Characterizing the Multiclass Learnability of Forgiving 0-1 Loss Functions

Anonymous authors
Paper under double-blind review

Abstract

In this paper we will give a characterization of the learnability of forgiving 0-1 loss functions in the multiclass setting with effectively finite cardinality of the output and label space. To do this, we create a new combinatorial dimension that is based off of the Natarajan Dimension (Natarajan, 1989)and we show that a hypothesis class is learnable in our setting if and only if this Generalized Natarajan Dimension is finite. We also show how this dimension characterizes other known learning settings such as a vast amount of instantiations of learning with set-valued feedback and a modified version of list learning.

1 Introduction

Classification is one of the most common tasks in machine learning. Within classification, there is normally a split between binary classification (only two possible outputs) and multiclass classification (more than two possible outputs). The theoretical analysis of these settings shares the same split. In binary classification and multiclass classification, the 0-1 loss has been extensively studied and characterized (Valiant, 1984; Ben-David et al., 1995; Brukhim et al., 2022). In binary classification, there are 16 possible loss functions ℓ such that $\ell(y, y') \in \{0, 1\}$, but only two where $\ell(y_1, y_2) \neq \ell(y_1, y_1)$ and $\ell(y_2, y_1) \neq \ell(y_2, y_2)$. Thus, a majority of these loss functions are uninteresting and the two that are interesting are the opposites of each other (the 0-1 loss and $1 - \ell_{0-1}$). However, in classification for an output space with cardinality k and a label space with cardinality t, there are $(2^t)^k$ different 0-1 loss functions. Most of these loss functions take the form of more "forgiving" losses (i.e. $\ell(z, y) \in \{0, 1\}$, but there can exist many z_i, y_j such that $\ell(z_i, y_j) = 0$). Even more so, many of these settings can have interesting utility; settings such as paraphrase generation in natural language processing (Zhou & Bhat, 2021), thresholding a metric, ranking with partial feedback (Raman et al., 2023), classifying graphs up to isomorphisms (such as is done for drug discovery (Yang et al., 2024)), and many others can all be seen as allowing for some tolerance on the output. This work attempts to characterize the learnability of a large class of learning problems with these "forgiving" multiclass 0-1 loss functions.

The contributions of this paper are as follows:

1. We give a new combinatorial dimension based off the Natarajan Dimension (Natarajan, 1989), which we call the Generalized Natarajan Dimension.
2. We show the Generalized Natarajan dimension characterizes learnability of for forgiving 0-1 loss functions in the multiclass setting with minimal extra assumptions on the loss function. We also show that the Generalized Natarajan dimension is incomparable to other known dimensions in the classification literature.
3. We show our characterization implies characterization of various other settings seen in the machine learning literature.

The paper is organized as follows. Section 2 discusses the necessary background information. Section 3 details the problem set-up. Section 4 shows the results needed to characterize the forgiving 0-1 loss functions along with how the Generalized Natarajan Dimension relates to other dimensions in the literature. Section 5 gives concrete applications of our characterization. Finally, Section 6 concludes and discusses possible future work.

1.1 Related Works

Under the PAC-learning model, binary classification learnability under the 0-1 loss is known to be characterized by the VC-dimension (Vapnik & Chervonenkis, 1974; Shalev-Shwartz & Ben-David, 2014). For multiclass classification, there has also been a further split between finite and infinite label cases. For example, it is known that Empirical Risk Minimization (ERM) is a valid learner in the finite label case (Shalev-Shwartz & Ben-David, 2014), however, Daniely et al. (2015) has shown that not every ERM is a learner in the infinite label case. The Natarajan Dimension has been known to characterize the learnability of the 0-1 loss in the finite label case for many years now (Natarajan, 1989; Ben-David et al., 1995), but only relatively recently has the DS-Dimension been shown to characterize the learnability of the 0-1 loss under the infinite label case (Daniely & Shalev-Shwartz, 2014; Brukhim et al., 2022).

For related work around PAC-learning general loss functions, Ben-David et al. (1995) show the finiteness of the Natarajan dimension is sufficient of learnability of any finite-label multiclass loss function. David et al. (2016) show that sample compression and multiclass learnability are equivalent, but stop short of creating any combinatorial dimension to characterize learnability. Hopkins et al. (2022) show conditions to go from realizable learning to agnostic learning. However, they require assumptions on loss functions that we do not assume. Recently, Hanneke et al. (2025) were able to upgrade Hopkins et al. (2022) to give an agnostic-to-realizable reduction for all multiclass loss functions with $\{0, 1\}$ output, however, their rates are of the order of $\frac{1}{\epsilon^3}$ in our setting.

Other ways people have generalized PAC-learnability deal with altering the definition. Alon et al. (2022) and Kalavasis et al. (2022) relax PAC-learnability to allow for partial concept classes, which allows the hypotheses to be undefined on parts of the input space. Bressan et al. (2025) alter realizable PAC-learnability to allow for more relaxed errors. We can see our realizable setting is similar to what they study when they have $z = 0$ and $w = \ell$. Using the reduction from Hanneke et al. (2025), one does get a characterization similar to the one arrived at in this paper. There are two things to note though. One is that their rates have at least an multiplicative factor of $\frac{1}{\epsilon^3}$, which is worse than ours of $\frac{1}{\epsilon^2}$. Second, is they assume there exists a hypothesis that, with probability 1, outputs the correct label. Our realizable case only assumes the infimum of the loss on our hypothesis class is 0. This makes their assumption noticeably stronger than ours. We also show their dimension is incomparable with ours; there exists hypothesis classes where our dimension and theirs can be arbitrarily (and in some cases infinitely) larger than each other. Thus, even if their characterization (along with the agnostic-to-realizable reduction) did characterize our setting, our dimension is a more clean, precise, and interpretable characterization of the setting. In their paper they state "it would be interesting to find characterizations beyond the J-cube dimensions that allow to fully recover known multiclass sample complexity bounds", the upper bound of which we do recover in our setting.

We also note the similarities of our setting to the setting of list learning. List learning is when the algorithm is allowed to output a list of possible outputs and the loss is given by whether or not the label is present in the learner's outputted list. We can see that, given a multiclass hypothesis h and a loss function ℓ whose output is in $\{0, 1\}$, we can create an equivalent h' and ℓ' where $h'(x) = \{y \mid \ell(h(x), y) = 0\}$ and

$$\ell'(\{y_1, \ldots, y_k\}, y) = \begin{cases} 0 & y \in \{y_1, \ldots, y_k\} \\ 1 & \text{otherwise} \end{cases}.$$

This shows that our setting is equivalent to a modified version of list-learning. List learning was originally studied in Brukhim et al. (2022) and was shown to be characterized by the k-DS dimension when output list size is bounded by k (Charikar & Pabbaraju, 2023). However, we note the k-DS dimension does not necessarily characterize our setting. In their setting, one is allowed to create a list learner that outputs *any* bounded list and then it is compared to the best 0-1 learner in the hypothesis class ($\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ but $\mathcal{A}(S) \in \left(2^{\mathcal{Y}}\right)^{\mathcal{X}}$). In our setting, we have to compare to other list learners, not 0-1 learners. We can also only output lists that correspond to sets $\{y' \mid \ell(y, y') = 0\}$. Hanneke et al. (2024) do show that agnostic/realizable list learning and uniform convergence are all equivalent in the setting where the the algorithms output is being compared to list learners whose lists are size k. However, they still allow the algorithm to output any list and our lists are not constrained to output a list of size exactly k. Thus, while there is this connection between the current list-learning literature and our setting, they are distinct settings.

For learning with set-valued labels, Liu & Dietterich (2014) have studied a similar problem where there is a list with a "true" label and other "distractor" labels. They see how well ERM works when trying to find the hypothesis that predicts the true label well. Our setting, where any value in our label set is considered a "true" label, has been characterized in the online setting (Raman et al., 2024b). The batch setting characterization, however, has remained open as far as we are aware.

2 Background and Notation

We will first set up the notion of learnability, which is known as *Probably Approximately Correct* (PAC) learnability (Valiant, 1984).

Definition 2.1 (Agnostic PAC-learnability)**.** *Let $\mathcal{X}$ be an input space, $\mathcal{Z}$ be our output space, $\mathcal{Y}$ be our label space, $\mathcal{H} \subset \mathcal{Z}^{\mathcal{X}}$ be a hypothesis class, and $\ell : \mathcal{Z} \times \mathcal{Y}$ be a loss function. Then, $\mathcal{H}$ is agnostic PAC-learnable with respect to ℓ if there exists a learning algorithm $\mathcal{A}$ and a sample function $m : \mathbb{R} \times \mathbb{R} \to \mathbb{N}$ such that $\forall \epsilon, \delta > 0$, for every distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, if we have samples S generated from $\mathcal{D}$, $|S| \geq m(\epsilon, \delta)$, then we have:*

$$\mathbb{P}\left(\mathbb{E}_{\mathcal{D}}\left[\ell(\hat{h}(x), y)\right] \leq \inf_{h \in \mathcal{H}} \mathbb{E}_{\mathcal{D}}\left[\ell(h(x), y)\right] + \epsilon\right) \geq 1 - \delta$$

where $\hat{h} = \mathcal{A}(S) \in \mathcal{H}$.

