Multiclass Online Learnability under Bandit Feedback

Ananth Raman ANANTHSRAMAN2007@GMAIL.COM
Bridgewater, New Jersey

Vinod Raman[*] VKRAMAN@UMICH.EDU
University of Michigan

Unique Subedi[*] SUBEDI@UMICH.EDU
University of Michigan

Idan Mehalel[*] IDANMEHALEL@GMAIL.COM
Technion

Ambuj Tewari TEWARIA@UMICH.EDU
University of Michigan

Editors: Claire Vernade and Daniel Hsu

Abstract

We study online multiclass classification under bandit feedback. We extend the results of Daniely and Helbertal (2013) by showing that the finiteness of the Bandit Littlestone dimension is necessary and sufficient for bandit online learnability even when the label space is unbounded. Moreover, we show that, unlike the full-information setting, sequential uniform convergence is necessary but not sufficient for bandit online learnability. Our result complements the recent work by Hanneke, Moran, Raman, Subedi, and Tewari (2023) who show that the Littlestone dimension characterizes online multiclass learnability in the full-information setting even when the label space is unbounded.

Keywords: Online Learnability, Bandit Feedback, Multiclass Classification

1. Introduction

In the standard online multiclass classification model, a learner plays a repeated game against an adversary. In each round $t \in [T]$, an adversary picks a labeled instance $(x_t, y_t) \in \mathcal{X} \times \mathcal{Y}$ and reveals x_t to the learner. Using access to a hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, the learner makes a possibly random prediction $\hat{y}_t \in \mathcal{Y}$. The adversary then reveals the true label y_t and the learner then suffers the loss $\mathbb{1}\{y_t \neq \hat{y}_t\}$. Overall, the goal of the learner is to output predictions such that its expected cumulative loss is not too much larger than the smallest cumulative loss amongst all fixed hypothesis in $\mathcal{H}$. This standard setting of online multiclass classification is commonly referred to as the *full-information* setting because the learner gets to observe the true label y_t at the end of each round. Perhaps a more practical setting is the *bandit* feedback setting, where the learner does not get to observe the true label at the end of each round, but only the indication $\mathbb{1}\{\hat{y}_t \neq y_t\}$ of whether its prediction was correct or not (Kakade, Shalev-Shwartz, and Tewari, 2008). One application of this setting is online advertising where the advertiser recommends an ad (label) to a user (instance), but only gets to observe whether the user clicked on the ad or not.

[*] Equal contribution

Unlike the full-information setting, where online learnability of a hypothesis class $\mathcal{H}$ has been fully characterized in both the realizable and agnostic settings, less is known about online learnability under bandit feedback. Indeed, the first work on characterizing bandit online learnability is due to Daniely, Sabato, Ben-David, and Shalev-Shwartz (2011). They introduce a dimension named the Bandit Littlestone dimension (BLdim), and show that it exactly characterizes the bandit online learnability of deterministic learners in the realizable setting. Even prior to that, Auer and Long (1999) related the Bandit Littlestone dimension (which is the optimal deterministic mistake bound with bandit feedback) to the multiclass extension of the Littlestone dimension (Ldim) (Littlestone, 1987; Daniely et al., 2011) and showed that $\mathrm{BL}(\mathcal{H}) = O(|\mathcal{Y}| \log(|\mathcal{Y}|)\mathrm{L}(\mathcal{H}))$, where $\mathrm{BL}(\mathcal{H})$ is the BLdim of $\mathcal{H}$, $\mathrm{L}(\mathcal{H})$ is the Ldim of $\mathcal{H}$, and $|\mathcal{Y}|$ denotes the size of the label space $\mathcal{Y}$. Following the work of Daniely et al. (2011), Daniely and Helbertal (2013) studies the price of bandit feedback by quantifying the ratio between optimal error rates of the two feedback models in the realizable and agnostic settings. Using the inequality $\mathrm{L}(\mathcal{H}) \leq \mathrm{BL}(\mathcal{H}) = O(|\mathcal{Y}| \log(|\mathcal{Y}|)\mathrm{L}(\mathcal{H}))$, they infer that BLdim characterizes realizable online learnability under bandit feedback when $|\mathcal{Y}|$ is finite. Later, Long (2017) and Geneson (2021) proved that this upperbound on BLdim is the best possible up to a leading constant. They also found the exact optimal leading constant.

Moving beyond the realizable setting, Daniely and Helbertal (2013) give an agnostic online learner whose expected regret, under bandit feedback, is at most $O\left(\sqrt{\mathrm{L}(\mathcal{H})|\mathcal{Y}|T \log(T|\mathcal{Y}|)}\right)$. As a corollary, when $|\mathcal{Y}|$ is finite, they infer that the BLdim qualitatively characterizes agnostic bandit online learnability. In addition, Daniely and Helbertal (2013) note a gap of $\tilde{O}(\sqrt{|\mathcal{Y}|})$ between their upperbound in the bandit setting and the known lowerbound of $\Omega(\sqrt{\mathrm{L}(\mathcal{H})\,T})$ in the full-information setting (Ben-David, Pál, and Shalev-Shwartz, 2009). Accordingly, they ask whether a tighter quantitative characterization of bandit learnability is possible in the agnostic setting. In fact, it is unclear whether BLdim characterizes bandit online learnability when $|\mathcal{Y}|$ is unbounded.

Along this direction, there has been a recent surge of interest in characterizing learnability when the label space is unbounded. For example, in a recent breakthrough result, Brukhim, Carmon, Dinur, Moran, and Yehudayoff (2022) show that the Daniely-Schwartz (DS) dimension, defined by Daniely and Shalev-Shwartz (2014), characterizes multiclass learnability in the PAC setting even when the label space in unbounded. Following this work, Hanneke, Moran, Raman, Subedi, and Tewari (2023) show that the multiclass extension of the Littlestone dimension, originally proposed by Daniely, Sabato, Ben-David, and Shalev-Shwartz (2011), continues to characterize online multiclass learnability under *full-information feedback* when the label space is unbounded. Motivated by these results, we ask whether the BLdim continues to characterize *bandit online learnability* even when the label space is unbounded. In particular, can the optimal expected regret in the realizable and agnostic settings, under bandit feedback, be expressed as a function of the BLdim without a dependence on $|\mathcal{Y}|$?

In this paper, we resolve this question by showing that the finiteness of BLdim is necessary and sufficient for bandit online learnability, in both the realizable and agnostic settings, even when the label space is unbounded.

Theorem 1 *Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ and $C_{\mathcal{H}} := \sup_{x \in \mathcal{X}} |\{h(x) : h \in \mathcal{H}\}|$. The following statements are equivalent:*

1. *$\mathcal{H}$ is bandit online learnable.*

2. $\mathrm{BL}(\mathcal{H}) < \infty$.

3. $C_{\mathcal{H}} < \infty$ *and* $\mathrm{L}(\mathcal{H}) < \infty$.

We prove $(2) \implies (1)$, $(3) \implies (2)$ in Section 3, and $(1) \implies (3)$ in Section 4. The proof of $(2) \implies (1)$ is given by an agnostic online learner whose expected regret under bandit feedback can be expressed as a function of BLdim without any dependence on $|\mathcal{Y}|$.

Theorem 2 *For any $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, there exists an agnostic online learner whose expected regret, under bandit feedback, is at most*

$$8\sqrt{\mathrm{L}(\mathcal{H})\mathrm{BL}(\mathcal{H})T\log(T)}.$$

Theorem 2 provides an improvement over the upperbound given by Daniely and Helbertal (2013) when $|\mathcal{Y}| \gg \mathrm{BL}(\mathcal{H})$. In fact, the gap between $|\mathcal{Y}|$ and $\mathrm{BL}(\mathcal{H})$ can be arbitrary. Consider the case where $\mathcal{Y} = \mathbb{N}$ but $|\mathcal{H}| < \infty$. Then, its not hard to see that $\mathrm{BL}(\mathcal{H}) \leq |\mathcal{H}|$ but $|\mathcal{Y}| = \infty$.

