Journal of Machine Learning Research 25 (2024) 1-54 Submitted 6/23 ; Revised 6/24; Published 10/24

A Characterization of Multioutput Learnability

Vinod Raman * VKRAMAN@UMICH.EDU
Department of Statistics
University of Michigan
Ann-Arbor, MI 48109, USA

Unique Subedi * SUBEDI@UMICH.EDU
Department of Statistics
University of Michigan
Ann-Arbor, MI 48109, USA

Ambuj Tewari TEWARIA@UMICH.EDU
Department of Statistics
Department of Electrical Engineering and Computer Science
University of Michigan
Ann-Arbor, MI 48109, USA

Editor: Gergely Neu

Abstract

We consider the problem of learning multioutput function classes in the batch and online settings. In both settings, we show that a multioutput function class is learnable if and only if each single-output restriction of the function class is learnable. This provides a complete characterization of the learnability of multilabel classification and multioutput regression in both batch and online settings. As an extension, we also consider multilabel learnability in the bandit feedback setting and show a similar characterization as in the full-feedback setting.

Keywords: Learnability, Online Learning, Multilabel Classification, Multioutput Regression

1 Introduction

Multioutput learning is a problem where an instance is labeled by a vector-valued target. This is a generalization of scalar-valued-target learning settings such as multiclass classification and regression. Multioutput learning has enjoyed a wide range of practical applications like image tagging, document categorization, recommender systems, an weather forecasting to name a few. This widespread applicability has motivated the development of several practical methods (Kapoor et al., 2012; Borchani et al., 2015; Yang et al., 2020; Xu et al., 2013; Nam et al., 2017), as well as theoretical analysis (Koyejo et al., 2015b; Liu and Tsang, 2015). However, the most fundamental question of learnability in a multioutput setting remains unanswered.

Characterizing learnability is the foundational step toward understanding any statistical learning problem. The fundamental theorem of statistical learning characterizes the learn-

∗. Equal contribution

ability of binary function class in terms of the finiteness of a combinatorial quantity called the Vapnik-Chervonenkis (VC) dimension (Vapnik and Chervonenkis, 1971, 1974). Extending VC theory, Natarajan (1989) proposed and studied the Natarajan dimension, which was later shown by Ben-David et al. (1995) to characterize learnability in multiclass settings with a finite number of labels. In fact, Ben-David et al. (1995) observed that the learnability of multiclass function class $\mathcal{F} \subseteq \{e_1, \ldots, e_K\}^{\mathcal{X}}$ can also be characterized in terms of the learnability of each component-wise binary function class $\mathcal{F}_k = \{x \mapsto \langle f(x), e_k \rangle : f \in \mathcal{F}\}$, where $\{e_1, \ldots, e_K\}$ is the standard basis of $\mathbb{R}^K$ and $\langle \cdot, \cdot \rangle$ is the Euclidean inner-product. Furthermore, they define an equivalence class of loss functions for the 0-1 loss and characterize the learnability of multiclass problems with respect to all losses in the equivalence class. We take a similar approach in this paper. Closing the question of learnability for multiclass problems, Brukhim et al. (2022) shows that the Daniely-Shwartz (DS) dimension, originally proposed by Daniely and Shalev-Shwartz (2014), characterizes multiclass learnability in the infinite labels setting. Similarly, in the online setting, the Littlestone dimension (Littlestone, 1987) characterizes the online learnability of a binary function class and a generalization of the Littlestone dimension (Daniely et al., 2011) characterizes online learnability in the multiclass setting with finite labels. As for scalar-valued regression, the fat shattering (Bartlett et al., 1996) and the sequential fat shattering (Rakhlin et al., 2015a) dimensions characterize batch and online learnability respectively. Surprisingly, to our best knowledge, no such characterization of the learnability of multioutput function classes exists in the literature.

In this paper, we close this gap by characterizing the learnability of function classes $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$, where $\mathcal{Y} \subseteq \mathbb{R}^K$ is vector-valued target space for some $K \in \mathbb{N}$. Let us define scalar-valued function classes $\mathcal{F}_k = \{x \mapsto \langle f(x), e_k \rangle : f \in \mathcal{F}\}$ for each $k \in [K]$, where $\{e_1, \ldots, e_K\}$ is the standard basis of $\mathbb{R}^K$. Similarly, define $\mathcal{Y}_k := \{\langle y, e_k \rangle : y \in \mathcal{Y}\}$. Our main result, informally stated below, asserts that $\mathcal{F}$ is learnable if and only if each coordinate restriction $\mathcal{F}_k$ is learnable.

Theorem. *(*Informal*) A multioutput function class $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ is learnable if and only if each restriction $\mathcal{F}_k \subseteq \mathcal{Y}_k^{\mathcal{X}}$ is learnable.*

We prove a version of this result in four canonical settings: batch classification, online classification, batch regression, and online regression. For the batch settings, we consider the PAC framework and for the online settings, we consider the fully adversarial model. In addition, our result holds for a wide family of loss functions. A unifying theme throughout all four learning settings is our ability to *constructively* convert a learning algorithm $\mathcal{A}$ for $\mathcal{F}$ into learning algorithm $\mathcal{A}_k$ for $\mathcal{F}_k$ for each $k \in \{1, ..., K\}$ and vice versa. We show that even for multioutput losses that tightly "couple" the K coordinates of a function class, their learnability still depends on the learnability of each coordinate. For the batch setting, our algorithmic techniques use the realizable-to-agnostic conversion introduced by Hopkins et al. (2022). In the online setting, we provide a new realizable-to-agnostic conversion similar in the spirit of Hopkins et al. (2022). In principle, both ours and Hopkins et al. (2022)'s realizable-to-agnostic conversion is based on the idea of using algorithms to construct a cover of function classes, originally introduced in the seminal work of Ben-David et al. (2009).

Our proof techniques, however, do not extend naturally to the case when K is infinite. So, characterizing learnability for an infinite-dimensional target space is an interesting open problem. Moreover, our reductions are computationally inefficient and lead to sub-optimal sample complexities and regret bounds. As such, we leave the construction of efficient and optimal multioutput learning algorithms as an interesting direction for the future work.

1.1 Related works

Multilabel classification has been extensively studied in the batch setting. We review a few works here and also refer the reader to the references therein. Dembczyński et al. (2010) quantify the 0-1 risk of multilabel classifiers trained by minimizing the Hamming loss and vice versa. Dembczyński et al. (2012) and Chekina et al. (2013) study how exploiting dependencies between labels can improve the predictive performance of multilabel classifiers and how such exploitation interacts with loss minimization. Jain et al. (2016) consider the case where the label set is extremely large and design new loss functions to handle these settings. Busa-Fekete et al. (2022) derive upper and lower bounds on the excess risk for non-parametric and parametric function classes for various loss functions assuming label sparsity. Gao and Zhou (2011) and Koyejo et al. (2015a) study the consistency of surrogate loss functions for multilabel classification. Finally, Gentile and Orabona (2012) consider online multilabel classification under partial feedback and present a novel algorithm based on second-order descent methods.

There is also a long history of studying least squares estimators for multioutput linear models in the statistical literature, see (Rao, 1965; Brown and Zidek, 1980) and references therein. The topic received widespread attention in learning theory following the seminal work of Micchelli and Pontil (2005) in RKHS methods for vector-valued regression. We refer the reader to a comprehensive review of kernel methods for vector-valued regression by Alvarez et al. (2012). An early work of Gnecco and Sanguineti (2008) provides estimation and approximation error of vector-valued functions using Rademacher complexity. Following the influential work of Maurer (2016) on Rademacher contraction inequalities for vector-valued functions, there have been works on the Rademacher analysis of vector-valued functions (see Cortes et al. (2016); Reeve and Kaban (2020); Yousefi et al. (2018); Foster and Rakhlin (2019)). Finally, we point out a recent work by Park and Muandet (2023) towards developing empirical process theory for vector-valued functions.

2 Preliminaries

Let $\mathcal{X}$ denote the instance space and $\mathcal{Y} \subseteq \mathbb{R}^K$ be the target space for some $K \in \mathbb{N}$. For a space $\mathcal{Z}$, we let $\mathcal{Z}^\star$ be the set of all finite sequences of elements from $\mathcal{Z}$. Consider a vector-valued function class $\mathcal{F} \subseteq \mathcal{Y}^\mathcal{X}$, where $\mathcal{Y}^\mathcal{X}$ denotes set of all functions from $\mathcal{X}$ to $\mathcal{Y}$. For an unlabeled sample $S_U \in \mathcal{X}^\star$, let $\mathcal{F}_{|S_U}$ denote the projection of $\mathcal{F}$ onto S_U. Let $\langle \cdot, \cdot \rangle$ denote the Euclidean inner-product on $\mathbb{R}^K$. Define scalar-valued function classes $\mathcal{F}_k = \{x \mapsto \langle f(x), e_k \rangle : f \in \mathcal{F}\}$ for each $k \in [K]$, where $\{e_1, \ldots, e_K\}$ is the standard basis of $\mathbb{R}^K$. Here, each $\mathcal{F}_k \subseteq \mathcal{Y}_k^\mathcal{X}$, where $\mathcal{Y}_k = \{\langle y, e_k \rangle : y \in \mathcal{Y}\}$ denotes the restriction of the target space to its k^{th} component.

For a function $f \in \mathcal{F}$, we use $f_k(x) := \langle f(x), e_k \rangle$ to denote the k^{th} coordinate output of $f(x)$. On the other hand, we use $y^k := \langle y, e_k \rangle$ to denote k^{th} coordinate of $y \in \mathcal{Y}$. Additionally, it is useful to distinguish between the range space $\mathcal{Y}$ and the image of functions $f \in \mathcal{F}$. We define the image of function class $\mathcal{F}$ as $\mathrm{im}(\mathcal{F}) := \cup_{f \in \mathcal{F}} \mathrm{im}(f)$, where $\mathrm{im}(f) = \{f(x) : x \in \mathcal{X}\}$. Finally, we take $[N] := \{1, 2, \ldots, N\}$.

In this work, we only consider bounded, non-negative loss functions $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ that satisfy the identity of the indiscernibles. For the remainder of the paper, we drop the adjectives "bounded" and "non-negative" when referring to loss functions.

Definition 1 (Identity of Indiscernibles)**.** *A loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ satisfies identity of indiscernibles whenever $\ell(y_1, y_2) = 0$ if and only if $y_1 = y_2$.*

Note that if ℓ_1 and ℓ_2 are two losses defined on $\mathcal{Y} \times \mathcal{Y}$ that satisfy the identity of indiscernibles, then $\ell_1(y_1, y_2) = 0$ if and only if $\ell_2(y_1, y_2) = 0$. We also define a notion of approximate subadditivity, although not all the loss functions we consider have this property.

Definition 2 (c-subadditive)**.** *A loss function ℓ is c-subadditive if there exists a constant $c > 0$ that only depends on the loss function ℓ such that $\ell(y_1, y_2) \leq c\,\ell(y_1, y) + \ell(y, y_2)$ for all $y, y_1, y_2 \in \mathcal{Y}$.*

If $|\mathcal{Y}| < \infty$, ℓ being c-subadditive is an immediate consequence of ℓ satisfying the identity of indiscernibles. In fact, the value of c in this case is $\frac{\max_{r \neq t} \ell(r,t)}{\min_{r \neq t} \ell(r,t)}$. To see why this is true, it suffices to only consider the case when $\ell(y_1, y_2) > \ell(y, y_2)$ because the inequality is trivially true otherwise. Since the loss values are distinct, we must have $y \neq y_1$. Using the identity of indiscernible, we obtain $\ell(y_1, y) \geq \min_{r \neq t} \ell(r, t)$, thus implying that $c\,\ell(y_1, y) \geq \max_{r \neq t} \ell(r, t) \geq \ell(y_1, y_2)$. The case when $|\mathcal{Y}| = \infty$ is a bit delicate because $\min_{r \neq t} \ell(r, t)$ may not exist. So, one needs extra structure in the loss function to infer c-subadditivity. For instance, if ℓ is a distance metric, then it is trivially 1-subadditive due to the triangle inequality.

2.1 Batch Setting

In the batch setting, we are interested in characterizing the learnability of $\mathcal{F}$ under the classical PAC models: both in the original realizable formulation (Valiant, 1984) and in the agnostic extension (Kearns et al., 1994).

Definition 3 (Agnostic Multioutput Learnability)**.** *A function class $\mathcal{F}$ is agnostic learnable with respect to loss $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$, if there exists a function $m : (0,1)^2 \to \mathbb{N}$ and a learning algorithm $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^\star \to \mathcal{Y}^{\mathcal{X}}$ with the following property: for every $\epsilon, \delta \in (0,1)$ and for every distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$, running algorithm $\mathcal{A}$ on $n \geq m(\epsilon, \delta)$ iid samples from $\mathcal{D}$ outputs a predictor $g = \mathcal{A}(S)$ such that with probability at least $1 - \delta$ over $S \sim \mathcal{D}^n$,*

$$\mathbb{E}_{\mathcal{D}}[\ell(g(x), y)] \leq \inf_{f \in \mathcal{F}} \mathbb{E}_{\mathcal{D}}[\ell(f(x), y)] + \epsilon.$$

Note that we do not require the output predictor $\mathcal{A}(S)$ to be in $\mathcal{F}$, but only require $\mathcal{A}(S)$ to compete with the best predictor in $\mathcal{F}$. If we restrict distribution $\mathcal{D}$ to a class such that $\inf_{f \in \mathcal{F}} \mathbb{E}_{\mathcal{D}}[\ell(f(x), y)] = 0$, then we get realizable learnability.

The learnability of a function class is generally characterized in terms of the complexity measure of the function class. As stated in the introduction, the VC dimension characterizes the learnability of binary function classes (Vapnik and Chervonenkis, 1974), and so does the Natarajan dimension for multiclass classification (Ben-David et al., 1995). In a scalar-valued regression problem, the fat-shattering dimension of the function class characterizes learnability with respect to the absolute-value and squared loss (Bartlett et al., 1996; Alon et al., 1997). In particular, a real-valued function class $\mathcal{G} \subseteq [0, B]^{\mathcal{X}}$ is learnable if and only if its fat-shattering dimension, denoted as $\text{fat}_\gamma(\mathcal{G})$, is finite for every scale $\gamma > 0$. We extend this characterization to a wide range of loss functions in Lemma 11. Finally, for a real-valued class $\mathcal{G}$, we also use a more general notion of complexity measure called Rademacher complexity, denoted as $\mathfrak{R}_n(\mathcal{G})$, that provides a sufficient condition for learnability (Bartlett and Mendelson, 2003). Precise definitions of all these complexity measures are provided in Appendix A.1.

One recurring theme in this work is to first construct a realizable multioutput learner and then convert it into an agnostic multioutput learner. It is well known that realizable learnability and agnostic learnability are equivalent for multiclass classification problems with respect to 0-1 loss, (see Ben-David et al. (1995), (Shalev-Shwartz and Ben-David, 2014, Theorem 6.7)). Lemma 4, which is an immediate consequence of (Hopkins et al., 2022, Theorem 18), extends this equivalence between realizable and agnostic learning to general loss functions and target spaces.

Lemma 4 (Hopkins et al. (2022))**.** *Consider a function class $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ such that $|im(\mathcal{F})| < \infty$ and a general loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ that is c-subadditive. Then, $\mathcal{F}$ is realizable PAC learnable with respect to ℓ if and only if $\mathcal{F}$ is agnostic PAC learnable with respect to ℓ.*

The result of (Hopkins et al., 2022, Theorem 18) is stated for the case when $|\mathcal{Y}| < \infty$. However, in the regression setting, we need a slightly general version of their result to handle α-discretized function classes $\mathcal{F}^\alpha \subseteq \mathcal{Y}^{\mathcal{X}}$ (see Proof of Theorem 12) where $|\text{im}(\mathcal{F}^\alpha)| < \infty$ but $|\mathcal{Y}| = |\,[0,1]^K| = \infty$. Nevertheless, the proof of Lemma 4 requires only a minor modification to that of (Hopkins et al., 2022, Theorem 18). Given the central role of this result in our characterization, we provide full proof of Lemma 4 in Section 3.1.2. Finally, we note that agnostic-to-realizable conversions in the batch setting are also possible via boosting and compression-based arguments (Montasser et al., 2019; Attias and Hanneke, 2023). We use the conversion of Hopkins et al. (2022) due to its generality and simplicity.

2.2 Online Setting

In the online setting, an adversary plays a sequential game with the learner over T rounds. In each round $t \in [T]$, an adversary selects a labeled instance $(x_t, y_t) \in \mathcal{X} \times \mathcal{Y}$ and reveals x_t to the learner. The learner makes a (potentially randomized) prediction $\hat{y}_t \in \mathcal{Y}$. Finally, the adversary reveals the true label y_t, and the learner suffers the loss $\ell(y_t, \hat{y}_t)$, where ℓ is some pre-specified loss function. Given a function class $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$, the goal of the learner is to output predictions $\hat{y}_t$ such that its cumulative loss is close to the best possible cumulative loss over functions in $\mathcal{F}$. A function class is online learnable if there exists an algorithm such that for any sequence of labeled examples $(x_1, y_1), ..., (x_T, y_T)$, the difference in cumulative loss between its predictions and the predictions of the best possible function in $\mathcal{F}$ is small.

Definition 5 (Online Multioutput Learnability)**.** *A multioutput function class $\mathcal{F}$ is online learnable with respect to loss ℓ, if there exists an (potentially randomized) algorithm $\mathcal{A}$ such that for any adaptively chosen sequence of labelled examples $(x_t, y_t) \in \mathcal{X} \times \mathcal{Y}$, the algorithm outputs $\mathcal{A}(x_t) \in \mathcal{Y}$ at every iteration $t \in [T]$ such that*

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell(\mathcal{A}(x_t), y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \ell(f(x_t), y_t)\right] \leq R(T)$$

where the expectation is taken with respect to the randomness of $\mathcal{A}$ and that of the possibly adaptive adversary, and $R(T) : \mathbb{N} \to \mathbb{R}^+$ is the additive regret: a non-decreasing, sub-linear function of T.

If it is guaranteed that the learner always observes a sequence of examples labeled by some function $f \in \mathcal{F}$, then we say we are in the *realizable* setting. On the other hand, if the true label y_t is not revealed to the learner in each round $t \in [T]$ and the adversary only reveals the learner's *loss* $\ell(\mathcal{A}(x_t), y_t)$ then we say we are in the *bandit* setting.

The online learnability of scalar-valued function classes $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ has been characterized. For example, when $\mathcal{Y}$ is finite (i.e. $\mathcal{Y} = [K]$ for some $K \in \mathbb{N}$), the Multiclass Littlestone Dimension (MCLdim) of $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ characterizes online learnability with respect to the 0-1 loss. A function class $\mathcal{H}$ is online learnable with respect to the 0-1 loss if and only if $\text{MCLdim}(\mathcal{H})$ is finite (Daniely et al., 2011). Moreover, $\text{MCLdim}(\mathcal{H})$ tightly captures the best achievable regret in both the realizable and agnostic settings (Daniely et al., 2011). When the label space $\mathcal{Y}$ is a bounded subset of $\mathbb{R}$, the *sequential* fat-shattering dimension of a real-valued function class $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{X}}$, denoted $\text{fat}^{\text{seq}}_{\gamma}(\mathcal{H})$, characterizes the online learnability of $\mathcal{H}$ with respect to the absolute value loss (Rakhlin et al., 2015a). Unlike the Littlestone dimension, note that the sequential fat-shattering dimension is a scale-sensitive dimension. That is, $\text{fat}^{\text{seq}}_{\gamma}(\mathcal{H})$ is defined at every scale $\gamma > 0$. Accordingly, a real-valued function class $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{X}}$ is learnable with respect to the absolute value loss if and only if its sequential fat-shattering dimension is finite at *every* scale $\gamma > 0$ (Rakhlin et al., 2015a). We extend this characterization to a wide range of loss functions in Lemma 22. Beyond scalar-valued learnability, for any label space $\mathcal{Y}$, function class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, and loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$, the *sequential* Rademacher complexity of the loss class $\ell \circ \mathcal{H}$, denoted $\mathfrak{R}^{\text{seq}}_T(\ell \circ \mathcal{H})$, is a useful tool for providing sufficient conditions for online learnability (Rakhlin et al., 2015a). See Appendix A.2 for complete definitions.

Like the batch setting, a key technique we use to prove online learnability is to first construct a realizable online learner and then convert it into an agnostic online learner. However, unlike the batch setting, there is no known generic algorithm that converts a (potentially randomized) realizable online learner into an agnostic online learner. Thus, one of the contributions of this work is Theorem 13, informally stated below, which provides an online analog of the realizable-to-agnostic conversion from Hopkins et al. (2022).

Theorem. *(Informal) Let $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a multioutput function class such that $|im(\mathcal{F})| < \infty$ and $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ be any bounded, c-subadditive loss function. If $\mathcal{F}$ is online learnable with respect to ℓ in the realizable setting, then $\mathcal{F}$ is online learnable with respect to ℓ in the agnostic setting.*

3 Batch Multioutput Learnability

3.1 Batch Multilabel Classification

In this section, we study the learnability of batch multilabel classification. Accordingly, let $\mathcal{Y} = \{-1, 1\}^K$. First, we consider the learnability of a natural decomposable loss. Then, we extend the result to more general non-decomposable losses that satisfy the identity of indiscernible. We want to point out that a multilabel classification with K labels can be viewed as a multiclass classification with 2^K labels. With this viewpoint, the Natarajan dimension of $\mathcal{F}$ continues to characterize the batch multilabel learnability for any loss satisfying the identity of indiscernibles (see (Ben-David et al., 1995, Section 4)). For the sake of completeness, we also provide proof of this characterization in Appendix B. However, it is more natural to view a multilabel classification as K different binary classification problems as opposed to a multi-class classification problem with 2^K labels. Exploiting this natural decomposability of a multilabel function class, we relate the learnability of $\mathcal{F}$ to the learnability of each component $\mathcal{F}_k$.

3.1.1 Characterizing Batch Learnability for the Hamming Loss

A canonical and natural loss function for multilabel classification is the Hamming loss, defined as $\ell_H(f(x), y) := \sum_{i=1}^K \mathbb{1}\left\{f_i(x) \neq y^i\right\}$, where $f(x) = (f_1(x), \ldots, f_K(x))$ and $y = (y^1, \ldots, y^K)$. The following result establishes an equivalence between the learnability of $\mathcal{F}$ with respect to Hamming loss and the learnability of each $\mathcal{F}_k$ with respect to 0-1 loss.

Theorem 6. *A function class $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ is agnostic PAC learnable with respect to ℓ_H if and only if each of $\mathcal{F}_k \subseteq \mathcal{Y}_k^{\mathcal{X}}$ is agnostic PAC learnable with respect to the* 0-1 *loss.*

Proof We first prove that learnability of each component is sufficient followed by the proof of necessity.

Part 1: Sufficiency. Here our goal is to prove that the agnostic PAC learnability of each $\mathcal{F}_k$ is sufficient for agnostic PAC learnability of $\mathcal{F}$. Our proof is constructive: given oracle access to agnostic PAC learners $\mathcal{A}_k$ for each $\mathcal{F}_k$ with respect to 0-1 loss, we construct an agnostic PAC learner $\mathcal{A}$ for $\mathcal{F}$ with respect to ℓ_H. Let $\mathcal{D}$ be arbitrary distribution on $\mathcal{X} \times \mathcal{Y}$ and $S = \{(x_i, y_i)\}_{i=1}^n \sim \mathcal{D}^n$ be iid samples from distribution $\mathcal{D}$. Denote $\mathcal{D}_k$ to be the marginal distribution of $\mathcal{D}$ restricted to $\mathcal{X} \times \mathcal{Y}_k$. Then, for all $k \in [K]$, the marginal samples $S_k = \{(x_i, y_i^k)\}_{i=1}^n$ with scalar-valued targets are iid samples from $\mathcal{D}_k$. For each $k \in [K]$, define $h_k = \mathcal{A}_k(S_k)$ to be the hypothesis returned by algorithm $\mathcal{A}_k$ when trained on S_k.

Let $m_k(\epsilon, \delta)$ denote the sample complexity of $\mathcal{A}_k$. Since $\mathcal{A}_k$ is an agnostic PAC learner for $\mathcal{F}_k$, we have that for $n \geq \max_k m_k(\frac{\epsilon}{K}, \frac{\delta}{K})$, with probability at least $1 - \delta/K$ over samples $S_k \sim \mathcal{D}_k^n$,

$$\mathbb{E}_{\mathcal{D}_k}\left[\mathbb{1}\left\{h_k(x) \neq y^k\right\}\right] \leq \inf_{f_k \in \mathcal{F}_k} \mathbb{E}_{\mathcal{D}_k}\left[\mathbb{1}\left\{f_k(x) \neq y^k\right\}\right] + \frac{\epsilon}{K}.$$

Summing these risk bounds over all coordinates k and using union bounds over probabilities, we get that with probability at least $1 - \delta$ over samples $S \sim \mathcal{D}^n$, we obtain $\sum_{k=1}^K \mathbb{E}_{\mathcal{D}_k}\left[\mathbb{1}\left\{h_k(x) \neq y^k\right\}\right] \leq \sum_{k=1}^K \inf_{f_k \in \mathcal{F}_k} \mathbb{E}_{\mathcal{D}_k}\left[\mathbb{1}\left\{f_k(x) \neq y^k\right\}\right] + \epsilon$. Now using the fact

that $\mathcal{F} \subseteq \mathcal{F}_1 \times \ldots \times \mathcal{F}_K$ followed by the linearity of expectation gives

$$\mathbb{E}_{\mathcal{D}}\left[\sum_{k=1}^{K} \mathbb{1}\left\{h_k(x) \neq y^k\right\}\right] \leq \inf_{f \in \mathcal{F}} \mathbb{E}_{\mathcal{D}}\left[\sum_{k=1}^{K} \mathbb{1}\left\{f_k(x) \neq y^k\right\}\right] + \epsilon.$$

This completes our proof as it shows that the learning rule that concatenates the predictors returned by each $\mathcal{A}_k$ on the marginalized samples S_k is an agnostic PAC learner for $\mathcal{F}$ with respect to ℓ_H with sample complexity at most $\max_k m_k(\epsilon/K, \delta/K)$.

Part 2: Necessity. Next, we show that if $\mathcal{F}$ is learnable with respect to ℓ_H, then each $\mathcal{F}_k$ is PAC learnable with respect to the 0-1 loss. Our proof is again based on reduction: given oracle access to agnostic PAC learner $\mathcal{A}$ for $\mathcal{F}$, we construct an agnostic PAC learner $\mathcal{A}_1$ for $\mathcal{F}_1$. A similar construction can be used for all other $\mathcal{F}_k$'s.

Let $\mathcal{D}_1$ be arbitrary distribution on $\mathcal{X} \times \mathcal{Y}_1$ and $S = \{(x_i, y_i^1)\}_{i=1}^n$ be iid samples from $\mathcal{D}_1$. In order to use the algorithm $\mathcal{A}$, we first augment the samples S to create samples with K-variate target. Define an augmented sample $\widetilde{S} = \{(x_i, (y_i^1, \ldots, y_i^K))\}_{i=1}^n$ such that $y_{ik} \sim \{-1, 1\}$ each with probability $1/2$ for all $i \in [n]$ and $k \in \{2, \ldots, K\}$. Next, we run $\mathcal{A}$ on $\widetilde{S}$ and obtain the hypothesis $h = (h_1, \ldots, h_K) = \mathcal{A}(\widetilde{S})$. We now show that h_1 obtains agnostic PAC bounds.