When $\inf_{h \in \mathcal{H}} \mathbb{E}_{\mathcal{D}}\left[\ell(h(x), y)\right] = 0$, it is called the *realizable* case. We believe this is the best way to generalize the realizable case to our setting as having to compete against a function that has 0 loss is what matters, not competing against a function that predicts the correct label. Were we to do the latter, there can exist cases where $\exists h \in \mathcal{H}$ such that $\mathbb{E}_{\mathcal{D}}\left[\ell(h(x), y)\right] = 0$, but we would not be in the realizable case. For example, let us have $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ and $\ell(\cdot, \cdot) = 0$. Then, for a distribution $\mathcal{D}$ on $\mathcal{X}$, where our samples are of the form $(x, c(x))$ with $c \in \mathcal{Y}^{\mathcal{X}} \setminus \mathcal{H}$, we can see this would be in the realizable case for our setting. However, this would not be the case for the stricter version.

One of the most important loss functions is the 0-1 loss function (Wang et al., 2022) and it is a common place to start when studying a new setting (or its equivalent in the setting) (Shalev-Shwartz & Ben-David, 2014; Brukhim et al., 2022; Charikar & Pabbaraju, 2023; Guermeur, 2007):

$$\ell_{0-1}(y, y') = \begin{cases} 0 & y = y' \\ 1 & \text{otherwise} \end{cases}$$

For multiclass classification with finite labels using the 0-1 loss, it is known that this setting is characterized by the Natarajan dimension (Natarajan, 1989; Shalev-Shwartz & Ben-David, 2014):

Definition 2.2 (Natarajan Dimension)**.** *We say that a hypothesis class $\mathcal{H}$ Natarajan shatters a set $S = \{s_1, ..., s_n\}$ if $\exists h_1, h_2 \in \mathcal{H}$ such that*

1. *$\forall s_i \in S$, $h_1(s_i) \neq h_2(s_i)$*
2. *$\forall S' \subseteq S$ $\exists h \in \mathcal{H}$ where $\forall s \in S'$ $h(s) = h_1(s)$ and for all $s \in S \setminus S'$ $h(s) = h_2(s)$*

The Natarajan dimension of a hypothesis class $\mathcal{H}$, denoted $Ndim(\mathcal{H})$, is the cardinality of the largest set that the hypothesis class Natarajan shatters.

One nice property of the 0-1 loss function is the *identity of indescernibles.*

Definition 2.3 (Identity of Indiscernibles)**.** *A loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ has the identity of indiscernibles property if*

$$\forall y_1, y_2 \in \mathcal{Y}, \quad \ell(y_1, y_2) = 0 \iff y_1 = y_2$$

This property is seen in many well-used loss functions such as the 0-1 loss and the squared loss. Hopkins et al. (2022) show that this property can be used to create an agnostic learner from a realizable learner and has

been used to characterize learnability problems (Raman et al., 2024a;b). In fact, Hopkins et al. (2022) prove this result by relating the loss between $h(x)$ and $h'(x)$ to the classification error $\mathbb{P}(h(x) \neq h'(x))$. Thus, just like we saw in the discussion of the definition the realizable setting, this notion of equality of labels and loss values has been extremely interlinked when the loss function satisfies the identity of indiscernibles.

In this work we will explicitly **not** assume the identity of indiscernibles or any generalization of it to different output and label spaces. In fact, our assumptions assume hardly any structure at all, allowing our results to generalize to many settings.

We shall now introduce some notation. Given a loss function $\ell : \mathcal{Z} \times \mathcal{Y} \to \{0, 1\}$, let $C := C(\ell) = \{(z, y) \mid \ell(z, y) = 0\}$. We will refer to this the equality set. We also define $\sigma(z) := \sigma_\ell(z) = \{y \mid \ell(z, y) = 0\}$. $\sigma(z)$ is all the labels y where (z, y) achieves 0 loss. We similarly define $\tau(y) := \tau_\ell(y) = \{z \mid \ell(z, y) = 0\}$. We will oftentimes omit the ℓ subscript from σ, τ and ℓ parameter for C, but there is always an implicit loss function these are referring to. Further, notice that from any one of C, ℓ, σ, or τ, we can reconstruct the others.

Note how σ, τ create equivalence relations on $\mathcal{Z}$ and $\mathcal{Y}$ by $z_1 \sim_\sigma z_2 \iff \sigma(z_1) = \sigma(z_2)$ and similar for $\mathcal{Y}$ using τ. Let $\sigma(\mathcal{Z})$ and $\tau(\mathcal{Y})$ be these equivalence classes. We note that $|\sigma(\mathcal{Z})| < \infty \implies |\tau(\mathcal{Y})| < 2^{|\sigma(\mathcal{Z})|}$ and $|\tau(\mathcal{Y})| < \infty \implies |\sigma(\mathcal{Z})| < 2^{|\tau(\mathcal{Y})|}$.

3 Setup

Let us have an output space $\mathcal{Z}$ and label space $\mathcal{Y}$. We analyze loss functions ℓ that satisfy the following constraints:

1. $\ell : \mathcal{Z} \times \mathcal{Y} \to \{0, 1\}$.
2. $|\sigma(\mathcal{Z})| < \infty$ and $\forall z \in \mathcal{Z}$, $|\sigma(z)| < \infty$
3. $\forall z_1, z_2 \in \mathcal{Z}$, $\sigma(z_1) \not\subset \sigma(z_2)$. We note $\subset$ means *strict* subset; we do allow equalities.

Hopkins et al. (2022) have showed that one can have a realizable learner that is not agnosticly learnable in multiclass settings that do not have the identity of indiscernibles. The example they show relies on the loss function taking 3 values, 0, 1, and c. Thus, assumption 1 allows our loss functions to only take values of 0 or 1, and our results show that this is sufficient to have both agnostic and realizable learnability.

Assumption 2 is needed to allow our generalization of the Natarajan dimension to work. The Natarajan dimension characterizes multiclass learnability for finite output spaces. We can generalize this to infinite output and label spaces, so long as σ partitions our output set into a finite amount of partitions and each output has 0 loss on a finite amount of labels. Since our loss can only distinguish over equivalence classes, this can be thought of as bounding the "effective" output and label space sizes. We will refer to this as *effectively finite*.

Assumption 3 is here as the problem does not make intuitive sense without it. The goal of our learning problem is to minimize the loss. Since we have full access to the loss values, if the outputs that z_1 takes no loss on is strictly contained in the outputs z_2 takes no loss on (i.e. $\sigma(z_1) \subset \sigma(z_2)$), why would one ever want to output z_1? It would always be objectively better to output z_2. Thus, without loss of generality, we assume this does not happen, otherwise we can just replace all instances of z_1 with z_2 in our hypothesis class. In our characterization, this assumption is used in the proof that the finiteness of our dimension is necessary for learning. In Appendix D.1 we show a counterexample to our main theorem of characterization if this assumption is broken as well.

3.1 Example: Multilabel Ranking

A ranking learning problem is one where a learner is required to output a permutation of $[k]$ that indicates a ranking of k elements. An example could be outputting what a learner believes to be the ranking of a person's favorite movies. However, one might only care about whether or not the learner got the person's top 10 movies correct. Thus, while there are a total of K outputs, the quality of the output need not depend on

all K values, but only on the top $p \leq K$. For more information on multilabel ranking, we refer the readers to Raman et al. (2023).

For this example, we will assume the output and label space are the same. Given an input space $\mathcal{X}$, output/label space $\mathcal{Y}$ (permutations of $[K]$), and a hypothesis class $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$, we can get the following losses:

$$\ell_{0-1}^{sum@p}(y, y') = \begin{cases} 0 & \forall i \in [p] \quad y_i = y'_i \\ 1 & \text{otherwise} \end{cases}$$

and

$$\ell_{0-1}^{prec@p}(y, y') = \begin{cases} 0 & \{y_1, \ldots, y_p\} = \{y'_1, \ldots, y'_p\} \\ 1 & \text{otherwise} \end{cases}.$$

The $\ell_{0-1}^{sum@p}$ loss cares about if the ranking is in the correct exact order for the top p values. Meanwhile, the $\ell_{0-1}^{prec@p}$ loss only cares about if the values in the top are correct and order does not matter. These losses then create equivalence classes over our output/label space by whether the top p are in correct order for $\ell_{0-1}^{sum@p}$ or whether the top p are the correct set for $\ell_{0-1}^{prec@p}$. Further, no one ranking output dominates another output since they only get 0 loss on the values in their equivalence class. Thus, we can see this setting will match the assumptions of our learning problem.

In Section 5.1.2 we show a characterization of this setting along with the characterization of a more general version of the above that allows for the labels to be relevance scores of labels and not themselves a ranking. This in turn, by the results of Raman et al. (2023), show that our characterization applies to many well-used multilabel ranking setups.

4 The Generalized Natarajan Dimension

With spaces $\mathcal{Z}$, $\mathcal{Y}$ and loss function ℓ as defined above, the goal of this paper is to characterize PAC-learnability of a hypothesis class.