In addition to characterizing learnability, there has been recent interest in showing a separation between uniform convergence and learnability. For example, Montasser, Hanneke, and Srebro (2019) show that while uniform convergence is sufficient for adverarsially robust PAC learnability, it is not necessary. Likewise, for online mutliclass learning under full-information feedback, Hanneke et al. (2023) give a class that is learnable, but the online analog of uniform convergence (Rakhlin, Sridharan, and Tewari, 2015b), termed Sequential Uniform Convergence (SUC), does not hold. Towards this end, we ask whether there is a separation between SUC and bandit online learnability. We answer this question affirmatively: while SUC is *necessary* for bandit learnability, it is not sufficient.

Theorem 3 *If a hypothesis class is online learnable under bandit feedback, then it enjoys the SUC property. However, there exists a class which satisfies the SUC property, but is not online learnable under bandit feedback.*

Theorem 3 is in contrast to the full information setting where SUC is sufficient, but not necessary for online learnability (Hanneke et al., 2023). We note that Theorem 3 along with Example 1 from Hanneke et al. (2023) also shows a separation in online learnability between the full-information and bandit feedback settings. Figure 1 visualizes the landscape of learnability for online multiclass problems.

2. Preliminaries

Let $\mathcal{X}$ denote the instance space, $\mathcal{Y}$ be the label space, and $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ denote a hypothesis class. In this paper, we place no assumptions on the size of the label space $\mathcal{Y}$. Given an instance $x \in \mathcal{X}$, we let $\mathcal{H}(x) := \{h(x) : h \in \mathcal{H}\}$ denote the projection of $\mathcal{H}$ onto x. As usual, $[N]$ is used to denote $\{1, 2, \ldots, N\}$.

2.1. Online Learning

In online multiclass classification with bandit feedback, 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 to the learner its loss $\mathbb{1}\{\hat{y}_t \neq y_t\}$, but not the true label y_t. Given a

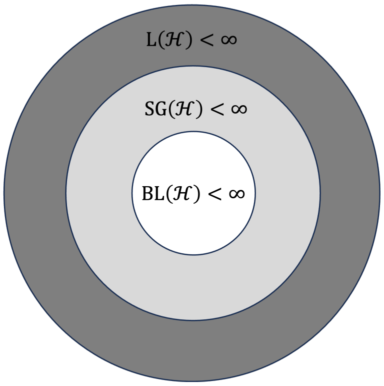
$L(\mathcal{H}) < \infty$
$SG(\mathcal{H}) < \infty$
$BL(\mathcal{H}) < \infty$

Figure 1: Landscape of multiclass online learnability. The Sequential Graph (SG) dimension (see Definition 9) characterizes SUC.

hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, the goal of the learner is to output predictions $\hat{y}_t$ under *bandit feedback* such that its *expected regret*

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\hat{y}_t \neq y_t\} - \inf_{h\in\mathcal{H}} \sum_{t=1}^{T} \mathbb{1}\{h(x_t) \neq y_t\}\right]$$

is small. A hypothesis class $\mathcal{H}$ is said to be bandit online learnable if there exists an algorithm such that for any sequence of labeled examples $(x_1, y_1), ..., (x_T, y_T)$, its expected regret, under bandit feedback, is a sublinear function of T. In this paper, we consider the oblivious setting where the adversary selects the entire sequence of labeled instances $(x_1, y_1), ..., (x_T, y_T)$ before the game begins. Thus, we treat the stream of labeled instances as a non-random, deterministic quantity.

Definition 4 (Bandit Online Learnability) *A hypothesis class $\mathcal{H}$ is bandit online learnable, if there exists an (potentially randomized) algorithm $\mathcal{A}$ such that its* expected regret,

$$R_{\mathcal{A}}(T, \mathcal{H}) := \sup_{(x_1,y_1),...,(x_T,y_T)} \left(\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\mathcal{A}(x_t) \neq y_t\}\right] - \inf_{h\in\mathcal{H}} \sum_{t=1}^{T} \mathbb{1}\{h(x_t) \neq y_t\}\right),$$

while only receiving bandit feedback*, is a non-decreasing sub-linear function of T.*

If it is guaranteed that the learner always observes a sequence of examples labeled by some hypothesis $h \in \mathcal{H}$, then we say that we are in the *realizable* setting.

Littlestone (1987) and Ben-David, Pál, and Shalev-Shwartz (2009) showed that a combinatorial parameter called the Littlestone dimension characterizes online learnability of binary hypothesis classes under full-information feedback, in both the realizable and agnostic settings, respectively. Later, Daniely et al. (2011) defined a multiclass extension of the Littlestone dimension and showed that it tightly characterizes online learnability of multiclass hypothesis classes under full-information feedback in both the realizable and agnostic settings. The Littlestone dimension, in both the binary and multiclass case, is defined in terms of trees, a combinatorial object that captures the temporal dependence inherent in online learning.

Given an instance space $\mathcal{X}$ and a set of objects $\mathcal{M}$, an $\mathcal{X}$-valued, $\mathcal{M}$-ary tree $\mathcal{T}$ of depth T is a complete rooted tree such that (1) each internal node v is labeled by an instance $x \in \mathcal{X}$ and (2) for every internal node v and object $m \in \mathcal{M}$, there is an outgoing edge e_v^m indexed by m. Such a tree can be identified by a sequence $(\mathcal{T}_1, ..., \mathcal{T}_T)$ of labeling functions $\mathcal{T}_t : \mathcal{M}^{t-1} \to \mathcal{X}$ which provide the labels for each internal node. A path of length T is given by a sequence of objects $m = (m_1, ..., m_T) \in \mathcal{M}^T$. Then, $\mathcal{T}_t(m_1, ..., m_{t-1})$ gives the label of the node by following the path $(m_1, ..., m_{t-1})$ starting from the root node, going down the edges indexed by the m_t's. We let $\mathcal{T}_1 \in \mathcal{X}$ denote the instance labeling the root node. For brevity, we define $m_{<t} = (m_1, ..., m_{t-1})$ and therefore write $\mathcal{T}_t(m_1, ..., m_{t-1}) = \mathcal{T}_t(m_{<t})$. Analogously, we let $m_{\leq t} = (m_1, ..., m_t)$. Using this notation, we define the extension of the Littlestone dimension to the multiclass setting proposed by Daniely et al. (2011).

Definition 5 (Littlestone dimension (Littlestone, 1987; Daniely et al., 2011)) *Let $\mathcal{T}$ be a complete, $\mathcal{X}$-valued, $\{\pm 1\}$-ary tree of depth d such that the edges from a single parent node to its child nodes are each labeled with a different element of $\mathcal{Y}$. The tree $\mathcal{T}$ is shattered by $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ if for every path $\sigma = (\sigma_1, ..., \sigma_d) \in \{\pm 1\}^d$, there exists a hypothesis $h_\sigma \in \mathcal{H}$ such that for all $t \in [d]$, $h_\sigma(\mathcal{T}_t(\sigma_{<t})) = y(\sigma_{\leq t})$, where $y(\sigma_{\leq t})$ is the label of the edge between the the nodes $(\mathcal{T}_t(\sigma_{<t}), (\mathcal{T}_{t+1}(\sigma_{\leq t})))$. The Littlestone dimension of $\mathcal{H}$, denoted $\mathrm{L}(\mathcal{H})$, is the maximal depth of a tree $\mathcal{T}$ that is shattered by $\mathcal{H}$. If there exist shattered trees of arbitrarily large depth, we say that $\mathrm{L}(\mathcal{H}) = \infty$.*

In the same work, Daniely et al. (2011) defined a combinatorial parameter called the Bandit Littlestone dimension (BLdim) and showed that it characterizes bandit online learnability of deterministic learners in the realizable setting.

