$ let $b_i = \sum_{j=1}^{i} \frac{1}{\alpha_j}$, and observe that $\alpha_i = \frac{1}{b_i - b_{i-1}}$ and $V_{i-1}^{\boldsymbol{\alpha}} = \frac{i-1}{b_i}$ when $i \geq 2$. Then notice that:

$$\alpha_i \geq V_i^{\boldsymbol{\alpha}} \iff \frac{1}{b_i - b_{i-1}} \geq \frac{i-1}{b_i} \iff \frac{1}{i-1} \geq 1 - \frac{b_{i-1}}{b_i} \iff \frac{b_{i-1}}{b_i} \geq 1 - \frac{1}{i-1} \,, \tag{16}$$

$$V_i^{\boldsymbol{\alpha}} \geq V_{i-1}^{\boldsymbol{\alpha}} \iff \frac{i-1}{b_i} \geq \frac{i-2}{b_{i-1}} \iff \frac{b_{i-1}}{b_i} \geq \frac{i-2}{i-1} \iff \frac{b_{i-1}}{b_i} \geq 1 - \frac{1}{i-1} \,. \tag{17}$$

Now, the ordering of $\boldsymbol{\alpha}$ implies that, for every fixed $|I| \leq |I_{\boldsymbol{\alpha}}|$, the maximum at the right-hand side of Equation (11) is attained by $I = \{1, \ldots, |I|\}$. By the fact above and the definition of $V_{\ell_{\boldsymbol{\alpha}}}^{\boldsymbol{\alpha}}$, this implies:

$$\max_{I \subseteq I_{\boldsymbol{\alpha}}} \frac{|I| - 1}{\sum_{i \in I} \frac{1}{\alpha_i}} = \max_{i=1,\ldots,|I_{\boldsymbol{\alpha}}|} \frac{i-1}{\sum_{j=1}^{j} \frac{1}{\alpha_j}} = \max_{i=1,\ldots,|I_{\boldsymbol{\alpha}}|} V_i^{\boldsymbol{\alpha}} = V_{\ell_{\boldsymbol{\alpha}}}^{\boldsymbol{\alpha}} \,. \tag{18}$$

Let $\boldsymbol{p} = \boldsymbol{\beta}$. Then for every $\boldsymbol{q} \in \Delta_{\mathcal{Y}}$:

$$\begin{aligned} w^{\mathrm{P}}(\boldsymbol{p}, \boldsymbol{q}) &= \sum_{i=1}^{\ell_{\boldsymbol{\alpha}}} q_i \alpha_i (1 - p_i) + \sum_{i=\ell_{\boldsymbol{\alpha}}+1}^{k} q_i \alpha_i (1 - p_i) &(19)\\ &= \sum_{i=1}^{\ell_{\boldsymbol{\alpha}}} q_i \alpha_i \frac{V_{\ell_{\boldsymbol{\alpha}}}^{\boldsymbol{\alpha}}}{\alpha_i} + \sum_{i=\ell_{\boldsymbol{\alpha}}+1}^{k} q_i \alpha_i &(20)\\ &\leq V_{\ell_{\boldsymbol{\alpha}}}^{\boldsymbol{\alpha}} \sum_{i=1}^{\ell_{\boldsymbol{\alpha}}} q_i \;+\; \alpha_{\ell_{\boldsymbol{\alpha}}+1} \sum_{i=\ell_{\boldsymbol{\alpha}}+1}^{|I_{\boldsymbol{\alpha}}|} q_i &(21) \end{aligned}$$

where the last inequality holds for the second sum by the ordering of $\boldsymbol{\alpha}$, and where we let $\alpha_{\ell_{\boldsymbol{\alpha}}+1} = 0$ if $\ell_{\boldsymbol{\alpha}} = k$ (in that case the second sum equals zero anyway). We conclude by showing that $\alpha_{\ell_{\boldsymbol{\alpha}}+1} \leq V_{\ell_{\boldsymbol{\alpha}}}^{\boldsymbol{\alpha}}$. This implies that the right-hand side of Equation (21) is bounded from above by $V_{\ell_{\boldsymbol{\alpha}}}^{\boldsymbol{\alpha}}$, and thus by $\max_{I \subseteq I_{\boldsymbol{\alpha}}} \frac{|I|-1}{\sum_{i \in I} \frac{1}{\alpha_i}}$ via Equation (18), which concludes the proof. If $\ell_{\boldsymbol{\alpha}} \geq |I_{\boldsymbol{\alpha}}|$ then $\alpha_{\ell_{\boldsymbol{\alpha}}+1} \leq V_{\ell_{\boldsymbol{\alpha}}}^{\boldsymbol{\alpha}}$ holds trivially. Suppose then $\ell_{\boldsymbol{\alpha}} < |I_{\boldsymbol{\alpha}}|$. Recall from above that for every $i = 2, \ldots, |I_{\boldsymbol{\alpha}}|$ we have $\alpha_i \geq V_i^{\boldsymbol{\alpha}}$ if and only if $V_i^{\boldsymbol{\alpha}} \geq V_{i-1}^{\boldsymbol{\alpha}}$. Since $\alpha_{\ell_{\boldsymbol{\alpha}}+1} < V_{\ell_{\boldsymbol{\alpha}}+1}^{\boldsymbol{\alpha}}$ by definition of $\ell_{\boldsymbol{\alpha}}$, we conclude that $V_{\ell_{\boldsymbol{\alpha}}+1}^{\boldsymbol{\alpha}} < V_{\ell_{\boldsymbol{\alpha}}}^{\boldsymbol{\alpha}}$, and therefore $\alpha_{\ell_{\boldsymbol{\alpha}}+1} < V_{\ell_{\boldsymbol{\alpha}}}^{\boldsymbol{\alpha}}$. We conclude that in every case $\alpha_{\ell_{\boldsymbol{\alpha}}+1} \leq V_{\ell_{\boldsymbol{\alpha}}}^{\boldsymbol{\alpha}}$. The proof is complete. ■