First, notice if $z \sim_\sigma z'$ and $y \sim_\tau y'$, then

$$\ell(z, y) = \ell(z', y) = \ell(z', y') = \ell(z, y')$$

This shows that our loss function respects our equivalence relation, and thus $\exists \ell^{\sigma,\tau} : \sigma(\mathcal{Z}) \times \tau(\mathcal{Y}) \to \{0, 1\}$ that agrees with ℓ. Thus:

Corollary 4.0.1. *Let $\mathcal{D}$ be a distribution on $\mathcal{X} \times \mathcal{Y}$, and $\mathcal{D}^{\sigma,\tau}$ be its quotient onto $\mathcal{X} \times \tau(\mathcal{Y})$. Further, let $h \in \mathcal{Z}^{\mathcal{X}}$, and h^σ be its quotient onto $(\sigma(\mathcal{Z}))^{\mathcal{X}}$. Then*

$$\mathbb{E}_{\mathcal{D}}[\ell(h(x), y)] = \mathbb{E}_{\mathcal{D}^{\sigma,\tau}}[\ell^{\sigma,\tau}(h^\sigma(x), y)]$$

One can also very easily convert a $(\mathcal{X}, \mathcal{Z}, \mathcal{Y}, \mathcal{H}, \ell)$ learning problem into a $(\mathcal{X}, \sigma(\mathcal{Z}), \tau(\mathcal{Y}), \sigma \circ \mathcal{H}, \ell^{\sigma,\tau})$ learning problem by creating the projection map $p : \mathcal{Z} \to \sigma(\mathcal{Z})$, and then using $p \circ h$ for predictions. This way it is also very easy to recover the original $h \in \mathcal{H}$. This can also be easily implemented in code by replacing the quotient spaces $\sigma(\mathcal{Z}), \tau(\mathcal{Y})$ with the set of representatives for the equivalence classes and then mapping each element in $\mathcal{Z}, \mathcal{Y}$ to its representative. Thus, we can see that the learnability of $(\mathcal{X}, \sigma(\mathcal{Z}), \tau(\mathcal{Y}), \sigma \circ \mathcal{H}, \ell^{\sigma,\tau})$ implies the learnability of $(\mathcal{X}, \mathcal{Z}, \mathcal{Y}, \mathcal{H}, \ell)$ while only using assumptions 1 and 2.

We also note that we use $(\mathcal{X}, \mathcal{Z}^C, \mathcal{Y}^C, \mathcal{H}^C, \ell^C)$ as alternative notation to $(\mathcal{X}, \sigma(\mathcal{Z}), \tau(\mathcal{Y}), \sigma \circ \mathcal{H}, \ell^{\sigma,\tau})$. When the context is clear, we will drop superscripts and abuse notation between the equivalence classes versions of $\mathcal{Z}$, $\mathcal{Y}$, $\mathcal{H}$, ℓ and themselves.

Now, we introduce our new dimension, the Generalized Natarajan dimension, based off Natarajan (1989).

Definition 4.1 (Generalized Natarajan Dimension)**.** *We say that a hypothesis class $\mathcal{H}$ and loss function ℓ Generalized Natarajan shatters a set $S = \{s_1, ..., s_n\}$ if $\exists h_1, h_2 \in \mathcal{H}$ such that*

1. $\forall s_i \in S$, $\sigma(h_1(s_i)) \neq \sigma(h_2(s_i))$
2. $\forall S' \subseteq S$ $\exists h \in \mathcal{H}$ *where* $\forall s \in S'$ $\sigma(h(s)) = \sigma(h_1(s))$ *and for all* $s \in S \setminus S'$ $\sigma(h(s)) = \sigma(h_2(s))$

The Generalized Natarajan dimension of a hypothesis class $\mathcal{H}$, denoted $GNdim(\mathcal{H}, \ell)$, is the cardinality of the largest set that the hypothesis class generalized Natarajan shatters.

From this we have the immediate corollary:

Corollary 4.0.2. *Given a hypothesis class $\mathcal{H}$ and loss function ℓ as studied in this paper, we have*

$$GNdim(\mathcal{H}, \ell) = Ndim(\sigma \circ \mathcal{H}) = GNdim(\sigma \circ \mathcal{H}, \ell^{\sigma,\tau})$$

The Generalized Natarajan Dimension effectively changes what it means to be equal. One implicit property relied on when a loss function has the identity of indiscernibles is the relationship between equality and loss value. Given $\mathcal{Z} = \mathcal{Y}$, we have that $y_1 = y_2$ if and only if $\ell(y_1, y_2) = 0$ and $y_1 \neq y_2$ if and only if $\ell(y_1, y_2) > 0$. This allows us to use equality of our labels to infer how the loss will act. This is now no longer the case; we can have $y_1 \neq y_2$ but still have $\ell(y_1, y_2) = 0$. Because of this, our dimension needs to explicitly use the loss values instead of being able to use the labels by proxy.

In Theorem 4.1 we show the Generalized Natarajan dimension characterizes our setting. Thus, the proxy for equality needed is for the equivalency of the set of labels where z_1 and z_2 achieve 0 loss for (i.e. $\sigma(z_1) = \sigma(z_2)$). This is a very strict type of equality. One can think of a setting where $|\mathcal{Z}|$ is very large (but finite) with $\sigma(z_1)$ and $\sigma(z_2)$ being almost equal. However, since they are not exactly equal, this dimension does not get any smaller than the Natarajan dimension. This is quite counter-intuitive as one would expect learnability to be much easier in this setting due to the vast amount of values that will output 0 loss, but it is not as you can always make a distribution that puts a lot of mass where $\sigma(z_1)$ and $\sigma(z_2)$ differ.

4.1 Characterizing Learnability of Forgiving 0-1 Loss Functions

With this setup, we state our main theorem:

Theorem 4.1. *A $(\mathcal{X}, \mathcal{Z}, \mathcal{Y}, \mathcal{H}, \ell)$ learning problem is PAC-learnable if and only if $GNdim(\mathcal{H}, \ell) < \infty$.*

We prove this through showing the necessity of $GNdim(\mathcal{H}, \ell) < \infty$ in Lemma 4.2 and then showing sufficiency in Lemma 4.3.

The only if direction is proven through an alteration of the No-Free-Lunch Theorem (Shalev-Shwartz & Ben-David, 2014). The if direction is done through a reduction into the $(\mathcal{X}, \sigma(\mathcal{Z}), \tau(\mathcal{Y}), \mathcal{H}, \ell^{\sigma,\tau})$ learning problem. Once this reduction is done, we can bound the the VC-dimension of the loss class by the Generalized Natarajan dimension. This then shows that ERM is a learner and gives an upper bound on the sample complexity.

Lemma 4.2. *Let us have a learning problem $(\mathcal{X}, \mathcal{Z}, \mathcal{Y}, \mathcal{H}, \ell)$ as has been studied in this paper. Then, $(\mathcal{X}, \mathcal{Z}, \mathcal{Y}, \mathcal{H}, \ell)$ is PAC-learnable $\implies$ the Generalized Natarajan Dimension of $\mathcal{H}$ is finite.*

Proof Sketch. We show that $GNdim(\mathcal{H}, \ell) = \infty \implies (\mathcal{X}, \mathcal{Z}, \mathcal{Y}, \mathcal{H}, \ell)$ is not learnable. To do this, we modify the No Free Lunch Theorem to change the realizable distributions we are looking over. In the No Free Lunch Theorem for the 0-1 loss, we have a shattered set $\mathcal{X}$ and create distributions over $\mathcal{X}$ where the labels are labeled by the $h_i \in \mathcal{H}$ that witness the shattering of $\mathcal{X}$. This does not work in our setting for two reasons: one is that our output space and label space need not be the same and the second is, even if they were, we can have $h_i(x) \neq f(x)$ but still have $\ell(f(x), h_i(x)) = 0$. To fix this, we change our realizable distribution to be uniform over the outputs in $\sigma(h_i(x))$ to make sure if $h \in \mathcal{H}$, $\sigma(h(x)) \neq \sigma(h_i(x))$, then $\exists y$ such that $\ell(h(x), y) \neq \ell(h_i(x), y)$. The rest of the proof is then modified to work with these changes. □

We give the full proof of the above in Appendix B.

Since we have shown necessity the of $GNdim(\mathcal{H}, \ell) < \infty$, we now show sufficiency.

Lemma 4.3. *Let us have a learning problem $(\mathcal{X}, \mathcal{Z}, \mathcal{Y}, \mathcal{H}, \ell)$ as has been studied in this paper. Then, $GNdim(\mathcal{H}, \ell)$ is finite $\implies$ Learnability of $(\mathcal{X}, \mathcal{Z}, \mathcal{Y}, \mathcal{H}, \ell)$.*

We give two proofs of this lemma in Appendix C. Appendix C.1 uses the VC-dimension of the loss class and Appendix C.2 shows a different proof of a special case of when $\mathcal{Z} = \mathcal{Y}$. While this specialized proof also gives a worse sample complexity bound than the VC-dimension proof, it is included as we believe a direct proof through uniform convergence adds intuition.

Given lemmas 4.2 and 4.3, we have now proved Theorem 4.1. We note the sufficiency proof does not rely assumption 3. Therefore, for any effectively finite learning problem with a loss function whose output is in $\{0, 1\}$, we have that finite $GNdim(\mathcal{H}, \ell)$ is sufficient for learnability.