Definition 6 (Bandit Littlestone dimension (Daniely et al., 2011)) *Let $\mathcal{T}$ be a complete, $\mathcal{X}$-valued, $\mathcal{Y}$-ary tree of depth d. The tree $\mathcal{T}$ is shattered by $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ if for every path $y = (y_1, ..., y_d) \in \mathcal{Y}^d$, there exists a hypothesis $h_y \in \mathcal{H}$ such that for all $t \in [d]$, $h_y(\mathcal{T}_t(y_{<t})) \neq y_t$. The Bandit Littlestone dimension of $\mathcal{H}$, denoted $\mathrm{BL}(\mathcal{H})$, is the maximal depth of a tree $\mathcal{T}$ that is shattered by $\mathcal{H}$. If there exist shattered trees of arbitrarily large depth, we say that $\mathrm{BL}(\mathcal{H}) = \infty$.*

In particular, Daniely et al. (2011) show a matching upper and lowerbound on the realizable error rate of deterministic learners in terms of the BLdim.

Theorem 7 (Realizable Learnability (Daniely et al., 2011)) *In the realizable setting, for any $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, there exists a deterministic online learner whose cumulative loss on the worst-case sequence, under bandit feedback, is at most $\mathrm{BL}(\mathcal{H})$. Also, the cumulative loss of any deterministic online learner on the worst-case sequence, under bandit feedback, is at least $\mathrm{BL}(\mathcal{H})$.*

In the agnostic setting, Daniely and Helbertal (2013) gave an upperbound on the expected regret under bandit feedback, in terms of $|\mathcal{Y}|$ and Ldim.

Theorem 8 (Agnostic Learnability (Daniely and Helbertal, 2013)) *For any $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, there exists an online learner $\mathcal{A}$ such that*

$$R_{\mathcal{A}}(T, \mathcal{H}) \leq e\sqrt{\mathrm{L}(\mathcal{H})|\mathcal{Y}|T \log(T|\mathcal{Y}|)}.$$

In Section 3, we show that $|\mathcal{Y}|$ in Theorem 8 can be replaced with $\mathrm{BL}(\mathcal{H})$. Note that this both *qualitatively* and *quantitatively* improves over Theorem 8. Qualitatively, it shows that finite $\mathrm{BL}(\mathcal{H})$ suffices for bandit online learnability, without any requirements on $|\mathcal{Y}|$. Furthermore, there is no better qualitative characterization, since as we show in Section 4, finite $\mathrm{BL}(\mathcal{H})$ is also *necessary* for learnability. A quantitative improvement is achieved in cases where $|\mathcal{Y}| \gg \mathrm{BL}(\mathcal{H})$.

2.2. Online Learnability and Uniform Convergence

The relationship between learnability and uniform convergence has a rich history in learning theory. For binary classification in the PAC setting, the seminal work by Vapnik and Chervonenkis (1974) shows that uniform convergence and PAC learnability are equivalent. Likewise, for online binary classification, an online analog of uniform convergence, termed Sequential Uniform Convergence (SUC), is equivalent to online learnability (Rakhlin, Sridharan, and Tewari, 2015b; Alon, Ben-Eliezer, Dagan, Moran, Naor, and Yogev, 2021). However, this equivalence between uniform convergence and learnability breaks down for multiclass classification. Indeed, in the PAC setting, it was shown that while uniform convergence suffices for multiclass learnability, it is not necessary (Natarajan, 1989). Recently, Hanneke et al. (2023) extended this separation to the online, full-information feedback setting by showing that SUC is sufficient but not necessary for multiclass learnability. Instead, they show that SUC is characterized by a different combinatorial parameter termed the Sequential Graph dimension (SGdim).

Definition 9 (Sequential Graph dimension (Hanneke et al., 2023)) *Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ and $\ell \circ \mathcal{H} = \{(x, y) \mapsto \mathbb{1}\{h(x) \neq y\} : h \in \mathcal{H}\}$ be its loss class. Then, the Sequential Graph dimension of $\mathcal{H}$, denoted $\mathrm{SG}(\mathcal{H})$, is defined as $\mathrm{SG}(\mathcal{H}) = \mathrm{L}(\ell \circ \mathcal{H})$.*

In particular, a hypothesis class $\mathcal{H}$ enjoys the SUC property if and only if its SGdim is finite.

Theorem 10 (Hanneke et al. (2023); Rakhlin et al. (2015a)) *For any hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, the SUC property holds for $\mathcal{H}$ if and only if $\mathrm{SG}(\mathcal{H}) < \infty$.*

In fact, there is a quantitative relation between SGdim and Ldim when the label space is bounded.

Theorem 11 (Hanneke et al. (2023)) *For any $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ such that $|\mathcal{Y}| < \infty$, we have $\mathrm{SG}(\mathcal{H}) = O(\mathrm{L}(\mathcal{H}) \log(|\mathcal{Y}|))$.*

A combination of Theorem 11 and a result due to Alon et al. (2021), Hanneke et al. (2023) derives a new upperbound on the best achievable expected regret under full-information feedback.

Theorem 12 (Hanneke et al. (2023); Alon et al. (2021)) *For any $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ such that $\mathrm{SG}(\mathcal{H}) < \infty$, there exists an online learner whose expected regret under full-information feedback is at most $O(\sqrt{\mathrm{SG}(\mathcal{H})\, T})$.*

In this work, we also investigate the relationship between bandit online learnablity and SUC. In Section 3, we show that finite BLdim implies finite SGdim, and more precisely that $\mathrm{SG}(\mathcal{H}) = O(\mathrm{L}(\mathcal{H}) \log(\mathrm{BL}(\mathcal{H})))$. On the other hand, in Section 4, we exhibit a class where SUC holds, but is not bandit online learnable. Together, these results imply that SUC is necessary, but not sufficient, for bandit online learnability.

3. BLdim is Sufficient for Bandit Online Learnability

In this section we prove Theorem 2, which implies direction $(2) \implies (1)$ in Theorem 1. We also prove $(3) \implies (2)$ towards the end of the section. The first ingredient of this proof is the following result which shows that the BLdim provides a uniform upperbound on the size of the projection of $\mathcal{H}$ on any instance $x \in \mathcal{X}$.

Lemma 13 *For any $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, we have $\sup_{x\in\mathcal{X}} |\mathcal{H}(x)| \leq \mathrm{BL}(\mathcal{H}) + 1$.*

Proof Suppose that $|\mathcal{H}(x)| \geq \mathrm{BL}(\mathcal{H}) + 2$ for some $x \in \mathcal{X}$. We will prove the lemma by contradiction, constructing a BLdim tree of depth $\mathrm{BL}(\mathcal{H}) + 1$ that is shattered by $\mathcal{H}$. Let $\mathcal{T}$ be a BLdim tree of depth $\mathrm{BL}(\mathcal{H}) + 1$ with every internal node labeled by x. Without loss of generality, suppose that $\mathcal{H}(x) = \{1, 2, ..., \mathrm{BL}(\mathcal{H}) + 2\}$, and let there be $h_1, ..., h_{\mathrm{BL}(\mathcal{H})+2} \in \mathcal{H}$ such that $h_i(x) = i$ for all $1 \leq i \leq \mathrm{BL}(\mathcal{H}) + 2$. We now show that $\mathcal{H}$ shatters $\mathcal{T}$. Consider any path down $\mathcal{T}$. Since $\mathcal{T}$ has depth $\mathrm{BL}(\mathcal{H}) + 1$, there can only be $\mathrm{BL}(\mathcal{H}) + 1$ different labels on that path. Since there are at least $\mathrm{BL}(\mathcal{H}) + 2$ hypotheses in $\mathcal{H}$, there is a hypothesis $h_i \in \mathcal{H}$ such that $h_i(x)$ is not equal to any of the labels on the path. Since the path is arbitrary, the tree is shattered by $\mathcal{H}$ according to Definition 6. By contradiction, $|\mathcal{H}(x)| \leq \mathrm{BL}(\mathcal{H}) + 1$ for all $x \in \mathcal{X}$. ■

A uniform upperbound C on the projection size of $\mathcal{H}$ is a strong property: it allows us to effectively reduce the label space from $\mathcal{Y}$ to $[C]$. Lemma 14 makes this precise. For a bandit algorithm $\mathcal{A}$, let $\mathcal{A}(x)$ be its prediction on x, given the history of the game so far (for the sake of readability, we omit the information received prior to instance x from the notation).