Consider a distribution $\widetilde{\mathcal{D}}$ on $\mathcal{X} \times \mathcal{Y}$ such that a sample $(x, (y^1, \ldots, y^K))$ from $\widetilde{\mathcal{D}}$ is obtained by first sampling $(x, y^1) \sim \mathcal{D}_1$ and appending y^k's sampled independently from uniform distribution on $\{-1, 1\}$ for each $k \in \{2, \ldots, K\}$. Let $m(\epsilon, \delta, K)$ denote the sample complexity of $\mathcal{A}$. Since $\mathcal{A}$ is an agnostic PAC learner for $\mathcal{F}$, for $n \geq m(\epsilon, \delta, K)$, with probability at least $1 - \delta$, we have

$$\mathbb{E}_{\widetilde{\mathcal{D}}}\left[\sum_{k=1}^{K} \mathbb{1}\left\{h_k(x) \neq y^k\right\}\right] \leq \inf_{f \in \mathcal{F}} \mathbb{E}_{\widetilde{\mathcal{D}}}\left[\sum_{k=1}^{K} \mathbb{1}\left\{f_k(x) \neq y^k\right\}\right] + \epsilon.$$

For $k \geq 2$, since the target is chosen uniformly at random from $\{-1, 1\}$, the 0-1 risk of any predictor is $1/2$. Therefore, the expression above can be written as

$$\mathbb{E}_{\mathcal{D}_1}\left[\mathbb{1}\left\{h_1(x) \neq y^1\right\}\right] + \sum_{k=2}^{K} 1/2 \leq \inf_{f \in \mathcal{F}} \left(\mathbb{E}_{\mathcal{D}_1}\left[\mathbb{1}\left\{f_1(x) \neq y^1\right\}\right] + \sum_{k=2}^{K} 1/2\right) + \epsilon,$$

which reduces to $\mathbb{E}_{\mathcal{D}_1}[\mathbb{1}\left\{h_1(x) \neq y^1\right\}] \leq \inf_{f_1 \in \mathcal{F}_1} \mathbb{E}_{\mathcal{D}_1}[\mathbb{1}\left\{f_1(x) \neq y^1\right\}] + \epsilon$. Therefore, $\mathcal{F}_1$ is agnostic PAC learnable with respect to 0-1 loss with sample complexity at most $m(\epsilon, \delta, K)$. ■

3.1.2 Characterizing Batch Learnability for General Losses

In this section, we characterize the learnability for general multilabel losses. Our main technical tool in characterizing the learnability for general loss functions is the equivalence between realizable and agnostic learning guaranteed by Lemma 4. Thus, we first provide the proof of that lemma before we proceed further.

Proof (of Lemma 4) Note that agnostic learnability implies realizable learnability by definition. So, it suffices to show that realizable learnability of $\mathcal{F}$ with respect to ℓ implies

agnostic learnability. Our proof here is constructive. That is, given a realizable algorithm $\mathcal{A}$ for $\mathcal{F}$, we provide an algorithm, stated as Algorithm 1, that constructs an agnostic learner for $\mathcal{F}$.

Algorithm 1 Agnostic learner for $\mathcal{F}$ with respect to ℓ

Input: Realizable learner $\mathcal{A}$ for $\mathcal{F}$ with respect to ℓ, unlabeled samples $S_U \sim \mathcal{D}_{\mathcal{X}}$, and different labeled samples $S_L \sim \mathcal{D}$ independent from S_U

1 Run $\mathcal{A}$ over all possible labelings of S_U by $\mathcal{F}$ to construct a concept class

$$C(S_U) := \left\{\mathcal{A}\big(S_U, f(S_U)\big) \mid f \in \mathcal{F}_{|S_U}\right\}.$$

2 Return $\hat{g} \in C(S_U)$ with the lowest empirical error over S_L with respect to ℓ.

Let $\mathcal{D}$ be an arbitrary distribution over $\mathcal{X} \times \mathcal{Y}$. Define

$$f^\star := \inf_{f \in \mathcal{F}} \mathbb{E}_{\mathcal{D}}[\ell(f(x), y)]$$

to be the optimal predictor in $\mathcal{F}$. Now consider a predictor $g = \mathcal{A}(S_U, f^\star(S_U)) \in C(S_U)$ returned by $\mathcal{A}$ when trained on samples S_U labeled by $f^\star$. Note that g exists in $C(S_U)$ because we consider all possible labelings of S_U by $\mathcal{F}$ and there must be a sample labeled by $f^\star$ as well. Let $m_{\mathcal{A}}(\epsilon, \delta, K)$ be the sample complexity of the algorithm $\mathcal{A}$. Since $\mathcal{A}$ is a realizable learner for $\mathcal{F}$, for $|S_U| \geq m_{\mathcal{A}}\left(\frac{\epsilon}{2c}, \frac{\delta}{2}, K\right)$ with probability $1 - \delta/2$ over sample $S_U \sim \mathcal{D}_{\mathcal{X}}$, we have

$$\mathbb{E}_{\mathcal{D}_{\mathcal{X}}}\left[\ell(g(x), f^\star(x))\right] \leq \frac{\epsilon}{2c},$$

where $\mathcal{D}_{\mathcal{X}}$ is the marginal distribution of $\mathcal{D}$ restricted to $\mathcal{X}$ and c is the subaddtivity constant of ℓ. Since the loss function is c-subadditive, for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$, we have $\ell(g(x), y) \leq \ell(f^\star(x), y) + c\,\ell(g(x), f^\star(x))$ pointwise. Taking expectation with respect to $(x, y) \sim \mathcal{D}$, we obtain

$$\mathbb{E}_{\mathcal{D}}[\ell(g(x), y)] \leq c\,\mathbb{E}_{\mathcal{D}_{\mathcal{X}}}[\ell(g(x), f^\star(x))] + \mathbb{E}_{\mathcal{D}}[\ell(f^\star(x), y)] \leq \mathbb{E}_{\mathcal{D}}[\ell(f^\star(x), y)] + \frac{\epsilon}{2},$$

where the inequality holds with probability $\geq 1 - \delta/2$. Thus, we have shown that there exists a predictor $g \in C(S_U)$ that achieves agnostic PAC bounds for $\mathcal{F}$ with respect to ℓ. Let $\ell(\cdot, \cdot) \leq b$ be the upper bound on ℓ. Recall that by Hoeffding's inequality and union bound, with probability at least $1 - \delta/2$ over sample $S_L \sim \mathcal{D}$, the empirical risk of every hypothesis in $C(S_U)$ on a sample of size $\geq \frac{8b^2}{\epsilon^2} \log \frac{4|C(S_U)|}{\delta}$ is at most $\epsilon/4$ away from its population risk. So, if $|S_L| \geq \frac{8b^2}{\epsilon^2} \log \frac{4|C(S_U)|}{\delta}$, then with probability at least $1 - \delta/2$ over sample $S_L \sim \mathcal{D}$, we have

$$\frac{1}{|S_L|} \sum_{(x,y) \in S_L} \ell(g(x), y) \leq \mathbb{E}_{\mathcal{D}}\left[\ell(g(x), y)\right] + \frac{\epsilon}{4} \leq \mathbb{E}_{\mathcal{D}}[\ell(f^\star(x), y)] + \frac{3\epsilon}{4}.$$

Next, consider the predictor $\hat{g}$ returned by Algorithm 1. Since it is an empirical risk minimizer, its empirical risk can be at most the empirical risk of g. Given that the population risk of $\hat{g}$ can be at most $\epsilon/4$ away from its empirical risk, we have that

$$\mathbb{E}_{\mathcal{D}}\left[\ell(\hat{g}(x), y)\right] \leq \mathbb{E}_{\mathcal{D}}[\ell(f^\star(x), y)] + \frac{3\epsilon}{4} + \frac{\epsilon}{4} \leq \inf_{f \in \mathcal{F}} \mathbb{E}_{\mathcal{D}}[\ell(f(x), y)] + \epsilon,$$

where the second inequality above uses the definition of $f^\star$. Note that this inequality holds with probability at least $1-\delta$, where the probability is taken over both samples S_U and S_L. Thus, we have shown that Algorithm 1 is an agnostic PAC learner for $\mathcal{F}$ with respect to ℓ.

We now upper bound the sample complexity of Algorithm 1, denoted $m(\epsilon, \delta, K)$ hereinafter. Note that $m_{\mathcal{A}}(\epsilon, \delta, K)$ is at most the number of unlabeled samples required for the realizable algorithm $\mathcal{A}$ to succeed plus the number of labeled samples for the ERM step to succeed. Thus,

$$\begin{aligned} m(\epsilon, \delta, K) &\leq m_{\mathcal{A}}\left(\frac{\epsilon}{2c}, \frac{\delta}{2}, K\right) + \frac{8b^2}{\epsilon^2} \log \frac{4|C(S_U)|}{\delta} \\ &\leq m_{\mathcal{A}}\left(\frac{\epsilon}{2c}, \frac{\delta}{2}, K\right) + \frac{8b^2}{\epsilon^2}\left(m_{\mathcal{A}}\left(\frac{\epsilon}{2c}, \frac{\delta}{2}, K\right) \log\left(|\mathrm{im}(\mathcal{F})|\right) + \log \frac{4}{\delta}\right), \end{aligned}$$

where the second inequality follows due to $|C(S_U)| \leq |\mathrm{im}(\mathcal{F})|^{|S_U|}$ and we need $|S_U|$ to be of size $m_{\mathcal{A}}\left(\frac{\epsilon}{2c}, \frac{\delta}{2}, K\right)$. ■

With Lemma 4 in hand, we can now relate the learnability of $\mathcal{H}$ with respect to any ℓ satisfying identity of indiscernibles to the learnability of $\mathcal{H}$ with respect to ℓ_H. To that end, we prove a result establishing the equivalence of learnability between any two loss functions satisfying the identity of indiscernibles.

Lemma 7. *Let ℓ and ℓ' be any two loss functions satisfying the identity of indiscernibles. Then, a function class $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ is agnostic PAC learnable with respect to ℓ if and only if $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ is agnostic PAC learnable with respect to ℓ'.*

The key idea behind the proof of Lemma 7 is to use a realizable learner for ℓ to construct a realizable learner for ℓ'. This is possible because $\ell'(y_1, y_2) = 0$ if and only if $\ell(y_1, y_2) = 0$ for any $y_1, y_2 \in \mathcal{Y}$. Given such realizable learner for ℓ', Lemma 4 guarantees the existence of an agnostic learner for ℓ'.

Proof Since ℓ and ℓ' are arbitrary, it suffices to prove only one direction. So, let us assume that $\mathcal{F}$ is learnable with respect to ℓ. We will now show that $\mathcal{F}$ is learnable with respect to ℓ' as well. First, we show this for any realizable distribution $\mathcal{D}$ with respect to ℓ'. Since, for any $y_1, y_2 \in \mathcal{Y}$, we have $\ell'(y_1, y_2) = 0$ if and only if $\ell(y_1, y_2) = 0$, the distribution $\mathcal{D}$ is also realizable with respect to ℓ. Furthermore, as there are at most 2^{2K} distinct possible inputs to $\ell'(\cdot, \cdot)$, the loss function can only take a finite number of values. So, we can always find universal constants $a > 0$ and $b > 0$ (that only depends on ℓ and ℓ') such that $a\ell \leq \ell' \leq b\ell$. Given that $\mathcal{F}$ is learnable with respect to ℓ, there exists a learning algorithm $\mathcal{A}$ with the following property: for any $\epsilon, \delta > 0$, for $S \sim \mathcal{D}^n$, such that $n = m_{\mathcal{A}}(\frac{\epsilon}{b}, \delta, K)$, the algorithm outputs a predictor $h = \mathcal{A}(S)$ such that, with probability $1-\delta$ over $S \sim \mathcal{D}^n$, we have $\mathbb{E}_{\mathcal{D}}[\ell(h(x), y)] \leq \frac{\epsilon}{b}$. This inequality upon using the fact that $\ell'(h(x), y) \leq b\ell(h(x), y)$ pointwise reduces to $\mathbb{E}_{\mathcal{D}}[\ell'(h(x), y)] \leq \epsilon$. Therefore, any realizable learner $\mathcal{A}$ for ℓ is also a realizable learner for ℓ'. Finally, as ℓ' satisfies the identity of indiscernibles and thus c-subadditivity, Lemma 4 guarantees the existence of agnostic PAC learner $\mathcal{B}$ for $\mathcal{F}$ with respect to ℓ'. In particular, one such agnostic PAC learner $\mathcal{B}$ is Algorithm 1 that has sample

complexity

$$m_{\mathcal{B}}(\epsilon, \delta, K) \leq m_{\mathcal{A}}\left(\frac{\epsilon}{2bc}, \frac{\delta}{2}, K\right) + \frac{b^2}{\epsilon^2}\, O\left(m_{\mathcal{A}}\left(\frac{\epsilon}{2bc}, \frac{\delta}{2}, K\right) K + \log\frac{1}{\delta}\right),$$

where $c > 1$ is the subadditivity constant of ℓ' and $m_{\mathcal{A}}(\cdot, \cdot, K)$ is the sample complexity of any realizable algorithm $\mathcal{A}$.

■

An immediate consequence of Lemma 7 is that learnability with respect to any loss ℓ satisfying the identity of indiscernibles is equivalent to learnability with respect to the Hamming loss ℓ_H. Thus, given Theorem 6, we can deduce the following result.

Theorem 8. *Let ℓ be any multilabel loss function satisfying the identity of indiscernibles. A function class $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ is agnostic PAC learnable with respect to ℓ if and only if each restriction $\mathcal{F}_k \subseteq \mathcal{Y}_k^{\mathcal{X}}$ is agnostic PAC learnable with respect to the* 0-1 *loss.*

Remark. Since the learnability of a binary function class with respect to the 0-1 loss is characterized by its VC dimension (Shalev-Shwartz and Ben-David, 2014, Theorem 6.7), Theorem 8 implies that $\mathcal{F}$ is learnable with respect to ℓ satisfying the identity of indiscernibles if and only if $\mathrm{VC}(\mathcal{F}_k) < \infty$ for each $k \in [K]$.

3.2 Batch Multioutput Regression

In this section, we consider the case when $\mathcal{Y} = [0,1]^K \subseteq \mathbb{R}^K$ for $K \in \mathbb{N}$. For bounded targets (with a known bound), this target space is without loss of generality because one can always normalize each $\mathcal{Y}_k$ to [0,1] by subtracting the lower bound and dividing by the upper bound of $\mathcal{Y}_k$. As usual, we consider an arbitrary multioutput function class $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$. Following our outline in classification, we first study the learnability of $\mathcal{F}$ under decomposable losses and then non-decomposable losses.

3.2.1 Characterizing Learnability for Decomposable Losses

A canonical loss for the scalar-valued regression is the absolute value metric, $d_1(f_k(x), y^k) := |f_k(x) - y^k|$. Analogously, we define $d_p(f_k(x), y^k) := |f_k(x) - y^k|^p$ for $p > 1$ are other natural scalar-valued losses. For multioutput regression, we consider decomposable losses that are natural multivariate extensions of the d_1 metric. In particular, we consider loss functions with the following properties.

Assumption 1. *The loss can be written as $\ell(f(x), y) = \sum_{k=1}^{K} \psi_k \circ d_1(f_k(x), y^k)$ where for each $k \in [K]$, the function $\psi_k : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is L-Lipschitz and satisfies $\psi_k(0) = 0$.*

Here, $\psi_k \circ d_1$ is a composition function defined as $\psi_k \circ d_1(f_k(x), y^k) := \psi_k\big(d_1(f_k(x), y^k)\big)$. Note that $\psi_k \circ d_1$ is a large family of loss functions that effectively contains all natural decomposable multioutput regression losses. For instance, taking $\psi_k(z) = |z|^p$ for $p \geq 1$ gives ℓ_p norms raised to their p-th power. Considering $\psi_k(z) = |z|^2/2\, \mathbb{1}[|z| \leq \delta] + \delta(|z| - \delta/2)\mathbb{1}[|z| > \delta]$ for some $1 > \delta > 0$ yields multivariate extension of Huber loss used for robust regression. One may also construct a multioutput loss by considering different scalar-valued

losses for each coordinate output. Next, we establish an equivalence between the learnability of $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ with respect to ℓ and the learnability of each $\mathcal{F}_k$ with respect to the loss $\psi_k \circ d_1$.

Theorem 9. *Let ℓ be any loss function that satisfies Assumption 1. Then, a function class $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ is agnostic learnable with respect to ℓ if and only if each of $\mathcal{F}_k \subseteq \mathcal{Y}_k^{\mathcal{X}}$ is agnostic learnable with respect to $\psi_k \circ d_1$.*

Proof The proof of the sufficiency direction is similar to that of Theorem 6, so we defer it to Appendix C.1. We now focus on the necessity direction. To that end, we show that if $\mathcal{F}$ is agnostic learnable with respect to ℓ, then each $\mathcal{F}_k$ is agnostic learnable with respect to $\psi_k \circ d_1$. In particular, given oracle access to agnostic learner $\mathcal{A}$ for $\mathcal{F}$, we construct agnostic learner $\mathcal{A}_1$ for $\mathcal{F}_1$. By symmetry, a similar reduction can then be used to construct an agnostic learner for each component $\mathcal{F}_k$.

Since we are given a sample with a single, univariate target, the main problem is to find the right way to augment samples to a K-variate target. In the proof of Theorem 6, we showed that randomly choosing $y_{ik} \sim \text{Uniform}(\{-1, 1\})$ for $k \geq 2$ results in all predictors having a constant $1/2$ risk–leaving only the risk of the first component on both sides. Unfortunately, in regression under general losses, no single augmentation works for every distribution on $\mathcal{X}$. Thus, we augment the samples by considering all possible behaviors of $(\mathcal{F}_2, \ldots, \mathcal{F}_K)$ on the sample. Since the function class maps to a potentially uncountably infinite space, we first discretize each component of the function class and consider all possible labelings over the discretized space. Fix $1 > \alpha > 0$. For $k \geq 2$, define the discretization

$$f_k^\alpha(x) = \left\lfloor \frac{f(x)}{\alpha} \right\rfloor \alpha \tag{1}$$

for every $f_k \in \mathcal{F}_k$ and the discretized component class $\mathcal{F}_k^\alpha = \{f_k^\alpha | f_k \in \mathcal{F}_k\}$. Note that a function in $\mathcal{F}_k$ maps to $\{0, \alpha, 2\alpha, \ldots, \lfloor 1/\alpha \rfloor \alpha\}$ and the size of the range of the discretized function class $\mathcal{F}_k^\alpha$ is $1 + \lfloor 1/\alpha \rfloor \leq (\alpha + 1)/\alpha \leq 2/\alpha$. For the convenience of exposition, let us define $\mathcal{F}_{2:K}^\alpha$ to be $\mathcal{F}^\alpha$ without the first component, and we denote $f_{2:K}^\alpha$ to be an element of $\mathcal{F}_{2:K}^\alpha$.

Algorithm 2 Agnostic learner for $\mathcal{F}_1$ with respect to $\psi_1 \circ d_1$

Input: Agnostic learner $\mathcal{A}$ for $\mathcal{F}$, samples $S = (x_{1:n}, y_{1:n}^1) \sim \mathcal{D}_1^n$, and another independent samples $\widetilde{S}$ from $\mathcal{D}_1$

1 Define $S_{\text{aug}} = \{(x_{1:n}, y_{1:n}^1, f_{2:K}^\alpha(x_{1:n}) \mid f_{2:K}^\alpha \in \mathcal{F}_{2:K}^\alpha\}$, all possible augmentations of S by $\mathcal{F}_{2:K}^\alpha$.

2 Run $\mathcal{A}$ over all possible augmentations to get $C(S) := \left\{\mathcal{A}(S_a) \mid S_a \in S_{\text{aug}}\right\}$.

3 Define $C_1(S) = \{g_1 \mid (g_1, \ldots, g_k) \in C(S)\}$, a restriction of $C(S)$ to its first coordinate output.

4 Return $\hat{g}_1$, the predictor in $C_1(S)$ with the lowest empirical error over $\widetilde{S}$ with respect to $\psi_1 \circ d_1$.

We now show that Algorithm 2 is an agnostic learner for $\mathcal{F}_1$. First, let us define

$$f_1^\star := \arg\min_{f_1 \in \mathcal{F}_1} \mathbb{E}_{\mathcal{D}_1}[\psi_1 \circ d_1(f_1(x), y^1)],$$

to be optimal predictor in $\mathcal{F}_1$ with respect to $\mathcal{D}_1$. By definition of $\mathcal{F}_1$, there must exist $f^\star_{2:K} \in \mathcal{F}_{2:K}$ such that $(f^\star_1, f^\star_{2:K}) \in \mathcal{F}$. We note that $f^\star_k$ need not be optimal predictors in $\mathcal{F}_k$ for $k \geq 2$, but we use the $\star$ notation just to associate these component functions with the first component function $f^\star_1$. Define $f^{\star,\alpha}_{2:K} \in \mathcal{F}^\alpha_{2:K}$ to be the corresponding discretization of $f^\star_{2:K}$. At a high level, the key idea of this proof is to show that the algorithm $\mathcal{A}$ when run on the sample $(x_{1:n}, y^1_{1:n}, f^{\star,\alpha}_{2:K}(x_{1:n})$ produces a predictor $g = \mathcal{A}(x_{1:n}, y^1_{1:n}, f^{\star,\alpha}_{2:K}(x_{1:n}))$ such that its restriction g_1 is a valid agnostic learner for $\mathcal{F}_1$. Although one such augmentation is enough to produce an agnostic learner for $\mathcal{F}_1$, all possible augmentations are required in step 2 of Algorithm 2 because $f^\star_1$ is not known to the learner apriori. We now make this argument precise.

Suppose $g = \mathcal{A}((x_{1:n}, y^1_{1:n}, f^{\star,\alpha}_{2:K}(x_{1:n}))$ is the predictor obtained by running $\mathcal{A}$ on the sample augmented by $f^{\star,\alpha}_{2:K}$. Note that $g \in C(S)$ by definition. Let $m_{\mathcal{A}}(\epsilon, \delta, K)$ be the sample complexity of $\mathcal{A}$. Since $\mathcal{A}$ is an agnostic learner for $\mathcal{F}$ with respect to ℓ, we have that for $n \geq m_{\mathcal{A}}(\epsilon/4, \delta/2, K)$, with probability at least $1 - \delta/2$,

$$\mathbb{E}_{\mathcal{D}_1}\left[\psi_1 \circ d_1(g_1(x), y^1)\right] + \sum_{k=2}^{K} \mathbb{E}_{\mathcal{D}_\mathcal{X}}\left[\psi_k \circ d_1(g_k(x), f^{\star,\alpha}_k(x))\right] \\ \leq \inf_{f \in \mathcal{F}} \left(\mathbb{E}_{\mathcal{D}_1}\left[\psi_1 \circ d_1(f_1(x), y^1)\right] + \sum_{k=2}^{K} \mathbb{E}_{\mathcal{D}_\mathcal{X}}\left[\psi_k \circ d_1(f_k(x), f^{\star,\alpha}_k(x))\right] \right) + \frac{\epsilon}{4}$$

Note that the quantity on the left is trivially lower bounded by the risk of the first component. To handle the right-hand side, we first note that the optimal risk is trivially upper bounded by the risk of $(f^\star_1, f^\star_{2:K})$, yielding

$$\mathbb{E}_{\mathcal{D}_1}\left[\psi_1 \circ d_1(g_1(x), y^1)\right] \leq \mathbb{E}_{\mathcal{D}_1}\left[\psi_1 \circ d_1(f^\star_1(x), y^1)\right] + \sum_{k=2}^{K} \mathbb{E}_{\mathcal{D}_\mathcal{X}}\left[\psi_k \circ d_1(f^\star_k(x), f^{\star,\alpha}_k(x))\right] + \frac{\epsilon}{4}.$$

Next, using the L-Lipschitzness of ψ_k and the fact that $\psi_k(0) = 0$ implies $\psi_k \circ d_1(f^\star_k(x), f^{\star,\alpha}_k(x)) \leq L\, d_1(f^\star_k(x), f^{\star,\alpha}_k(x)) \leq L\alpha$ for all $k \geq 2$. Thus, picking $\alpha = \frac{\epsilon}{4LK}$ and using the definition of $f^\star_1$, we obtain

$$\mathbb{E}_{\mathcal{D}_1}\left[\psi_1 \circ d_1(g_1(x), y^1)\right] \leq \inf_{f_1 \in \mathcal{F}_1} \mathbb{E}_{\mathcal{D}_1}[\psi_1 \circ d_1(f_1(x), y^1)] + \frac{\epsilon}{2}.$$

Therefore, we have shown the existence of one predictor $g \in C(S)$ such that its restriction to the first component, g_1, obtains the agnostic bound. Note that since ψ_1 is L-Lipschitz and satisfies $\psi_1(0) = 0$, we obtain that $\psi_1 \circ d_1(\cdot, \cdot) \leq L$. The upper bound also uses the fact that $|f_1(x) - y^1| \leq 1$. Now recall that by Hoeffding's Inequality and union bound, with probability at least $1 - \delta/2$, the empirical risk of every hypothesis in $C_1(S)$ on a sample of size $\geq \frac{8L^2}{\epsilon^2} \log \frac{4|C_1(S)|}{\delta}$ is at most $\epsilon/4$ away from its true error. So, if $|\widetilde{S}| \geq \frac{8L^2}{\epsilon^2} \log \frac{4|C_1(S)|}{\delta}$, then with probability at least $1 - \delta/2$, the empirical risk of the predictor g_1 is

$$\frac{1}{|\widetilde{S}|} \sum_{(x,y^1) \in \widetilde{S}} \psi_1 \circ d_1(g_1(x), y^1) \leq \mathbb{E}_{\mathcal{D}_1}\left[\psi_1 \circ d_1(g_1(x), y^1)\right] + \frac{\epsilon}{4} \leq \inf_{f_1 \in \mathcal{F}_1} \mathbb{E}_{\mathcal{D}_1}[\psi_1 \circ d_1(f_1(x), y^1)] + \frac{3\epsilon}{4}.$$

Since $\hat{g}_1$, the output of Algorithm 2 is the ERM on $\widetilde{S}$ over $C_1(S)$, its empirical risk can be at most the empirical risk of g_1, which is at most $\inf_{f_1 \in \mathcal{F}_1} \mathbb{E}_{\mathcal{D}_1}[\psi_1 \circ d_1(f_1(x), y^1)] + \frac{3\epsilon}{4}$. Given that the population risk of $\hat{g}_1$ is at most $\epsilon/4$ away from its empirical risk, we can conclude that the population risk of $\hat{g}_1$ is

$$\mathbb{E}_{\mathcal{D}_1}[\psi_1 \circ d_1(\hat{g}_1(x), y^1)] \leq \inf_{f_1 \in \mathcal{F}_1} \mathbb{E}_{\mathcal{D}_1}[\psi_1 \circ d_1(f_1(x), y^1)] + \epsilon.$$

Applying union bounds, the entire process, algorithm $\mathcal{A}$ on the dataset augmented by $f^{\star,\alpha}_{2:K}$ and ERM in step 4, succeeds with probability $1 - \delta$. The sample complexity of Algorithm 2 is the sample complexity of Algorithm $\mathcal{A}$ and the sample complexity of ERM in step 4, which is

$$\begin{aligned}
&\leq m_{\mathcal{A}}(\epsilon/4, \delta/2, K) + \frac{8L^2}{\epsilon^2} \log \frac{4|C_1(S)|}{\delta} \\
&\leq m_{\mathcal{A}}(\epsilon/4, \delta/2, K) + \frac{8KL^2}{\epsilon^2} \left(m_{\mathcal{A}}(\epsilon/4, \delta/2, K) \log \left(\frac{4K\,L}{\epsilon} \right) + \log \frac{4}{\delta} \right),
\end{aligned}$$

where the second inequality follows due to $|C_1(S)| \leq (2/\alpha)^{m_{\mathcal{A}}(\epsilon/4, \delta/2, K)\,K}$ is the required size of $C_1(S)$ and our choice of $\alpha = \epsilon/(4KL)$. This completes the proof as we have shown that Algorithm 2 is an agnostic learner for $\mathcal{F}_1$ with respect to $\psi_1 \circ d_1$. ■

3.2.2 A More General Characterization of Learnability for Decomposable Losses

Unlike Theorems 6 and 8 in classification setting where we connected the learnability of $\mathcal{F}$ to the learnability of $\mathcal{F}_k$'s with respect to 0-1 loss, Theorem 9 relates the learnability of $\mathcal{F}$ to the learnability of $\mathcal{F}_k$ with respect to $\psi_k \circ d_1$ instead of the more canonical loss d_1. In this section, we complete that final step to characterizing learnability in terms of d_1 with an additional assumption on ψ_k.