To find sample complexity, we use the VC-dimension of the loss class. Since we have now shown that $GNdim(\mathcal{H}, \ell)$ characterizes our setting, Lemma A.1 shows that the VC-dimension of the loss class does as well. From this, we use the well-known sample complexity bounds of binary classification get the following result:

Corollary 4.3.1. *Given an $(\mathcal{X}, \mathcal{Z}, \mathcal{Y}, \mathcal{H}, \ell)$ learning problem studied in this paper, we have the following bounds on the agnostic PAC-learning sample complexity:*

$$m(\epsilon, \delta) \leq O\left(\frac{GNdim(\mathcal{H}, \ell) \log\left(|\sigma(\mathcal{Z})|\right) + \log(1/\delta)}{\epsilon^2}\right).$$

Bressan et al. (2025) state that a class is not learnable when there is at least a pair of outputs that are difficult to distinguish. Our results show that, while true, this does not tell the full picture. Two outputs being indistinguishable is also the only way we can get a smaller upper bound for the sample complexity as well. In fact, over all learning settings that satisfy our assumptions, the only way we can have our hypothesis class go from unlearnable to learnable is through changing the loss function to have more equivalent outputs.

However, this can be a blessing or a curse. If a loss function ℓ and hypothesis class $\mathcal{H}$ pair has Generalized Natarajan dimension d, then creating an ℓ' such that $C(\ell) \subset C(\ell')$ can lead to a larger Generalized Natarajan dimension, and thus possibly a larger sample complexity required to learn. We show an example of this, albeit contrived, in the proof of Proposition 2.

We can see from above then how the "forgiveness" of a loss function is hypothesis class-dependent. We also see that, for a given hypothesis class, all loss functions that induce the same equivalence classes on $\mathcal{Z}$ will have the same Generalized Natarajan dimension. For example, if a loss does not induce any equivalent outputs, then, by Corollary 4.0.1, we have that our Generalized Natarajan dimension of this hypothesis class/loss pair is equal to the Natarajan dimension of the hypothesis class. Thus, even if a loss function looks to be more forgiving, in terms of learnability it might not be. Therefore, the colloquial meaning of a "forgiving" loss function and what loss functions are truly more forgiving are at odds with each other.

4.2 Relation to Other Dimensions

In Corollary A.1.1 of the appendix we show that the VC-dimension of the loss class also characterizes our setting. The fact that the Generalized Natarajan dimension and the VC-dimension of the loss class both rely on both the hypothesis class and the loss function is crucial to them being able to characterize the setting. Below, we show that any dimension that relies solely on comparing hypotheses through the normal notions of "shattering" will not characterize our setting.

Here we give the incomparability results of the Natarajan Dimension and the d_J dimension (Bressan et al., 2025).

Proposition 1. *Let $\mathcal{X}$ be our input space and $\mathcal{Y}$, $|\mathcal{Y}| < \infty$ be our label space. There exists a hypothesis class $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ such that $\forall J \subset \mathcal{Y}$, $Ndim(\mathcal{H}) = d_J(\mathcal{H}) = \infty$, but there exists a loss function ℓ such that $GNdim(\mathcal{H}, \ell) = 0$.*

Proof. Let $|\mathcal{X}| = \infty$, $\mathcal{H} = \mathcal{Y}^{\mathcal{X}}$ and let $\ell(y, y') = 0$ for all $y, y' \in \mathcal{Y}$. Note that, since $\mathcal{H}$ is all functions from inputs to outputs, then, $Ndim(\mathcal{H}) = \infty$, and $\forall J \subset \mathcal{Y}$, $d_J(\mathcal{H}) = \infty$. However, since our loss partitions $\mathcal{Y}$ to all the same equivalence class, $\sigma \circ \mathcal{H}$ contains only 1 element, thus $GNdim(\mathcal{H}, \ell) = 0$. □

Proposition 2. *For all $q \in \mathbb{N}$, there exists a finite input space $\mathcal{X}$, finite output space $\mathcal{Y}$, finite hypothesis class $\mathcal{H}$, and loss function ℓ such that $Ndim(\mathcal{H}) = 1$, and $\forall J \subset \mathcal{Y}$, $d_J(\mathcal{H}) \leq 1$, but $GNdim(\mathcal{H}, \ell) = q$.*

Proof. Let $q \in \mathbb{N}$, $\mathcal{X} = \{x_1, \ldots, x_q\}$, $\mathcal{Y} = \{y_1, \ldots, y_{q^{q+1}}\}$, let

$$h_i(x_j) = y_{q(i-1)+j},$$

and let $\mathcal{H} = \{h_i \mid i \in [q^q]\}$. Notice how $Ndim(\mathcal{H}) = 1$ as no hypotheses have any overlapping outputs. In Bressan et al. (2025), they state that $Ndim(\mathcal{H}) \geq \max_J d_J(\mathcal{H})$, thus $\forall J \subset \mathcal{Y}$, $d_J(\mathcal{H}) \leq 1$. Now, let $\mathcal{H}' = [q]^{\mathcal{X}}$. Note there are $|\mathcal{H}'| = q^q$, thus create an arbitrary bijection $f : \mathcal{H}' \to \mathcal{H}$. Partition $\mathcal{Y}$ on the following:

$$y \sim y' \iff \exists h', h^\dagger \in \mathcal{H}', \exists x_i \in \mathcal{X} \text{ where } h'(x_i) = h^\dagger(x_i) \text{ and } f(h')(x_i) = y, f(h^\dagger)(x_i) = y'$$

Notice this partitions $\mathcal{Y}$ such that each partition is equivalent to a specific $c \in [q]$ for the outputs of $\mathcal{H}'$. Thus, let

$$\ell(y, y') = \begin{cases} 0 & y \sim y' \\ 1 & \text{otherwise} \end{cases}.$$

From the definition of the equivalence class, we can see that $\sigma \circ \mathcal{H}$ is then equivalent to all functions from $\{x_1, \ldots, x_q\}$ to $[q]$, which has a Natarajan dimension of q. □

We do note that, due to our characterization results along with results in Ben-David et al. (1995) and Bressan et al. (2025), we have that $GNdim(\mathcal{H}, \ell) = \infty \implies Ndim(\mathcal{H}) = \infty \implies \max_J d_J(\mathcal{H}) = \infty$. Thus, while the Generalized Natarajan dimension can be be infinitely smaller and arbitrarly larger than these dimensions, it can not be infintely larger.

Since the Natarajan dimension characterizes the 0-1 loss in our setting, is is not surprising to see cases where $GNdim(\mathcal{H}, \ell) < Ndim(\mathcal{H})$, but it might be surprising to see $GNdim(\mathcal{H}, \ell) > Ndim(\mathcal{H})$. However, we believe that looking through the angle of *discernment* gives the intuition for why this can happen. The job of the algorithm is to, through the samples given, discern which hypothesis are useful and which are not. Forgiving losses, while bringing the actual loss values of all hypotheses down, does not need to add information on which hypothesis might be better than others. For example, it might help some hypotheses more than others, but the only way to tell is to find points on which these hypotheses differ. Now though, this job is harder since hypotheses that did disagree before now agree on more places. Due to this difficulty in discernment, it has become harder for the algorithm to find a hypothesis that is close to the infimum, even though all the loss values are lowered.

We also note that the same proof examples and ideas as above can be used to show the incomparability of the Generalized Natarajan dimension to the DS-dimension (Daniely & Shalev-Shwartz, 2014; Brukhim et al., 2022) and the k-DS-dimension (Charikar & Pabbaraju, 2023).

Corollary 4.3.2. *There exist input spaces $\mathcal{X}, \mathcal{X}'$, output spaces $\mathcal{Y}, \mathcal{Y}'$, hypothesis classes $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}, \mathcal{H}' \subset \mathcal{Y}'^{\mathcal{X}'}$, and loss functions ℓ, ℓ' such that $GNdim(\mathcal{H}, \ell) = 0$ and $DS(\mathcal{H}) =$$k$-$DS(\mathcal{H}) = \infty$ and $GNdim(\mathcal{H}', \ell') = q \in \mathbb{N}$ and $DS(\mathcal{H}') = k$-$DS(\mathcal{H}') = 0$ for any $k \in \mathbb{N}$, $k \geq 1$.*

One can see that $Ndim(\mathcal{H}) = \infty \implies DS(\mathcal{H}) = \infty$ since all distributions over $\mathcal{X}$ and a finite subset of $\mathcal{Y}$ is included is a subset of distributions over all $\mathcal{X} \times \mathcal{Y}$. Thus, we have as well $GNdim(\mathcal{H}, \ell) = \infty \implies DS(\mathcal{H}) = \infty$. It is known that for any $k < k'$, k-$DS(\mathcal{H}) >$$k'$-$DS(\mathcal{H})$ (Charikar & Pabbaraju, 2023), thus we posit a conjecture that $GNdim(\mathcal{H}, \ell) = \infty$ does **not** imply k-$DS(\mathcal{H}) = \infty$ for $k > 1$. This conjecture comes from the ideas stated in Section 1.1 where the modified list-learning setting can be both easier and harder than the normal list-learning setting.

5 Applications of Characterization

Given the minimal assumptions made on our learning setting, the Generalized Natarajan dimension is able to characterize many machine learning scenarios. In this section we detail some of these settings.