Lemma 14 *Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ such that $\sup_{x\in\mathcal{X}} |\mathcal{H}(x)| \leq C$. Then, there exists a hypothesis class $\bar{\mathcal{H}} \subseteq [C]^{\mathcal{X}}$ such that*

(i) $\mathrm{L}(\bar{\mathcal{H}}) = \mathrm{L}(\mathcal{H})$.

(ii) $\mathrm{SG}(\bar{\mathcal{H}}) = \mathrm{SG}(\mathcal{H})$.

(iii) *For every bandit algorithm $\bar{\mathcal{A}}$ for $\bar{\mathcal{H}}$ such that $\bar{\mathcal{A}}(x) \in \bar{\mathcal{H}}(x)$ at all times, there exists a bandit algorithm $\mathcal{A}$ for $\mathcal{H}$ such that $R_{\mathcal{A}}(T, \mathcal{H}) = R_{\bar{\mathcal{A}}}(T, \bar{\mathcal{H}})$ for all T. Furthermore, if $\bar{\mathcal{A}}$ is deterministic, then so is $\mathcal{A}$.*

(iv) $\mathrm{BL}(\bar{\mathcal{H}}) = \mathrm{BL}(\mathcal{H})$.

Proof Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be such that $\sup_{x\in\mathcal{X}} |\mathcal{H}(x)| \leq C$. For every $x \in \mathcal{X}$, define a function $\phi_x : \mathcal{H}(x) \to [|\mathcal{H}(x)|]$ such that ϕ_x is one-to-one. Finally, consider the following hypothesis class $\bar{\mathcal{H}} = \{x \mapsto \phi_x(h(x)) : h \in \mathcal{H}\}$. Clearly, we have that $\bar{\mathcal{H}} \subseteq [C]^{\mathcal{X}}$ and we now show that $\bar{\mathcal{H}}$ also satisfies the four properties above.

Property (i) follows from observing that any non-empty shattered Ldim tree for $\bar{\mathcal{H}}$ can be transformed into a shattered Ldim tree for $\mathcal{H}$, since the the outgoing edges of any internal node labeled by x must be labeled using elements of $[|\mathcal{H}(x)|]$. Thus, the inverse mapping $\phi_x^{-1} : [|\mathcal{H}(x)|] \to \mathcal{H}(x)$ can be used to transform this tree into an Ldim tree of the same depth for $\mathcal{H}$. Likewise, one can use the forward mapping ϕ_x to transform any non-empty shattered Ldim tree for $\mathcal{H}$ into a non-empty shattered Ldim tree for $\bar{\mathcal{H}}$. The equality trivially holds if no non-empty Ldim tree exists.

Property (ii) follows from the fact that every internal node in a non-empty shattered Ldim tree for the loss class $\{(x, y) \mapsto \mathbb{1}\{h(x) \neq y\} : h \in \mathcal{H}\}$ must be labeled using elements of $\{(x, y) : y \in \mathcal{H}(x)\}$. Thus the mapping function ϕ_x can be used to transform any non-empty shattered Ldim tree for the loss class $\{(x, y) \mapsto \mathbb{1}\{h(x) \neq y\} : h \in \mathcal{H}\}$ into a non-empty shattered Ldim tree for the loss class $\{(x, y) \mapsto \mathbb{1}\{\bar{h}(x) \neq y\} : \bar{h} \in \bar{\mathcal{H}}\}$. The reverse direction follows analogously by using the inverse mapping function ϕ_x^{-1}.

To prove property (iii), suppose that $\bar{\mathcal{A}}$ is a bandit algorithm for $\bar{\mathcal{H}}$ such that on any instance $x \in \mathcal{X}$, the prediction of $\bar{\mathcal{A}}$ always lies in $\bar{\mathcal{H}}(x)$. Algorithm 1 uses $\bar{\mathcal{A}}$ in a black-box fashion to construct a bandit learner $\mathcal{A}$ for $\mathcal{H}$.

Algorithm 1 Bandit algorithm $\mathcal{A}$

```
Input: Hypothesis class $\mathcal{H}$, bandit algorithm $\bar{\mathcal{A}}$ for $\bar{\mathcal{H}}$
for $t = 1, ..., T$ do
    Receive example $x_t$
    Query $\bar{y}_t = \bar{\mathcal{A}}(x_t)$
    Predict $\hat{y}_t = \phi_{x_t}^{-1}(\bar{y}_t)$
    Observe loss $\mathbb{1}\{y_t \neq \hat{y}_t\}$ and pass along the indication to $\bar{\mathcal{A}}$
end
```

We claim that the expected regret of Algorithm $\mathcal{A}$ is $R_{\bar{\mathcal{A}}}(T, \bar{\mathcal{H}})$. To see this, fix $T \in \mathbb{N}$ and let $S = (x_1, y_1), \ldots, (x_T, y_T) \in (\mathcal{X} \times \mathcal{Y})^T$ be the sequence of examples to be passed to $\mathcal{A}$. We show that there exists a sequence of examples $S' \in (\mathcal{X} \times [C] \cup \{\star\})^T$ for $\bar{\mathcal{A}}$ such that

$$\min_{h \in \mathcal{H}} \sum_{(x_t, y_t) \in S} \mathbb{1}\{h(x_t) \neq y_t\} = \min_{\bar{h} \in \bar{\mathcal{H}}} \sum_{(x_t, y_t') \in S'} \mathbb{1}\{\bar{h}(x_t) \neq y_t'\}, \tag{1}$$

$$\mathbb{E}\left[\sum_{(x_t, y_t) \in S} \mathbb{1}\{\mathcal{A}(x_t) \neq y_t\}\right] = \mathbb{E}\left[\sum_{(x_t, y_t') \in S'} \mathbb{1}\{\bar{\mathcal{A}}(x_t) \neq y_t'\}\right], \tag{2}$$

and (3) the feedback that $\mathcal{A}$ provides to $\bar{\mathcal{A}}$ matches the feedback that $\bar{\mathcal{A}}$ would have received if it was executed on S'.

Combining (1), (2), and (3) and the regret guarantee $R_{\bar{\mathcal{A}}}(T, \bar{\mathcal{H}})$ for $\bar{\mathcal{A}}$ immediately implies property (iii). It remains to construct S' for which all three statements hold. For every $t \in [T]$, let $y_t' = \phi_{x_t}(y_t)\mathbb{1}\{y_t \in \mathcal{H}(x_t)\} + \star\mathbb{1}\{y_t \notin \mathcal{H}(x_t)\}$. Consider the following stream $S' = (x_1, y_1'), ..., (x_T, y_T') \in (\mathcal{X} \times [C] \cup \{\star\})^T$. To see that (1) holds, observe that for every $h \in \mathcal{H}$ we have that

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{1}\{h(x_t) \neq y_t\} &= \sum_{t: y_t \in \mathcal{H}(x_t)} \mathbb{1}\{h(x_t) \neq y_t\} + \sum_{t: y_t \notin \mathcal{H}(x_t)} \mathbb{1}\{h(x_t) \neq y_t\} \\
&= \sum_{t: y_t \in \mathcal{H}(x_t)} \mathbb{1}\{\phi_{x_t}(h(x_t)) \neq \phi_{x_t}(y_t)\} + \sum_{t: y_t \notin \mathcal{H}(x_t)} \mathbb{1}\{\phi_{x_t}(h(x_t)) \neq \star\} \\
&= \sum_{t=1}^{T} \mathbb{1}\{\bar{h}(x_t) \neq y_t'\}.
\end{aligned}$$