Assumption 2. *For all $k \in [K]$, the function $\psi_k : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is monotonic.*

Under these assumptions, Theorem 10 provides a more general characterization than Theorem 9.

Theorem 10. *Let ℓ be any loss function that satisfies Assumptions 1 and 2. Then, a multioutput function class $\mathcal{F} \subseteq ([0,1]^K)^{\mathcal{X}}$ is agnostic learnable with respect to ℓ if and only if each $\mathcal{F}_k \subseteq [0,1]^{\mathcal{X}}$ is agnostic learnable with respect to d_1.*

Since the fat-shattering dimension of a real-valued function class characterizes its learnability with respect to d_1 loss, Theorem 10 implies that a multioutput function class $\mathcal{F}$ is learnable with respect to ℓ satisfying Assumptions 1 and 2 if and only if $\mathrm{fat}_\gamma(\mathcal{F}) < \infty$ for every fixed scale $\gamma > 0$. Theorem 10 is an immediate consequence of Theorem 9 and the following lemma, the proof of which is deferred to Appendix C.2.

Lemma 11. *Let $\mathcal{Y} = [0,1]$ be the label space and $\psi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ be any Lipschitz and monotonic function that satisfies $\psi(0) = 0$. A scalar-valued function class $\mathcal{G} \subseteq [0,1]^{\mathcal{X}}$ is agnostic learnable with respect to $\psi \circ d_1$ if and only if $\mathcal{G}$ is agnostic learnable with respect to d_1.*

The part of the lemma showing that d_1 learnability implies $\psi \circ d_1$ is trivial using the Rademacher-based argument and Talagrand's contraction lemma. However, proving $\psi \circ d_1$ learnability implies d_1 learnability is non-trivial. The case $\psi(z) = |z|^2$ is considered in (Anthony and Bartlett, 1999, Theorem 19.5), but their proof requires a mismatch between the label space and the prediction space, namely $\mathcal{Y} = [-1, 2]$ but $\mathcal{G} \subseteq [0, 1]^{\mathcal{X}}$. In this work, we improve their result by showing the equivalence between $\psi \circ d_1$ learnability and d_1 learnability without requiring extended label space.

An application of Lemma 11 is that the learnability of a real-valued function class $\mathcal{G}$ with respect to losses d_1 and d_p are equivalent for any $p > 1$.

3.2.3 Characterizing Learnability for Non-Decomposable Losses

Next, we study the learnability of function class $\mathcal{F}$ with respect to non-decomposable losses. In the regression setting, the natural non-decomposable loss to consider is ℓ_p norm, which is defined as $\ell_p(f(x), y) := \left(\sum_{k=1} |f_k(x) - y^k|^p\right)^{1/p}$ for $1 \leq p < \infty$. For $p = \infty$, the p-norm is defined as $\ell_\infty(f(x), y) := \max_k |f_k(x) - y^k|$. One might be interested in ℓ_p norms instead of their decomposable counterparts ℓ_p^p losses discussed in the previous section mainly for robustness to outliers. The following result characterizes the agnostic learnability of $\mathcal{F}$ with respect to ℓ_p norms.

Theorem 12. *Fix $p \geq 1$. The function class $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ is agnostic learnable with respect to ℓ_p norm if and only if each of $\mathcal{F}_k \subseteq \mathcal{Y}_k^{\mathcal{X}}$ is agnostic learnable with respect to the absolute value loss, d_1.*

Using $\psi_k(z) = |z|$ in Theorem 9 implies that $\mathcal{F}$ is learnable with respect to ℓ_1 if and only if each $\mathcal{F}_k$ is learnable with respect to d_1. Thus, to prove Theorem 12, it suffices to show that for all $p > 1$, $\mathcal{F}$ is learnable with respect to ℓ_p if and only if $\mathcal{F}$ is learnable with respect to ℓ_1 norm.

Proof We begin by proving the sufficiency direction. As discussed above, the learnability of each $\mathcal{F}_k$ with respect to d_1 implies that $\mathcal{F}$ is learnable with respect to ℓ_1. Then, the high-level idea of the proof is to use an agnostic learner for $\mathcal{F}$ with respect to ℓ_1 to construct a realizable learner for $\mathcal{F}^\alpha$ with respect to ℓ_1. Using the fact $\ell_1(y_1, y_2) = 0$ if and only if $\ell_p(y_1, y_2) = 0$ for any $y_1, y_2 \in \mathcal{Y}$, the realizable learner for ℓ_1 is also a realizable learner for ℓ_p. Finally, as $|\mathrm{im}(\mathcal{F}^\alpha)| < \infty$ for every $\alpha > 0$, Lemma 4 guarantees the existence of an agnostic learner for $\mathcal{F}^\alpha$ with respect to ℓ_p. Then, a simple application of the triangle inequality shows that an agnostic learner for $\mathcal{F}^\alpha$ is also an agnostic learner for $\mathcal{F}$ with respect to ℓ_p. We now proceed with the formal proof.

Part 1: Sufficiency. Fix $p > 1$. Recall that the learnability of each $\mathcal{F}_k$ with respect to d_1 implies that $\mathcal{F}$ is learnable with respect to ℓ_1. Let $\mathcal{D}$ be arbitrary distribution on $\mathcal{X} \times \mathcal{Y}$ and $\mathcal{A}$ be an agnostic PAC learner for $\mathcal{F}$ with respect to ℓ_1 with sample complexity $m(\epsilon, \delta)$. For any $\epsilon, \delta > 0$, with n sufficiently larger than $m(\epsilon/2, \delta, K)$, with probability at least $1 - \delta$ over $S \sim \mathcal{D}^n$, we have

$$\mathbb{E}_{\mathcal{D}}[\ell_1(g(x), y)] \leq \inf_{f \in \mathcal{F}} \mathbb{E}_{\mathcal{D}}[\ell_1(f(x), y)] + \frac{\epsilon}{2},$$

where $g = \mathcal{A}(S)$. Define $\mathcal{F}^\alpha$ to be a discretized function class obtained by discretizing $\mathcal{F}$ component-wise using the scheme (1). Consider $\mathcal{D}$ to be a realizable distribution

with respect to $\mathcal{F}^\alpha$. Note that the triangle inequality implies $\ell_1(f(x), y) \leq \ell_1(f^\alpha(x), y) + \ell_1(f(x), f^\alpha(x)) \leq \ell_1(f^\alpha(x), y) + K\alpha$. Taking $\alpha = \frac{\epsilon}{2K}$ and using the fact that $\inf_{f\in\mathcal{F}} \mathbb{E}[\ell_1(f^\alpha(x), y)] = 0$, we obtain $\mathbb{E}_{\mathcal{D}}[\ell_1(g(x), y)] \leq \epsilon$. Next, using the inequality $\ell_p(g(x), y) \leq \ell_1(g(x), y)$ pointwise yields

$$\mathbb{E}_{\mathcal{D}}[\ell_p(g(x), y)] \leq \epsilon.$$

Therefore, $\mathcal{A}$ is a realizable learner for $\mathcal{F}^\alpha$ with respect to ℓ_p with sample complexity $m(\epsilon/2, \delta, K)$. Since $|\mathrm{im}(\mathcal{F}^\alpha)| < \infty$ and $\ell_p(\cdot,\cdot)$ are 1-subadditive, Lemma 4 implies that $\mathcal{F}^\alpha$ is agnostic learnable with respect to ℓ_p via Algorithm 1, referred to as algorithm $\mathcal{B}$ henceforth. Thus, for any $\epsilon, \delta > 0$, there exists a $n \geq m_{\mathcal{B}}(\epsilon/2, \delta/2)$, for any distribution $\widetilde{\mathcal{D}}$ on $\mathcal{X} \times \mathcal{Y}$, running $\mathcal{B}$ on $S \sim \mathcal{D}^n$ outputs a predictor $\tilde{g} \in \mathcal{Y}^{\mathcal{X}}$ such that with probability at least $1 - \delta$ over $S \sim \widetilde{\mathcal{D}}^n$, we have

$$\mathbb{E}_{\widetilde{\mathcal{D}}}[\ell_p(\tilde{g}(x), y)] \leq \inf_{f^\alpha \in \mathcal{F}^\alpha} \mathbb{E}_{\widetilde{\mathcal{D}}}[\ell_p(f^\alpha(x), y)] + \frac{\epsilon}{2} = \inf_{f\in\mathcal{F}} \mathbb{E}_{\widetilde{\mathcal{D}}}[\ell_p(f^\alpha(x), y)] + \frac{\epsilon}{2}.$$

Using triangle inequality, we have $\ell_p(f^\alpha(x), y) \leq \ell_p(f(x), y) + \ell_p(f^\alpha(x), f(x)) \leq \ell_p(f(x), y) + \alpha K$ pointwise. Again taking $\alpha = \frac{\epsilon}{2K}$, we obtain

$$\mathbb{E}_{\widetilde{\mathcal{D}}}[\ell_p(\tilde{g}(x), y)] \leq \inf_{f\in\mathcal{F}} \mathbb{E}_{\widetilde{\mathcal{D}}}[\ell_p(f(x), y)] + \epsilon.$$

Therefore, we have shown that $\mathcal{F}$ is agnostic PAC learnable with respect to ℓ_p.

The sample complexity of agnostic learner $\mathcal{B}$ can be made precise using the sample complexity of Algorithm 1. In particular, the sample complexity of $\mathcal{B}$ is the sample complexity of the realizable learner $\mathcal{A}$ and the sample complexity of the ERM in step 2 of Algorithm 1. Proof of Lemma 4 shows that the sample complexity of $\mathcal{B}$ must be

$$m_{\mathcal{B}}(\epsilon, \delta, K) \leq m_{\mathcal{A}}\left(\frac{\epsilon}{2c}, \frac{\delta}{2}, K\right) + \frac{8b^2}{\epsilon^2}\left(m_{\mathcal{A}}\left(\frac{\epsilon}{2c}, \frac{\delta}{2}, K\right) \log|(\mathrm{im}(\mathcal{F}^\alpha)|) + \log\frac{4}{\delta}\right),$$

where b is the upperbound on the loss and c is the subadditivity constant. Since $b \leq K$, and $c = 1$ for all ℓ_p norms with $p \geq 1$, we obtain

$$m_{\mathcal{B}}(\epsilon, \delta, K) \leq m_{\mathcal{A}}(\epsilon/2, \delta/2, K) + \frac{8K^3}{\epsilon^2}\left(m_{\mathcal{A}}(\epsilon/2, \delta/2, K) \log\left(\frac{4K}{\epsilon}\right) + \log\frac{4}{\delta}\right),$$

where we also use the fact that $|\mathrm{im}(\mathcal{F}^\alpha)| \leq (2/\alpha)^K$ for our choice of $\alpha = \frac{\epsilon}{2K}$.

Part 2: Necessity. Fix $p > 1$. We now prove that $\mathcal{F}$ being learnable with respect to ℓ_p implies $\mathcal{F}$ is learnable with respect to ℓ_1. The proof is identical to the proof of sufficiency, so we only provide a sketch of the argument here. Our proof strategy follows a similar route through realizable learnability of the discretized class $\mathcal{F}^\alpha$ and then the use of Lemma 4.

Recall that any agnostic learner $\mathcal{A}$ for $\mathcal{F}$ with respect to ℓ_p is a realizable learner for $\mathcal{F}^\alpha$ with respect to ℓ_p. Using the inequality $\ell_1(\cdot,\cdot) \leq K\ell_p(\cdot,\cdot)$ pointwise, we can deduce that $\mathcal{A}$ is also a realizable learner for $\mathcal{F}^\alpha$ with respect to ℓ_1. Since $|\mathrm{im}(\mathcal{F}^\alpha)| \leq (2/\alpha)^K < \infty$, Lemma 4 guarantees existence of an agnostic learner for $\mathcal{F}^\alpha$ with respect to ℓ_1. Using triangle inequality, we obtain $\ell_1(f^\alpha(x), y) \leq \ell_1(f(x), y) + \ell_1(f^\alpha(x), f(x)) \leq \ell_1(f(x), y) + \alpha K$ pointwise, and choosing appropriate discretization scale allows us to turn agnostic bound for $\mathcal{F}^\alpha$ into an agnostic bound for $\mathcal{F}$. ■

Remark. We note that Theorem 12 holds for any norm on $\mathbb{R}^K$, but we only focus on ℓ_p norms here due to their practical significance. As the fat-shattering dimension of a real-valued function class characterizes its learnability with respect to d_1 loss (Bartlett et al., 1996), Theorem 12 implies that a multioutput function class $\mathcal{F}$ is learnable with respect to ℓ_p for $1 \leq p \leq \infty$ if and only if $\text{fat}_\gamma(\mathcal{F}_k) < \infty$ for all $k \in [K]$ at every fixed scale $\gamma > 0$.

Since we are only concerned with the question of learnability in this work, our focus is not on optimal sample complexity rates. However, we point out that for any $p \geq 1$, if each $\mathcal{F}_k$ is learnable with respect to d_1, then $\mathcal{F}$ is learnable with respect to ℓ_p via ERM with a better sample complexity than Algorithm 1. The proof of this claim is based on Rademacher complexity and is provided in Appendix D.

4 Online Multioutput Learnability

Here, we study the online learnability of multioutput function classes. Throughout this section, we give regret bounds assuming an *oblivious* adversary. A standard reduction (Chapter 4 in Cesa-Bianchi and Lugosi (2006)) allows us to convert oblivious regret bounds to adaptive regret bounds in the full-information setting. A key requirement allowing an oblivious regret bound to generalize to an adaptive regret bound is that the learner's predictions on round t should not depend on any of its past predictions from previous rounds. This is true for all of the online learning algorithms in this section.

4.1 Online Agnostic-to-Realizable Reduction

Our strategy for constructively characterizing the learnability of general losses in both the batch classification and regression setting required the ability to convert a realizable learner to an agnostic learner in a black-box fashion. In this section, we provide an analog of this conversion for the *online* setting. More specifically, we focus on the setting where $\mathcal{F} \subseteq \mathcal{Y}^\mathcal{X}$ is a multioutput function class but $|\text{im}(\mathcal{F})| < \infty$ is finite. Then, for any c-subadditive loss function ℓ, we constructively convert a (potentially randomized) realizable online learner for $\mathcal{F}$ with respect to ℓ into an agnostic online learner for $\mathcal{F}$ with respect to ℓ. Theorem 13 formalizes the main result of this subsection.

Theorem 13. *Let $\mathcal{F} \subseteq \mathcal{Y}^\mathcal{X}$ be a multioutput function class such that $|im(\mathcal{F})| < \infty$ and $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ be any c-subadditive loss function such that $\ell(\cdot, \cdot) \leq M$. If $\mathcal{A}$ is a realizable online learner for $\mathcal{F}$ with respect to ℓ with sub-linear expected regret $R(T, |im(\mathcal{F})|)$, then for every $\beta \in (0, 1)$, there exists an agnostic online learner for $\mathcal{F}$ with respect to ℓ with expected regret*

$$\frac{cT}{T^\beta} \overline{R}(T^\beta, |im(\mathcal{F})|) + M\sqrt{2T^{1+\beta} \ln(|im(\mathcal{F})|)},$$

where $\overline{R}(T, |im(\mathcal{F})|)$ is any concave, sublinear upperbound on $R(T, |im(\mathcal{F})|)$.

Note that if $\overline{R}(T^\beta, |\text{im}(\mathcal{F})|)$ is sublinear in its first argument, then $\frac{cT}{T^\beta}\overline{R}(T^\beta, |\text{im}(\mathcal{F})|) + M\sqrt{2T^{1+\beta} \ln(|\text{im}(\mathcal{F})|)}$ is sublinear in T for any $\beta \in (0, 1)$. By Lemma 14, we are guaranteed the existence of $\overline{R}(T, |\text{im}(\mathcal{F})|)$.

Lemma 14. (Ceccherini-Silberstein et al., 2017, Lemma 5.17) *Let g be a positive sublinear function. Then, g is bounded from above by a concave sublinear function.*

Therefore, Theorem 13 and Lemma 14 show that for any function class $\mathcal{F}$ with finite image space and any c-subadditive loss function, realizable and agnostic online learnability are equivalent. We now begin the proof of Theorem 13.

Proof Let $\mathcal{A}$ be a (potentially randomized) online realizable learner for $\mathcal{F}$ with respect to ℓ. By definition, this means that for any (realizable) sequence $(x_1, f(x_1)), ..., (x_T, f(x_T))$ labeled by a function $f \in \mathcal{F}$, we have

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell(\mathcal{A}(x_t), f(x_t))\right] \leq R(T, |\text{im}(\mathcal{F})|),$$

where $R(T, |\text{im}(\mathcal{F})|)$ is a sub-linear function of T. We now use $\mathcal{A}$ to construct an agnostic online learner $\mathcal{Q}$ for $\mathcal{F}$ with respect to ℓ. Since we are assuming an oblivious adversary, let $(x_1, y_1), ..., (x_T, y_T) \in (\mathcal{X} \times \mathcal{Y})^T$ denote the stream of points to be observed by the online learner and $f^\star = \arg\min_{f \in \mathcal{F}} \sum_{t=1}^{T} \ell(f(x_t), y_t)$ to be the optimal function in hindsight.

Our high-level strategy is to construct a large set of Experts that approximately cover all possible labelings of the instances $x_1, ..., x_T$ by functions in $\mathcal{F}$. In particular, each Expert uses an independent copy of $\mathcal{A}$ to make predictions, but update $\mathcal{A}$ using *different* sequences of labeled instances. Together, our set of Experts update $\mathcal{A}$ using all possible sequences of labeled instances. In order to ensure that the number of Experts is not too large, we construct such a set of Experts over a sufficiently small *sub-sample* of the stream. Finally, we run the celebrated Randomized Exponential Weights Algorithm (REWA) (Cesa-Bianchi and Lugosi, 2006) using our set of experts and the scaled loss function $\frac{\ell}{M}$ over the original stream of points $(x_1, y_1), ..., (x_T, y_T)$. We now formalize this idea below.

For any bitstring $b \in \{0,1\}^T$, let $\phi : \{t : b_t = 1\} \to \text{im}(\mathcal{F})$ denote a function mapping time points where $b_t = 1$ to elements in the image space $\text{im}(\mathcal{F})$. Let $\Phi_b \subseteq (\text{im}(\mathcal{F}))^{\{t:b_t=1\}}$ denote all such functions ϕ. For every $f \in \mathcal{F}$, let $\phi_b^f \in \Phi_b$ be the mapping such that for all $t \in \{t : b_t = 1\}$, $\phi_b^f(t) = f(x_t)$. Let $|b| = |\{t : b_t = 1\}|$. For every $b \in \{0,1\}^T$ and $\phi \in \Phi_b$, define an Expert $E_{b,\phi}$. Expert $E_{b,\phi}$, formally presented in Algorithm 3, uses $\mathcal{A}$ to make predictions in each round. However, $E_{b,\phi}$ only updates $\mathcal{A}$ on those rounds where $b_t = 1$, using ϕ to compute a labeled instance $(x_t, \phi(t))$. For every $b \in \{0,1\}^T$, let $\mathcal{E}_b = \bigcup_{\phi \in \Phi_b} \{E_{b,\phi}\}$ denote the set of all Experts parameterized by functions $\phi \in \Phi_b$. If b is the all zeros bitstring, then $\mathcal{E}_b$ is empty. Therefore, we actually define $\mathcal{E}_b = \{E_0\} \cup \bigcup_{\phi \in \Phi_b} \{E_{b,\phi}\}$, where E_0 is the expert that never updates $\mathcal{A}$ and plays $\hat{y}_t = \mathcal{A}(x_t)$ for all $t \in [T]$. Note that $1 \leq |\mathcal{E}_b| \leq (|\text{im}(\mathcal{F})|)^{|b|}$.

With this notation in hand, we are now ready to present Algorithm 4, our main agnostic online learner $\mathcal{Q}$ for $\mathcal{F}$ with respect to ℓ. Our goal is to show that $\mathcal{Q}$ enjoys sublinear expected regret. There are three main sources of randomness: the randomness involved in sampling B, the internal randomness of $\mathcal{A}$, and the internal randomness of REWA. Let B, A and P denote the random variables associated with each source of randomness respectively. By construction, B, A, and P are independent.

Algorithm 3 Expert(b, ϕ)

Input: Independent copy of realizable online learner $\mathcal{A}$ for $\mathcal{F}$ with respect to ℓ

for $t = 1, ..., T$ **do**
  Receive example x_t
  Predict $\hat{y}_t = \mathcal{A}(x_t)$
  if $b_t = 1$ **then**
    Update $\mathcal{A}$ by passing $(x_t, \phi(t))$
end

Algorithm 4 Agnostic online learner $\mathcal{Q}$ for $\mathcal{F}$ with respect to ℓ

Input: Parameter $0 < \beta < 1$

Let $B \in \{0,1\}^T$ such that $B_t \overset{\text{iid}}{\sim} \text{Bernoulli}(\frac{T^\beta}{T})$
Construct the set of experts $\mathcal{E}_B = \{E_0\} \cup \bigcup_{\phi \in \Phi_B} \{E_{B,\phi}\}$ according to Algorithm 3
Run REWA $\mathcal{P}$ using $\mathcal{E}_B$ and the loss function $\frac{\ell}{M}$ over the stream $(x_1, y_1), ..., (x_T, y_T)$

Using Theorem 21.11 in Shalev-Shwartz and Ben-David (2014) and the fact that B, A and P are independent, REWA guarantees almost surely that

$$\sum_{t=1}^{T} \mathbb{E}\left[\ell(\mathcal{P}(x_t), y_t) | B, A\right] \leq \inf_{E \in \mathcal{E}_B} \sum_{t=1}^{T} \ell(E(x_t), y_t) + M\sqrt{2T \ln(|\mathcal{E}_B|)}.$$

Taking an outer expectation gives

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell(\mathcal{P}(x_t), y_t)\right] \leq \mathbb{E}\left[\inf_{E \in \mathcal{E}_B} \sum_{t=1}^{T} \ell(E(x_t), y_t)\right] + \mathbb{E}\left[M\sqrt{2T \ln(|\mathcal{E}_B|)}\right].$$

Therefore,

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \ell(\mathcal{Q}(x_t), y_t)\right] &= \mathbb{E}\left[\sum_{t=1}^{T} \ell(\mathcal{P}(x_t), y_t)\right] \\
&\leq \mathbb{E}\left[\inf_{E \in \mathcal{E}_B} \sum_{t=1}^{T} \ell(E(x_t), y_t)\right] + \mathbb{E}\left[M\sqrt{2T \ln(|\mathcal{E}_B|)}\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T} \ell(E_{B,\phi_B^{f\star}}(x_t), y_t)\right] + M\mathbb{E}\left[\sqrt{2T \ln(|\mathcal{E}_B|)}\right].
\end{aligned}$$

In the last step, we used the fact that for all $b \in \{0,1\}^T$ and $f \in \mathcal{F}$, we have $E_{b,\phi_b^f} \in \mathcal{E}_b$.

It now suffices to upperbound $\mathbb{E}\left[\sum_{t=1}^{T} \ell(E_{B,\phi_B^{f\star}}(x_t), y_t)\right]$. To do so, we need some additional notation. Given the realizable online learner $\mathcal{A}$, an instance $x \in \mathcal{X}$, and an ordered finite sequence of labeled examples $L \in (\mathcal{X} \times \mathcal{Y})^*$, let $\mathcal{A}(x|L)$ be the random variable denoting the prediction of $\mathcal{A}$ on the instance x after running and updating on L. For any $b \in \{0,1\}^T$, $f \in \mathcal{F}$, and $t \in [T]$, let $L^f_{b<t} = \{(x_i, f(x_i)) : i < t \text{ and } b_i = 1\}$ denote

the *subsequence* of the sequence of labeled instances $\{(x_i, f(x_i))\}_{i=1}^{t-1}$ where $b_i = 1$. Using this notation, we can write

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \ell(E_{B,\phi_B^{f^\star}}(x_t), y_t)\right] &= \mathbb{E}\left[\sum_{t=1}^{T} \ell(\mathcal{A}(x_t|L_{B_{<t}}^{f^\star}), y_t)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \ell(\mathcal{A}(x_t|L_{B_{<t}}^{f^\star}), y_t)\frac{\mathbb{P}\left[B_t = 1\right]}{\mathbb{P}\left[B_t = 1\right]}\right] \\
&= \frac{T}{T^\beta}\sum_{t=1}^{T} \mathbb{E}\left[\ell(\mathcal{A}(x_t|L_{B_{<t}}^{f^\star}), y_t)\mathbb{P}\left[B_t = 1\right]\right] \\
&= \frac{T}{T^\beta}\sum_{t=1}^{T} \mathbb{E}\left[\ell(\mathcal{A}(x_t|L_{B_{<t}}^{f^\star}), y_t)\mathbb{1}\{B_t = 1\}\right].
\end{aligned}$$

To see the last equality, note that the prediction $\mathcal{A}(x_t|L_{B_{<t}}^{f^\star})$ only depends on bitstring $(B_1, \ldots, B_{t-1})$ and the internal randomness of A, both of which are independent of B_t. Thus, we have

$$\begin{aligned}
\mathbb{E}\left[\ell(\mathcal{A}(x_t|L_{B_{<t}}^{f^\star}), y_t)\,\mathbb{1}\{B_t = 1\}\right] &= \mathbb{E}\left[\ell(\mathcal{A}(x_t|L_{B_{<t}}^{f^\star}), y_t)\right]\,\mathbb{E}\left[\mathbb{1}\{B_t = 1\}\right] \\
&= \mathbb{E}\left[\ell(\mathcal{A}(x_t|L_{B_{<t}}^{f^\star}), y_t)\right]\mathbb{P}[B_t = 1]
\end{aligned}$$

as needed. Continuing onwards,

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \ell(E_{B,\phi_B^{f^\star}}(x_t), y_t)\right] &= \frac{T}{T^\beta}\mathbb{E}\left[\sum_{t=1}^{T} \ell(\mathcal{A}(x_t|L_{B_{<t}}^{f^\star}), y_t)\mathbb{1}\{B_t = 1\}\right] \\
&= \frac{T}{T^\beta}\mathbb{E}\left[\sum_{t:B_t=1} \ell(\mathcal{A}(x_t|L_{B_{<t}}^{f^\star}), y_t)\right] \\
&\leq \frac{cT}{T^\beta}\mathbb{E}\left[\sum_{t:B_t=1} \ell(\mathcal{A}(x_t|L_{B_{<t}}^{f^\star}), f^\star(x_t))\right] + \frac{T}{T^\beta}\mathbb{E}\left[\sum_{t:B_t=1} \ell(f^\star(x_t), y_t)\right] \\
&= \frac{cT}{T^\beta}\mathbb{E}\left[\sum_{t:B_t=1} \ell(\mathcal{A}(x_t|L_{B_{<t}}^{f^\star}), f^\star(x_t))\right] + \sum_{t=1}^{T} \ell(f^\star(x_t), y_t) \\
&= \frac{cT}{T^\beta}\mathbb{E}\left[\sum_{t:B_t=1} \ell(\mathcal{A}(x_t|L_{B_{<t}}^{f^\star}), f^\star(x_t))\right] + \inf_{f\in\mathcal{F}}\sum_{t=1}^{T} \ell(f(x_t), y_t)
\end{aligned}$$