5.1 Characterizing Learnability of Set Learning

We define set learning as receiving set-valued feedback and checking whether or not the output is in the set. More precisely, let $\mathcal{X}$ be our input space, let $\mathcal{Z}$, $2 < |\mathcal{Z}| < \infty$, be our output space, let our label space be $\mathcal{Y} \subset 2^{\mathcal{Z}}$, such that no $y, y' \in \mathcal{Y}$ have $y \subseteq y'$ or $y' \subseteq y$, and let us have a hypothesis class $\mathcal{H} \subset \mathcal{Z}^{\mathcal{X}}$. The samples we receive are of the form $(x, v) \in \mathcal{X} \times \mathcal{Y}$. Finally, our loss function becomes for any $z \in \mathcal{Y}$ and for any $y \in \mathcal{Y}$:

$$\ell(z, y) = \begin{cases} 0 & z \in y \\ 1 & \text{otherwise} \end{cases}$$

We give an example of what this loss function looks like for $\mathcal{Z} = \{1, 2, 3\}$ and $\mathcal{Y} = 2^{\mathcal{Z}} \setminus \{\emptyset\}$ in Figure 1.

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccccccc} \{1\} & \{2\} & \{3\} & \{1,2\} & \{1,3\} & \{2,3\} & \{1,2,3\} \end{array} \\ \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

Figure 1: An example loss matrix for the set learning setup where $\mathcal{Z} = \{1, 2, 3\}$ $\mathcal{Y} = 2^{\{1,2,3\}} \setminus \{\emptyset\}$. Since each $z \in \mathcal{Z}$ is only equivalent to themselves, we see this is characterized by the Natarajan dimension. Note this would still be true if $\mathcal{Y} = 2^{\{1,2,3\}} \setminus \{\emptyset, \{1\}, \{2\}, \{3\}\}$ as well.

Note how these problems fit the assumptions in this paper. Thus, we have the following corollary:

Corollary 5.0.1. *For all instantiations of set learning that meet our assumptions, the learnability of the learning problem is characterized by the Natarajan dimension.*

This problem has been characterized in the online setting (Raman et al., 2024b), but for a majority of these instantiations the batch setting has remained open as far as we are aware until now.

This is a very useful result as many problems can be formulated as a set-learning problem. For example, any setting that is classifying up-to equivalence classes is included under this setting. This is because, for any $y, y' \in \mathcal{Y}$, $\ell(y, y') = \ell([y], [y'])$ where $[y], [y']$ are the equivalence classes y, y' are in, respectively. In the following subsections we give two concrete examples.

5.1.1 Classifying Graphs up to Isomorphism

Classifying graphs is well-used in fields such as drug discovery (Yang et al., 2024). In cases such as molecules, it does not matter which graph is outputted, so long as it is isomorphic to the correct label graph. It is known that graph isomorphisms form a partition over the graph space, thus we have the following:

Corollary 5.0.2. *Given an input space $\mathcal{X}$ and output/label space of graphs of with finite amount of edges/nodes, the Generalized Natarajan dimension characterizes learnability over any 0-1 loss function that respects graph isomorphisms (i.e. $\forall y, y', y^\dagger \in \mathcal{Y}$, y isomorphic to y', $\ell(y, y^\dagger) = \ell(y', y^\dagger)$).*

5.1.2 Multilabel Ranking

Since the example in Section 3.1 meets the assumptions of our learning problem, we can see our characterization applies there.

To generalize this example to more real-life scenarios, we note that, as discussed in Raman et al. (2023), labels are normally given by a relevance score, not an exact ranking. A relevance score is an ordered set $y \in \{0, 1, \ldots, B\}^K$ where the ith value denotes how relevant the ith label is (a higher value means more relevant). They create two losses:

$$\ell^{sum@p}(z, y) = \sum_{i=1}^{K} \min\left(z_i, p+1\right) y^i - C_y^p$$

and

$$\ell^{prec@p}(z, y) = \sum_{i=1}^{K} 1_{\{z_i \leq p\}} y^i - C_y^p,$$

where y^i is the relevance score of the label ranked at z_i and the C_y^p values are a normalization constant that depend only on p and y that ensure

$$\min_{z \in \mathcal{Z}} \ell^{prec@p}(z, y) = \min_{z \in \mathcal{Z}} \ell^{sum@p}(z, y) = 0.$$

The loss $\ell^{sum@p}(z, y)$ essentially checks if the values and the order are correct for the top p values and $\ell^{sum@p}(z, y)$ checks whether or not the set of the top p values are correct. From these, we can see there is no z, z' such that $\sigma(z) \subset \sigma(z')$ for either loss.

Raman et al. (2023) also show that, for either of the following sets, that if one loss is learnable, then any loss in that set is learnable:

$$\begin{aligned}
&\mathcal{L}(\ell^{sum@p}) = \\
&\{\ell \in \mathbb{R}^{\mathcal{Z} \times \mathcal{Y}} \mid \ell = 0 \iff \ell^{sum@p} = 0\} \cap \{\ell \in \mathbb{R}^{\mathcal{Z} \times \mathcal{Y}} \mid z \overset{[p]}{=} z' \implies \ell(z, y) = \ell(z', y)\} \\
&\mathcal{L}(\ell^{prec@p}) = \\
&\{\ell \in \mathbb{R}^{\mathcal{Z} \times \mathcal{Y}} \mid \ell = 0 \iff \ell^{prec@p} = 0\} \cap \{\ell \in \mathbb{R}^{\mathcal{Z} \times \mathcal{Y}} \mid z \overset{p}{=} z' \implies \ell(z, y) = \ell(z', y)\}.
\end{aligned}$$

where $z \overset{[p]}{=} z'$ denotes the order of the top p are the same and $z \overset{p}{=} z'$ denotes the set of the top p are the same. Raman et al. (2023) show that many well-used ranking loss functions are captured in these sets of losses.

Now, let us create the following loss functions:

$$\begin{aligned}
\ell_{0-1}^{sum@p}(z, y) &= 1_{\{\ell^{sum@p}(z,y) > 0\}} \\
\ell_{0-1}^{prec@p}(z, y) &= 1_{\{\ell^{prec@p}(z,y) > 0\}}.
\end{aligned}$$

Notice how these are forgiving $0-1$ loss functions that satisfy the assumptions for our learning problem. Thus, the Generalized Natarajan Dimension characterizes these functions. Further, notice $\ell_{0-1}^{sum@p}(z, y) \in \mathcal{L}(\ell^{sum@p})$ and $\ell_{0-1}^{prec@p}(z, y) \in \mathcal{L}(\ell^{prec@p})$. Therefore, we get the following corollary:

Corollary 5.0.3. *$GNdim(\mathcal{H}, \ell_{0-1}^{prec@p})$ characterizes all loss functions in $\mathcal{L}(\ell^{prec@p})$ and $GNdim(\mathcal{H}, \ell_{0-1}^{sum@p})$ characterizes all loss functions in $\mathcal{L}(\ell^{sum@p})$.*

This is the first characterization that we are aware of that gives a dimension on the entire hypothesis class and not on individual indices.

5.2 Characterizing Learnability of Modified List Learning

As noted in Section 1.1, we characterize a modified version of list learning for many possible set-ups. This version is the "flip" of set learning where now our hypotheses output lists and we want to minimize $\mathbb{P}\left(y \notin h(x)\right)$.

Thus, given possible labels $\mathcal{Y}$, $|\mathcal{Y}| < \infty$, a set of lists $\mathcal{Z} \subset 2^{\mathcal{Y}}$ where $\forall z, z' \in \mathcal{Z}$, $z \not\subset z'$, hypothesis class $\mathcal{H} \subset \mathcal{Z}^{\mathcal{X}}$ and loss:

$$\ell(\{y_{i_1}, \ldots, y_{i_n}\}, y) = \begin{cases} 0 & y \in \{y_{i_1}, \ldots, y_{i_n}\} \\ 1 & \text{otherwise} \end{cases}$$

for $|i| = n < |\mathcal{Y}|$, this gives us a list learning problem that satisfies our constraints. Thus, we have the following corollary:

Corollary 5.0.4. *The version of list learning described above is characterized by the Generalized Natarajan dimension.*

6 Conclusion and Future Work

In this paper we have characterized a large set of more forgiving 0-1 loss functions in the effectively finite label multiclass setting.

Through the properties of our Generalized Natarajan dimension, we can see that the Natarajan dimension will still characterize many of these loss functions as well. Many of these loss functions could be mostly zeros, but if the quotient space of the outputs is the same as the output space, the learnability is the same as in the 0-1 loss. While this is unintuitive, we believe this is due to how stringent PAC-learnability is. Since PAC-learnability requires learnability over all distributions, one can see how, given a hypothesis class and a loss function that does not reduce the output space, one can adversarially create a distribution that puts a lot of mass on the areas where the sets $\sigma(h(x))$ disagree. This will then nullify the "forgiving" portion of the loss function. Overall, thinking through discernment shows that forgiving losses might or might not actually live up to their name.

Further work can be done by finding a way to remove the assumption on the loss function to allow for outputs where one output is "dominated" by the other. Also, seeing whether or not extensions to effectively infinite output/label spaces would mirror the Natarajan/DS dimension split as in the 0-1 loss case would be an interesting result. Finally, further work of tightening the sample complexity bounds can be useful to give quantitative results on forgiveness. For example, deriving a lower bound might show that forgiveness can help a lot on specific distributions. Also, since our analysis of the upper bound of the sample complexity goes through the VC-dimension of the loss class, there is potential work to show the tightness of this bound.

References

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

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

Marco Bressan, Nataly Brukhim, Nicolò Cesa-Bianchi, Emmanuel Esposito, Yishay Mansour, Shay Moran, and Maximilian Thiessen. A fine-grained characterization of pac learnability. In Nika Haghtalab and Ankur Moitra (eds.), *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 641–676. PMLR, 30 Jun–04 Jul 2025. URL `https://proceedings.mlr.press/v291/bressan25b.html`.