To see (2), note that

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\{\mathcal{A}(x_t)\neq y_t\}\right] &= \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\{\phi_{x_t}^{-1}(\bar{y}_t)\neq y_t\}\right]\\
&= \mathbb{E}\left[\sum_{t:y_t\in\mathcal{H}(x_t)}\mathbb{1}\{\phi_{x_t}^{-1}(\bar{y}_t)\neq y_t\}+\sum_{t:y_t\notin\mathcal{H}(x_t)}\mathbb{1}\{\phi_{x_t}^{-1}(\bar{y}_t)\neq y_t\}\right]\\
&= \mathbb{E}\left[\sum_{t:y_t\in\mathcal{H}(x_t)}\mathbb{1}\{\bar{\mathcal{A}}(x_t)\neq \phi_{x_t}(y_t)\}+\sum_{t:y_t\notin\mathcal{H}(x_t)}\mathbb{1}\{\bar{\mathcal{A}}(x_t)\neq \star\}\right]\\
&= \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\{\bar{\mathcal{A}}(x_t)\neq y_t'\}\right].
\end{aligned}$$

Finally, to prove (3), it suffices to show that $\mathbb{1}\{y_t\neq\hat{y}_t\}=\mathbb{1}\{y_t'\neq\bar{\mathcal{A}}(x_t)\}$. If $\mathbb{1}\{y_t\neq\hat{y}_t\}=0$, then $y_t=\phi_{x_t}^{-1}(\bar{y}_t)$ and $\bar{\mathcal{A}}(x_t)=\phi_{x_t}(y_t)=y_t'$ as needed. If $\mathbb{1}\{y_t\neq\hat{y}_t\}=1$ and $y_t\in\mathcal{H}(x_t)$, then $\bar{y}_t\neq\phi_{x_t}(y_t)=y_t'$. Lastly, if $\mathbb{1}\{y_t\neq\hat{y}_t\}=1$ and $y_t\notin\mathcal{H}(x_t)$, we get that $\bar{y}_t\neq y_t'=\star$ since $\bar{\mathcal{A}}(x_t)\in\bar{\mathcal{H}}(x_t)$. The "furthermore" part of the property is straightforward by the construction of $\mathcal{A}$.

Let us move on to Property (iv). The direction $\mathrm{BL}(\mathcal{H})\leq\mathrm{BL}(\bar{\mathcal{H}})$ follows from Property (iii). Indeed, let $\bar{\mathcal{A}}$ be the optimal BSOA deterministic learner defined in (Daniely et al., 2011) for $\bar{\mathcal{H}}$ under the assumption of realizability. For every round t, the algorithm $\bar{\mathcal{A}}$ never predicts $y\notin\bar{\mathcal{H}}(x_t)$ by its definition. Therefore, by Property (iii) there exists a deterministic learner $\mathcal{A}$ for $\mathcal{H}$ having the same guarantees as of $\bar{\mathcal{A}}$. Therefore $\mathrm{BL}(\mathcal{H})\leq\mathrm{BL}(\bar{\mathcal{H}})$. The reverse direction $\mathrm{BL}(\mathcal{H})\geq\mathrm{BL}(\bar{\mathcal{H}})$ follows by considering the realizable setting and the bandit algorithm for $\bar{\mathcal{H}}$ that, given any instance x_t, passes x_t to the BSOA for $\mathcal{H}$, receives its prediction $\bar{y}_t\in\mathcal{Y}$, makes the prediction $\hat{y}_t=\phi_{x_t}(\bar{y}_t)\in[C]$, and upon receiving the feedback $\mathbb{1}\{\hat{y}_t\neq y_t\}$, passes the same feedback to the BSOA. The same analysis as in Property (iii) can be used to show that this algorithm makes at most $\mathrm{BL}(\mathcal{H})$ mistakes on any realizable stream. ■

In order to use Property (iii) of Lemma 14, we need to construct a bandit learner $\bar{\mathcal{A}}$ which on any instance $x\in\mathcal{X}$, makes a prediction that lies in $\mathcal{H}(x)$ and achieves a sublinear regret bound whenever $\mathrm{BL}(\mathcal{H})<\infty$. Unfortunately, the generic bandit learner witnessing the proof of Theorem 8 does not guarantee that its predictions always lie in the projection of $\mathcal{H}$. Fortunately, the following lemma, whose proof can be found in Appendix A, shows that a slight modification of the bandit learner used to prove Theorem 8 can achieve the same regret bound, while ensuring that the predictions always lie in the projection of $\mathcal{H}$.

Lemma 15 *For any $\mathcal{H}\subseteq\mathcal{Y}^{\mathcal{X}}$, there exists an online learner $\mathcal{A}$ such that*

$$R_{\mathcal{A}}(T,\mathcal{H})\leq e\sqrt{\mathrm{L}(\mathcal{H})|\mathcal{Y}|T\log(T|\mathcal{Y}|)},$$

while ensuring that $\mathcal{A}(x_t)\in\mathcal{H}(x_t)$ almost surely.

We are now ready to prove Theorem 2, which implies that finitness of BLdim is sufficient for bandit online learnability even when the label space is unbounded. This proves direction $(2)\implies(1)$ in Theorem 1.

Proof (of Theorem 2) We first prove a stronger result and then show that Theorem 2 follows. Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be such that $\text{BL}(\mathcal{H}) < \infty$. Then, by Lemmas 14 and 15, there exists an online learner whose expected regret in the agnostic setting under bandit feedback is at most $e\sqrt{\text{L}(\mathcal{H})CT\log(TC)}$ where $C = \sup_{x \in \mathcal{X}} |\mathcal{H}(x)|$. Since Lemma 13 states that $C \leq \text{BL}(\mathcal{H}) + 1 \leq 2\text{BL}(\mathcal{H})$, we can further upperbound the expected regret by $2e\sqrt{\text{L}(\mathcal{H})\text{BL}(\mathcal{H})T\log(T\,\text{BL}(\mathcal{H}))}$. There are now two cases to consider. If $T \leq \text{BL}(\mathcal{H})$, the expected regret of any bandit online learner can be trivially upperbounded by T. Noting that $8\sqrt{\text{L}(\mathcal{H})\text{BL}(\mathcal{H})T\log(T)} \geq T$ when $T \leq \text{BL}(\mathcal{H})$ completes this case. If $T > \text{BL}(\mathcal{H})$, then we can upperbound the expected regret of the bandit online learner by

$$2e\sqrt{\text{L}(\mathcal{H})\text{BL}(\mathcal{H})T\log(T\,\text{BL}(\mathcal{H}))} \leq 2e\sqrt{2\text{L}(\mathcal{H})\text{BL}(\mathcal{H})T\log(T)} \leq 8\sqrt{\text{L}(\mathcal{H})\text{BL}(\mathcal{H})T\log(T)},$$

matching the upperbound given in the statement of Theorem 2. This completes the proof. ■

In online learning theory, upperbounds on the minimax expected regret are traditionally derived in terms of the single combinatorial dimension that characterizes learnability. However, our upperbound in Theorem 2 is in terms of both the Ldim and BLdim. To get a bound depending only on the BLdim, one can trivially use the fact that $\text{L}(\mathcal{H}) \leq \text{BL}(\mathcal{H})$ to get a suboptimal upperbound of $8\,\text{BL}(\mathcal{H})\sqrt{T\log(T)}$ on the minimax expected regret. However, as an intermediate step to our upperbound in Theorem 2, we show that the minimax expected regret can actually be upperbounded by $e\sqrt{\text{L}(\mathcal{H})CT\log(TC)}$, and thus it is natural to ask whether there is an upperbound on $\sqrt{\text{L}(\mathcal{H})C}$ that is significantly better than $\text{BL}(\mathcal{H})$. Unfortunately, the following example shows that this is not the case.