The inequality follows from the fact that ℓ is a c-subadditive and the last equality follows from the definition of $f^\star$. We now need to bound $\frac{cT}{T^\beta}\mathbb{E}\left[\sum_{t:B_t=1} \ell(\mathcal{A}(x_t|L_{B_{<t}}^{f^\star}), f^\star(x_t))\right]$. Using the fact that $\mathcal{A}^\star$ is a realizable online learner and gets updated on a stream of

instances labeled by $f^\star$ only on rounds where $B_t = 1$, we get

$$\frac{cT}{T^\beta}\mathbb{E}\left[\sum_{t:B_t=1}\ell(\mathcal{A}(x_t|L^{f^\star}_{B_{<t}}), f^\star(x_t))\right] = \frac{cT}{T^\beta}\mathbb{E}\left[\mathbb{E}\left[\sum_{t:B_t=1}\ell(\mathcal{A}(x_t|L^{f^\star}_{B_{<t}}), f^\star(x_t))\Big|B\right]\right]$$
$$\leq \frac{cT}{T^\beta}\mathbb{E}\left[R(|B|, |\mathrm{im}(\mathcal{F})|)\right].$$

Putting things together, we find that,

$$\mathbb{E}\left[\sum_{t=1}^{T}\ell(\mathcal{Q}(x_t), y_t)\right] \leq \mathbb{E}\left[\sum_{t=1}^{T}\ell(E_{B,\phi^{f^\star}_B}(x_t), y_t)\right] + M\mathbb{E}\left[\sqrt{2T\ln(|\mathcal{E}_B|)}\right]$$
$$\leq \inf_{f\in\mathcal{F}}\sum_{t=1}^{T}\ell(f(x_t), y_t) + \frac{cT}{T^\beta}\mathbb{E}\left[R(|B|, |\mathrm{im}(\mathcal{F})|)\right] + M\mathbb{E}\left[\sqrt{2T\ln(|\mathcal{E}_B|)}\right]$$
$$\leq \inf_{f\in\mathcal{F}}\sum_{t=1}^{T}\ell(f(x_t), y_t) + \frac{cT}{T^\beta}\mathbb{E}\left[R(|B|, |\mathrm{im}(\mathcal{F})|)\right] + M\mathbb{E}\left[\sqrt{2T|B|\ln(|\mathrm{im}(\mathcal{F})|)}\right],$$

where the last inequality follows from the fact that that $|\mathcal{E}_B| \leq (|\mathrm{im}(\mathcal{F})|)^{|B|}$. By Jensen's inequality, we further get that, $\mathbb{E}\left[\sqrt{2T|B|\ln(|\mathrm{im}(\mathcal{F})|)}\right] \leq \sqrt{2T^{\beta+1}\ln(|\mathrm{im}(\mathcal{F})|)}$, which implies that

$$\mathbb{E}\left[\sum_{t=1}^{T}\ell(\mathcal{Q}(x_t), y_t)\right] \leq \inf_{f\in\mathcal{F}}\sum_{t=1}^{T}\ell(f(x_t), y_t) + \frac{cT}{T^\beta}\mathbb{E}\left[R(|B|, |\mathrm{im}(\mathcal{F})|)\right] + M\sqrt{2T^{\beta+1}\ln(|\mathrm{im}(\mathcal{F})|)}.$$

Next, by Lemma 14, there exists a concave sublinear function $\overline{R}(|B|, |\mathrm{im}(\mathcal{F})|)$ that upper-bounds $R(|B|, |\mathrm{im}(\mathcal{F})|)$. By Jensen's inequality, we obtain $\mathbb{E}[\overline{R}(|B|, |\mathrm{im}(\mathcal{F})|)] \leq \overline{R}(T^\beta, |\mathrm{im}(\mathcal{F})|)$, which yields

$$\mathbb{E}\left[\sum_{t=1}^{T}\ell(\mathcal{Q}(x_t), y_t)\right] \leq \inf_{f\in\mathcal{F}}\sum_{t=1}^{T}\ell(f(x_t), y_t) + \frac{cT}{T^\beta}\,\overline{R}(T^\beta, |\mathrm{im}(\mathcal{F})|) + M\sqrt{2T^{\beta+1}\ln(|\mathrm{im}(\mathcal{F})|)}.$$

This completes the proof as we have shown that $\mathcal{Q}$ is an agnostic online learner for $\mathcal{F}$ with respect to ℓ with the stated regret bound. ■

4.2 Online Multilabel Classification

Let $\mathcal{Y} = \{-1, 1\}^K$. We provide analogs of Theorem 6 and 8 in the online setting. We begin by characterizing the learnability of the Hamming loss and then move to give a characterization of learnability for all losses satisfying the identity of indiscernibles. Similar to the batch setting, we can show that the MCLdim of $\mathcal{F}$ characterizes online multilabel learnability (see Appendix F), but here, we give a characterization that better exploits the multilabel structure of the problem.

4.2.1 Characterizing Online Learnability for the Hamming Loss

Theorem 15 characterizes the online learnability of a multilabel function class $\mathcal{F}$ with respect to ℓ_H.

Theorem 15. *A function class $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ is online learnable with respect to the Hamming loss if and only if each restriction $\mathcal{F}_k \subseteq \mathcal{Y}_k^{\mathcal{X}}$ is online learnable with respect to the 0-1 loss.*

The proof of Theorem 15 is similar to that of Theorem 6, so we defer the full proof to Appendix E and only provide a sketch here. The proof of sufficiency direction is based on a reduction: given oracle access to online learners $\{\mathcal{A}_k\}_{k=1}^K$ for $\{\mathcal{F}_k\}_{k=1}^K$ with respect to $\ell_{0\text{-}1}$, we construct an online learner $\mathcal{A}$ for $\mathcal{F}$ with respect to ℓ_H. In fact, similar to the batch setting, the online multilabel learning algorithm $\mathcal{A}$ is simple: in each round $t \in [T]$, receive x_t, query the predictions $\mathcal{A}_1(x_t), ..., \mathcal{A}_K(x_t)$, and finally predict the concatenation $\hat{y}_t = (\mathcal{A}_1(x_t), ..., \mathcal{A}_K(x_t))$. Once the true label $y_t = (y_t^1, ..., y_t^K)$ is revealed, update each online learner $\mathcal{A}_k$ by passing (x_t, y_t^k) for $k \in [K]$. Using some algebra, one can show that this prediction rule achieves sublinear regret for $\mathcal{F}$.

For the necessity direction, given oracle access to an online learner $\mathcal{A}$ for $\mathcal{F}$ with respect to ℓ_H, we construct an online learner $\mathcal{B}$ for $\mathcal{F}_1$ with respect to $\ell_{0\text{-}1}$. A similar reduction can be used to construct online learners for each restriction $\mathcal{F}_k$. Similar to the batch setting, the online learning algorithm $\mathcal{B}$ is simple: in each round $t \in [T]$, receive x_t, query $\hat{y}_t = \mathcal{A}(x_t)$ and predict $\hat{y}_t^1 = \mathcal{A}_1(x_t)$. Once the true label y_t^1 is revealed, update $\mathcal{A}$ by passing (x_t, y_t) where $y_t = (y_t^1, \sigma_t^2, ..., \sigma_t^K)$ and $\{\sigma_t^i\}_{i=2}^K$ is an i.i.d sequence of Rademacher random variables. A straightforward analysis shows that such a prediction rule achieves sublinear regret for $\mathcal{F}_1$.

4.2.2 Characterizing Online Learnability for General Losses

Using Theorem 13 and Theorem 15, we now characterize the learnability of arbitrary multilabel loss functions ℓ as long as they satisfy the identity of indiscernibles. The key idea is that since there are only finite number of possible inputs to ℓ, for any ℓ satisfying the identity of indiscernibles, there must exist universal constants a and b such that $a\ell_H(y_1, y_2) \leq \ell(y_1, y_2) \leq b\ell_H(y_1, y_2)$. Then, we can characterize the learnability of ℓ by relating it to the learnability of ℓ_H. In fact, we prove a slightly more general result, showing an equivalence between the learnability of any two arbitrary losses satisfying the identity of indiscernibles.

Lemma 16. *Let ℓ and ℓ' be any two loss functions satisfying the identity of indiscernibles. A function class $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ is online learnable with respect to ℓ if and only if $\mathcal{F}$ is online learnable with respect to ℓ'.*

The proof of Lemma 16 is similar to that of Lemma 7 with the main difference being the use of Theorem 13 instead of Lemma 4. Since Lemma 16 is our first application of Theorem 13, we provide the full proof here.

Proof Since ℓ and ℓ' are arbitrary, it suffices to prove only one direction. To that end, suppose $\mathcal{F}$ is online learnable with respect to ℓ. We now show that $\mathcal{F}$ is online learnable with respect to ℓ' as well.

Let a and b be the universal constants such that for all $y_1, y_2 \in \mathcal{Y}$, $a\ell(y_1, y_2) \leq \ell'(y_1, y_2) \leq b\ell(y_1, y_2)$. Let $c = \frac{\max_{r\neq t} \ell'(r,t)}{\min_{r\neq t} \ell'(r,t)}$. Since $|\text{im}(\mathcal{F})| = 2^K < \infty$ and ℓ' is a c-subadditive, by Theorem 13, it suffices to give a realizable online learner for $\mathcal{F}$ with respect to ℓ'. Since $\mathcal{F}$ is online learnable with respect to ℓ, there exists an algorithm $\mathcal{A}$ such that for any sequence $(x_1, y_1), ..., (x_T, y_T)$, we have

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell(\mathcal{A}(x_t), y_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^{T} \ell(f(x_t), y_t)\right] \leq R(T, 2^K)$$

where $R(T, 2^K)$ is a sublinear function of T. In the realizable setting, we are guaranteed that for any sequence $(x_1, y_1), ..., (x_T, y_T)$ that the online learner may observe, there exists a $f \in \mathcal{F}$ s.t $f(x_t) = y_t$ for all $t \in [T]$. Since ℓ satisfies the identity of indiscernibles, we have that for any realizable sequence $(x_1, y_1), ..., (x_T, y_T)$, $\inf_{f\in\mathcal{F}} \sum_{t=1}^{T} \ell(f(x_t), y_t) = 0$. Thus, we have that $\mathbb{E}\left[\sum_{t=1}^{T} \ell(\mathcal{A}(x_t), y_t)\right] \leq R(T, 2^K)$. Noting that $\ell(\mathcal{A}(x_t), y_t) \geq \frac{\ell'(\mathcal{A}(x_t), y_t)}{b}$ implies that $\mathbb{E}\left[\sum_{t=1}^{T} \ell'(\mathcal{A}(x_t), y_t)\right] \leq bR(T, 2^K)$, showing that $\mathcal{A}$ is also a realizable online learner for $\mathcal{F}$ with respect to ℓ'. For any $\beta \in (0, 1)$, the construction in Theorem 13 can then be used to convert $\mathcal{A}$ into an agnostic online learner for $\mathcal{F}$ with respect to ℓ' with expected regret bound

$$\frac{cbT}{T^\beta}\overline{R}(T^\beta, 2^K) + M\sqrt{4KT^{1+\beta}}$$

where M is such that $\ell \leq M$ and $\overline{R}(T^\beta, 2^K)$ is any concave sublinear upperbound of $R(T^\beta, 2^K)$. This completes our proof.

■

As an immediate consequence of Lemma 16 and Theorem 15, we get the following theorem characterizing the online learnability of general multilabel losses.

Theorem 17. *Let ℓ be any multilabel loss function that satisfies the identity of indiscernibles. A function class $\mathcal{F} \subseteq \mathcal{Y}^\mathcal{X}$ is online learnable with respect to ℓ if and only if each restriction $\mathcal{F}_k \subseteq \mathcal{Y}_k^\mathcal{X}$ is online learnable with respect to the 0-1 loss.*

Remark. Since the Littlestone dimension characterizes online learnability for binary classification under the 0-1 loss (Ben-David et al., 2009), Theorem 17 also implies that finiteness of $\text{Ldim}(\mathcal{F}_k)$ for all $k \in [K]$ is a necessary and sufficient condition for online multilabel learnability.

Moreover, if $\text{Ldim}(\mathcal{F}_k) < \infty$ for all $k \in [K]$, then we have $\text{MCLdim}(\mathcal{F}) < \infty$. This follows from the fact that $\text{MCLdim}(\mathcal{F}) \leq \sum_{k=1}^{K} \text{Ldim}(\mathcal{F}_k)$. To see this, note that $\text{MCLdim}(\mathcal{F})$ is the lowerbound on the number of mistakes of any deterministic multiclass learner in the realizable setting (Daniely et al., 2011, Theorem 17). On the other hand, one can construct a deterministic realizable learner for $\mathcal{F}$ using K different Standard Optimal Algorithms (SOA) for binary function classes $\mathcal{F}_k$'s. Namely, define an algorithm $\mathcal{A}$ such that $\mathcal{A}(x) := (\text{SOA}(\mathcal{F}_1)(x), \ldots, \text{SOA}(\mathcal{F}_K)(x)) \in \{-1, 1\}^K$. Since each $\text{SOA}(\mathcal{F}_k)$ makes at most $\text{Ldim}(\mathcal{F}_k)$ number of mistakes, $\mathcal{A}$ makes no more than $\sum_{k=1}^{K} \text{Ldim}(\mathcal{F}_k)$ mistakes. We can use this fact to give an improved version of Theorem 13 for classes $\mathcal{F}$

with $\text{MCLdim}(\mathcal{F}) < \infty$. In particular, when $\mathcal{Y} = \{-1, 1\}^K$, any $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ that is learnable in the realizable setting with respect to ℓ is also learnable in the agnostic setting with regret $O\Big(B\sqrt{T\text{MCLdim}(\mathcal{F})\ln(T)}\Big) \leq O\Big(B\sqrt{T\sum_{k=1}^{K}\text{Ldim}(\mathcal{F}_k)\ln(T)}\Big)$. Here, B is the maximum value ℓ can take. The improved regret bound can be found in the sufficiency proof of Theorem 34 in Appendix F.

4.3 Bandit Online Multilabel Classification

We extend the results in the previous subsection to the online setting where the learner only observes *bandit* feedback in each round. Theorem 18 gives a characterization of bandit online learnability of a function class $\mathcal{F}$ in terms of the online learnability of each restriction.

Theorem 18. *Let ℓ be any loss function that satisfies the identity of indiscernibles. A function class $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ is bandit online learnable with respect to ℓ if and only if each restriction $\mathcal{F}_k \subseteq \mathcal{Y}_k^{\mathcal{X}}$ is online learnable with respect to the 0-1 loss.*

Similar to the full-feedback setting, Theorem 18 also gives that the finiteness of $\text{Ldim}(\mathcal{F}_k)$ for all $k \in [K]$ is a necessary and sufficiency condition for bandit online multilabel learnability. The proof of Theorem 18 uses the realizable-to-agnostic conversion for *bandit* feedback setting when the label space $\mathcal{Y}$ is finite. The following Theorem makes this argument precise.

Theorem 19. *Let $\mathcal{Y}$ be a finite label space, $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ a multioutput function class, and $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ be any c-subadditive loss function such that $\ell(\cdot, \cdot) \leq M$. If $\mathcal{A}$ is a realizable online learner for $\mathcal{F}$ with respect to ℓ under full-feedback with sub-linear expected regret $R(T, |\mathcal{Y}|)$, then for every $\beta \in (0, 1)$, there exists an online learner for $\mathcal{F}$ with respect to ℓ with expected regret*

$$\frac{cT}{T^{\beta}}\overline{R}(T^{\beta}, |\mathcal{Y}|) + eM\sqrt{2T^{1+\beta}|\mathcal{Y}|\ln(|\mathcal{Y}|)},$$

under bandit feedback*, where $\overline{R}(T, |\mathcal{Y}|)$ is any concave, sublinear upperbound on $R(T, |\mathcal{Y}|)$.*

The proofs for Theorem 18 and Theorem 19 are provided in Appendix G.

4.4 Online Multioutput Regression

In this section, we characterize the online learnability of multioutput function classes. Similar to the batch setting, we consider, without loss of generality, the case when $\mathcal{Y} = [0, 1]^K \subset \mathbb{R}^K$ for $K \in \mathbb{N}$. In addition, we consider the same set of decomposable and non-decomposable loss functions as in the batch setting. Namely, our decomposable loss functions satisfy Assumptions 1 and 2, and our non-decomposable loss functions are ℓ_p norms. Informally, our main result asserts that a multioutput function class $\mathcal{F} \subset \mathcal{Y}^{\mathcal{X}}$ is online learnable if and only if each restriction $\mathcal{F}_k$ is online learnable.

4.4.1 Characterizing Learnability for Decomposable Losses

In this subsection, we characterize the online learnability of multioutput function classes with respect to decomposable losses satisfying Assumption 1. Our main theorem is presented below.

Theorem 20. *Let ℓ be a decomposable loss function satisfying* Assumption 1. *A multioutput function class $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ is online learnable with respect to ℓ if and only if each $\mathcal{F}_k \subseteq \mathcal{Y}_k^{\mathcal{X}}$ is online learnable with respect to $\psi_k \circ d_1$.*

Proof As usual, we prove Theorem 20 in two parts: first sufficiency and then necessity. The sufficiency proof is similar to that of the proof for Hamming loss in Theorems 6 and 15. The necessity direction is more involved, but the main idea is to combine the augmentation technique used in the proof of Theorem 9 with the algorithmic conversion technique developed in the proof of Theorem 13.

Part 1: Sufficiency. We first prove that online learnability of each restriction $\mathcal{F}_k$ with respect to $\psi_k \circ d_1$ is sufficient for online learnability of $\mathcal{F}$ with respect to ℓ. Since $\ell(f(x), y) = \sum_{k=1}^{K} \psi_k \circ d_1(f_k(x), y_k)$ is decomposable, we can use the exact same strategy as in Section 4.2.1 to convert online learners $\mathcal{A}_1$, ..., $\mathcal{A}_K$ for $\mathcal{F}_1, ..., \mathcal{F}_K$ with respect to $\psi_k \circ d_1$ to an online learner $\mathcal{A}$ for $\mathcal{F}$ with respect to ℓ. More specifically, in each round $t \in [T]$, receive x_t, query the predictions $\mathcal{A}_1(x_t), ..., \mathcal{A}_K(x_t)$, and finally predict the concatenation $\hat{y}_t = (\mathcal{A}_1(x_t), ..., \mathcal{A}_K(x_t))$. Once the true label $y_t = (y_t^1, ..., y_t^K)$ is revealed, update each online learner $\mathcal{A}_k$ by passing (x_t, y_t^k) for $k \in [K]$. Using the exact same proof as in Section 4.2.1, it follows that the expected regret of $\mathcal{A}$ is $\sum_{k=1}^{K} R_k(T)$ where $R_k(T)$ is the regret of online algorithm $\mathcal{A}_k$. Since K is finite, the regret of $\mathcal{A}$ is sublinear in T when evaluated using ℓ.

Part 2: Necessity. Similar to the batch setting, we prove the necessity direction of Theorem 20 constructively. That is, given oracle access to an online learner $\mathcal{A}$ for $\mathcal{F}$ with respect to ℓ, we construct an online learner $\mathcal{Q}$ for $\mathcal{F}_1$ with respect to $\psi_1 \circ d_1$. By symmetry, a similar reduction can be used to construct online learners for each restriction $\mathcal{F}_k$. As mentioned before, we assume an oblivious adversary, and therefore the stream of points to be observed by the online learner, denoted $(x_1, y_1), ..., (x_T, y_T) \in (\mathcal{X} \times [0, 1])^T$, is fixed beforehand. Let $f_1^\star = \arg\min_{f_1 \in \mathcal{F}_1} \sum_{t=1}^{T} \psi_k \circ d_1(f_1(x_t), y_t)$ denote the optimal function in hindsight and $f^\star \in \mathcal{F}$ its completion.

Since we are trying to construct an online learner for $\mathcal{F}_1$, the targets $y_1, ..., y_T$ are *scalar-valued*. However, $\mathcal{A}$ is an online learner for $\mathcal{F}$ and therefore can only processes *vector-valued* targets. Thus, we need to figure out how to augment the scalar-valued targets $y_1, ..., y_T$ in a way that allows us to use $\mathcal{A}$ to construct an online learner $\mathcal{Q}$ for $\mathcal{F}_1$. Following a similar strategy as in the proof of Theorem 13, we can construct a set of Experts that simulate online games with $\mathcal{A}$ by augmenting, in all possible ways, the scalar-valued targets of a *sub-sample* of the stream into vector-valued targets using vectors in $\mathcal{Y}_{2:k}^{\alpha}$, the discretized label space for components 2 through K. In particular, our high-level strategy is to:

1. Randomly *sub-sample* points from the stream
2. Construct a set of Experts, each of which:
   (a) Uses an independent copy of $\mathcal{A}$ to make predictions $\hat{y}_t = \mathcal{A}_1(x_t)$
   (b) Augments the scalar-valued targets of each labeled instance in the *sub-sampled* stream to vector-valued targets using vectors in the discretized image space $\text{im}(\mathcal{F}_{2:K}^{\alpha})$
   (c) Simulates an online game with its independent copy of $\mathcal{A}$ over only the augmented *sub-sampled* stream with vector-valued targets

3. Run REWA using the set of experts in Step 2 and the $\psi_1 \circ d_1$ loss function over the *original* stream of points.

We now formalize this idea. For any bitstring $b \in \{0,1\}^T$, let $\phi : \{t : b_t = 1\} \to \text{im}(\mathcal{F}^{\alpha}_{2:K})$ denote a function mapping time points where $b_t = 1$ to vectors in the discretized image space $\text{im}(\mathcal{F}^{\alpha}_{2:K})$. Let $\Phi_b \subseteq (\text{im}(\mathcal{F}^{\alpha}_{2:K}))^{\{t:b_t=1\}}$ denote all such functions ϕ. For every $f \in \mathcal{F}$, let $\phi^f_b \in \Phi_b$ be the mapping such that for all $t \in \{t : b_t = 1\}$, $\phi^f_b(t) = f^{\alpha}_{2:K}(x_t)$. Let $|b| = |\{t : b_t = 1\}|$. For every $b \in \{0,1\}^T$ and $\phi \in \Phi_b$, define an Expert $E_{b,\phi}$. Expert $E_{b,\phi}$, formally presented in Algorithm 5, uses $\mathcal{A}$ to make predictions in each round. However, $E_{b,\phi}$ only updates $\mathcal{A}$ on those rounds where $b_t = 1$, using ϕ to augment the scalar-valued labeled instance (x_t, y_t) to the vector-valued labeled instance $(x_t, (y_t, \phi(t)))$. For every $b \in \{0,1\}^T$, let $\mathcal{E}_b = \bigcup_{\phi \in \Phi_b} \{E_{b,\phi}\}$ denote the set of all Experts parameterized by functions $\phi \in \Phi_b$. If b is the all zeros bitstring, then $\mathcal{E}_b$ is empty. Therefore, we actually define $\mathcal{E}_b = \{E_0\} \cup \bigcup_{\phi \in \Phi_b} \{E_{b,\phi}\}$, where E_0 is the expert that never updates $\mathcal{A}$ and plays $\mathcal{A}_1(x_t)$ for all $t \in [T]$. Note that $1 \leq |\mathcal{E}_b| \leq (\frac{2}{\alpha})^{K|b|}$.

Algorithm 5 Expert(b, ϕ)

Input: Independent copy of Online Learner $\mathcal{A}$ for ℓ

for $t = 1, ..., T$ **do**
  Receive example x_t
  Predict $\tilde{y}_t = \mathcal{A}_1(x_t)$
  Receive y_t
  if $b_t = 1$ **then**
    Update $\mathcal{A}$ by passing $(x_t, (y_t, \phi(t)))$
end

With this notation in hand, we are now ready to present Algorithm 6, our main online learner $\mathcal{Q}$ for $\mathcal{F}_1$.

Algorithm 6 Online learner $\mathcal{Q}$ for $\mathcal{F}_1$ with respect to $\psi_1 \circ d_1$

Input: Parameters $0 < \beta < 1$ and $0 < \alpha < 1$

Let $B \in \{0,1\}^T$ such that $B_t \overset{\text{iid}}{\sim} \text{Bernoulli}(\frac{T^\beta}{T})$
Construct the set of experts $\mathcal{E}_B = \{E_0\} \cup \bigcup_{\phi \in \Phi_B} \{E_{B,\phi}\}$ according to Algorithm 5
Run REWA $\mathcal{P}$ using $\mathcal{E}_B$ and the loss function $\psi_1 \circ d_1$ over the stream $(x_1, y_1), ..., (x_T, y_T)$

Our goal now is to show that $\mathcal{Q}$ enjoys sublinear expected regret. There are three main sources of randomness: the randomness involved in sampling B, the internal randomness of each independent copy of the online learner $\mathcal{A}$, and the internal randomness of REWA. Let B, A and P denote the random variable associated with these sources of randomness respectively. By construction, B, A, and P are independent.

Using Theorem 21.11 in Shalev-Shwartz and Ben-David (2014) and the fact that A, P, and B are independent, REWA guarantees

$$\mathbb{E}\left[\sum_{t=1}^{T} \psi_1 \circ d_1(\mathcal{P}(x_t), y_t)\right] \leq \mathbb{E}\left[\inf_{E \in \mathcal{E}_B} \sum_{t=1}^{T} \psi_1 \circ d_1(E(x_t), y_t)\right] + \mathbb{E}\left[\sqrt{2T \ln(|\mathcal{E}_B|)}\right].$$

Thus,

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\psi_1\circ d_1(\mathcal{Q}(x_t),y_t)\right] &= \mathbb{E}\left[\sum_{t=1}^{T}\psi_1\circ d_1(\mathcal{P}(x_t),y_t)\right]\\ &\le \mathbb{E}\left[\inf_{E\in\mathcal{E}_B}\sum_{t=1}^{T}\psi_1\circ d_1(E(x_t),y_t)\right]+\mathbb{E}\left[\sqrt{2T\ln(|\mathcal{E}_B|)}\right]\\ &\le \mathbb{E}\left[\sum_{t=1}^{T}\psi_1\circ d_1(E_{B,\phi_B^{f^\star}}(x_t),y_t)\right]+\mathbb{E}\left[\sqrt{2T\ln(|\mathcal{E}_B|)}\right].\end{aligned}$$

In the last step, we used the fact that for all $b\in\{0,1\}^T$ and $f\in\mathcal{F}$, $E_{b,\phi_b^f}\in\mathcal{E}_b$.