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

Moses Charikar and Chirag Pabbaraju. A characterization of list learnability. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pp. 1713–1726, 2023.

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

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

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

Yann Guermeur. Vc theory of large margin multi-category classifiers. *The Journal of Machine Learning Research*, 8:2551–2594, 2007.

Steve Hanneke, Shay Moran, and Waknine Tom. List sample compression and uniform convergence. In *The Thirty Seventh Annual Conference on Learning Theory*, pp. 2360–2388. PMLR, 2024.

Steve Hanneke, Qinglin Meng, and Amirreza Shaeiri. Representation preserving multiclass agnostic to realizable reduction. In *Forty-second International Conference on Machine Learning*, 2025.

Max Hopkins, Daniel M Kane, Shachar Lovett, and Gaurav Mahajan. Realizable learning is all you need. In *Conference on Learning Theory*, pp. 3015–3069. PMLR, 2022.

Alkis Kalavasis, Grigoris Velegkas, and Amin Karbasi. Multiclass learnability beyond the pac framework: Universal rates and partial concept classes. *Advances in Neural Information Processing Systems*, 35: 20809–20822, 2022.

Liping Liu and Thomas Dietterich. Learnability of the superset label learning problem. In *International conference on machine learning*, pp. 1629–1637. PMLR, 2014.

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

Vinod Raman, Unique Subedi, and Ambuj Tewari. On the learnability of multilabel ranking. *Advances in Neural Information Processing Systems*, 36:11988–11999, 2023.

Vinod Raman, Unique Subedi, and Ambuj Tewari. A characterization of multioutput learnability. *Journal of Machine Learning Research*, 25(342):1–54, 2024a.

Vinod Raman, Unique Subedi, and Ambuj Tewari. Online learning with set-valued feedback. In *The Thirty Seventh Annual Conference on Learning Theory*, pp. 4381–4412. PMLR, 2024b.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

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

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

Qi Wang, Yue Ma, Kun Zhao, and Yingjie Tian. A comprehensive survey of loss functions in machine learning. *Annals of Data Science*, 9(2):187–212, 2022.

Nianzu Yang, Huaijin Wu, Kaipeng Zeng, Yang Li, Siyuan Bao, and Junchi Yan. Molecule generation for drug design: a graph learning perspective. *Fundamental Research*, 2024.

Jianing Zhou and Suma Bhat. Paraphrase generation: A survey of the state of the art. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5075–5086, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.414. URL `https://aclanthology.org/2021.emnlp-main.414/`.

A Proving the Relationship Between $VC(\ell \circ \mathcal{H})$ and $GNdim(\mathcal{H}, \ell)$

Here we show that the VC-dimenion of the loss class and the Generalized Natarajan Dimension are bounded by each other.

Lemma A.1. *Given a $(\mathcal{X}, \mathcal{Z}, \mathcal{Y}, \mathcal{H}, \ell)$ learning problem, we have*

$$GNdim(\mathcal{H}, \ell) \leq VC(\ell \circ \mathcal{H}) \leq 4.67 GNdim(\mathcal{H}, \ell) \log_2(|\sigma(Z)| + 1).$$

Proof. First we will prove $GNdim(\mathcal{H}, \ell) \leq VC(\ell \circ \mathcal{H})$. Let S, $|S| = n$ GN-shatter $\mathcal{H}, \ell$. Then $\exists f_1, f_2 \in \mathcal{H}$ such that $\forall x \in S$, $\sigma(f_1(x)) \neq \sigma(f_2(x))$ and $\forall T \subset S$, $\exists h_T \in \mathcal{H}$ where $\forall x \in T$, $\sigma(h_T(x)) = \sigma(f_1(x))$ and $\forall x \in S \setminus T$, $\sigma(h_T(x)) = \sigma(f_2(x))$. Let $S' = \{(x, y) \mid x \in S, y \in \sigma(f_1(x)) \triangle \sigma(f_2(x))\}$ where $\triangle$ denotes the symmetric difference of the two sets. Using f_1, f_2 and h_T as above we can see that $\ell \circ \mathcal{H}$ is VC-shattered by S'.

Now we will prove $VC(\ell \circ \mathcal{H}) \leq 4.67 GNdim(\mathcal{H}, \ell) \log_2(|\sigma(Z)| + 1))$. It is easy to see that for any set $S \subset \mathcal{X} \times \mathcal{Y}$ we have $|(\ell \circ \mathcal{H})(S)| \leq |\sigma(\mathcal{H})(S_{\mathcal{X}})|$. This is because if $\ell \circ h$ and $\ell \circ h'$ differ on a point, then h, h' must be different on that point as well.

Note how $\sigma(\mathcal{H})$ only has $|\sigma(\mathcal{Z})|$ different possible outputs. By Ben-David et al. (1995), we have for m samples:

$$|\sigma \circ \mathcal{H}| \leq \left(\frac{me(|\sigma(\mathcal{Z})| + 1)^2}{2Ndim(\sigma \circ \mathcal{H})} \right)^{Ndim(\sigma \circ \mathcal{H})}$$

Which, by 4.0.2, we have

$$\left(\frac{me(|\sigma(\mathcal{Z})| + 1)^2}{2Ndim(\sigma \circ \mathcal{H})} \right)^{Ndim(\sigma \circ \mathcal{H})} = \left(\frac{me(|\sigma(\mathcal{Z})| + 1)^2}{2GNdim(\mathcal{H}, \ell)} \right)^{GNdim(\mathcal{H}, \ell)}$$

Thus, for a sample of size $VC(\ell \circ \mathcal{H})$ to shatter $\ell \circ \mathcal{H}$, we have:

$$2^{VC(\ell \circ \mathcal{H})} \leq |\ell \circ \mathcal{H}| \leq |\sigma \circ \mathcal{H}| \leq \left(\frac{VC(\ell \circ \mathcal{H}) e(|\sigma(\mathcal{Z})| + 1)^2}{2GNdim(\mathcal{H}, \ell)} \right)^{GNdim(\mathcal{H}, \ell)}$$

This is the same inequality set-up as seen in Ben-David et al. (1995), thus we get the same bounds, except we substitute in $GNdim(\mathcal{H}, \ell)$ and $|\sigma(\mathcal{Z})|$ to get

$$VC(\ell \circ \mathcal{H}) \leq 4.67 GNdim(\mathcal{H}, \ell) \log_2(|\sigma(Z)| + 1)$$

□

Given Thoerem 4.1 and our assumption of $|\sigma(\mathcal{Z})| < \infty$ we have the following corollary:

Corollary A.1.1. *Given a $(\mathcal{X}, \mathcal{Z}, \mathcal{Y}, \mathcal{H}, \ell)$ learning problem studied in this chapter, we have*

$$(\mathcal{X}, \mathcal{Z}, \mathcal{Y}, \mathcal{H}, \ell) \text{ is learnable } \iff VC(\ell \circ \mathcal{H}) < \infty$$

B Finite $GNdim(\mathcal{H}, \ell)$ is Necessary for Learnability Proof

B.1 Proof by Adapting The No Free Lunch Theorem

Proof. We will show the contrapositive. Let us have a $(\mathcal{X}, \mathcal{Z}, \mathcal{Y}, \mathcal{H}, \ell)$ learning problem with $GNdim(\mathcal{H}, \ell) = \infty$. Then, by shattering, for any m we can find a set of $2m$ points where $f_1, f_2, h \in \mathcal{H}$ such that $\sigma(f_1(x)) \neq \sigma(f_2(x))$ for all $x \in \{x_1, \ldots, x_{2m}\}$ but $\sigma(h(x)) = \sigma(f_1(x))$ for $x \in \{x_1, \ldots, x_m\}$ and $\sigma(h(x)) = \sigma(f_2(x))$ for the others. Let $T = 2^{2m}$, and $h_1, ..., h_T$ be all of the functions that are guaranteed to exists by the shattering condition.

Let us define the distribution D_{f_1} as follows:

$$\mathbb{P}_{D_{f_1}}(x) = \begin{cases} 1/2m & x \in \{x_1, \ldots, x_{2m}\} \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbb{P}_{D_{f_1}}(y \mid x) = \begin{cases} 1/|\sigma(f_1(x))| & y \in \sigma(f_1(x)) \\ 0 & \text{otherwise} \end{cases}$$

Notice that $\mathbb{P}_{D_{f_1}}$ is a uniform distribution over our $2m$ inputs ($D_{f1}^{\mathcal{X}}$), and the output values are also uniformly distributed over the possible values where f_1 would get 0 loss on ($D_{f1}^{\mathcal{Y}|\mathcal{X}}$). Let us define D_i, $D_i^{\mathcal{Y}|\mathcal{X}}$, and $D_i^{\mathcal{X}}$ the same as above, but for h_i instead of f_1. Note each of these D_i distributions are realizable under $\mathcal{H}$.