Example 1. Fix $d, C \in \mathbb{N}$. Define the instance space $\mathcal{X} = \{x_0, ..., x_d\}$ and the label space $\mathcal{Y} = \{0, ..., C-1\}$. Let $\mathcal{H}_1 = \{0,1\}^{\{x_1,...,x_d\}}$ and $\mathcal{H}_2 = \{x \mapsto y\,\mathbb{1}\{x = x_0\} : y \in \mathcal{Y}\}$. Consider the hypothesis class $\mathcal{H} = \mathcal{H}_1 \cup \mathcal{H}_2$. Clearly, $\text{L}(\mathcal{H}) \geq \text{L}(\mathcal{H}_1) = \text{BL}(\mathcal{H}_1) = d$. Moreover, $\text{BL}(\mathcal{H}_2) \leq C - 1$. We now give an upperbound on $\text{BL}(\mathcal{H})$ by constructing a deterministic learner for $\mathcal{H}$. Consider the learning algorithm that predicts 0 until its first mistake, removes inconsistent hypotheses, and plays the Bandit Standard Optimal Algorithm (BSOA) from Daniely et al. (2011) on future rounds. We now show that this algorithm makes at most $1 + \max\{d, C-1\}$ mistakes on any realizable stream. There are two cases to consider. Suppose the algorithm makes its first mistake on x_0. Then, by construction of $\mathcal{H}$, the true hypothesis must be in $\mathcal{H}_2$ and thus the BSOA makes no more than $\text{BL}(\mathcal{H}_2) \leq C - 1$ mistakes in all future rounds. On the other hand, if the algorithm makes its first mistake on $x \in \{x_1, ..., x_d\}$, then the true hypothesis must be in $\mathcal{H}_1$ and thus the BSOA makes at most $\text{BL}(\mathcal{H}_1) = d$ mistakes on all future rounds. Overall, the algorithm makes at most $1 + \max\{d, C-1\}$ mistakes. Since the BLdim lowerbounds the number of mistakes made by any deterministic learner under bandit feedback, we must have that $\text{BL}(\mathcal{H}) \leq 1 + \max\{d, C-1\}$. Taking $C = d + 1$, we have that $\text{BL}(\mathcal{H}) \leq 1 + d \leq 1 + \sqrt{\text{L}(\mathcal{H})C}$, which completes the example.

We leave it as an interesting open question to derive optimal lower and upper bounds on the minimax expected regret in terms of only the BLdim (see Section 5). Lemma 14 can also be used to sharpen the relationship between BLdim and Ldim. In particular, due to (Auer and Long, 1999; Daniely and Helbertal, 2013; Long, 2017), there exists a *deterministic* online learner in the realizable setting whose number of mistakes, under bandit feedback, is at most $O(\text{L}(\mathcal{H})|\mathcal{Y}|\log(|\mathcal{Y}|))$. Since the BLdim lowerbounds the number of mistakes made by any deterministic online learner in the realizable setting, Lemma 14 immediately implies that when $\sup_{x \in \mathcal{X}} |\mathcal{H}(x)| \leq C$, we have

$\mathrm{BL}(\mathcal{H}) = O(\mathrm{L}(\mathcal{H})C \log C)$, proving direction $(3) \implies (2)$ in Theorem 1. In Section 4, we show that finiteness of both C and $\mathrm{L}(\mathcal{H})$ is also necessary for learnability (direction $(1) \implies (3)$).

We end this section with Corollary 16, which shows that SUC is necessary for a hypothesis class to be bandit online learnable.

Corollary 16 *If* $\mathrm{BL}(\mathcal{H}) < \infty$, *then* $\mathrm{SG}(\mathcal{H}) = O(\mathrm{L}(\mathcal{H}) \log(\mathrm{BL}(\mathcal{H})))$.

Proof Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ such that $\mathrm{BL}(\mathcal{H}) < \infty$. Then by Lemmas 13 and 14, there exists a class $\bar{\mathcal{H}} \subseteq [\mathrm{BL}(\mathcal{H}) + 1]^{\mathcal{X}}$ such that $\mathrm{L}(\bar{\mathcal{H}}) = \mathrm{L}(\mathcal{H})$ and $\mathrm{SG}(\bar{\mathcal{H}}) = \mathrm{SG}(\mathcal{H})$. Since $\mathrm{BL}(\mathcal{H}) + 1 < \infty$, Theorem 11 implies that $\mathrm{SG}(\bar{\mathcal{H}}) = O(\mathrm{L}(\bar{\mathcal{H}}) \log(\mathrm{BL}(\bar{\mathcal{H}}))) = O(\mathrm{L}(\mathcal{H}) \log(\mathrm{BL}(\mathcal{H})))$. ■

Since $\mathrm{BL}(\mathcal{H}) < \infty$ implies that $\mathrm{L}(\mathcal{H}) < \infty$, Corollary 16 and Theorem 10 taken together prove the first half of Theorem 3, showing that $\mathcal{H}$ enjoys SUC when $\mathrm{BL}(\mathcal{H}) < \infty$. Moreover, when $\mathrm{BL}(\mathcal{H}) < \infty$, Corollary 16 along with Theorem 12 implies a slightly sharper upperbound on the optimal expected regret in the agnostic setting under *full-information* feedback.

Corollary 17 *Let* $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ *such that* $\mathrm{BL}(\mathcal{H}) < \infty$. *Then, there exists an agnostic online learner whose expected regret, under* full-information feedback, *is at most*

$$O\left(\sqrt{\mathrm{L}(\mathcal{H}) T \log(\mathrm{BL}(\mathcal{H}))}\right).$$

Proof Let $\mathcal{H}$ be such that $\mathrm{BL}(\mathcal{H}) < \infty$. Then, by Corollary 16, $\mathrm{SG}(\mathcal{H}) = O(\mathrm{L}(\mathcal{H}) \log(\mathrm{BL}(\mathcal{H})))$. Also, by Theorem 12, we have that under full-information feedback, there exists a online learner whose expected regret is at most $O(\sqrt{T\,\mathrm{SG}(\mathcal{H})})$. Combining these two results gives the stated claim. ■

Namely, Corollary 17 improves upon the upperbound on expected regret given by (Hanneke et al., 2023, Theorem 1) by replacing the $\log(\frac{T}{\mathrm{L}(\mathcal{H})})$ factor with $\log(\mathrm{BL}(\mathcal{H}))$.

4. Finite BLdim is Necessary for Bandit Online Learnability

In this section, we complement the results of Section 3, and deduce that finiteness of BLdim is necessary for bandit online learnability in the realizable setting even when the label space is unbounded. Since agnostic learnability implies realizable learnability, this also implies that finiteness of the BLdim is necessary for agnostic learnability, completing the proof of the direction $(1) \implies (2)$ in Theorem 1. This will also imply $(1) \implies (3)$, which completes the proof of Theorem 1.