It now suffices to upperbound $\mathbb{E}\left[\sum_{t=1}^{T}\psi\circ d_1(E_{B,\phi_B^{f^\star}}(x_t),y_t)\right]$. We use the same notation used to prove Theorem 13, but for the sake of completeness, we restate it here. Given an online learner $\mathcal{A}$ for ℓ, an instance $x\in\mathcal{X}$, and an ordered sequence of labeled examples $L\in(\mathcal{X}\times[0,1]^K)^*$, let $\mathcal{A}(x|L)$ be the random variable denoting the prediction of $\mathcal{A}$ on the instance x after running and updating on L. For any $b\in\{0,1\}^T$, $f^\alpha_{2:K}\in\mathcal{F}^\alpha_{2:K}$, and $t\in[T]$, let $L^f_{b_{<t}}=\{(x_i,(y_i,f^\alpha_{2:K}(x_i))):i<t \text{ and } b_i=i\}$ denote the *subsequence* of the sequence of labeled instances $\{(x_i,(y_i,f^\alpha_{2:K}(x_i)))\}_{i=1}^{t-1}$ where $b_i=1$. Using this notation, we can write

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\psi_1\circ d_1(E_{B,\phi_B^{f^\star}}(x_t),y_t)\right] &= \mathbb{E}\left[\sum_{t=1}^{T}\psi_1\circ d_1(\mathcal{A}_1(x_t|L^{f^\star}_{B_{<t}}),y_t)\right]\\ &= \mathbb{E}\left[\sum_{t=1}^{T}\psi_1\circ d_1(\mathcal{A}_1(x_t|L^{f^\star}_{B_{<t}}),y_t)\frac{\mathbb{P}\left[B_t=1\right]}{\mathbb{P}\left[B_t=1\right]}\right]\\ &= \frac{T}{T^\beta}\sum_{t=1}^{T}\mathbb{E}\left[\psi_1\circ d_1(\mathcal{A}_1(x_t|L^{f^\star}_{B_{<t}}),y_t)\mathbb{P}\left[B_t=1\right]\right]\\ &= \frac{T}{T^\beta}\sum_{t=1}^{T}\mathbb{E}\left[\psi_1\circ d_1(\mathcal{A}_1(x_t|L^{f^\star}_{B_{<t}}),y_t)\mathbb{1}\{B_t=1\}\right].\end{aligned}$$

To see the last equality, note that the prediction $\mathcal{A}(x_t|L^{f^\star}_{B_{<t}})$ (and therefore $\mathcal{A}_1(x_t|L^{f^\star}_{B_{<t}})$) only depends on bitstring $(B_1,\ldots,B_{t-1})$ and the internal randomness of A, both of which are independent of B_t. Thus, we have

$$\begin{aligned}\mathbb{E}\left[\psi_1\circ d_1(\mathcal{A}_1(x_t|L^{f^\star}_{B_{<t}}),y_t)\mathbb{1}\{B_t=1\}\right] &= \mathbb{E}\left[\psi_1\circ d_1(\mathcal{A}_1(x_t|L^{f^\star}_{B_{<t}}),y_t)\right]\mathbb{E}\left[\mathbb{1}\{B_t=1\}\right]\\ &= \mathbb{E}\left[\psi_1\circ d_1(\mathcal{A}_1(x_t|L^{f^\star}_{B_{<t}}),y_t)\right]\mathbb{P}[B_t=1]\end{aligned}$$

as needed. Continuing onwards,

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\psi_1\circ d_1(E_{B,\phi_B^{f^\star}}(x_t),y_t)\right] &= \frac{T}{T^\beta}\mathbb{E}\left[\sum_{t=1}^{T}\psi_1\circ d_1(\mathcal{A}_1(x_t|L^{f^\star}_{B_{<t}}),y_t)\mathbb{1}\{B_t=1\}\right]\\
&= \frac{T}{T^\beta}\mathbb{E}\left[\sum_{t:B_t=1}\psi_1\circ d_1(\mathcal{A}_1(x_t|L^{f^\star}_{B_{<t}}),y_t)\right]\\
&\le \frac{T}{T^\beta}\mathbb{E}\left[\sum_{t:B_t=1}\ell(\mathcal{A}(x_t|L^{f^\star}_{B_{<t}}),(y_t,f^{\star,\alpha}_{2:K}(x_t)))\right]\\
&= \frac{T}{T^\beta}\mathbb{E}\left[\mathbb{E}\left[\sum_{t:B_t=1}\ell(\mathcal{A}(x_t|L^{f^\star}_{B_{<t}}),(y_t,f^{\star,\alpha}_{2:K}(x_t)))\middle|B\right]\right]\\
&\le \frac{T}{T^\beta}\mathbb{E}\left[\sum_{t:B_t=1}\ell(f^\star(x_t),(y_t,f^{\star,\alpha}_{2:K}(x_t)))+R_{\mathcal{A}}(|B|)\right]\\
&= \frac{T}{T^\beta}\mathbb{E}\left[\sum_{t:B_t=1}\ell(f^\star(x_t),(y_t,f^{\star,\alpha}_{2:K}(x_t)))\right]+\frac{T}{T^\beta}\mathbb{E}\left[R_{\mathcal{A}}(|B|)\right]
\end{aligned}$$

The first inequality follows from the definition of ℓ. The second inequality follows from the fact that $\mathcal{A}$ is an online learner for ℓ with regret bound $R_{\mathcal{A}}(T)$ and is updated on the stream labeled by $f^{\star,\alpha}_{2:K}$ only when $B_t = 1$. Now, we can upperbound the first term as follows:

$$\begin{aligned}
\frac{T}{T^\beta}\mathbb{E}\left[\sum_{t:B_t=1}\ell(f^\star(x_t),(y_t,f^{\star,\alpha}_{2:K}(x_t)))\right] &= \frac{T}{T^\beta}\mathbb{E}\left[\sum_{t:B_t=1}\left(\psi_1\circ d_1(f^\star_1(x_t),y_t)+\sum_{k=2}^{K}\psi_k\circ d_1(f^\star_k(x_t),f^{\star,\alpha}_k(x_t))\right)\right]\\
&\le \frac{T}{T^\beta}\mathbb{E}\left[\sum_{t:B_t=1}\psi_1\circ d_1(f^\star_1(x_t),y_t)+\sum_{t:B_t=1}KL\alpha\right]\\
&\le \frac{T}{T^\beta}\mathbb{E}\left[\sum_{t=1}^{T}\psi_1\circ d_1(f^\star_1(x_t),y_t)\mathbb{1}\{B_t=1\}\right]+\frac{T}{T^\beta}\mathbb{E}\left[|B|KL\alpha\right]\\
&= \frac{T}{T^\beta}\sum_{t=1}^{T}\psi_1\circ d_1(f^\star_1(x_t),y_t)\frac{T^\beta}{T}+\frac{T}{T^\beta}T^\beta KL\alpha\\
&= \sum_{t=1}^{T}\psi_1\circ d_1(f^\star_1(x_t),y_t)+KTL\alpha.
\end{aligned}$$

The first inequality follows from the fact that ψ_k is L-Lipschitz and $d_1(f^\star_k(x_t), f^{\star,\alpha}_k(x_t)) \le \alpha$. Putting things together, we find that,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\psi_1 \circ d_1(\mathcal{Q}(x_t), y_t)\right] \\
&\quad\leq \mathbb{E}\left[\sum_{t=1}^{T}\psi_1 \circ d_1(E_{B,\phi_B^{f^\star}}(x_t), y_t)\right] + \mathbb{E}\left[\sqrt{2T\ln(|\mathcal{E}_B|)}\right] \\
&\quad\leq \sum_{t=1}^{T}\psi_1 \circ d_1(f_1^\star(x_t), y_t) + KTL\alpha + \frac{T}{T^\beta}\mathbb{E}\left[R_{\mathcal{A}}(|B|)\right] + \mathbb{E}\left[\sqrt{2T\ln(|\mathcal{E}_B|)}\right] \\
&\quad\leq \inf_{f_1\in\mathcal{F}_1}\sum_{t=1}^{T}\psi_1 \circ d_1(f_1(x_t), y_t) + KTL\alpha + \frac{T}{T^\beta}\mathbb{E}\left[R_{\mathcal{A}}(|B|)\right] + \mathbb{E}\left[\sqrt{2TK|B|\ln(\frac{2}{\alpha})}\right].
\end{aligned}$$

where the last inequality follows from the fact that that $|\mathcal{E}_B| \leq (\frac{2}{\alpha})^{K|B|}$ and the definition of $f^\star$. By Jensen's inequality, we further get that, $\mathbb{E}\left[\sqrt{2TK|B|\ln(\frac{2}{\alpha})}\right] \leq \sqrt{2T^{\beta+1}K\ln(\frac{2}{\alpha})}$, which implies that

$$\mathbb{E}\left[\sum_{t=1}^{T}\psi_1 \circ d_1(\mathcal{Q}(x_t), y_t)\right] \leq \inf_{f_1\in\mathcal{F}_1}\sum_{t=1}^{T}\psi_1 \circ d_1(f_1(x_t), y_t) + KTL\alpha + \frac{T}{T^\beta}\mathbb{E}\left[R_{\mathcal{A}}(|B|)\right] + \sqrt{2T^{\beta+1}K\ln(\frac{2}{\alpha})}.$$

Next, by Lemma 14, there exists a concave sublinear function $\overline{R}_{\mathcal{A}}(|B|)$ of $|B|$ that upperbounds $R_{\mathcal{A}}(|B|)$. By Jensen's inequality, we obtain $\mathbb{E}[\overline{R}_{\mathcal{A}}(|B|)] \leq \overline{R}_{\mathcal{A}}(T^\beta)$, which yields

$$\mathbb{E}\left[\sum_{t=1}^{T}\psi_1 \circ d_1(\mathcal{Q}(x_t), y_t)\right] \leq \inf_{f_1\in\mathcal{F}_1}\sum_{t=1}^{T}\psi_1 \circ d_1(f_1(x_t), y_t) + KTL\alpha + \frac{T}{T^\beta}\overline{R}_{\mathcal{A}}(T^\beta) + \sqrt{2T^{\beta+1}K\ln(\frac{2}{\alpha})}.$$

Picking $\alpha = \frac{1}{KTL}$ and $\beta \in (0,1)$, gives that $\mathcal{Q}$ enjoys sublinear expected regret:

$$\mathbb{E}\left[\sum_{t=1}^{T}\psi_1 \circ d_1(\mathcal{Q}(x_t), y_t)\right] - \inf_{f_1\in\mathcal{F}_1}\sum_{t=1}^{T}\psi_1 \circ d_1(f_1(x_t), y_t) \leq 1 + \frac{T}{T^\beta}\overline{R}_{\mathcal{A}}(T^\beta) + \sqrt{4T^{\beta+1}K\ln(KTL)}.$$

This completes the proof as we have shown that $\mathcal{Q}$ is an online learner for $\mathcal{F}_1$ with respect to $\psi_1 \circ d_1$.

■

4.4.2 A More General Characterization of Learnability for Decomposable Losses

Theorem 20 characterizes the learnability of multioutput function classes $\mathcal{F}$ with respect to decomposable loss functions ℓ in terms of the learnability of $\mathcal{F}_k$ with respect to $\psi_k \circ d_1$.

Similar to the batch setting, we can remove ψ_k, and characterize the learnability of $\mathcal{F}$ with respect to ℓ in terms of the learnability of $\mathcal{F}_k$'s with respect to d_1. However, to do so, we need to an place additional assumption on the decomposable loss function ℓ. Theorem 21 below summarizes the main result of this section.

Theorem 21. *Let ℓ be any decomposable loss function satisfying Assumptions 1 and 2. A multioutput function class $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ is online learnable with respect to ℓ if and only if each $\mathcal{F}_k \subseteq \mathcal{Y}_k^{\mathcal{X}}$ is online learnable with respect to d_1.*

The main tool needed to prove Theorem 21 is Lemma 22, which relates the online learnability of a scalar-output function class $\mathcal{H} \subset [0,1]^{\mathcal{X}}$ with respect to $\psi \circ d_1$ to its online learnability with respect to d_1, where ψ is any monotonic, Lipschitz function such that $\psi(0) = 0$. The proof of Lemma 22 can be found in Appendix H.

Lemma 22. *Let $\psi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ be any monotonic and Lipschitz function such that $\psi(0) = 0$. A scalar-valued function class $\mathcal{G} \subset [0,1]^{\mathcal{X}}$ is online learnable with respect to $\psi \circ d_1$, if and only if $\mathcal{G}$ is online learnable with respect to d_1.*

Note that for every $1 \leq p < \infty$ and $x \geq 0$, $\psi(x) = x^p$ is a monotonic increasing, Lipschitz function. Therefore, Lemma 22 shows that online learnability with respect to d_p is equivalent to online learnability with respect to d_1. Combining Assumption 2, Theorem 20, and Lemma 22 immediately gives Theorem 21. Since the Sequential Fat Shattering dimension characterizes online learnability with respect to the absolute loss, Theorem 21 further implies that for any decomposable loss satisfying Assumptions 1 and 2, the finiteness of $\mathrm{fat}_\gamma^{\mathrm{seq}}(\mathcal{F}_k)$ for all $k \in [K]$ and $\gamma > 0$ is a sufficient and necessary condition for online multioutput learnability.

4.4.3 Characterizing Learnability of Non-Decomposable Losses

In this section, we characterize the online learnability of multioutput function classes $\mathcal{F}$ for a natural family of non-decomposable losses, ℓ_p norms for $1 \leq p \leq \infty$. We prove an analogous theorem to Theorem 12, by relating the online learnability of $\mathcal{F}$ with respect to ℓ_p to the online learnability of each $\mathcal{F}_k$ with respect to d_1. We note that the proof of only uses the fact that ℓ_p norms are equivalent (up to a K dependent constant) to the ℓ_1 norm. Since any two norms in a finite dimensional space are equivalent, Theorem 23 actually holds true for *any* norm in $\mathbb{R}^K$. But we only consider ℓ_p norms here due to their practical importance.

Theorem 23. *Let $1 \leq p \leq \infty$. A function class $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ is online learnable with respect to ℓ_p if and only if each $\mathcal{F}_k \subseteq \mathcal{Y}_k^{\mathcal{X}}$ is online learnable with respect to d_1.*

By Theorem 20, $\mathcal{F}$ is online learnable with respect to ℓ_1 if and only if each restriction $\mathcal{F}_k$ is online learnable with respect to d_1. Thus, to prove Theorem 23 it suffices to show that $\mathcal{F}$ is online learnable with respect to ℓ_p if and only if $\mathcal{F}$ is online learnable with respect to ℓ_1 for $p > 1$. At a high-level, the proof Theorem 23 follows a similar route as the proof of Theorem 12: convert an agnostic learner for $\mathcal{F}$ into a realizable learner for $\mathcal{F}^\alpha$ and then use realizable-to-agnostic conversion for $\mathcal{F}^\alpha$.

Proof Fix $p > 1$. By the argument above, it suffices to show that $\mathcal{F}$ is online learnable with respect to ℓ_p if and only if $\mathcal{F}$ is online learnable with respect to ℓ_1. We begin by

proving sufficiency - if $\mathcal{F}$ is online learnable with respect to ℓ_1 then $\mathcal{F}$ is online learnable with respect to ℓ_p.

Let $\mathcal{A}$ be an online learner for $\mathcal{F}$ with respect to ℓ_1. Our goal is to construct an online learner $\mathcal{Q}$ for $\mathcal{F}$ with respect to ℓ_p. We assume an oblivious adversary, and therefore the stream of points to be observed by the online learner $\mathcal{Q}$, denoted $(x_1, y_1), ..., (x_T, y_T) \in (\mathcal{X} \times [0,1]^K)^T$, is fixed beforehand. Let $f^\star = \arg\min_{f \in \mathcal{F}} \sum_{t=1}^T \ell_p(f(x_t), y_t)$ also denote the optimal function in hindsight with respect to the ℓ_p loss.

Our strategy follows three steps. First, we show that $\mathcal{A}$ is a realizable online learner for $\mathcal{F}^\alpha$ with respect to ℓ_p. Then, since $|\text{im}(\mathcal{F}^\alpha)| \leq (\frac{2}{\alpha})^K < \infty$ is finite and ℓ_p is a 1-subadditive (by triangle inequality), Theorem 13 allows to convert the realizable online learner $\mathcal{A}$ for $\mathcal{F}^\alpha$ with respect to ℓ_p into an agnostic online learner $\mathcal{Q}$ for $\mathcal{F}^\alpha$ with respect to ℓ_p. Finally, for an appropriately selected discretization parameter α, we show that $\mathcal{Q}$ is also an agnostic online for $\mathcal{F}$ with respect to ℓ_p, which completes the proof. To that end, let $(x_1, f^{\star,\alpha}(x_1)), ..., (x_T, f^{\star,\alpha}(x_T)$ denote a (realizable) sequence of instances labeled by some function $f^{\star,\alpha} \in \mathcal{F}^\alpha$. Since $||\cdot||_q \leq ||\cdot||_p$ for $q > p$, we have that

$$\mathbb{E}\left[\sum_{t=1}^T \ell_p(\mathcal{A}(x_t), f^{\star,\alpha}(x_t))\right] \leq \mathbb{E}\left[\sum_{t=1}^T \ell_1(\mathcal{A}(x_t), f^{\star,\alpha}(x_t))\right]$$

Since $\mathcal{A}$ is an online learner for $\mathcal{F}$ with respect to ℓ_1, we get,

$$\mathbb{E}\left[\sum_{t=1}^T \ell_1(\mathcal{A}(x_t), f^{\star,\alpha}(x_t))\right] \leq \inf_{f \in \mathcal{F}} \sum_{t=1}^T \ell_1(f(x_t), f^{\star,\alpha}(x_t)) + R_{\mathcal{A}}(T) \leq \alpha K T + R_{\mathcal{A}}(T)$$

where $R_{\mathcal{A}}(T)$ is the regret of online learner $\mathcal{A}$. Combining things together, we have that

$$\mathbb{E}\left[\sum_{t=1}^T \ell_p(\mathcal{A}(x_t), f^{\star,\alpha}(x_t))\right] \leq \alpha K T + R_{\mathcal{A}}(T),$$

showing that $\mathcal{A}$ is a realizable online learner for $\mathcal{F}^\alpha$ with respect to ℓ_p for a small enough α. Now, since $|\text{im}(\mathcal{F}^\alpha)| \leq (\frac{2}{\alpha})^K < \infty$ is a finite, ℓ_p is a 1- subadditive loss function, and $\mathcal{A}$ is a realizable online learner, for any $\beta \in (0,1)$, Theorem 13 gives an agnostic online learner $\mathcal{Q}$ for $\mathcal{F}^\alpha$ with respect to ℓ_p with the following regret guarantee over the original stream $(x_1, y_1), ..., (x_T, y_T)$:

$$\mathbb{E}\left[\sum_{t=1}^T \ell_p(\mathcal{Q}(x_t), y_t)\right] \leq \inf_{f^\alpha \in \mathcal{F}^\alpha} \sum_{t=1}^T \ell_p(f^\alpha(x_t), y_t) + \frac{T}{T^\beta}(\alpha K T^\beta + \overline{R}_{\mathcal{A}}(T^\beta)) + K\sqrt{2T^{\beta+1} K \ln(\frac{2}{\alpha})},$$

where $\alpha K T^\beta + \overline{R}_{\mathcal{A}}(T^\beta)$ is any concave sublinear upperbound of $\alpha K T^\beta + R_{\mathcal{A}}(T^\beta)$. We also use the fact that the function $T \mapsto \alpha K T^\beta$ is a concave sublinear function of T and the sum of two concave functions is itself a concave function. Noting that $\ell_p(f^\alpha(x_t), y_t) \leq \ell_p(f(x_t), y_t) + \ell_p(f^\alpha(x_t), f(x_t)) \leq \ell_p(f(x_t), y_t) + \alpha K$, gives

$$\mathbb{E}\left[\sum_{t=1}^T \ell_p(\mathcal{Q}(x_t), y_t)\right] \leq \inf_{f \in \mathcal{F}} \sum_{t=1}^T \ell_p(f(x_t), y_t) + \alpha K T + \frac{T}{T^\beta}(\alpha K T^\beta + \overline{R}_{\mathcal{A}}(T^\beta)) + K\sqrt{2T^{\beta+1} K \ln(\frac{2}{\alpha})}.$$

Combining like terms together, we have

$$\mathbb{E}\left[\sum_{t=1}^{T}\ell_p(\mathcal{Q}(x_t),y_t)\right] \leq \inf_{f\in\mathcal{F}}\sum_{t=1}^{T}\ell_p(f(x_t),y_t) + 2\alpha KT + \frac{T}{T^\beta}\overline{R}_{\mathcal{A}}(T^\beta) + K\sqrt{2T^{\beta+1}K\ln(\frac{2}{\alpha})}.$$

Finally, picking $\alpha = \frac{1}{2KT}$ gives that

$$\mathbb{E}\left[\sum_{t=1}^{T}\ell_p(\mathcal{Q}(x_t),y_t)\right] - \inf_{f\in\mathcal{F}}\sum_{t=1}^{T}\ell_p(f(x_t),y_t) \leq 1 + \frac{T}{T^\beta}\overline{R}_{\mathcal{A}}(T^\beta) + K\sqrt{2T^{\beta+1}K\ln(4KT)}.$$

Since $\overline{R}_{\mathcal{A}}(T^\beta)$ is sublinear in T^β and $\beta \in (0,1)$, $\mathcal{Q}$ enjoys sublinear expected regret. Thus, we have shown that $\mathcal{Q}$ is also an agnostic online learner for $\mathcal{F}$ with respect to ℓ_p.

The reverse direction follows identically and uses the fact that for any $p > 1$, $||\cdot||_p \leq ||\cdot||_1 \leq K||\cdot||_p$. In particular, using the exact same argument, we can show that if $\mathcal{A}$ is an online learner for $\mathcal{F}$ with respect to ℓ_p, then $\mathcal{A}$ is also a realizable online learner for $\mathcal{F}^\alpha$ with respect to ℓ_1 with expected regret bound:

$$\mathbb{E}\left[\sum_{t=1}^{T}\ell_1(\mathcal{A}(x_t),f^{\star,\alpha}(x_t))\right] \leq K\mathbb{E}\left[\sum_{t=1}^{T}\ell_p(\mathcal{A}(x_t),f^{\star,\alpha}(x_t))\right] \leq \alpha K^2T + KR_{\mathcal{A}}(T).$$

Using $\mathcal{A}$ as the realizable learner in Theorem 13, for any $\beta \in (0,1)$, picking $\alpha = \frac{1}{(K+K^2)T}$ gives a regret bound:

$$\mathbb{E}\left[\sum_{t=1}^{T}\ell_1(\mathcal{Q}(x_t),y_t)\right] - \inf_{f\in\mathcal{F}}\sum_{t=1}^{T}\ell_1(f(x_t),y_t) \leq 1 + \frac{KT}{T^\beta}\overline{R}_{\mathcal{A}}(T^\beta) + K\sqrt{2T^{\beta+1}K\ln(4K^2T)},$$

where $\overline{R}_{\mathcal{A}}(T^\beta)$ is any concave sublinear upperbound of $R_{\mathcal{A}}(T^\beta)$. Since $\beta \in (0,1)$, $\mathcal{Q}$ is an online learner for $\mathcal{F}$ with respect to ℓ_1 as needed. This completes the proof of Theorem 23. ■

Remark. As with decomposable losses, Theorem 23 also implies that for any ℓ_p norm loss, the finiteness of $\mathrm{fat}^{\mathrm{seq}}_\gamma(\mathcal{F}_k)$, for all $k \in [K]$ and fixed $\gamma > 0$, is a sufficient and necessary condition for online multioutput learnability.

5 Discussion

In this work, we give a characterization of multioutput learnability in four settings: batch classification, online classification, batch regression, and online regression. In all four settings, we show that a multioutput function class is learnable if and only if each restriction is learnable. All of our bounds in this paper scale with K, preventing our current techniques from extending to the case when K is infinite. Accordingly, we pose it as an open question to characterize multioutput learnability when K is infinite (e.g. function-space valued regression). Furthermore, we also leave it open to find combinatorial dimensions that provide tight quantitative characterizations of batch and online multioutput learnability.

Acknowledgements

We acknowledge the support of NSF via grants IIS-2007055 (AT) and DMS-2413089 (AT, US). VR acknowledges the support of the NSF Graduate Research Fellowship.

References

Noga Alon, Shai Ben-David, Nicolò Cesa-Bianchi, and David Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *J. ACM*, 44(4), 1997. ISSN 0004-5411. doi: 10.1145/263867.263927.

Mauricio A Alvarez, Lorenzo Rosasco, Neil D Lawrence, et al. Kernels for vector-valued functions: A review. *Foundations and Trends® in Machine Learning*, 4(3):195–266, 2012.

Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999. doi: 10.1017/CBO9780511624216.

Idan Attias and Steve Hanneke. Adversarially robust pac learnability of real-valued functions, 2023.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.*, 3:463–482, 2003.

Peter L. Bartlett, Philip M. Long, and Robert C. Williamson. Fat-shattering and the learnability of real-valued functions. *Journal of Computer and System Sciences*, 52(3): 434–452, 1996. ISSN 0022-0000. doi: https://doi.org/10.1006/jcss.1996.0033.

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

Shai Ben-David, Dávid Pál, and Shai Shalev-Shwartz. Agnostic online learning. In *COLT*, volume 3, page 1, 2009.

Hanen Borchani, Gherardo Varando, Concha Bielza, and Pedro Larranaga. A survey on multi-output regression. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 5(5):216–233, 2015.

Philip J Brown and James V Zidek. Adaptive multivariate ridge regression. *The Annals of Statistics*, 8(1):64–74, 1980.

Nataly Brukhim, Daniel Carmon, Irit Dinur, Shay Moran, and Amir Yehudayoff. A characterization of multiclass learnability, 2022. URL `https://arxiv.org/abs/2203.01550`.

Róbert Busa-Fekete, Heejin Choi, Krzysztof Dembczynski, Claudio Gentile, Henry Reeve, and Balazs Szorenyi. Regret bounds for multilabel classification in sparse label regimes. *Advances in Neural Information Processing Systems*, 35:5404–5416, 2022.

T. Ceccherini-Silberstein, M. Salvatori, and E. Sava-Huss. *Groups, Graphs and Random Walks*. London Mathematical Society Lecture Note Series. Cambridge University Press, 2017. doi: 10.1017/9781316576571.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

Lena Chekina, Dan Gutfreund, Aryeh Kontorovich, Lior Rokach, and Bracha Shapira. Exploiting label dependencies for improved sample complexity. *Machine learning*, 91:1–42, 2013.

Corinna Cortes, Vitaly Kuznetsov, Mehryar Mohri, and Scott Yang. Structured prediction theory based on factor graph complexity. *Advances in Neural Information Processing Systems*, 29, 2016.

Amit Daniely and Tom Helbertal. The price of bandit information in multiclass online classification. In *Conference on Learning Theory*, pages 93–104. PMLR, 2013.

Amit Daniely and Shai Shalev-Shwartz. Optimal learners for multiclass problems. In Maria Florina Balcan, Vitaly Feldman, and Csaba Szepesvári, editors, *Proceedings of The 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pages 287–316, Barcelona, Spain, 13–15 Jun 2014. PMLR.

Amit Daniely, Sivan Sabato, Shai Ben-David, and Shai Shalev-Shwartz. Multiclass learnability and the erm principle. In Sham M. Kakade and Ulrike von Luxburg, editors, *Proceedings of the 24th Annual Conference on Learning Theory*, volume 19 of *Proceedings of Machine Learning Research*, pages 207–232, Budapest, Hungary, 09–11 Jun 2011. PMLR.

Krzysztof Dembczyński, Willem Waegeman, Weiwei Cheng, and Eyke Hüllermeier. Regret analysis for performance metrics in multi-label classification: the case of hamming and subset zero-one loss. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010, Proceedings, Part I 21*, pages 280–295. Springer, 2010.

Krzysztof Dembczyński, Willem Waegeman, Weiwei Cheng, and Eyke Hüllermeier. On label dependence and loss minimization in multi-label classification. *Machine Learning*, 88:5–45, 2012.

Dylan J. Foster and Alexander Rakhlin. ℓ_∞ vector contraction for rademacher complexity, 2019.

Wei Gao and Zhi-Hua Zhou. On the consistency of multi-label learning. In *Proceedings of the 24th annual conference on learning theory*, pages 341–358. JMLR Workshop and Conference Proceedings, 2011.

Claudio Gentile and Francesco Orabona. On multilabel classification and ranking with partial feedback. *Advances in Neural Information Processing Systems*, 25, 2012.

Giorgio Gnecco and Marcello Sanguineti. Estimates of the approximation error using rademacher complexity: learning vector-valued functions. *Journal of Inequalities and Applications*, 2008:1–16, 2008.

Max Hopkins, Daniel M. Kane, Shachar Lovett, and Gaurav Mahajan. Realizable learning is all you need. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 3015–3069. PMLR, 02–05 Jul 2022.

Himanshu Jain, Yashoteja Prabhu, and Manik Varma. Extreme multi-label loss functions for recommendation, tagging, ranking & other missing label applications. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 935–944, 2016.