Let $S_{X_1}, ..., S_{X_k}$ denote all of the samples of size m sampled from $D_i^{\mathcal{X}}$. Note k is the same for every i as this is just all samples of size m from $\mathcal{X}$. We will denote S^i_{Y,X_j} as the full (x, y) samples and S^i_Y will be the $\mathcal{Y}$ values from sample S^i_{Y,X_j}. While there is an implicit X_j needed for S^i_Y, we drop it to ease notation as we do not ever look at two S^i_Y with different X_js at the same time. Since each sample of S_{X_j} of size m is equally likely to be chosen, we can use the law of total expectation to get

$$\mathbb{E}_{S\sim D_i}[L_{D_i}(A(S))] = \frac{1}{k}\sum_{j=1}^{k} \mathbb{E}_{S^i_Y \sim D_i^{\mathcal{Y}|S_{X_j}}}[L_{D_i}(A(S^i_{Y,X_j}))]$$

where the expectation on the right hand side is over all possible label values we could get from our input sample X_j. Thus we get

$$\max_{i\in[T]} \mathbb{E}_{S\sim D_i}[L_{D_i}(A(S))] = \max_{i\in[T]} \frac{1}{k}\sum_{j=1}^{k} \mathbb{E}_{S^i_Y \sim D_i^{\mathcal{Y}|S_{X_j}}}[L_{D_i}(A(S^i_{Y,X_j}))] \geq$$
$$\frac{1}{T}\sum_{i=1}^{T}\frac{1}{k}\sum_{j=1}^{k} \mathbb{E}_{S^i_Y \sim D_i^{\mathcal{Y}|S_{X_j}}}[L_{D_i}(A(S^i_{Y,X_j}))] =$$
$$\frac{1}{k}\sum_{j=1}^{k}\frac{1}{T}\sum_{i=1}^{T} \mathbb{E}_{S^i_Y \sim D_i^{\mathcal{Y}|S_{X_j}}}[L_{D_i}(A(S^i_{Y,X_j}))] \geq$$
$$\min_{j\in[k]} \frac{1}{T}\sum_{i=1}^{T} \mathbb{E}_{S^i_Y \sim D_i^{\mathcal{Y}|S_{X_j}}}[L_{D_i}(A(S^i_{Y,X_j}))]$$

Let us fix a $j \in [k]$, and let $\{s_1, ...s_p\} = \mathcal{X} \setminus S^i_{X_j}$. Notice $p \geq m$ as $|\mathcal{X}| = 2m$ and $|S^i_{X_j}| \leq m$. We then get for any function $f : \mathcal{X} \to \mathcal{Z}$

$$L_{D_i}(f) = \frac{1}{2m}\sum_{r=1}^{2m} \mathop{\mathbb{E}}_{v\sim Unif(\sigma(h_i(x_r)))}\left[1_{f(x_r)\notin\tau(v)}\right] \geq$$
$$\frac{1}{2p}\sum_{r=1}^{p} \mathop{\mathbb{E}}_{v\sim Unif(\sigma(h_i(s_r)))}\left[1_{f(s_r)\notin\tau(v)}\right]$$

Thus, as expectation is a linear operator:

$$\frac{1}{T}\sum_{i=1}^{T} \mathbb{E}_{S^i_Y \sim D_i^{\mathcal{Y}|S_{X_j}}}[L_{D_i}(A(S^i_{Y,X_j}))]$$
$$\geq \frac{1}{T}\sum_{i=1}^{T}\frac{1}{2p}\sum_{r=1}^{p} \mathbb{E}_{S^i_Y \sim D_i^{\mathcal{Y}|S_{X_j}}}\left[\mathop{\mathbb{E}}_{v\sim Unif(\sigma(h_i(s_r)))}\left[1_{A(S^i_{Y,X_j})(s_r)\notin\tau(v)}\right]\right].$$

Continuing, we see

$$\frac{1}{T}\sum_{i=1}^{T}\frac{1}{2p}\sum_{r=1}^{p}\mathbb{E}_{S_Y^i\sim D_i^{\mathcal{Y}|S_{X_j}}}\left[\mathop{\mathbb{E}}_{v\sim Unif(\sigma(h_i(s_r)))}\left[1_{A(S^i_{Y,X_j})(s_r)\notin\tau(v)}\right]\right]=$$
$$\frac{1}{2p}\sum_{r=1}^{p}\frac{1}{T}\sum_{i=1}^{T}\mathbb{E}_{S_Y^i\sim D_i^{\mathcal{Y}|S_{X_j}}}\left[\mathop{\mathbb{E}}_{v\sim Unif(\sigma(h_i(s_r)))}\left[1_{A(S^i_{Y,X_j})(s_r)\notin\tau(v)}\right]\right]\geq$$
$$\frac{1}{2}\min_{r\in[p]}\frac{1}{T}\sum_{i=1}^{T}\mathbb{E}_{S_Y^i\sim D_i^{\mathcal{Y}|S_{X_j}}}\left[\mathop{\mathbb{E}}_{v\sim Unif(\sigma(h_i(s_r)))}\left[1_{A(S^i_{Y,X_j})(s_r)\notin\tau(v)}\right]\right].$$

Fix $r\in[p]$. Partition $h_1,...,h_T$ into disjoint tuples $(h_i,h_{i'})$ by $\sigma(h_i(x))\neq\sigma(h_{i'}(x))$ iff $x=s_r$. Since s_r is not in S_{X_j}, we can see that each pair has the same probabilities of getting the same S_Y^i given S_{X_j}. When $S^i_{Y,X_j}=S^{i'}_{Y,X_j}$, for each tuple we have that for exactly at most one of them (without loss of generality, assume its h_i)

$$\mathop{\mathbb{E}}_{v\sim Unif(\sigma(h_i(s_r)))}\left[1_{A(S^i_{Y,X_j})(s_r)\notin\tau(v)}\right]=0.$$

We have this for at most one i as we know that there are no strict subsets in $\sigma(\mathcal{Z})$. Thus, without loss of generality, if $\sigma(A(S^i_{Y,X_j})(s_r))=\sigma(h_i(s_r))$, then $\exists v\in\sigma(h_{i'}(s_r))$ where $v\notin\sigma(A(S^i_{Y,X_j})(s_r))$.

Therefore:

$$\mathop{\mathbb{E}}_{v\sim Unif(\sigma(h_{i'}(s_r)))}\left[1_{A(S^i_{Y,X_j})(s_r)\notin\tau(v)}\right]=\frac{|\sigma(h_{i'}(s_r)))\setminus\sigma(A(S^i_j)(s_r))|}{|\sigma(h_{i'}(s_r)))|}\geq\frac{1}{|\sigma(h_{i'}(s_r)))|}$$
$$\geq\frac{1}{\max_{z\in\mathcal{Z}^C}|\sigma(z)|}$$

With this, we see that over every tuple

$$\mathop{\mathbb{E}}_{v\sim Unif(\sigma(h_i(s_r)))}\left[1_{A(S^i_{Y,X_j})(s_r)\notin\tau(v)}\right]+\mathop{\mathbb{E}}_{v\sim Unif(\sigma(h_{i'}(s_r)))}\left[1_{A(S^i_{Y,X_j})(s_r)\notin\tau(v)}\right]\geq\frac{1}{\max_{z\in\mathcal{Z}^C}|\sigma(z)|}.$$

We note again that S^i_{Y,X_j} has the same probability as $S^{i'}_{Y,X_j}$ in each respective tuple. Thus, we can substitute the expectations over $D_i^{\mathcal{Y}|S_{X_j}}$ and $D_{i'}^{\mathcal{Y}|S_{X_j}}$ into just one $D_i^{\mathcal{Y}|S_{X_j}}$ to yield

$$\mathbb{E}_{S_Y^i\sim D_i^{\mathcal{Y}|S_{X_j}}}\left[\mathop{\mathbb{E}}_{v\sim Unif(\sigma(h_i(s_r)))}\left[1_{A(S^i_{Y,X_j})(s_r)\notin\tau(v)}\right]\right]+$$
$$\mathbb{E}_{S_Y^{i'}\sim D_{i'}^{\mathcal{Y}|S_{X_j}}}\left[\mathop{\mathbb{E}}_{v\sim Unif(\sigma(h_{i'}(s_r)))}\left[1_{A(S^{i'}_{Y,X_j})(s_r)\notin\tau(v)}\right]\right]=$$
$$\mathbb{E}_{S_{j,Y}^i\sim D_i^{\mathcal{Y}|S_{X_j}}}\left[\mathop{\mathbb{E}}_{v\sim Unif(\sigma(h_i(s_r)))}\left[1_{A(S^i_{Y,X_j})(s_r)\notin\tau(v)}\right]+\mathop{\mathbb{E}}_{v\sim Unif(\sigma(h_{i'}(s_r)))}\left[1_{A(S^i_{Y,X_j})(s_r)\notin\tau(v)}\right]\right]$$
$$\geq\frac{1}{\max_{z\in\mathcal{Z}^C}|\sigma(z)|}.$$

By summing each partition together we see

$$\frac{1}{T}\sum_{i=1}^{T}\mathbb{E}_{S_{j,Y}^i\sim D_i^{\mathcal{Y}|S^i_{j,X}}}\left[\mathop{\mathbb{E}}_{v\sim Unif(\sigma(h_i(s_r)))}\left[1_{A(S^i_{Y,X_j})(s_r)\notin\tau(v)}\right]\right]$$
$$\geq\frac{1}{T}\left(\frac{T}{2}\frac{1}{\max_{z\in\mathcal{Z}^C}|\sigma(z)|}\right)=\frac{1}{2\max_{z\in\mathcal{Z}^C}|\sigma(z)|}.$$