Proof of Theorem 6. The two characterizations are easily seen to be equivalent by rewriting. We shall then prove the second one:

$$\begin{aligned}
\mathscr{L}_{\boldsymbol{w}^{\mathrm{P}}}(\mathcal{H}) &= \bigcap_{J\subseteq\mathcal{Y}\,:\,d_J(\mathcal{H})=\infty} D_J(\boldsymbol{w}^{\mathrm{P}}) && \text{(by Theorem 4)}\\
&= \bigcap_{J\subseteq\mathcal{Y}\,:\,d_J(\mathcal{H})=\infty} \left\{\boldsymbol{z}\in[0,1]^k \,:\, \forall\{i,j\}\in\binom{J}{2},\ \sqrt{z_i}+\sqrt{z_j}\geq 1\right\} && \text{(by Lemma 14)}\\
&= \bigcap_{\{i,j\}\subseteq J\subseteq\mathcal{Y}\,:\,d_J(\mathcal{H})=\infty} \left\{\boldsymbol{z}\in[0,1]^k \,:\, \sqrt{z_i}+\sqrt{z_j}\geq 1\right\} && \text{(by rewriting)}\\
&= \bigcap_{J=\{i,j\}\subseteq\mathcal{Y}\,:\,d_J(\mathcal{H})=\infty} \left\{\boldsymbol{z}\in[0,1]^k \,:\, \sqrt{z_i}+\sqrt{z_j}\geq 1\right\}\ . && (22)
\end{aligned}$$

For the last equality, observe that the set $\{\{i,j\} : \{i,j\}\subseteq J\subseteq\mathcal{Y}, d_J(\mathcal{H})=\infty\}$ equals $\{\{i,j\}\subseteq\mathcal{Y} : d_{\{i,j\}}(\mathcal{H})=\infty\}$. Indeed, if $\{i,j\}\subseteq J$ and $d_J(\mathcal{H})=\infty$ then $d_{\{i,j\}}(\mathcal{H})=\infty$; conversely, if $d_{\{i,j\}}(\mathcal{H})=\infty$ then in particular $J=\{i,j\}$ satisfies $\{i,j\}\subseteq J$ and $d_J(\mathcal{H})=\infty$. ■

Proof of Theorem 7. Consider $\boldsymbol{z}\in\mathscr{L}_{\boldsymbol{w}^{\mathrm{P}}}(\mathcal{H})$ and let $J\subseteq\mathcal{Y}$ with $|J|\geq 2$. By Lemma 14, if $\boldsymbol{z}\notin D_J(\boldsymbol{w}^{\mathrm{P}})$ then there exists $\{i,j\}\subseteq J$ such that $\sqrt{z_i}+\sqrt{z_j}<1$, and by Theorem 6 then $d_{\{i,j\}}(\mathcal{H})<\infty$. In turn, by Definition 8 this implies the existence of a PAC $\{i,j\}$-eliminator $A_{\{i,j\}}$ w.r.t. $\mathcal{H}$. We conclude that the set

$$\mathcal{R}=\left\{\mathcal{A}_J : J=\{i,j\}\subseteq\mathcal{Y}, \sqrt{z_i}+\sqrt{z_j}<\infty\right\}$$

is sufficient for $(\boldsymbol{w}^{\mathrm{P}},\boldsymbol{z})$ w.r.t. $\mathcal{H}$ (see Definition 11). Applying Theorem 12 and the sample complexity bounds of Definition 8 yields the claimed sample complexity bounds. ■

Appendix E. Additional Claims

Proposition 41 *Let $\mathcal{Y}$ be finite and $\mathcal{H}\subseteq\mathcal{Y}^{\mathcal{X}}$ be a class with $d_N(\mathcal{H})=\infty$. Then there exists a $J\subseteq\mathcal{Y}$ with $|J|=2$ and $d_J(\mathcal{H})=\infty$.*

Proof Let S be an N-shattered set with witnesses f, f'. Let $y\in\mathcal{Y}$ be the most frequent label in $\{f(x) : x\in S\}$ (ties broken arbitrarily) and let $y'=f'(s)$ be the most frequent label among all $s\in S$ with $f(s)=y$. Now, if $S'=\{s\in S : f(s)=y,\, f'(s)=y'\}$, then $|S'|\geq\frac{|S|}{|\mathcal{Y}|^2}$ by construction. Note that this set is J-shattered with $J=\{y,y'\}$. As S can be chosen arbitrarily large and $|\mathcal{Y}|$ is fixed, there are arbitrarily large J-shattered sets S', implying $d_J(\mathcal{H})=\infty$. ■