Ashish Kapoor, Raajay Viswanathan, and Prateek Jain. Multilabel classification using bayesian compressed sensing. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper/2012/file/e1d5be1c7f2f456670de3d53c7b54f4a-Paper.pdf.

Michael J Kearns, Robert E Schapire, and Linda M Sellie. Toward efficient agnostic learning. *Machine Learning*, 17:115–141, 1994.

Oluwasanmi O Koyejo, Nagarajan Natarajan, Pradeep K Ravikumar, and Inderjit S Dhillon. Consistent multilabel classification. *Advances in Neural Information Processing Systems*, 28, 2015a.

Oluwasanmi O Koyejo, Nagarajan Natarajan, Pradeep K Ravikumar, and Inderjit S Dhillon. Consistent multilabel classification. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015b. URL https://proceedings.neurips.cc/paper/2015/file/85f007f8c50dd25f5a45fca73cad64bd-Paper.pdf.

Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2:285–318, 1987.

Weiwei Liu and Ivor Tsang. On the optimality of classifier chain for multi-label classification. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper/2015/file/854d9fca60b4bd07f9bb215d59ef5561-Paper.pdf.

Andreas Maurer. A vector-contraction inequality for rademacher complexities. In *Algorithmic Learning Theory: 27th International Conference, ALT 2016, Bari, Italy, October 19-21, 2016, Proceedings 27*, pages 3–17. Springer, 2016.

Charles A Micchelli and Massimiliano Pontil. On learning vector-valued functions. *Neural computation*, 17(1):177–204, 2005.

Omar Montasser, Steve Hanneke, and Nathan Srebro. Vc classes are adversarially robustly learnable, but only improperly. In *Conference on Learning Theory*, pages 2512–2530. PMLR, 2019.

Jinseok Nam, Eneldo Loza Mencía, Hyunwoo J Kim, and Johannes Fürnkranz. Maximizing subset accuracy with recurrent neural networks in multi-label classification. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper/2017/file/2eb5657d37f474e4c4cf01e4882b8962-Paper.pdf`.

B. K. Natarajan. On learning sets and functions. *Mach. Learn.*, 4(1):67–97, oct 1989. ISSN 0885-6125. doi: 10.1023/A:1022605311895. URL `https://doi.org/10.1023/A:1022605311895`.

Junhyung Park and Krikamol Muandet. Towards empirical process theory for vector-valued functions: Metric entropy of smooth function classes. In *Proceedings of The 34th International Conference on Algorithmic Learning Theory*, 2023.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning via sequential complexities. *J. Mach. Learn. Res.*, 16(1):155–186, 2015a.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Sequential complexities and uniform martingale laws of large numbers. *Probability theory and related fields*, 161: 111–153, 2015b.

C Radhakrishna Rao. The theory of least squares when the parameters are stochastic and its application to the analysis of growth curves. *Biometrika*, 52(3/4):447–458, 1965.

Henry Reeve and Ata Kaban. Optimistic bounds for multi-output learning. In *International Conference on Machine Learning*, pages 8030–8040. PMLR, 2020.

Mark Rudelson and Roman Vershynin. Combinatorics of random processes and sections of convex bodies. *Annals of Mathematics*, pages 603–648, 2006.

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

Leslie G. Valiant. A theory of the learnable. In *Symposium on the Theory of Computing*, 1984.

V. Vapnik and A. Chervonenkis. *Theory of Pattern Recognition [in Russian]*. 1974.

Vladimir Naumovich Vapnik and Aleksei Yakovlevich Chervonenkis. On uniform convergence of the frequencies of events to their probabilities. *Teoriya Veroyatnostei i ee Primeneniya*, 16(2):264–279, 1971.

Shuo Xu, Xin An, Xiaodong Qiao, Lijun Zhu, and Lin Li. Multi-output least-squares support vector regression machines. *Pattern Recognition Letters*, 34(9):1078–1084, 2013.

Liang Yang, Xi-Zhu Wu, Yuan Jiang, and Zhi-Hua Zhou. Multi-label learning with deep forest. In *ECAI*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pages 1634–1641. IOS Press, 2020.

Niloofar Yousefi, Yunwen Lei, Marius Kloft, Mansooreh Mollaghasemi, and Georgios C Anagnostopoulos. Local rademacher complexity-based learning guarantees for multi-task learning. *Journal of Machine Learning Research*, 19(38):1–47, 2018.

Appendix A. Complexity Measures

A.1 Complexity Measures for Batch Learning

In binary classification, the Vapnik-Chervonenkis (VC) dimension of a function class characterizes its learnability.

Definition 24 (Vapnik-Chervonenkis Dimension). *A set $S = \{x_1, \ldots, x_d\}$ is shattered by a binary function class $\mathcal{H} \subseteq \{-1, 1\}^d$ if for every $\sigma \in \{-1, 1\}^d$, there exists a hypothesis $h_\sigma \in \mathcal{H}$ such that for all $i \in [d]$, we have $h_\sigma(x_i) = \sigma_i$. The VC dimension of $\mathcal{H}$, denoted VC($\mathcal{H}$), is the size of the largest shattered set $S \subseteq \mathcal{X}$. If the size of the shattered set can be arbitrarily large, we say that VC($\mathcal{H}$) $= \infty$.*

The learnability of a multiclass function class is characterized by its Natarajan dimension.

Definition 25 (Natarajan Dimension). *A set $S = \{x_1, \ldots, x_d\}$ is shattered by a multiclass function class $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ if there exist two witness functions $f, g : S \to \mathcal{Y}$ such that $f(x_i) \neq g(x_i)$ for all $i \in [d]$, and for every $\sigma \in \{-1, 1\}^d$, there exists a function $h_\sigma \in \mathcal{H}$ such that for all $i \in [d]$, we have*

$$h_\sigma(x_i) = \begin{cases} f(x_i) & \text{if } \sigma_i = 1 \\ g(x_i) & \text{if } \sigma_i = -1 \end{cases} .$$

The Natarajan dimension of $\mathcal{H}$, denoted Ndim($\mathcal{H}$), is the size of the largest shattered set $S \subseteq \mathcal{X}$. If the size of the shattered set can be arbitrarily large, we say that Ndim($\mathcal{H}$) $= \infty$.

For real-valued regression problems, the learnability is characterized in terms of the fat-shattering dimension of a function class.

Definition 26 (Fat-Shattering Dimension). *A real-valued function class $\mathcal{G} \subseteq [0, 1]^{\mathcal{X}}$ shatters points $S = \{x_1, x_2, \ldots, x_d\}$ at scale $1 > \gamma > 0$, if there exists witness functions $r : S \to [0, 1]$ such that, for every $\sigma \in \{\pm 1\}^d$, there exists $g_\sigma \in \mathcal{G}$ such that $\forall i \in [d]$, $\sigma_i(g_\sigma(x_i) - r(x_i)) \geq \gamma$. The fat-shattering dimension of $\mathcal{G}$ at scale γ, denoted $fat_\gamma(\mathcal{G})$, is the size of the largest set that can be γ-shattered by $\mathcal{G}$. If the size of the shattered set can be arbitrarily large, then we say that $fat_\gamma(\mathcal{G}) = \infty$.*

We also define a general notion of complexity called Rademacher complexity that provides a sufficient and necessary condition of uniform convergence. Since uniform convergence implies learnability, we use Rademacher complexity to argue sufficient conditions for learnability.

Definition 27 (Empirical Rademacher Complexity). *Let $\mathcal{D}$ be a distribution over $\mathcal{X} \times \mathcal{Y}$. For a bounded loss function ℓ, define the loss class to be $\ell \circ \mathcal{F} = \{(x, y) \mapsto \ell(f(x), y)\}$. If $S = \{(x_1, y_1), ..., (x_n, y_n)\}$ be a set of i.i.d samples drawn from $\mathcal{D}$, then the empirical Rademacher complexity of $\ell \circ \mathcal{F}$ is defined as*

$$\mathfrak{R}_n(\ell \circ \mathcal{F}) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=1}^{n} \sigma_i \ell(f(x_i), y_i) \right) \right],$$

where $\sigma \in \{\pm 1\}^n$ is a sequence of n i.i.d. Rademacher random variables.

A.2 Complexity Measures for Online Learning

For the complexity measures below, it is useful to define a $\mathcal{Z}$-valued binary tree (Rakhlin et al., 2015b). A binary tree $\mathcal{T}$ of depth d is $\mathcal{Z}$-valued if each of its internal nodes are labelled by elements of $\mathcal{Z}$. Such a tree can be identified by a sequence $(\mathcal{T}_1, ..., \mathcal{T}_d)$ labelling functions $\mathcal{T}_i : \{\pm 1\}^{i-1} \to \mathcal{Z}$ which provide labels for each internal node. A path of length d is given by a sequence $\sigma = (\sigma_1, ..., \sigma_d) \in \{\pm 1\}^d$. Then, $\mathcal{T}_i(\sigma_1, ..., \sigma_{i-1})$ gives the label of node following the path $(\sigma_1, ..., \sigma_{i-1})$ starting from the root, going "right" if $\sigma_j = +1$ and "left" if $\sigma_j = -1$. Note that, $\mathcal{T}_1 \in \mathcal{Z}$ is the label for the root node. For brevity, we slightly abuse notation by letting $\mathcal{T}_i(\sigma_1, ..., \sigma_{i-1}) = \mathcal{T}_i(\sigma_{<i})$, but it is understood that $\mathcal{T}_i$ only depends on the prefix $(\sigma_1, ..., \sigma_{i-1})$. We are now ready to formally define complexity measures in the online setting.

When $\mathcal{Y} = \{-1, +1\}$ is binary, the Littlestone Dimension (Littlestone (1987)) tightly characterizes the online learnability of a function class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ with respect to the 0-1 loss.

Definition 28 (Littlestone Dimension)**.** *Let $\mathcal{T}$ denote a complete binary tree of depth d whose internal nodes are labeled by elements $\mathcal{X}$ and two edges from parent to child nodes are labeled by -1 and $+1$. The tree is shattered by a binary hypothesis class $\mathcal{G} \subseteq \{-1, +1\}^{\mathcal{X}}$ if for every $\sigma \in \{-1, +1\}^d$, there exists a hypothesis $g_\sigma \in \mathcal{G}$ such that the root to leaf path $(x_1, \ldots, x_d)$ obtained by taking left when $\sigma_t = -1$ and right when $\sigma_t = +1$ satisfies $g_\sigma(x_t) = \sigma_t$ for all $1 \le t \le d$. The Littlestone Dimension of $\mathcal{G}$, denoted Ldim($\mathcal{G}$), is the maximal depth of the complete binary tree shattered by $\mathcal{G}$. If $\mathcal{G}$ can shatter a tree of arbitrary depth, we say that Ldim($\mathcal{G}$) $= \infty$.*

For finite label spaces $\mathcal{Y}$, the Multiclass Littlestone Dimension (Daniely et al., 2011) tightly characterizes the online learnability of a function class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ with respect to the 0-1 loss.

Definition 29 (Multiclass Littlestone Dimension)**.** *Let $\mathcal{T}$ denote a $\mathcal{X}$-valued binary tree of depth d whose edges are labelled by elements from $\mathcal{Y}$, such that the edges from a single parent to its child-nodes are each labeled with a different label. The tree $\mathcal{T}$ is shattered by a function class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ if, for every path $\sigma \in \{\pm 1\}^d$, there is a function $h_\sigma \in \mathcal{H}$ such that $h_\sigma(\mathcal{T}_i(\sigma_{<i})) = y(\sigma_i)$, where $y(\sigma_i)$ is the label of the edge between nodes $(\mathcal{T}_i(\sigma_{<i}), \mathcal{T}_{i+1}(\sigma_{<i+1}))$. The Multiclass Littlestone Dimension (MCLdim) of $\mathcal{H}$, denoted MCLdim($\mathcal{H}$), is the maximal depth of a complete binary tree that is shattered by $\mathcal{H}$. If MCLdim $= \infty$, then there exists shattered trees of arbitrarily large depth.*

When $\mathcal{Y}$ is a bounded subset of $\mathbb{R}$, the sequential fat-shattering dimension (Rakhlin et al., 2015a) at scale γ characterizes the learnability of $\mathcal{H} \subseteq [0, 1]^{\mathcal{X}}$ with respect to to the absolute loss d_1.

Definition 30 (Sequential Fat-Shattering Dimension)**.** *Let $\mathcal{T}$ denote a $\mathcal{X}$-valued binary tree of depth d. The tree $\mathcal{T}$ is γ-shattered by a function class $\mathcal{H} \subseteq [0, 1]^{\mathcal{X}}$ if there exists an $\mathbb{R}$-valued binary tree $\mathcal{R}$ of depth d such that for all $\sigma \in \{\pm 1\}^d$, there exists $h_\sigma \in \mathcal{H}$ such that for all $t \in [d]$,*

$$\sigma_t(h_\sigma(\mathcal{T}_t(\sigma_{<t})) - \mathcal{R}_t(\sigma_{<t})) \ge \gamma$$

The tree $\mathcal{R}$ is called the witness to shattering. The sequential fat shattering dimension of $\mathcal{H}$ at scale γ, denoted $\mathrm{fat}^{seq}_{\gamma}(\mathcal{H})$, is the maximal depth of a complete binary tree that is γ-shattered by $\mathcal{H}$. If there exists γ-shattered trees of arbitrarily large depth, then $\mathrm{fat}^{seq}_{\gamma}(\mathcal{H}) = \infty$.

Beyond both finite and bounded label spaces, the sequential Rademacher complexity (Rakhlin et al., 2015a) provides a useful tool for giving sufficient conditions for learnability.

Definition 31 (Sequential Rademacher Complexity)**.** *Let $\sigma = \{\sigma_i\}_{i=1}^{T}$ be a sequence of independent Rademacher random variables. Let $\mathcal{T}$ be a $\mathcal{Z}$-valued binary tree of depth d. The sequential Rademacher complexity of a function class $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{Z}}$ on $\mathcal{T}$ is defined as*

$$\mathfrak{R}^{seq}_{T}(\mathcal{H}; \mathcal{T}) = \mathbb{E}_{\sigma \sim \{\pm 1\}^n} \left[\sup_{h \in \mathcal{H}} \frac{1}{T} \sum_{t=1}^{T} \sigma_t h(\mathcal{T}_t(\sigma_{<t})) \right].$$

Then, the worst-case sequential Rademacher complexity is defined as $\mathfrak{R}^{seq}_{T}(\mathcal{H}) = \sup_{\mathcal{T}} \mathfrak{R}^{seq}_{T}(\mathcal{H}; \mathcal{T})$.

Appendix B. Natarajan Dimension Characterizes Batch Multilabel Learnability

A multilabel classification problem where labels (i.e. bitstrings) in $\mathcal{Y}$ are of length K can also be viewed as multiclass classification on the target space with 2^K labels. Given this observation, the Natarajan dimension of the function class $\mathcal{F}$ continues to characterize the multilabel learnability with respect to any loss function ℓ satisfying the identity of indiscernibles.

Theorem 32 (Ben-David et al. (1995))**.** *Let ℓ be any loss function satisfying the identity of indiscernibles. A function class $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ is agnostic learnable with respect to ℓ in the batch setting if and only if $Ndim(\mathcal{F}) < \infty$.*

The proof in Ben-David et al. (1995) involves arguments based on growth function. Here, we provide proof that uses realizable and agnostic learnability due to Hopkins et al. (2022).

Proof (of sufficiency) We first show that the finiteness of $\mathrm{Ndim}(\mathcal{F})$ is sufficient for learnability. Suppose $\mathrm{Ndim}(\mathcal{F}) < \infty$. Then, we know that $\mathcal{F}$ is agnostic learnable with respect to 0-1 loss (Ben-David et al., 1995). Since the target space $\mathcal{Y}$ as well as the range space of $\mathcal{F}$ is finite, for every loss ℓ satisfying the identity of indiscernibles, there exists an $a > 0$ such that $a\,\ell(h(x), y) \leq \mathbb{1}\{h(x) \neq y\}$ for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and function $h \in \mathcal{Y}^{\mathcal{X}}$. Let $\mathcal{D}$ be a realizable distribution to $\mathcal{F}$ with respect to ℓ. Since $\ell(y_1, y_2) = 0$ if and only if $\mathbb{1}\{y_1 \neq y_2\} = 0$, the distribution $\mathcal{D}$ is also realizable with respect to 0-1 loss. Since $\mathcal{F}$ is learnable with respect to 0-1 loss, there exists a learning algorithm $\mathcal{A}$ with the following property: for any $\epsilon, \delta > 0$, for a sufficiently large $S \sim \mathcal{D}^n$, the algorithm outputs a predictor $h = \mathcal{A}(S)$ such that, with probability $1 - \delta$ over $S \sim \mathcal{D}^n$, we have $\mathbb{E}_{\mathcal{D}}[\mathbb{1}\{h(x) \neq y\}] \leq a\,\epsilon$. Using the inequality stated above pointwise, the predictor h also satisfies $\mathbb{E}_{\mathcal{D}}[\ell(h(x), y)] \leq \epsilon$. Therefore, $\mathcal{A}$ is also a realizable algorithm with respect to ℓ. Since ℓ satisfies the identity of indiscernible, Lemma 4 guarantees the existence of agnostic PAC learner $\mathcal{B}$ for $\mathcal{F}$ with

respect to ℓ. ■

Proof (of necessity) Suppose $\mathcal{F}$ is learnable with respect to ℓ. Since the target space is finite, there must exist a constant $b > 0$ such that $\mathbb{1}\{h(x) \neq y\} \leq b\,\ell(h(x), y)$ for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and any function $h \in \mathcal{Y}^{\mathcal{X}}$. Let $\mathcal{D}$ be a realizable distribution with respect to ℓ. Due to the 0 alignment property, $\mathcal{D}$ is also realizable with respect to 0-1 loss. Since $\mathcal{F}$ is learnable with respect to ℓ loss, there exists a learning algorithm $\mathcal{A}$ with the following property: for any $\epsilon, \delta > 0$, for a sufficiently large $S \sim \mathcal{D}^n$, the algorithm outputs a predictor $h = \mathcal{A}(S)$ such that, with probability $1 - \delta$ over $S \sim \mathcal{D}^n$, we have $\mathbb{E}_{\mathcal{D}}[\ell(h(x), y)] \leq b\,\epsilon$. In particular, using the inequality stated above pointwise, we obtain $\mathbb{E}_{\mathcal{D}}[\mathbb{1}\{h(x) \neq y\}] \leq \epsilon$. Therefore, $\mathcal{F}$ is learnable with respect to 0-1 loss in the realizable setting. As the finiteness of the Natarajan dimension is necessary for the learnability of $\mathcal{F}$ under the 0-1 loss (Natarajan, 1989), we must have $\mathrm{Ndim}(\mathcal{F}) < \infty$. ■

Appendix C. Proofs for Batch Multioutput Regression

C.1 Proof of Sufficiency in Theorem 9

Proof We first prove that the agnostic learnability of each $\mathcal{F}_k$ is sufficient for the agnostic learnability of $\mathcal{F}$. As in the classification setting, the proof here is based on a reduction. That is, given oracle access to agnostic learners $\mathcal{A}_k$ for each $\mathcal{F}_k$ with respect to $\psi_k \circ d_1$ loss, we construct an agnostic learner $\mathcal{A}$ for $\mathcal{F}$ with respect to loss ℓ.

Denote $\mathcal{D}_k$ to be the marginal distribution of $\mathcal{D}$ restricted to $\mathcal{X} \times \mathcal{Y}_k$. Let us use $m_k(\epsilon, \delta)$ to denote the sample complexity of $\mathcal{A}_k$. Then, for all $k \in [K]$, the marginal samples $S_k = \{(x_i, y_i^k)\}_{i=1}^n$ with scalar-valued targets are iid samples form $\mathcal{D}_k$. For each $k \in [K]$, define $g_k = \mathcal{A}_k(S_k)$ to be the predictor returned by algorithm $\mathcal{A}_k$ when trained on S_k. Since $\mathcal{A}_k$ is an agnostic learner for $\mathcal{F}_k$, we have that for sample size $n \geq \max_k m_k(\frac{\epsilon}{K}, \frac{\delta}{K})$, with probability at least $1 - \delta/K$ over samples $S_k \sim \mathcal{D}_k^n$,

$$\mathbb{E}_{\mathcal{D}_k}[\psi_k \circ d_1(g_k(x), y^k)] \leq \inf_{f_k \in \mathcal{F}_k} \mathbb{E}_{\mathcal{D}_k}[\psi_k \circ d_1(f_k(x), y^k)] + \frac{\epsilon}{K}.$$

Summing these risk bounds over all k coordinates and union bounding over the success probabilities, we get that with probability at least $1 - \delta$ over samples $S \sim \mathcal{D}^n$,

$$\sum_{k=1}^K \mathbb{E}_{\mathcal{D}_k}[\psi_k \circ d_1(g_k(x), y^k)] \leq \sum_{k=1}^K \inf_{f_k \in \mathcal{F}_k} \mathbb{E}_{\mathcal{D}_k}[\psi_k \circ d_1(f_k(x), y^k)] + \epsilon.$$

Using the fact that the sum of infimums over individual coordinates is at most the overall infimum of sums followed by the linearity of expectation, we can write the expression above as

$$\mathbb{E}_{\mathcal{D}}\left[\sum_{k=1}^K \psi_k \circ d_1(g_k(x), y^k)\right] \leq \inf_{f \in \mathcal{F}} \mathbb{E}_{\mathcal{D}}\left[\sum_{k=1}^K \psi_k \circ d_1(f_k(x), y^k)\right] + \epsilon.$$

This shows that the learning rule that runs $\mathcal{A}_k$ on marginal samples S_k and concatenates the resulting scalar-valued predictors to get a vector-valued predictor is an agnostic learner

for $\mathcal{F}$ with respect to loss ℓ with sample complexity at most $\max_k m_k(\epsilon/K, \delta/K)$. This completes our proof of sufficiency. ■

C.2 Equivalence of d_1 and $\psi \circ d_1$ Learnability in Batch Regression

In this section, we provide proof of Lemma 11, which establishes the equivalence of d_1 and $\psi \circ d_1$ learnability in scalar-valued batch regression.

Proof (of Lemma 11) To prove sufficiency first, let $\mathcal{G}$ be agnostically learnable with respect to d_1. This implies that the fat-shattering dimension of $\mathcal{G}$ is finite at every scale (Bartlett et al., 1996) and uniform convergence holds over the loss class $d_1 \circ \mathcal{G}$. Since ψ is a Lipschitz function, a simple application of Talagrand's contraction lemma on Rademacher complexity (Bartlett and Mendelson, 2003) implies that uniform convergence holds over the loss class $\psi \circ d_1 \circ \mathcal{G}$ as well. Thus, $\mathcal{G}$ is learnable with respect to $\psi \circ d_1$ via ERM.

Next, we show that if $\mathcal{G} \subseteq [0,1]^{\mathcal{X}}$ is learnable with respect to $\psi \circ d_1$, then $\mathcal{G}$ is learnable with respect to d_1. Since the fat-shattering dimension of $\mathcal{G}$ characterizes d_1 learnability of $\mathcal{G}$, it suffices to show that $\mathcal{G}$ being learnable with respect to $\psi \circ d_1$ implies $\mathrm{fat}_\gamma(\mathcal{G}) < \infty$ for every $\gamma \in (0,1)$.

Suppose, for the sake of contradiction, $\mathcal{G}$ is learnable with respect to $\psi \circ d_1$ but there exists a scale $\gamma \in (0,1)$ such that $\mathrm{fat}_\gamma(\mathcal{G}) = \infty$. Then, for every $d \in \mathbb{N}$, there exists $X = \{x_1, \ldots, x_d\} \subseteq \mathcal{X}$ and a witness function $r : \mathcal{X} \to [0,1]$ such that for every $\sigma \in \{-1,1\}^d$, there exists a $g_\sigma \in \mathcal{G}$ such that $\sigma_i(g_\sigma(x_i) - r(x_i)) \geq \gamma$ for all $i \in [d]$. Define $\mathcal{G}_X = \{g_\sigma \in \mathcal{G} \mid \sigma \in \{-1,1\}^d\}$ be the set of functions that shatters X. Define $\mathcal{H} = \{-1,1\}^X$ to be a set of all functions from X to $\{-1,1\}$. By definition of $\mathcal{H}$, we must have $\mathrm{VC}(\mathcal{H}) = d$. We use an agnostic learner for $\mathcal{G}$ with respect to $\psi \circ d_1$ to construct an agnostic learner for $\mathcal{H}$ whose sample complexity, for large enough d, is smaller than the known lower bound for VC classes. Since $\mathrm{fat}_\gamma(\mathcal{G}) = \infty$, d can be made arbitrarily large and thus we derive a contradiction.

Let $\mathcal{A}$ be the promised agnostic learner for $\mathcal{G}$ with respect to $\psi \circ d_1$ with sample complexity $m(\epsilon, \delta)$. For all $f \in [0,1]^X$, define a threshold function $h_f : X \to \{-1,1\}$ as $h_f(x) = 2\,\mathbb{1}\{f(x) \geq r(x)\} - 1$. Let $\mathcal{D}$ be an arbitrary distribution on $X \times \{-1,1\}$ and $\mathcal{D}_X$ be its marginal on X.

Algorithm 7 Agnostic PAC learner for $\mathcal{H}$

Input: Agnostic learner $\mathcal{A}$ for $\mathcal{G}$, unlabeled samples $S_U \sim \mathcal{D}_X$, and another independent labeled samples $S_L \sim \mathcal{D}$

1 Define $S_{\mathrm{aug}} = \{(S_U, g^\alpha(S_U)) \mid g \in \mathcal{G}_X\}$, all possible augmentations of S_U by α-discretization of functions in $\mathcal{G}_X$ for $\alpha \leq \gamma/2$.