Lemma 18 *Let* $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ *and* $C = \sup_{x \in \mathcal{X}} |\mathcal{H}(x)|$. *Then, for every bandit online learner* $\mathcal{A}$:

1. *There exists a realizable stream with expected regret at least* $\frac{\mathrm{BL}(\mathcal{H})}{4C \log C}$ *if* $T \geq \mathrm{L}(\mathcal{H})$ *and at least* $T/2$ *otherwise.*

2. *There exists a realizable stream with expected regret at least* $\frac{C-1}{2}$ *if* $T \geq C$, *and at least* $\frac{T-1}{2}$ *otherwise.*

Proof Let us start with the first item. A well-known result by Ben-David et al. (2009) states that there exists a realizable stream of length $T = \mathrm{L}(\mathcal{H})$ such that in expectation, $\mathcal{A}$ makes at least $\mathrm{L}(\mathcal{H})/2$ mistakes under full information feedback. On the other hand, by Long (2017) and Lemma 14 we have $\mathrm{BL}(\mathcal{H}) \leq 2\mathrm{L}(\mathcal{H})C \log C$, implying the item for the case $T \geq \mathrm{L}(\mathcal{H})$. If $T < \mathrm{L}(\mathcal{H})$, we employ the lower bound on T instead of on $\mathrm{L}(\mathcal{H})$, concluding this item. The second item follows immediately from (Daniely and Helbertal, 2013, Claim 2). ■

Lemma 18 implies that finiteness of BLdim is necessary for bandit online learnability in the realizable setting. Recall that $\mathrm{BL}(\mathcal{H}) \geq C - 1$ due to Lemma 13. Now, if $C = \infty$ (where $C := \sup_{x \in \mathcal{X}} |\mathcal{H}(x)|$), then Lemma 18 implies that the expected regret of any online learner under bandit feedback and in the realizable setting, is at least $\frac{T-1}{2}$, a linear function of T. On the other hand, if $\mathrm{BL}(\mathcal{H}) = \infty$ and $C < \infty$, then the bound $\mathrm{BL}(\mathcal{H}) = O(\mathrm{L}(\mathcal{H})C \log C)$ implies that $\mathrm{L}(\mathcal{H}) = \infty$, and then Lemma 18 implies a lowerbound of $\frac{T}{2}$ on the expected regret. This proves the direction (1) $\implies$ (2) in Theorem 1. Using the fact that $\mathrm{BL}(\mathcal{H}) \geq \mathrm{L}(\mathcal{H})$ and Lemma 13 shows that (2) $\implies$ (3), completing the proof of Theorem 1.

Furthermore, if C is a constant, then taken together with Theorem 7, Lemma 18 implies that the BLdim characterizes the optimal expected mistake bound of randomized learners in the realizable setting up to constant factors. In the agnostic setting, the full-information lowerbound of $\sqrt{\frac{\mathrm{L}(\mathcal{H})T}{8}}$ on the expected regret can also be a tight lowerbound under bandit feedback up to logarithmic factors in T. For example, for every class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ such that $\sup_{x \in \mathcal{X}} |\mathcal{H}(x)| \leq 2$, Theorem 8 and Lemma 14 imply the existence of a bandit online learner whose expected regret is at most $8\sqrt{\mathrm{L}(\mathcal{H})T \log(T)}$.

Finally, Lemma 18 together with Lemma 19 shows that neither the finitness of Ldim nor the finiteness of SGdim is sufficient for bandit online learnability.

Lemma 19 *Let* $\mathcal{X} = \{0\}$, $\mathcal{Y} = \mathbb{N}$ *and* $\mathcal{H} = \{h_a : a \in \mathbb{N}\}$ *where* $h_a(0) = a$. *Then,* $\mathrm{L}(\mathcal{H}) = \mathrm{SG}(\mathcal{H}) = 1$ *but* $\mathrm{BL}(\mathcal{H}) = \infty$.

Proof The equality $\mathrm{BL}(\mathcal{H}) = \infty$ follows from the fact that $|\mathcal{H}(0)| = \infty$ and Lemma 13. We have $\mathrm{L}(\mathcal{H}) = 1$ because for any labeled instance $(0, y) \in \mathcal{X} \times \mathcal{Y}$, there only exists one hypothesis $h \in \mathcal{H}$ such that $h(0) = y$ (namely h_y). Lastly, $\mathrm{SG}(\mathcal{H}) = 1$ because for any labeled instance $(0, y) \in \mathcal{X} \times \mathcal{Y}$ there exists only one function in the loss class $\{(0, y) \mapsto \mathbb{1}\{h(0) \neq y\} : h \in \mathcal{H}\}$ that achieves loss 0. ■

Lemma 19 completes the proof of Theorem 3, since we have exhibited a class for which SUC holds but is not bandit online learnable.

5. Discussion and Open Questions

In this paper, we revisited multiclass online learnability under bandit feedback and showed that, when $\mathcal{Y}$ is unbounded: (1) the Bandit Littlestone dimension, originally proposed by Daniely et al. (2011), continues to characterize bandit online learnability, and (2) while SUC is necessary for bandit online learnability, it is not sufficient.

Moving forward, there are still many interesting open questions. By Theorem 2, in the agnostic setting there is a gap of $\sqrt{\mathrm{BL}(\mathcal{H}) \log(T)}$ between the upper and lowerbounds on the optimal expected regret under bandit feedback. Is this gap between the upper and lowerbound unavoidable?

Using the fact that $\text{BL}(\mathcal{H}) \leq 4C \log(C)\text{L}(\mathcal{H})$, one can get a lowerbound of $\Omega\left(\sqrt{\frac{\text{BL}(\mathcal{H})\,T}{C\log(C)}}\right)$ on the expected regret in the agnostic setting, where $C = \sup_{x \in \mathcal{X}} |\mathcal{H}(x)|$. Is it possible to remove the dependence on C and improve this lowerbound to $\Omega(\sqrt{\text{BL}(\mathcal{H})\,T})$?

While the BLdim provides a sharp quantitative characterization of deterministic learnability in the realizable setting, it is unclear whether it provides a tight *quantitative* characterization of randomized learnability in both the realizable and agnostic settings. Recently, Filmus, Hanneke, Mehalel, and Moran (2023) gave a combinatorial parameter called the Randomized Littlestone dimension and showed that it exactly quantifies the optimal expected mistake bound for randomized learners in the realizable setting under full-information feedback. Is there a modification of this dimension that can exactly quantify the optimal expected mistake bound for randomized learners in the realizable setting under *bandit feedback*? Can such a dimension also be used to give a sharper upperbound on the expected regret in the agnostic setting?

Acknowledgments

AT acknowledges the support of NSF via grant IIS-2007055. VR acknowledges the support of the NSF Graduate Research Fellowship.

References

Noga Alon, Omri Ben-Eliezer, Yuval Dagan, Shay Moran, Moni Naor, and Eylon Yogev. Adversarial laws of large numbers and optimal regret in online classification. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 447–455, 2021.

Peter Auer and Philip M Long. Structural results about on-line learning models with and without queries. *Machine Learning*, 36:147–181, 1999.

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

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

Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

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 *Conference on Learning Theory*, pages 287–316. PMLR, 2014.

Amit Daniely, Sivan Sabato, Shai Ben-David, and Shai Shalev-Shwartz. Multiclass learnability and the erm principle. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 207–232. JMLR Workshop and Conference Proceedings, 2011.

Yuval Filmus, Steve Hanneke, Idan Mehalel, and Shay Moran. Optimal prediction using expert advice and randomized littlestone dimension. In *COLT*, volume 195 of *Proceedings of Machine Learning Research*, pages 773–836. PMLR, 2023.

Jesse Geneson. A note on the price of bandit feedback for mistake-bounded online learning. *Theoretical Computer Science*, 874:42–45, 2021.

Steve Hanneke, Shay Moran, Vinod Raman, Unique Subedi, and Ambuj Tewari. Multiclass online learning and uniform convergence. *Proceedings of the 36th Annual Conference on Learning Theory (COLT)*, 2023.

Sham M Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. Efficient bandit algorithms for online multiclass prediction. In *Proceedings of the 25th international conference on Machine learning*, pages 440–447, 2008.

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

Philip M Long. New bounds on the price of bandit feedback for mistake-bounded online multiclass learning. In *International Conference on Algorithmic Learning Theory*, pages 3–10. PMLR, 2017.

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.

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.

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.