Therefore,

$$\max_{i\in[T]} \mathbb{E}_{S\sim D_i}[L_{D_i}(A(S))] \geq \frac{1}{2}\min_{r\in[p]}\frac{1}{T}\sum_{i=1}^{T}\mathbb{E}_{S^i_{j,Y}\sim D_i^{\mathcal{Y}|S_{X_j}}}\left[\mathop{\mathbb{E}}_{v\sim Unif(\sigma(h_i(s_r)))}\left[1_{A(S^i_{Y,X_j})(s_r)\notin\tau(v)}\right]\right] \geq$$
$$\frac{1}{2}\min_{r\in[p]}\frac{1}{2\max_{z\in\mathcal{Z}^C}|\sigma(z)|} = \frac{1}{4\max_{z\in\mathcal{Z}^C}|\sigma(z)|}.$$

Thus, we have shown there exists a realizable distribution where $\mathcal{A}$ does not do well, thereby showing this is not PAC-learnable, which is what we needed to show for the contrapositive. □

C Finite $GNdim(\mathcal{H},\ell)$ is Sufficient for Learnability Proof

C.1 Using VC Dimension of Loss Class

Proof. Suppose we have an $(\mathcal{X},\mathcal{Z},\mathcal{Y},\mathcal{H},\ell)$ learning scenario where $GNdim(\mathcal{H},\ell)$ is finite. By Lemma A.1, we have that

$$VC(\ell\circ\mathcal{H}) \leq 4.67GNdim(\mathcal{H},\ell)\log_2(|\sigma(Z)|+1).$$

Since $|\sigma(\mathcal{Z})|<\infty$, we know that $VC(\ell\circ\mathcal{H})$ must be finite. It is known that the upper bound of the sample complexity of $\ell\circ\mathcal{H}$ is

$$O\left(\frac{VC(\ell\circ\mathcal{H})+\log(1/\delta)}{\epsilon^2}\right)$$

due to it being a binary classifier. Notice for any distribution on $\mathcal{X}\times\mathcal{Y}$, there is an equivalent distribution over $\mathcal{X}\times\mathcal{Y}\times\{0\}$. Therefore, learnability of the 0-1 loss of $\ell\circ\mathcal{H}$ on the latter distribution implies learnability of $\mathcal{H}$ with ℓ on the former. This gives us an upper bound on the sample complexity needed to learn $(\mathcal{X},\mathcal{Z},\mathcal{Y},\mathcal{H},\ell)$ by replacing $VC(\ell\circ\mathcal{H})$ with $4.67GNdim(\mathcal{H},\ell)\log_2(|\sigma(Z)|+1)$. Thus, we have learnability of $(\mathcal{X},\mathcal{Z},\mathcal{Y},\mathcal{H},\ell)$ if $GNdim(\mathcal{H},\ell)$ is finite.

□

C.2 Special Case where $\mathcal{Z}=\mathcal{Y}$

Proof. Much like Lemma 4.2, we shall use the $(\mathcal{X},\mathcal{Y}^C,\mathcal{H}^C,\ell^C)$ equivalent learning problem.

Suppose $(\mathcal{H},\ell)$ has finite generalized Natarajan dimension. For each $y\in\mathcal{Y}^C$, put an arbitrary ordering on $\sigma(y)$, and then for all $k>|\sigma(y)|$, let $\sigma(y)_k$ be some label not in $\mathcal{Y}^C$ (from here on out denoted 0). Let the maps $\sigma_i:\mathcal{Y}^C\cup\{0\}\to\mathcal{Y}^C\cup\{0\}$ denote the function $\sigma_i(y)=\sigma(y)_i$, where $\sigma_i(0)=0\ \forall i$. We can extend this map (abusing notation) to $\sigma_i:\mathcal{X}\times(\mathcal{Y}^C\cup\{0\})\to\mathcal{X}\times(\mathcal{Y}^C\cup\{0\})$, $\sigma_i(x,y)=(x,\sigma_i(y))$. Now, given a distribution D, let us define a new distribution $D^i:=D\circ\sigma_i^{-1}$ by the pushforward of D through σ_i. Thus D^i can be thought of as the distribution where each y "becomes" $\sigma(y)_i$. Notice then

$$1-L_{\mathcal{D}}(h) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[1_{h(x)\in\sigma(y)}\right] = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\sum_{i=1}^{|\sigma(y)|}1_{h(x)=\sigma(y)_i}\right] =$$
$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\sum_{i=1}^{|\sigma(y)|}1_{h(x)=\sigma(y)_i}+\sum_{i=|\sigma(y)|+1}^{k}0\right] = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\sum_{i=1}^{|\sigma(y)|}1_{h(x)=\sigma(y)_i}+\sum_{i=|\sigma(y)|+1}^{k}1_{h(x)=0}\right] =$$
$$\sum_{i=1}^{k}\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[1_{h(x)=\sigma(y)_i}\right] = \sum_{i=1}^{k}\mathop{\mathbb{E}}_{(x,y')\sim\mathcal{D}^i}\left[1_{h(x)=y'}\right] = \sum_{i=1}^{k}1-L_{\mathcal{D}^i}(h)$$

Where

$$L_{\mathcal{D}^i}(h) := \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}^i}\left[1_{h(x)\neq y}\right]$$

is the 0-1 loss for a multiclass learning problem on the distribution of D^i, and where the penultimate equality comes from

$$\begin{aligned}
&D(\{(x,y) \mid h(x) = \sigma_i(y)\}) = \\
&D(\{(x,y) \mid h(x) = y' \wedge y' = \sigma_i(y)\}) = \\
&D(\{(x,y) \mid h(x) = y' \wedge \sigma_i(x,y) = (x,y')\}) = \\
&D \circ \sigma_i^{-1}(\{(x,y') \mid h(x) = y'\}) = \\
&D^i(\{(x,y') \mid h(x) = y'\})
\end{aligned}$$

For the sample loss of a set of size m, we can do a similar decomposition.

$$\begin{aligned}
&1 - L_S(h) = \frac{|\{(y_j, h(x_j)) \in C\}|}{m} = \frac{\sum_{j=1}^m 1_{h(x_j)\in\sigma(y_j)}}{m} = \\
&\frac{\sum_{j=1}^m \sum_{i=1}^k 1_{h(x_j)=\sigma(y_j)_i}}{m} = \frac{\sum_{i=1}^k \sum_{j=1}^m 1_{h(x_j)=\sigma(y_j)_i}}{m} = \sum_{i=1}^k 1 - L_{S^i}(h)
\end{aligned}$$

Using these decompositions, we get

$$\begin{aligned}
&|L_D(h) - L_S(h)| = |L_D(h) - 1 + 1 - L_S(h)| = \\
&\left| -\sum_{i=1}^k 1 - L_{D^i}(h) + \sum_{i=1}^k 1 - L_{S^i}(h) \right| = \left| \sum_{i=1}^k L_{D^i}(h) - L_{S^i}(h) \right| \leq \sum_{i=1}^k |L_{D^i}(h) - L_{S^i}(h)|
\end{aligned}$$

Since we have finite generalized Natarajan dimension, by corollary 4.0.2 we have that $\mathcal{H}^C$ is multiclass learnable, and thus has the uniform convergence property with function $m^{UC}_{\mathcal{H}^C}$ (Shalev-Shwartz & Ben-David (2014)). Since each L_{D^i}, L_{S^i} are equivalent to a multiclass problem on a different distribution, pick sample S with $|S| \geq m^{UC}_{\mathcal{H}^C}(\epsilon/k, \delta/k)$. By union bound we get:

$$\begin{aligned}
&\mathbb{P}\left(\sup_{h\in\mathcal{H}} |L_D(h) - L_S(h)| > \epsilon\right) \leq \mathbb{P}\left(\sup_{h\in\mathcal{H}} \sum_{i=1}^k \left|L^i_D(h) - L^i_S(h)\right| > \epsilon\right) \leq \\
&\mathbb{P}\left(\sum_{i=1}^k \sup_{h\in\mathcal{H}} \left|L^i_D(h) - L^i_S(h)\right| > \epsilon\right) \leq \mathbb{P}\left(\bigcup_{i=1}^k \left(\sup_{h\in\mathcal{H}} \left|L^i_D(h) - L^i_S(h)\right| > \epsilon/k\right)\right) \leq k\delta/k = \delta
\end{aligned}$$

Therefore, we have that $\mathcal{H}^C$ has the uniform convergence property and thus ERM is a valid learner for the problem. □

D Miscellaneous

D.1 Example of a Learning Problem where $GNdim(\mathcal{H}, \ell) = \infty$ but it is Learnable

Let $\mathcal{X}$ be an infinite input space, $\mathcal{Y} = \{1, 2, 3\}$, $\mathcal{H} = \{1, 2\}^{\mathcal{X}} \cup \{3\}^{\mathcal{X}}$, and let the loss matrix be as below:

$$L = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \end{array}$$

where $\ell(i, j) = L_{ij}$. Note how $GNdim(\mathcal{H}, \ell) = VC(\ell \circ \mathcal{H}) = \infty$ as we have all functions from $\mathcal{X}$ to $\{1, 2\}$, but the algorithm that always chooses $h \in \mathcal{H}$ such that $\forall x \in \mathcal{X}$, $h(x) = 3$ will always be a valid PAC-learner.