2 Run $\mathcal{A}$ over all possible augmentations to get $C(S_U) := \{\mathcal{A}(S) \mid S \in S_{\mathrm{aug}}\}$.

3 Define $C_{\pm 1}(S_U) = \{h_f \mid f \in C(S_U)\}$, a thresholded class of $C(S_U)$.

4 Return the predictor in $C_{\pm 1}(S_U)$ with the lowest empirical 0-1 risk over S_L.

We now show that Algorithm 7 is an agnostic learner for $\mathcal{H}$. Consider $d \gg S_U + S_L$. Then, $|C_{\pm 1}(S_U)| = |S_{\mathrm{aug}}| \leq (2/\alpha)^{|S_U|}$ can be much smaller than 2^d. Let $h^\star := \arg\min_{h \in \mathcal{H}} \mathbb{E}_{\mathcal{D}}[\mathbb{1}\{h(x) \neq y\}]$ be the optimal hypothesis for $\mathcal{D}$. Note that, by definition of

shattering, for every $h \in \mathcal{H}$, there exists a $g \in \mathcal{G}_X$ such that $h(x) = h_g(x)$ for all $x \in X$. In particular, there must exist $g^\star \in \mathcal{G}_X$ such that $h^\star(x) = h_{g^\star}(x) := 2\,\mathbb{1}\{g^\star(x) \geq r(x)\} - 1$ for all $x \in X$. Let g^α denote the α-discretization of g as defined in Equation (1) for some $\alpha \leq \gamma/2$. Now, consider a sample $(S_U, g^{\star,\alpha}(S_U)) \in S_{\text{aug}}$. Let $\hat{g} = \mathcal{A}((S_U, g^{\star,\alpha}(S_U)))$ be the predictor returned by the algorithm when run on a sample labeled by $g^{\star,\alpha}$. Define $h_{\hat{g}} = 2\,\mathbb{1}\{\hat{g}(x) \geq r(x)\} - 1$ to be its thresholded function. Then, using the triangle inequality on the indicator function, we have

$$\mathbb{E}_{\mathcal{D}}[\mathbb{1}\{h_{\hat{g}}(x) \neq y\}] \leq \mathbb{E}_{\mathcal{D}}[\mathbb{1}\{h^\star(x) \neq y\}] + \mathbb{E}_{\mathcal{D}_X}[\mathbb{1}\{h_{\hat{g}}(x) \neq h^\star(x)\}]. \tag{2}$$

Note that $\mathbb{1}\{h_{\hat{g}}(x) \neq h^\star(x)\} = \mathbb{1}\{h_{\hat{g}}(x) \neq h_{g^\star}(x)\} \leq \mathbb{1}\{|\hat{g}(x) - g^\star(x)| \geq \gamma\}$. To see why the last inequality is true, we only have to consider the case where the indicator on the left is 1, otherwise, the inequality is trivial. Recall that $\mathbb{1}\{h_{\hat{g}}(x) \neq h_{g^\star}(x)\} = 1$ whenever $\hat{g}(x)$ and $g^\star(x)$ lie on the opposite side of witness $r(x)$. Since $g^\star$ has to be at least γ away from the witness $r(x)$, we obtain $\mathbb{1}\{|\hat{g}(x) - g^\star(x)| \geq \gamma\} = 1$. Next, using the fact that $\alpha \leq \gamma/2$, we have $\mathbb{1}\{|\hat{g}(x) - g^\star(x)| \geq \gamma\} \leq \mathbb{1}\{|\hat{g}(x) - g^{\star,\alpha}(x)| \geq \gamma/2\}$ because discretization can decrease the distance between these functions by at most $\gamma/2$. Furthermore, using monotonicity of ψ, we get $\mathbb{1}\{|\hat{g}(x) - g^{\star,\alpha}(x)| \geq \gamma/2\} \leq \mathbb{1}\{\psi(|\hat{g}(x) - g^{\star,\alpha}(x)|) \geq \psi(\gamma/2)\}$. Combining everything, we get a pointwise inequality

$$\mathbb{1}\{h_{\hat{g}}(x) \neq h^\star(x)\} \leq \mathbb{1}\{\psi(|\hat{g}(x) - g^{\star,\alpha}(x)|) \geq \psi(\gamma/2)\} \leq \frac{1}{\psi(\gamma/2)}\psi(|\hat{g}(x) - g^{\star,\alpha}(x)|).$$

Using this inequality gives an upperbound on the risk of $h_{\hat{g}}$, namely

$$\mathbb{E}_{\mathcal{D}_X}[\mathbb{1}\{h_{\hat{g}}(x) \neq h^\star(x)\}] \leq \frac{1}{\psi(\gamma/2)}\,\mathbb{E}_{\mathcal{D}}[\psi(|\hat{g}(x) - g^{\star,\alpha}(x)|)]. \tag{3}$$

Since $\hat{g} = \mathcal{A}((S_U, g^{\star,\alpha}(S_U)))$, we can use the algorithm's guarantee to get a further upperbound on the expectation above. In particular, if $|S_U| \geq m(\frac{\epsilon\psi(\gamma/2)}{4}, \delta/2)$, then with probability at least $1 - \delta/2$ over sampling $S_U \sim \mathcal{D}_X$, we have

$$\mathbb{E}_{\mathcal{D}_X}[\psi(|\hat{g}(x) - g^{\star,\alpha}(x)|)] \leq \inf_{g \in \mathcal{G}} \mathbb{E}_{\mathcal{D}_X}[\psi(|g(x) - g^{\star,\alpha}(x)|)] + \frac{\epsilon\psi(\gamma/2)}{4}.$$

Note that $\inf_{g\in\mathcal{G}} \mathbb{E}_{\mathcal{D}_X}[\psi(|g(x) - g^{\star,\alpha}(x)|)] \leq \mathbb{E}_{\mathcal{D}_X}[\psi(|g^\star(x) - g^{\star,\alpha}(x)|)] \leq \psi(\alpha)$, where the last step uses the fact that $|g^\star(x) - g^{\star,\alpha}(x)| \leq \alpha$ and ψ is monotonic. Using L-Lipschitzness of ψ and the fact that $\psi(0) = 0$, we get $\psi(\alpha) \leq L\alpha$. Picking $\alpha = \min(\gamma/2, \frac{\epsilon\psi(\gamma/2)}{4L})$, we get $\inf_{g\in\mathcal{G}} \mathbb{E}_{\mathcal{D}_X}[\psi(|g(x) - g^{\star,\alpha}(x)|)] \leq \frac{\epsilon\psi(\gamma/2)}{4}$. Plugging this back to the inequality in the display above, we get to $\mathbb{E}_{\mathcal{D}_X}[\psi(|\hat{g}(x) - g^{\star,\alpha}(x)|)] \leq \frac{\epsilon\psi(\gamma/2)}{2}$. Using this guarantee on (3), we obtain

$$\mathbb{E}_{\mathcal{D}_X}[\mathbb{1}\{h_{\hat{g}}(x) \neq h^\star(x)\}] \leq \frac{\epsilon}{2}.$$

This bound applied to (2) yields

$$\mathbb{E}_{\mathcal{D}}[\mathbb{1}\{h_{\hat{g}}(x) \neq y\}] \leq \inf_{h\in\mathcal{H}} \mathbb{E}_{\mathcal{D}}[\mathbb{1}\{h(x) \neq y\}] + \frac{\epsilon}{2}.$$

Thus, we have shown the existence of a predictor $h_{\hat{g}} \in C_{\pm 1}(S_U)$ that achieves agnostic PAC bounds for $\mathcal{H}$. Let $\hat{h}$ be the predictor returned by step 4 of the algorithm. Next, we show that for sufficiently large S_L, the predictor $\hat{h}$ also attains agnostic PAC bounds. Recall that by Hoeffding's Inequality and union bound, with probability at least $1 - \delta/2$, the empirical risk of every hypothesis in $C_{\pm 1}(S_L)$ on a sample of size $\geq \frac{8}{\epsilon^2} \log \frac{4|C_{\pm 1}(S_U)|}{\delta}$ is at most $\epsilon/4$ away from its true error. So, if $|S_L| \geq \frac{8}{\epsilon^2} \log \frac{4|C_{\pm 1}(S_U)|}{\delta}$, then with probability at least $1 - \delta/2$, the empirical risk of the predictor $h_{\hat{g}}(x)$ is

$$\frac{1}{|S_L|} \sum_{(x,y) \in S_L} \mathbb{1}\{h_{\hat{g}}(x) \neq y\} \leq \mathbb{E}_{\mathcal{D}}[\mathbb{1}\{h_{\hat{g}}(x) \neq y\}] + \frac{\epsilon}{4} \leq \inf_{h \in \mathcal{H}} \mathbb{E}_{\mathcal{D}}[\mathbb{1}\{h(x) \neq y\}] + \frac{3\epsilon}{4},$$

where the last inequality follows from the risk guarantee of $h_{\hat{g}}$ established above. Since $\hat{h}$ is the empirical risk minimizer over S_L, we must have

$$\frac{1}{|S_L|} \sum_{(x,y) \in S_L} \mathbb{1}\{\hat{h}(x) \neq y\} \leq \frac{1}{|S_L|} \sum_{(x,y) \in S_L} \mathbb{1}\{h_{\hat{g}}(x) \neq y\} \leq \inf_{h \in \mathcal{H}} \mathbb{E}_{\mathcal{D}}[\mathbb{1}\{h(x) \neq y\}] + \frac{3\epsilon}{4}.$$

Finally, as the population risk of $\hat{h}$ is at most $\epsilon/4$ away from its empirical risk, we have

$$\mathbb{E}_{\mathcal{D}}[\mathbb{1}\{\hat{h}(x) \neq y\}] \leq \inf_{h \in \mathcal{H}} \mathbb{E}_{\mathcal{D}}[\mathbb{1}\{h(x) \neq y\}] + \epsilon,$$

which is the agnostic PAC guarantee for $\mathcal{H}$. Applying union bounds, the entire process, running algorithm $\mathcal{A}$ on the dataset augmented by $g^{\star,\alpha}$ and the ERM in step 4, succeeds with probability $1 - \delta$. This establishes that the Algorithm 7 is an agnostic PAC learner for $\mathcal{H}$. The sample complexity of Algorithm 7 is the number of samples required for Algorithm $\mathcal{A}$ to succeed and the ERM in step 4 to succeed. Thus, the overall sample complexity of Algorithm 7, denoted $m_{\mathcal{H}}(\epsilon, \delta)$, can be bounded as

$$\begin{aligned} m_{\mathcal{H}}(\epsilon, \delta) &\leq m_{\mathcal{A}}\left(\frac{\epsilon\, \psi(\gamma/2)}{4}, \frac{\delta}{2}\right) + \frac{8}{\epsilon^2} \log \frac{4|C_{\pm 1}(S_U)|}{\delta} \\ &\leq m_{\mathcal{A}}\left(\frac{\epsilon\, \psi(\gamma/2)}{4}, \frac{\delta}{2}\right) \left(1 + \frac{8}{\epsilon^2} \log\left(\frac{2}{\min(\gamma/2, \epsilon\psi(\gamma/2)/4)}\right)\right) + \frac{8}{\epsilon^2} \log \frac{4}{\delta} \end{aligned}$$

where the second inequality follows because $|C_{\pm 1}(S_U)| = |S_{\text{aug}}| \leq (2/\alpha)^{|S_U|}$ and we need $|S_U|$ to be of size $m_{\mathcal{A}}\left(\frac{\epsilon\, \psi(\gamma/2)}{4}, \frac{\delta}{2}\right)$. We also use the fact that $\alpha = \min(\frac{\gamma}{2}, \frac{\epsilon\psi(\frac{\gamma}{2})}{4})$.

However, it is well known (Shalev-Shwartz and Ben-David, 2014, Theorem 6.8) that the sample complexity of learning $\mathcal{H}$ in agnostic setting is

$$C \frac{d + \log(2/\delta)}{\epsilon^2}$$

for some $C > 0$. Thus, we must have $m_{\mathcal{H}}(\epsilon, \delta) \geq C(d + \log(2/\delta))/\epsilon^2$. However, this is a contradiction because d can be arbitrarily large but $m_{\mathcal{H}}(\epsilon, \delta)$ must have a finite upper bound for every fixed ϵ, δ. Therefore, the function class $\mathcal{G}$ cannot be learnable with respect to $\psi \circ d_1$ whenever there exists a scale $\gamma \in (0, 1)$ such that $\text{fat}_\gamma(\mathcal{G}) = \infty$. ■

Appendix D. Rademacher Based Proof for Batch Regression

To show that the learnability of each $\mathcal{F}_k$ with respect to d_1 is sufficient for the learnability of $\mathcal{F}$ with respect to ℓ_p norms for $p \geq 1$, we use the fact that $\ell_p(f(x), y)$ is a K-Lipschitz in its first argument with respect to $\|\cdot\|_\infty$ norm, that is $|\ell_p(f(x), y) - \ell_p(g(x), y)| \leq K \|f(x) - g(x)\|_\infty$, and use the following bound on the Rademacher complexity of the loss class $\ell \circ \mathcal{F} = \{(x, y) \mapsto \ell(f(x), y)\} \,|\, f \in \mathcal{F}\}$.

Lemma 33 (Foster and Rakhlin (2019)). *Let $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a multioutput function class. For any $\delta \in (0, 1)$, there exists a constants $0 < c < 1$ and $C > 0$ such that*

$$\mathfrak{R}_n(\ell_p \circ \mathcal{F}) \leq K \inf_{\alpha > 0} \left\{ 4\alpha + \frac{C}{\sqrt{n}} \sum_{k=1}^{K} \int_\alpha^1 \sqrt{\mathrm{fat}_{c\epsilon}(\mathcal{F}_k) \log^{1+\delta}\left(\frac{e\, n}{\epsilon}\right)} d\epsilon \right\}.$$

The result presented here is in fact the intermediate result in Foster and Rakhlin (2019), and we provide a sketch of how their argument can be adapted to our setting.

Proof Note that for $f, g \in \mathcal{F}$, we have $|\ell_p(f(x), y) - \ell_p(g(x), y)| \leq |\ell_p(f(x), g(x))| \leq \ell_1(f(x), g(x)) \leq K \|f(x) - g(x)\|_\infty$. Furthermore, we have that $|f_k(x) - y^k| \leq 1$, so we obtain $|\ell_p(f(x), y)| \leq K$. Define the normalized ℓ_p loss as $\bar{\ell}_p(f(x), y) := \ell_p(f(x), y)/K$. By standard chaining argument, we know that

$$\mathfrak{R}_n(\ell \circ \mathcal{F}) = K\, \mathfrak{R}_n(\bar{\ell} \circ \mathcal{F}) \leq K \inf_{\alpha > 0} \left\{ 4\alpha + \frac{12}{\sqrt{n}} \int_\alpha^1 \sqrt{\log \mathcal{N}_2(\bar{\ell}_p \circ \mathcal{F}, \epsilon, n)} d\epsilon \right\}.$$

Since a cover with $||\cdot||_\infty$ norm is also a cover with respect to $||\cdot||_2$ norm, we have that $\log \mathcal{N}_2(\bar{\ell}_p \circ \mathcal{F}, \epsilon, n) \leq \log \mathcal{N}_\infty(\bar{\ell} \circ \mathcal{F}, \epsilon, n)$. Since $\bar{\ell}_p(f(x), y)$ is $1-$Lipschitz with respect to $||\cdot||_\infty$ norm, following Lemma 1 of Foster and Rakhlin (2019), we obtain $\log \mathcal{N}_\infty(\bar{\ell}_p \circ \mathcal{F}, \epsilon, n) \leq \sum_{k=1}^K \log \mathcal{N}_\infty(\mathcal{F}_k, \epsilon, n)$.

A result due to Rudelson and Vershynin (2006) states that for any $\delta \in (0, 1)$, there exists constants $0 < c_k < 1$ and $C_k > 0$ such that

$$\log \mathcal{N}_\infty(\mathcal{F}_k, \epsilon, n) \leq C_k\, \mathrm{fat}_{c_k \epsilon}(\mathcal{F}_k)\, \log^{1+\delta}(en/\epsilon).$$

Picking $C = \max_k C_k$ and $c = \min_k c_k$, we obtain the contraction inequality

$$\mathfrak{R}_n(\ell \circ \mathcal{F}) \leq K \inf_{\alpha > 0} \left\{ 4\alpha + \frac{12\, C}{\sqrt{n}} \int_\alpha^1 \sqrt{\sum_{k=1}^{K} \mathrm{fat}_{c\epsilon}(\mathcal{F}_k) \log^{1+\delta}\left(\frac{e\, n}{\epsilon}\right)} d\epsilon \right\}.$$

Using $\sqrt{\sum_{k=1}^K \mathrm{fat}_{c\epsilon}(\mathcal{F}_k) \log^{1+\delta}\left(\frac{e\, n}{\epsilon}\right)} \leq \sum_{k=1}^K \sqrt{\mathrm{fat}_{c\epsilon}(\mathcal{F}_k) \log^{1+\delta}\left(\frac{e\, n}{\epsilon}\right)}$ yields the desired contraction inequality.

■

With Lemma 33 in our repertoire, the sufficiency proof is a routine uniform convergence argument.

Proof (of sufficiency in Theorem 12) Suppose each restriction $\mathcal{F}_k$ is learnable with respect to d_1. Then, we know that for all $k \in [K]$ and for all $1 > \gamma > 0$, we have $\mathrm{fat}_\gamma(\mathcal{F}_k) < \infty$

(Bartlett et al., 1996), (Anthony and Bartlett, 1999, Chatper 19)). Using Lemma 33, for $\delta = 1/2$, we can find constants c, C such that

$$\mathfrak{R}_n(\ell_p \circ \mathcal{F}) \leq K \inf_{\alpha > 0} \left\{ 4\alpha + \frac{C}{\sqrt{n}} \sum_{k=1}^{K} \int_{\alpha}^{1} \sqrt{\mathrm{fat}_{c\epsilon}(\mathcal{F}_k) \log^{3/2}\left(\frac{e\,n}{\epsilon}\right)} d\epsilon \right\}.$$

Fix $\alpha > 0$. The second term inside infimum vanishes as $n \to \infty$, yielding $\mathfrak{R}_n(\ell \circ \mathcal{F}) \leq 4\alpha K$. As $\alpha > 0$ is arbitrary, the Rademacher complexity $\mathfrak{R}_n(\ell \circ \mathcal{F})$ goes to 0 as $n \to \infty$. This argument can be readily turned into non-asymptotic bounds on $\mathfrak{R}_n(\ell_p \circ \mathcal{F})$ if the precise form of $\mathrm{fat}_\gamma(\mathcal{F}_k)$ as a function of γ is known. Since the empirical Rademacher complexity vanishes, uniform convergence holds over the loss class $\ell_p \circ \mathcal{F}$ and thus $\mathcal{F}$ is learnable with respect to ℓ_p via empirical risk minimization. ■

Appendix E. Online Multilabel Learnability with respect to Hamming Loss

In this section, we provide the proof of Theorem 15.

Proof We first prove that the online learnability of each restriction is sufficient for the online learnability of ℓ_H.

Part 1: Sufficiency. Our proof is based on a reduction: given oracle access to online learners $\{\mathcal{A}_k\}_{k=1}^K$ for $\{\mathcal{F}_k\}_{k=1}^K$ with respect to $\ell_{0\text{-}1}$, we construct an online learner $\mathcal{A}$ for $\mathcal{F}$ with respect to ℓ_H. In fact, similar to the batch setting, the online multilabel learning algorithm $\mathcal{A}$ is simple: in each round $t \in [T]$, receive x_t, query the predictions $\mathcal{A}_1(x_t), ..., \mathcal{A}_K(x_t)$, and finally predict the concatenation $\hat{y}_t = (\mathcal{A}_1(x_t), ..., \mathcal{A}_K(x_t))$. Once the true label $y_t = (y_t^1, ..., y_t^K)$ is revealed, update each online learner $\mathcal{A}_k$ by passing (x_t, y_t^k) for $k \in [K]$. It suffices to show that the expected regret of $\mathcal{A}$ is sublinear in T with respect to ℓ_H. By Definition 5, we have that for all $k \in [K]$,

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\mathcal{A}_k(x_t) \neq y_t^k\} - \inf_{f_k \in \mathcal{F}_k} \sum_{t=1}^{T} \mathbb{1}\{f_k(x_t) \neq y_t^k\}\right] \leq R_k(T)$$

where $R_k(T)$ is some sublinear function in T. Summing the regret bounds across all $k \in [K]$ splitting up the expectations, and using linearity of expectation, we get that $\mathbb{E}\left[\sum_{k=1}^K \sum_{t=1}^T \mathbb{1}\{\mathcal{A}_k(x_t) \neq y_t^k\}\right] - \mathbb{E}\left[\sum_{k=1}^K \inf_{f_k \in \mathcal{F}_k} \sum_{t=1}^T \mathbb{1}\{f_k(x_t) \neq y_t^k\}\right] \leq \sum_{k=1}^K R_k(T)$. Noting that $\sum_{k=1}^K \inf_{f_k \in \mathcal{F}_k} \sum_{t=1}^T \mathbb{1}\{f_k(x_t) \neq y_t^k\} \leq \inf_{f \in \mathcal{F}} \sum_{k=1}^K \sum_{t=1}^T \mathbb{1}\{f_k(x_t) \neq y_t^k\}$, swapping the order of summations, and using the definition of ℓ_H we have that,

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_H(\mathcal{A}(x_t), y_t)\right] - \mathbb{E}\left[\inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \ell_H(f(x_t), y_t)\right] \leq \sum_{k=1}^{K} R_k(T),$$

where $\mathcal{A}(x_t) = (\mathcal{A}_1(x_t), ..., \mathcal{A}_K(x_t))$. This concludes the proof of this direction since $\sum_{k=1}^K R_k(T)$ is still a sublinear function in T.

Part 2: Necessity. Next we prove that if $\mathcal{F}$ is online learnable with respect to to ℓ_H, then each $\mathcal{F}_k$ is online learnable with respect to $\ell_{0\text{-}1}$. Namely, given oracle access to an online

learner $\mathcal{A}$ for $\mathcal{F}$ with respect to ℓ_H, we construct an online learner $\mathcal{B}$ for $\mathcal{F}_1$ with respect to $\ell_{0\text{-}1}$. A similar reduction can be used to construct online learners for each restriction $\mathcal{F}_k$. Similar to the batch setting, the online learning algorithm $\mathcal{B}$ is simple: in each round $t \in [T]$, receive x_t, query $\hat{y}_t = \mathcal{A}(x_t)$ and predict $\hat{y}_t^1 = \mathcal{A}_1(x_t)$. Once the true label y_t^1 is revealed, update $\mathcal{A}$ by passing (x_t, y_t) where $y_t = (y_t^1, \sigma_t^2, ..., \sigma_t^K)$ and $\{\sigma_t^i\}_{i=2}^K$ is an i.i.d sequence of Rademacher random variables.

It suffices to show that the expected regret of $\mathcal{B}$ is sublinear in T with respect to $\ell_{0\text{-}1}$. As previously mentioned, we assume that the sequence $(x_1, y_1^1), ..., (x_T, y_T^1)$ is chosen by an oblivious adversary, and thus is not random. Let $y_t = (y_t^1, \sigma_t^2, ..., \sigma_t^K)$. By Definition 5, we have that,

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_H(\mathcal{A}(x_t), y_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^{T} \ell_H(f(x_t), y_t)\right] \leq R(T, 2^K)$$

where the expectation is over both the randomness of $\mathcal{A}(x_t)$ and $(\sigma_t^2, ..., \sigma_t^K)$ and $R(T, 2^K)$ is a sub-linear function of T. Splitting up the expectation, using the definition of the Hamming loss, and by the linearity of expectation, we have that

$$\sum_{t=1}^{T}\sum_{k=1}^{K} \mathbb{E}\left[\mathbb{1}\{\mathcal{A}_k(x_t) \neq y_t^k\}\right] - \inf_{f\in\mathcal{F}} \sum_{t=1}^{T}\sum_{k=1}^{K} \mathbb{E}\left[\mathbb{1}\{f_k(x_t) \neq y_t^k\}\right] \leq R(T, 2^K).$$

Next, observe that for every $t \in [T]$, for every $k \in \{2, ..., K\}$, the randomness of $y_t^k = \sigma_t^k$ implies $\mathbb{E}\left[\mathbb{1}\{\mathcal{A}_k(x_t) \neq y_t^k\}\right] = \mathbb{E}\left[\mathbb{1}\{f(x_t) \neq y_t^k\}\right] = \frac{1}{2}$. Thus,

$$\sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}\{\mathcal{A}_1(x_t) \neq y_t^1\} + \frac{K-1}{2}\right] - \inf_{f\in\mathcal{F}} \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}\{f_1(x_t) \neq y_t^1\} + \frac{K-1}{2}\right] \leq R(T, 2^K).$$

Canceling constant factors gives, $\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\mathcal{A}_1(x_t) \neq y_t^1\}\right] - \inf_{f_1\in\mathcal{F}_1} \sum_{t=1}^{T} \mathbb{1}\{f_1(x_t) \neq y_t^1\} \leq R(T, 2^K)$, showing that $\mathcal{B}$ is an online agnostic learner for $\mathcal{F}_1$ with respect to $\ell_{0\text{-}1}$. ■

Appendix F. MCLdim Characterizes Online Multilabel Learnability

In this section, we show that the MCLdim characterizes the online learnability of a multilabel function class $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ with respect to to any loss ℓ that satisfies the identity of indiscernibles. Theorem 34 makes this more precise.

Theorem 34. *Let ℓ be any loss function satisfying the identity of indiscernibles. A function class $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ is online learnable with respect to ℓ if and only if MCLdim$(\mathcal{F}) < \infty$.*

Proof (of sufficiency) We first show that finiteness of MCLdim is sufficient for online learnability. The proof follows exactly like the proof of Lemma 16. We include it here again for completeness sake. Let ℓ be any loss function satisfying the identity of indiscernibles and $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a multilabel function class such that MCLdim$(\mathcal{F}) = d < \infty$. Since $\mathcal{F}$ has finite MCLdim, the deterministic Multiclass Standard Optimal Algorithm for $\mathcal{F}$,

hereinafter denoted MCSOA($\mathcal{F}$), achieves mistake-bound d in the realizable setting (Daniely et al., 2011). Therefore, following the same procedure as in Daniely et al. (2011), we can construct a finite set of experts $\mathcal{E}$ of size $|\mathcal{E}| = \sum_{j=0}^{d} \binom{T}{j} |\text{im}(\mathcal{F})|^j \leq (2^K T)^d$ such that for any (oblivious) sequence of instances $x_1, ..., x_T$, for any function $f \in \mathcal{F}$, there exists an expert $E_f \in \mathcal{E}$, such that $f(x_t) = E(x_t)$ for all $t \in [T]$. Finally, running the celebrated Randomized Exponential Weights Algorithm (REWA) using $\mathcal{E}$ as the set of experts and the scaled loss function $\frac{\ell}{B} \in [0, 1]$ guarantees that for any labelled sequence $(x_1, y_1), ..., (x_T, y_T)$,

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) - \inf_{E \in \mathcal{E}} \sum_{t=1}^{T} \ell(E(x_t), y_t)\right] &\leq \mathbb{E}\left[\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \ell(f(x_t), y_t)\right] \\ &\leq O\left(B\sqrt{T \ln(|\mathcal{E}|)}\right) \leq O\left(B\sqrt{dTK \ln(T)}\right)\end{aligned}$$

where $\hat{y}_t$ is the prediction of REWA in the t'th round. Thus, running REWA over the set of experts $\mathcal{E}$ using $\frac{\ell}{B}$ gives an online learner for $\mathcal{F}$ with respect to ℓ. ■

Proof (of necessity) To prove necessity, we need to show that if $\mathcal{F}$ is online learnable with respect to ℓ, then MCLdim($\mathcal{F}$) $< \infty$. To do so, we show that if $\mathcal{F}$ is online learnable with respect to ℓ, then $\mathcal{F}$ is online learnable with respect to $\ell_{0\text{-}1}$ in the realizable setting. Let $\mathcal{A}$ be an online learner for $\mathcal{F}$ with respect to ℓ. Then, by definition,

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell(\mathcal{A}(x_t), y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \ell(f(x_t), y_t)\right] \leq R(T, 2^K)$$

where $R(T, 2^K)$ is a sublinear function in T. Since ℓ satisfies the identity of indiscernibles, in the realizable setting, $\inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \ell(f(x_t), y_t) = 0$. Therefore, under realizability,

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell(\mathcal{A}(x_t), y_t)\right] \leq R(T, 2^K).$$

Because there are only a finite number of inputs to ℓ, there must exist a universal constant a such that $a\ell_{0\text{-}1} \leq \ell$. Substituting in gives that,

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_{0\text{-}1}(\mathcal{A}(x_t), y_t)\right] \leq \frac{R(T, 2^K)}{a}.$$

Since a is a universal constant that does not depend on T, $\frac{R(T,2^K)}{a}$ is still a sublinear function in T, implying that $\mathcal{A}$ is also a realizable online learner for $\mathcal{F}$ with respect to $\ell_{0\text{-}1}$. This completes the proof as MCLdim characterizes realizable learnability and so we must have MCLdim($\mathcal{F}$) $< \infty$. ■

Appendix G. Proofs for Bandit Online Multilabel Classification

In this section, we provide proofs for the characterization of online multilabel classification under bandit feedback.