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

Appendix A. Proof of Lemma 15

To prove Lemma 15, we slightly modify the generic agnostic learner witnessing the proof of Theorem 8. Recall that the agnostic learner in Theorem 8 first constructs a sufficiently small set of experts E such that for every hypothesis $h \in \mathcal{H}$, there exists an expert $\mathcal{E}_h \in E$ whose predictions exactly match h over the stream. Then, the learner runs the non-mixing version of EXP4 (see Figure 4.1 and Theorem 4.2 in Bubeck and Cesa-Bianchi (2012)) with this set of experts E on the stream, for an appropriately chosen learning rate. Unfortunately, in all rounds $t \in [T]$, some of the experts constructed by this learner output predictions lying outside of $\mathcal{H}(x_t)$. Thus, EXP4 with this set of experts does not satisfy the constraint imposed by Lemma 15, that its predictions on x_t must lie in $\mathcal{H}(x_t)$. To fix this issue, we modify each expert $\mathcal{E} \in E$ such that for every $t \in [T]$, we have that $\mathcal{E}(x_t) \in \mathcal{H}(x_t)$ while still maintaining the property of the expert set of Daniely and Helbertal (2013): for every $h \in \mathcal{H}$, there exists an expert $\mathcal{E}_h \in E$ that predicts exactly like h over the stream. Our modification is simple: in contrast to the experts constructed by Daniely and Helbertal (2013), our experts may predict using the "covering function" ϕ (as defined in Daniely and Helbertal (2013)) only if its value lies in $\mathcal{H}(x_t)$. Running the EXP4 algorithm using this new set of experts gives the claimed regret guarantee. We now formalize this construction.

Let $(x_1, y_1), ..., (x_T, y_T) \in (\mathcal{X} \times \mathcal{Y})^T$ denote the stream of instances to be observed by the learner and $h^\star \in \arg\min_{h \in \mathcal{H}} \sum_{t=1}^{T} \mathbb{1}\{h(x_t) \neq y_t\}$ denote the optimal hypothesis in hindsight. As stated before, our high-level strategy will be to construct a set of experts E and then run EXP4 using E over the stream. Crucially, we will guarantee that $\mathcal{E}(x_t) \in \mathcal{H}(x_t)$ for every $\mathcal{E} \in E$.

Given the time horizon T, let $L_T = \{L \subset [T]; |L| \leq \mathrm{L}(\mathcal{H})\}$ denote the set of all possible subsets of $[T]$ of size at most $\mathrm{L}(\mathcal{H})$. For every $L \in L_T$, let $\phi : L \to \mathcal{Y}$ denote a function mapping time points in L to a label in $\mathcal{Y}$. Let $\Phi_L = \mathcal{Y}^L$ denote all such functions ϕ. For each $L \in L_T$ and $\phi \in \Phi_L$, we define an expert $\mathcal{E}_{L,\phi}$. As presented below in Algorithm 2, expert $\mathcal{E}_{L,\phi}$ uses the Standard Optimal Algorithm (SOA) (Littlestone, 1987) to make its prediction in rounds t where $t \notin L$. When $t \in L$, there are two cases. If $\phi(t) \in \mathcal{H}(x_t)$, the expert $\mathcal{E}_{L,\phi}$ uses the function ϕ to compute a labeled instance to predict and update the SOA with. Otherwise, the expert chooses an arbitrary label in $\mathcal{H}(x_t)$ to predict and update SOA with. Let $E = \bigcup_{L \in L_T} \bigcup_{\phi \in \Phi_L} \mathcal{E}_{L,\phi}$ denote the set of all Experts parameterized by subsets $L \in L_T$ and $\phi \in \Phi_L$. Crucially, observe that by definition of SOA, for every time point $t \in [T]$ and expert $\mathcal{E} \in E$, it holds that $\mathcal{E}(x_t) \in \mathcal{H}(x_t)$. Finally, note that $|E| \leq (T|\mathcal{Y}|)^{\mathrm{L}(\mathcal{H})}$.

Algorithm 2 Expert $\mathcal{E}_{L,\phi}$

```
Input: Independent copy of SOA
for $t = 1, ..., T$ do
    Receive example $x_t$
    Let $\tilde{y}_t = \mathrm{SOA}(x_t)$
    if $t \in L$ and $\phi(t) \in \mathcal{H}(x_t)$ then
        Predict $\hat{y}_t = \phi(t)$
    else if $t \in L$ and $\phi(t) \notin \mathcal{H}(x_t)$ then
        Predict arbitrary label $\hat{y}_t \in \mathcal{H}(x_t)$
    else
        Predict $\hat{y}_t = \tilde{y}_t$
    Update SOA by passing $(x_t, \hat{y}_t)$
end
```

We claim that there exists an expert $\mathcal{E}_{L^\star,\phi^\star} \in E$ such that $h^\star(x_t) = \mathcal{E}_{L^\star,\phi^\star}(x_t)$ for all $t \in [T]$. To see this, consider the hypothetical stream of instances labeled by the optimal hypothesis $S^\star = (x_1, h^\star(x_1)), ..., (x_T, h^\star(x_T))$. Let $L^\star = \{t_1, t_2, ...\}$ be the indices on which the SOA algorithm would have made a mistake had it run on S^*. By the guarantees of the SOA (Littlestone, 1987), we have that $|L^\star| \leq \mathrm{L}(\mathcal{H})$. Consider the function $\phi^\star : L^\star \to \mathcal{Y}$ such that for all $t \in L^\star$, we have $\phi^\star(t) = h^\star(x_t)$. By construction of E, there exists an expert $\mathcal{E}_{L^\star,\phi^\star} \in E$ parameterized by $L^\star$ and $\phi^\star$. We claim that for all $t \in [T]$, we have $\mathcal{E}_{L^\star,\phi^\star}(x_t) = h^\star(x_t)$. This follows by observing that $\mathcal{E}_{L^\star,\phi^\star}$ predicts and updates its copy of SOA using exactly the stream of instances labeled by $h^\star$. Since by definition of $L^\star$ the predictions of SOA match that of $h^\star$ outside of $L^\star$, we have $\mathcal{E}_{L^\star,\phi^\star}(x_t) = \mathrm{SOA}(x_t) = h^\star(x_t)$ for all $t \notin L^\star$. Moreover, for those time points $t \in L^\star$, we have that $\mathcal{E}_{L^\star,\phi^\star}(x_t) = \phi^\star(t) = h^\star(x_t)$ by definition of $\phi^\star(t)$. Thus, for all $t \in [T]$, we have that $\mathcal{E}_{L^\star,\phi^\star}(x_t) = h^\star(x_t)$.

Consider the agnostic online learner $\mathcal{A}$ that runs the non-mixing version of EXP4 (see Fig. 4.1 and Theorem 4.2 in Bubeck and Cesa-Bianchi (2012)) using the set of experts E with learning rate $\eta = \sqrt{\frac{\ln |E|}{T|\mathcal{Y}|}}$. By the guarantees of the EXP4 algorithm, it follows that

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\mathcal{A}(x_t) \neq y_t\}\right] &\leq \inf_{\mathcal{E}\in E} \sum_{t=1}^{T} \mathbb{1}\{\mathcal{E}(x_t) \neq y_t\} + e\sqrt{T|\mathcal{Y}| \ln |E|} \\ &\leq \sum_{t=1}^{T} \mathbb{1}\{\mathcal{E}_{L^\star,\phi^\star}(x_t) \neq y_t\} + e\sqrt{\mathrm{L}(\mathcal{H})|\mathcal{Y}|T \ln(T|\mathcal{Y}|)} \\ &= \sum_{t=1}^{T} \mathbb{1}\{h^\star(x_t) \neq y_t\} + e\sqrt{\mathrm{L}(\mathcal{H})|\mathcal{Y}|T \ln(T|\mathcal{Y}|)}.\end{aligned}$$

Finally, observing that for all $t \in [T]$, $\cup_{\mathcal{E}\in E}\{\mathcal{E}(x_t)\} \subseteq \mathcal{H}(x_t)$ together with the fact that EXP4 algorithm in Figure 4.1 of Bubeck and Cesa-Bianchi (2012) samples a label using a distribution supported only over $\cup_{\mathcal{E}\in E}\{\mathcal{E}(x_t)\}$ ensures that $\mathcal{A}(x_t) \in \mathcal{H}(x_t)$ almost surely (equivalently, the EXP4 algorithm samples an expert $\mathcal{E} \in E$ and uses its prediction). This completes the proof of Lemma 15.