Proof (of Theorem 19) The proof of Theorem 19 is nearly identical to the proof of Theorem 13. The only difference is that in Algorithm 4, we now need use the bandit Expert's algorithm EXP4 from Auer et al. (2002) instead of REWA. Similar to REWA, based on Theorem 2.3 in Daniely and Helbertal (2013) and the fact that A, B and P are independent, EXP4 guarantees that

$$\mathbb{E}\left[\sum_{t=1}^{T}\ell(\mathcal{P}(x_t),y_t)\right] \leq \mathbb{E}\left[\inf_{E\in\mathcal{E}_B}\sum_{t=1}^{T}\ell(E(x_t),y_t)\right] + eM\mathbb{E}\left[\sqrt{2T|\mathcal{Y}|\ln(|\mathcal{E}_B|)}\right],$$

where $\mathcal{P}(x_t)$ denotes EXP4's prediction in round t. The remaining proof for deriving the upper bound

$$\mathbb{E}\left[\inf_{E\in\mathcal{E}_B}\sum_{t=1}^{T}\ell(E(x_t),y_t)\right] \leq \inf_{f\in\mathcal{F}}\sum_{t=1}^{T}\ell(f(x_t),y_t) + \frac{cT}{T^\beta}\overline{R}(T^\beta,|\mathcal{Y}|)$$

is identical to that in Theorem 13, so we omit it here. Putting these pieces together gives the stated guarantee. ■

Proof (of Theorem 18) Let $c = \frac{\max_{r\neq t}\ell(r,t)}{\min_{r\neq t}\ell(r,t)}$. We first show necessity: if $\mathcal{F}$ is bandit online learnable with respect to ℓ, then each restriction $\mathcal{F}_k$ is online learnable with respect to $\ell_{0\text{-}1}$. This follows trivially from the fact that if $\mathcal{A}$ is a bandit online learner for $\mathcal{F}$, then $\mathcal{A}$ is also an online learner for $\mathcal{F}$ under full-feedback. Thus, by Theorem 17, online learnability of $\mathcal{F}$ with respect to ℓ implies online learnability of restriction $\mathcal{F}_k$ with respect to the 0-1 loss.

We now focus on showing sufficiency: if for all $k \in [K]$, $\mathcal{F}_k$ is online learnable with respect to 0-1 loss, then $\mathcal{F}$ is *bandit* online learnable with respect to loss ℓ. Since $|\mathcal{Y}| = 2^K < \infty$ and ℓ is a c-subadditive, by Theorem 19, it suffices to show that there exists a realizable online learner for $\mathcal{F}$ with respect to ℓ. However, using Theorem 17, online learnability of each restriction $\mathcal{F}_k$ with respect to 0-1 the loss implies (agnostic) online learnability of $\mathcal{F}$ with respect to ℓ. Since an agnostic online learner is trivially a realizable online learner, the proof is complete. ■

Appendix H. Equivalence of d_1 and $\psi \circ d_1$ Online Learnability

Proof (of Lemma 22) Since ψ is a Lipschitz function, the proof of sufficiency follows immediately from Corollary 5 in Rakhlin et al. (2015a), a contraction Lemma for the sequential Rademacher complexity. Thus, we focus on proving necessity - if $\mathcal{G}$ is online learnable with respect to $\psi \circ d_1$, then $\mathcal{G}$ is online learnable with respect to d_1.

Since the sequential fat shattering dimension of $\mathcal{G}$ characterizes d_1 learnability (Rakhlin et al., 2015a), it suffices to show that $\mathcal{G}$ being online learnable with respect to $\psi \circ d_1$ implies $\text{fat}_\gamma^{\text{seq}}(\mathcal{G}) < 0$ for every $\gamma \in (0,1)$. Like in the batch setting, we prove this via contradiction.

Suppose, for the sake of contradiction, $\mathcal{G}$ is online learnable with respect to $\psi \circ d_1$ but there exists a scale $\gamma \in (0,1)$ such that $\mathrm{fat}^{\mathrm{seq}}_\gamma(\mathcal{G}) = \infty$. Then, for every $T \in \mathbb{N}$, there exists a $\mathcal{X}$-valued binary tree $\mathcal{T}$ and a $[0,1]$-valued binary witness tree $\mathcal{R}$ both of depth T such that for all $\sigma \in \{-1,1\}^T$, there exists $g_\sigma \in \mathcal{G}$ such that for all $t \in [T]$, $\sigma_t(g_\sigma(\mathcal{T}_t(\sigma_{<t})) - \mathcal{R}_t(\sigma_{<t})) \geq \gamma$. Without loss of generality, assume that for any path $\sigma \in \{-1,1\}^T$, the set of instances $\{\mathcal{T}_t(\sigma_{<t})\}_{t=1}^T$ are distinct. This is true because we can construct a γ-shattered tree of much bigger depth and prune away repeated instances along a path to get a tree of depth T. Define $\mathcal{G}_T = \{g_\sigma \in \mathcal{G} | \sigma \in \{-1,1\}^T\}$ to be the set of functions that shatter $\mathcal{T}$ with witness $\mathcal{R}$. Let $X \subseteq \mathcal{X}$ denote the set of examples that label the internal nodes of $\mathcal{T}$. Consider the binary hypothesis class $\mathcal{H} = \{-1,1\}^X$ which contains all possible functions from X to $\{-1,1\}$. By definition of $\mathcal{H}$, we must have $\mathrm{Ldim}(\mathcal{H}) \geq T$. Therefore, $\mathcal{T}$, with left and right edges labeled by -1 and $+1$ respectively, is shattered by $\mathcal{H}$. Let $\mathcal{T}_\pm$ denote such a tree. Note that for all $t \in [T]$, we have $\mathcal{T}_t(\sigma_{<t}) = \mathcal{T}_{\pm,t}(\sigma_{<t})$. Since $\mathrm{Ldim}(\mathcal{H}) \geq T$, any realizable online learner for $\mathcal{H}$ must make at least $\frac{T}{2}$ mistakes in expectation for an adversary that plays according to a root-to-leaf path in $\mathcal{T}_\pm$ chosen *uniformly at random*. However, using an agnostic online learner for $\mathcal{G}$ with respect to $\psi \circ d_1$, we construct a realizable online learner for $\mathcal{H}$ that achieves a *sublinear* regret bound when an adversary plays according to a root-to-leaf path in $\mathcal{T}_\pm$ chosen uniformly at random. Since $\mathrm{fat}^{\mathrm{seq}}_\gamma(\mathcal{G}) = \infty$, T can be made arbitrarily large, eventually giving us a contradiction.

To that end, let $\mathcal{A}$ be an online learner for $\mathcal{G}$ with respect to $\psi \circ d_1$ with regret $R_{\mathcal{A}}(T)$. Let $\sigma \sim \{-1,1\}^T$ denote a sequence of T i.i.d Rademacher random variables and $\{(\mathcal{T}_{\pm,t}(\sigma_{<t}), \sigma_t)\}_{t=1}^T$ the associated sequence of labeled instances determined by traversing $\mathcal{T}_\pm$ using σ. Note that $\{(\mathcal{T}_{\pm,t}(\sigma_{<t}), \sigma_t)\}_{t=1}^T$ is a sequence of labeled instances corresponding to a root-to-leaf path in $\mathcal{T}_\pm$ chosen *uniformly at random*. By construction of $\mathcal{H}$, there exists a $h^\star_\sigma \in \mathcal{H}$ such that $h^\star_\sigma(\mathcal{T}_{\pm,t}(\sigma_{<t})) = \sigma_t$ for all $t \in [T]$ and therefore the stream is realizable by $\mathcal{H}$. Let $g^\star_\sigma \in \mathcal{G}$ be the function at the end of the root-to-leaf path corresponding to σ in $\mathcal{T}$, the original tree shattered by $\mathcal{G}$.

We now use $\mathcal{A}$ to construct an agnostic online learner for $\mathcal{H}$ with sublinear regret on the stream $\{(\mathcal{T}_{\pm,t}(\sigma_{<t}), \sigma_t)\}_{t=1}^T$. Our algorithm is very similar to realizable-to-agnostic conversion in Theorem 13. Namely, we construct a finite set of experts, each of which uses $\mathcal{A}$ to make predictions, but only updates $\mathcal{A}$ on certain rounds. Finally, we run REWA using this set of Experts over our stream. For completeness' sake, we provide the full description below.

For any bitstring $b \in \{0,1\}^T$, let $\phi : \{t : b_t = 1\} \to \mathrm{im}(\mathcal{G}^\alpha)$ denote a function mapping time points where $b_t = 1$ to elements in the discretized image space $\mathrm{im}(\mathcal{G}^\alpha)$. Let $\Phi_b : (\mathrm{im}(\mathcal{G}^\alpha))^{\{t:b_t=1\}}$ denote all such functions ϕ. For every $g \in \mathcal{G}$, let $\phi^g_b \in \Phi_b$ be the mapping such that for all $t \in \{t : b_t = 1\}$, $\phi^g_b(t) = g^\alpha(\mathcal{T}_t(\sigma_{<t}))$. Let $|b| = |\{t : b_t = 1\}|$. For every $b \in \{0,1\}^T$ and $\phi \in \Phi_b$, define an Expert $E_{b,\phi}$. Expert $E_{b,\phi}$, formally presented in Algorithm 8, uses $\mathcal{A}$ to make predictions in each round. However, $E_{b,\phi}$ only updates $\mathcal{A}$ on those rounds where $b_t = 1$, using ϕ to produce a labeled instance $(\mathcal{T}_t(\sigma_{<t}), \phi(t))$. For every $b \in \{0,1\}^T$, let $\mathcal{E}_b = \bigcup_{\phi \in \Phi_b} \{E_{b,\phi}\}$ denote the set of all Experts parameterized by functions $\phi \in \Phi_b$. If b is the all zeros bitstring, then $\mathcal{E}_b$ is empty. Therefore, we actually define $\mathcal{E}_b = \{E_0\} \cup \bigcup_{\phi \in \Phi_b} \{E_{b,\phi}\}$, where E_0 is the expert that never updates $\mathcal{A}$. Note that $1 \leq |\mathcal{E}_b| \leq (\frac{2}{\alpha})^{|b|}$.

Algorithm 8 Expert(b, ϕ)

Input: Independent copy of Online Learner $\mathcal{A}$ for $\psi \circ d_1$

for $t = 1, ..., T$ **do**
  Receive example $\mathcal{T}_{\pm,t}(\sigma_{<t})$
  Predict $\hat{y}_t = 2\,\mathbb{1}\{\mathcal{A}(\mathcal{T}_{\pm,t}(\sigma_{<t})) \geq \mathcal{R}_t(\sigma_{<t})\} - 1$
  Receive $y_t = \sigma_t$
  if $b_t = 1$ **then**
    Update $\mathcal{A}$ by passing $(\mathcal{T}_{\pm,t}(\sigma_{<t}), \phi(t))$
end

With this notation in hand, we are now ready to present Algorithm 9, our main online learner $\mathcal{Q}$ for $\mathcal{H}$ with respect to 0-1 loss. The analysis is similar to the one before, but we include it below for completeness sake.

Algorithm 9 Online learner $\mathcal{Q}$ for $\mathcal{H}$ with respect to 0-1 loss

Input: Parameters $0 < \beta < 1$ and $0 < \alpha < \frac{\gamma}{2}$

Let $B \in \{0,1\}^T$ such that $B_t \overset{\text{iid}}{\sim} \text{Bernoulli}(\frac{T^\beta}{T})$

Construct the set of experts $\mathcal{E}_B = \{E_0\} \cup \bigcup_{\phi \in \Phi_B} \{E_{B,\phi}\}$ according to Algorithm 8.

Run REWA $\mathcal{P}$ using $\mathcal{E}_B$ and the 0-1 loss over the stream $(\mathcal{T}_{\pm,1}(\sigma_{<1}), \sigma_1), ..., (\mathcal{T}_{\pm,T}(\sigma_{<T}), \sigma_T)$

Our goal now is to show that $\mathcal{Q}$ enjoys sublinear expected regret. There are three main sources of randomness: the randomness involved in sampling B, the internal randomness of each independent copy of the online learner $\mathcal{A}$, and the internal randomness of REWA. Let B, A and P denote the random variable associated with these sources of randomness respectively. By construction, A, B, and P are independent.

Using Theorem 21.11 in Shalev-Shwartz and Ben-David (2014) and the fact that A, B and P, are independent, REWA guarantees,

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\mathcal{P}(\mathcal{T}_{\pm,t}(\sigma_{<t})) \neq \sigma_t\}\right] \leq \mathbb{E}\left[\inf_{E \in \mathcal{E}_B} \sum_{t=1}^{T} \mathbb{1}\{E(\mathcal{T}_{\pm,t}(\sigma_{<t})) \neq \sigma_t\}\right] + \mathbb{E}\left[\sqrt{2T \ln(|\mathcal{E}_B|)}\right].$$

Therefore,

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\mathcal{Q}(\mathcal{T}_{\pm,t}(\sigma_{<t})) \neq \sigma_t\}\right] &= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\mathcal{P}(\mathcal{T}_{\pm,t}(\sigma_{<t})) \neq \sigma_t\}\right] \\
&\leq \mathbb{E}\left[\inf_{E \in \mathcal{E}_B} \sum_{t=1}^{T} \mathbb{1}\{E(\mathcal{T}_{\pm,t}(\sigma_{<t})) \neq \sigma_t\}\right] + \mathbb{E}\left[\sqrt{2T \ln(|\mathcal{E}_B|)}\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{E_{B,\phi_B^{g_\sigma^\star}}(\mathcal{T}_{\pm,t}(\sigma_{<t})) \neq \sigma_t\}\right] + \mathbb{E}\left[\sqrt{2T \ln(|\mathcal{E}_B|)}\right].
\end{aligned}$$

In the last step, we used the fact that for all $b \in \{0,1\}^T$ and $g \in \mathcal{G}$, $E_{b,\phi_b^g} \in \mathcal{E}_b$.

It now suffices to upperbound $\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{E_{B,\phi_B^{g_\sigma^\star}}(\mathcal{T}_{\pm,t}(\sigma_{<t})) \neq \sigma_t\}\right]$. We use the same notation used to prove Theorem 13, but for the sake of completeness, we restate it here. Given an online learner $\mathcal{A}$ for $\psi \circ d_1$, an instance $x \in \mathcal{X}$, and an ordered sequence of labeled examples $L \in (\mathcal{X} \times [0,1])^*$, let $\mathcal{A}(x|L)$ be the random variable denoting the prediction of $\mathcal{A}$ on the instance x after running and updating on L. For any $b \in \{0,1\}^T$, $g^\alpha \in \mathcal{G}^\alpha$, and $t \in [T]$, let $L^g_{b_{<t}} = \{(\mathcal{T}_{\pm,t}(\sigma_{<i}), g^\alpha(\mathcal{T}_{\pm,t}(\sigma_{<i}))) : i < t \text{ and } b_i = 1\}$ denote the *subsequence* of the sequence of labeled instances $\{(\mathcal{T}_{\pm,t}(\sigma_{<i}), g^\alpha(\mathcal{T}_{\pm,t}(\sigma_{<i})))\}_{i=1}^{t-1}$ where $b_i = 1$. Using this notation, we can write

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{E_{B,\phi_B^{g_\sigma^\star}}(\mathcal{T}_{\pm,t}(\sigma_{<t})) \neq \sigma_t\}\right] &= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\Big\{2\,\mathbb{1}\{\mathcal{A}(\mathcal{T}_{\pm,t}(\sigma_{<t})|L^{g_\sigma^\star}_{B_{<t}}) \geq \mathcal{R}_t(\sigma_{<t})\} - 1 \neq \sigma_t\Big\}\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\Big\{2\,\mathbb{1}\{\mathcal{A}(\mathcal{T}_{\pm,t}(\sigma_{<t})|L^{g_\sigma^\star}_{B_{<t}}) \geq \mathcal{R}_t(\sigma_{<t})\} - 1 \neq h^\star_\sigma(\mathcal{T}_{\pm,t}(\sigma_{<t}))\Big\}\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{|\mathcal{A}(\mathcal{T}_{\pm,t}(\sigma_{<t})|L^{g_\sigma^\star}_{B_{<t}}) - g^\star_\sigma(\mathcal{T}_{\pm,t}(\sigma_{<t}))| \geq \gamma\right\}\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{|\mathcal{A}(\mathcal{T}_{\pm,t}(\sigma_{<t})|L^{g_\sigma^\star}_{B_{<t}}) - g^{\star,\alpha}_\sigma(\mathcal{T}_{\pm,t}(\sigma_{<t}))| \geq \frac{\gamma}{2}\right\}\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{\psi \circ d_1(\mathcal{A}(\mathcal{T}_{\pm,t}(\sigma_{<t})|L^{g_\sigma^\star}_{B_{<t}}), g^{\star,\alpha}_\sigma(\mathcal{T}_{\pm,t}(\sigma_{<t}))) \geq \psi(\frac{\gamma}{2})\right\}\right] \\
&\leq \frac{1}{\psi(\frac{\gamma}{2})}\mathbb{E}\left[\sum_{t=1}^{T} \psi \circ d_1(\mathcal{A}(\mathcal{T}_{\pm,t}(\sigma_{<t})|L^{g_\sigma^\star}_{B_{<t}}), g^{\star,\alpha}_\sigma(\mathcal{T}_{\pm,t}(\sigma_{<t})))\right]
\end{aligned}$$

The first inequality follows from γ-shattering. Indeed, if $2\mathbb{1}\{\mathcal{A}(\mathcal{T}_{\pm,t}(\sigma_{<t})|L^{g_\sigma^\star}_{B_{<t}}) \geq \mathcal{R}_t(\sigma_{<t})\} - 1 \neq h^\star_\sigma(\mathcal{T}_{\pm,t}(\sigma_{<t}))$, then $\mathcal{A}(\mathcal{T}_{\pm,t}(\sigma_{<t})|L^{g_\sigma^\star}_{B_{<t}})$ and $g^\star_\sigma$ must lie on opposite sides of the witness $\mathcal{R}_t(\sigma_{<t})$. The second inequality stems from the choice of $\alpha < \frac{\gamma}{2}$. The third inequality follows from the monotonicity of ψ. The last inequality follows from Markov's. Now, we can continue like before.

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} \psi \circ d_1(\mathcal{A}(\mathcal{T}_{\pm,t}(\sigma_{<t})|L^{g^\star}_{B_{<t}}), g^{\star,\alpha}_\sigma(\mathcal{T}_{\pm,t}(\sigma_{<t})))\right] \\
&\qquad = \mathbb{E}\left[\sum_{t=1}^{T} \psi \circ d_1(\mathcal{A}(\mathcal{T}_{\pm,t}(\sigma_{<t})|L^{g^\star}_{B_{<t}}), g^{\star,\alpha}_\sigma(\mathcal{T}_{\pm,t}(\sigma_{<t})))\frac{\mathbb{P}\left[B_t = 1\right]}{\mathbb{P}\left[B_t = 1\right]}\right] \\
&\qquad = \frac{T}{T^\beta}\mathbb{E}\left[\sum_{t=1}^{T} \psi \circ d_1(\mathcal{A}(\mathcal{T}_{\pm,t}(\sigma_{<t})|L^{g^\star}_{B_{<t}}), g^{\star,\alpha}_\sigma(\mathcal{T}_{\pm,t}(\sigma_{<t})))\mathbb{1}\{B_t = 1\}\right]
\end{aligned}$$

To see the last equality, note that the prediction $\mathcal{A}(\mathcal{T}_{\pm,t}(\sigma_{<t})|L^{g^\star}_{B_{<t}})$ only depends on bit-string $(B_1,\ldots,B_{t-1})$, the string $(\sigma_1,...,\sigma_{t-1})$, and the internal randomness of A, all of which are independent of B_t. Thus, we have

$$\begin{aligned}\mathbb{E}\left[\psi\circ d_1(\mathcal{A}(\mathcal{T}_{\pm,t}(\sigma_{<t})|L^{g^\star}_{B_{<t}}),g^{\star,\alpha}_\sigma(\mathcal{T}_{\pm,t}(\sigma_{<t})))\mathbb{1}\{B_t=1\}\right]&=\mathbb{E}\left[\psi_1\circ d_1(\mathcal{A}_1(x_t|L^{g^\star_\sigma}_{B_{<t}}),y_t)\right]\mathbb{E}\left[\mathbb{1}\{B_t=1\}\right]\\&=\mathbb{E}\left[\psi_1\circ d_1(\mathcal{A}_1(x_t|L^{g^\star_\sigma}_{B_{<t}}),y_t)\right]\mathbb{P}[B_t=1]\end{aligned}$$

as needed. Continuing onwards,

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=1}^{T}\psi\circ d_1(\mathcal{A}(\mathcal{T}_{\pm,t}(\sigma_{<t})|L^{g^\star_\sigma}_{B_{<t}}),g^{\star,\alpha}_\sigma(\mathcal{T}_{\pm,t}(\sigma_{<t})))\right]\\&\qquad=\frac{T}{T^\beta}\mathbb{E}\left[\sum_{t=1}^{T}\psi\circ d_1(\mathcal{A}(\mathcal{T}_{\pm,t}(\sigma_{<t})|L^{g^\star_\sigma}_{B_{<t}}),g^{\star,\alpha}_\sigma(\mathcal{T}_{\pm,t}(\sigma_{<t})))\mathbb{1}\{B_t=1\}\right]\\&\qquad=\frac{T}{T^\beta}\mathbb{E}\left[\sum_{t:B_t=1}\psi\circ d_1(\mathcal{A}(\mathcal{T}_{\pm,t}(\sigma_{<t})|L^{g^\star_\sigma}_{B_{<t}}),g^{\star,\alpha}_\sigma(\mathcal{T}_{\pm,t}(\sigma_{<t})))\right]\\&\qquad=\frac{T}{T^\beta}\mathbb{E}\left[\mathbb{E}\left[\sum_{t:B_t=1}\psi\circ d_1(\mathcal{A}(\mathcal{T}_{\pm,t}(\sigma_{<t})|L^{g^\star_\sigma}_{B_{<t}}),g^{\star,\alpha}_\sigma(\mathcal{T}_{\pm,t}(\sigma_{<t})))\middle|B\right]\right]\\&\qquad\le\frac{T}{T^\beta}\mathbb{E}\left[\sum_{t:B_t=1}\psi\circ d_1(g^\star_\sigma(\mathcal{T}_{\pm,t}(\sigma_{<t})),g^{\star,\alpha}_\sigma(\mathcal{T}_{\pm,t}(\sigma_{<t})))+R_{\mathcal{A}}(|B|)\right]\end{aligned}$$

The last inequality follows from the fact that $\mathcal{A}$ is an online learner for $\psi\circ d_1$ with regret bound $R_{\mathcal{A}}(T)$ and is updated using a stream labeled by $g^{\star,\alpha}$ only when $B_t=1$. Now, we can upperbound:

$$\begin{aligned}&\frac{T}{T^\beta}\mathbb{E}\left[\sum_{t:B_t=1}\psi\circ d_1(g^\star(\mathcal{T}_{\pm,t}(\sigma_{<t})),g^{\star,\alpha}_\sigma(\mathcal{T}_{\pm,t}(\sigma_{<t})))\right]+\frac{T}{T^\beta}\mathbb{E}\left[R_{\mathcal{A}}(|B|)\right]\\&\qquad\le\frac{T}{T^\beta}\mathbb{E}\left[\sum_{t:B_t=1}\psi(\alpha)\right]+\frac{T}{T^\beta}\mathbb{E}\left[R_{\mathcal{A}}(|B|)\right]\\&\qquad\le\frac{T}{T^\beta}\mathbb{E}\left[\alpha L|B|\right]+\frac{T}{T^\beta}\mathbb{E}\left[R_{\mathcal{A}}(|B|)\right]\\&\qquad=\alpha LT+\frac{T}{T^\beta}\mathbb{E}\left[R_{\mathcal{A}}(|B|)\right]\end{aligned}$$

The first two inequalities follow from the fact that ψ is monotonic, L-Lipschitz, $\psi(0) = 0$, and $d_1(g^\star(\mathcal{T}_{\pm,t}(\sigma_{<t})), g_\sigma^{\star,\alpha}(\mathcal{T}_{\pm,t}(\sigma_{<t})) \leq \alpha$. Putting things together, we find that

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\mathcal{Q}(\mathcal{T}_{\pm,t}(\sigma_{<t})) \neq \sigma_t\}\right] &\leq \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{E_{B,\phi_B^{g_\sigma^\star}}(\mathcal{T}_{\pm,t}(\sigma_{<t})) \neq \sigma_t\}\right] + \mathbb{E}\left[\sqrt{2T\ln(|\mathcal{E}_B|)}\right] \\
&\leq \frac{\alpha LT + \frac{T}{T^\beta}\mathbb{E}\left[R_{\mathcal{A}}(|B|)\right]}{\psi(\frac{\gamma}{2})} + \mathbb{E}\left[\sqrt{2T\ln(|\mathcal{E}_B|)}\right] \\
&\leq \frac{\alpha LT + \frac{T}{T^\beta}\mathbb{E}\left[R_{\mathcal{A}}(|B|)\right]}{\psi(\frac{\gamma}{2})} + \mathbb{E}\left[\sqrt{2T|B|\ln(\frac{2}{\alpha})}\right].
\end{aligned}$$

where the last inequality follows from the fact that that $|\mathcal{E}_B| \leq (\frac{2}{\alpha})^{|B|}$. By Jensen's inequality, we further get that, $\mathbb{E}\left[\sqrt{2T|B|\ln(\frac{2}{\alpha})}\right] \leq \sqrt{2T^{\beta+1}\ln(\frac{2}{\alpha})}$, which implies that

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\mathcal{Q}(\mathcal{T}_{\pm,t}(\sigma_{<t})) \neq \sigma_t\}\right] \leq \frac{\alpha LT + \frac{T}{T^\beta}\mathbb{E}\left[R_{\mathcal{A}}(|B|)\right]}{\psi(\frac{\gamma}{2})} + \sqrt{2T^{\beta+1}\ln(\frac{2}{\alpha})}.$$

Next, by Lemma 14, there exists a concave sublinear function $\overline{R}_{\mathcal{A}}(|B|)$ that upperbounds $R_{\mathcal{A}}(|B|)$. By Jensen's inequality, we obtain $\mathbb{E}[\overline{R}_{\mathcal{A}}(|B|)] \leq \overline{R}_{\mathcal{A}}(T^\beta)$, which yields

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\mathcal{Q}(\mathcal{T}_{\pm,t}(\sigma_{<t})) \neq \sigma_t\}\right] \leq \frac{\alpha LT + \frac{T}{T^\beta}\overline{R}_{\mathcal{A}}(T^\beta)}{\psi(\frac{\gamma}{2})} + \sqrt{2T^{\beta+1}\ln(\frac{2}{\alpha})}.$$

Picking $\alpha = \frac{1}{LT}$ and $\beta \in (0,1)$, gives that $\mathcal{Q}$ enjoys sublinear expected regret

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\mathcal{Q}(\mathcal{T}_{\pm,t}(\sigma_{<t})) \neq \sigma_t\}\right] \leq \frac{1}{\psi(\frac{\gamma}{2})} + \frac{T}{\psi(\frac{\gamma}{2})T^\beta}\overline{R}_{\mathcal{A}}(T^\beta) + \sqrt{4T^{\beta+1}\ln(LT)}.$$

Since $\overline{R}_{\mathcal{A}}(T^\beta)$ is sublinear in T^β, $\mathcal{Q}$ is a realizable online learner for $\mathcal{H}$ with *sublinear* regret. Thus, for a sufficiently large T, $\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\mathcal{Q}(\mathcal{T}_{\pm,t}(\sigma_{<t})) \neq \sigma_t\}\right] < \frac{T}{2}$. This is a contradiction because $\{(\mathcal{T}_{\pm,t}(\sigma_{<t}), \sigma_t)\}_{t=1}^T$ is a realizable sequence of instances corresponding to a root-to-leaf path in $\mathcal{T}_\pm$ chosen *uniformly at random* and thus any realizable online learner must suffer expected regret at least $\frac{T}{2}$. Thus, if there exists a scale $\gamma > 0$ such that $\mathrm{fat}_\gamma^{\mathrm{seq}}(\mathcal{G}) = \infty$, there cannot exist an online learner for $\mathcal{G}$ with respect to $\psi \circ d_1$. ■