On Computable Online Learning

Niki Hasrati NHASRATI@UWATERLOO.CA
Cheriton School of Computer Science, University of Waterloo

Shai Ben-David SHAI@UWATERLOO.CA
Cheriton School of Computer Science, University of Waterloo and Vector Institute

Editors: Shipra Agrawal and Francesco Orabona

Abstract

We initiate a study of computable online (c-online) learning, which we analyze under varying requirements for "optimality" in terms of the mistake bound. Our main contribution is to give a necessary and sufficient condition for optimal c-online learning and show that the Littlestone dimension no longer characterizes the optimal mistake bound of c-online learning. Furthermore, we introduce anytime optimal (a-optimal) online learning, a more natural conceptualization of "optimality" and a generalization of Littlestone's Standard Optimal Algorithm. We show the existence of a computational separation between a-optimal and optimal online learning, proving that a-optimal online learning is computationally more difficult. Finally, we consider online learning with no requirements for optimality, and show, under a weaker notion of computability, that the finiteness of the Littlestone dimension no longer characterizes whether a class is c-online learnable with finite mistake bound. A potential avenue for strengthening this result is suggested by exploring the relationship between c-online and CPAC learning, where we show that c-online learning is as difficult as improper CPAC learning.

Keywords: computability, online learning, Littlestone dimension

1. Introduction

Motivated by recent work on computable PAC (CPAC) learning (Agarwal et al., 2020, 2021; Sterkenburg, 2022), we initiate a study of computable online (c-online) learning, where learners and their output hypotheses are required to be computable. As stated in Littlestone's seminal paper (1988, p. 289), the original definition of online learning was limited to finite domains and hypothesis classes to avoid "computability issues." Although Littlestone's results are easily extendable to the infinite setting (see Shalev-Shwartz and Ben-David, 2014, Chapter 21), an implicit assumption is that learners are functions, not necessarily computable, that map input samples to output hypotheses. Indeed, this assumption is implicit in many recent advances in online learning—for example, the equivalence between online learning and differentially private PAC learning (Alon et al., 2022) and the characterizations of proper online learning (Chase and Freitag, 2020; Hanneke et al., 2021) and agnostic online learning (Ben-David et al., 2009). A further motivation for the study of computable learning stems from recent work on the undecidability of learning, where the authors state that "the source of the problem is in defining learnability as the existence of a learning function rather than the existence of a learning algorithm" (Ben-David et al., 2019, p. 48).

A key result in online learning is that the Littlestone dimension characterizes the mistake bound of optimal online learners (Littlestone, 1988, Theorem 3). It is therefore natural to ask whether this fundamental result still holds in the computable setting. In this work, we formalize and investigate computable online learning under different notions of "optimality" in terms of the mistake bound.

© 2023 N. Hasrati & S. Ben-David.

Our main contribution is to give a necessary and sufficient condition for optimal c-online learning (Section 5.2), the proof of which relies on expanding the concept of significant points introduced by Frances and Litman (1998). Using this condition, we show that the Littlestone dimension no longer characterizes the optimal mistake bound of c-online learning (Section 5.3). In particular, we construct a class with finite Littlestone dimension for which no optimal online learner is computable. We also provide a positive result for the learnability of Littlestone dimension 1 classes in the computable setting (Section 5.2).

Additionally, we introduce a notion of anytime optimal (a-optimal) online learning which captures the optimality property displayed by Littlestone's Standard Optimal Algorithm (Sections 3.1, 4.1). Although optimal and a-optimal online learning are equivalent in the standard online learning model, we prove a computational separation between the two, showing that a-optimal online learning is computationally more difficult than optimal online learning. Specifically, we construct a class that is optimally but not a-optimally c-online learnable (Section 4.2).

A corollary of Theorem 3 from Littlestone (1988) is that the finiteness of the Littlestone dimension characterizes whether a class is online learnable at all—that is, whether it is online learnable with finite mistake bound. However, we show the existence of a "weakly computable" class with finite Littlestone dimension for which no computable online learner achieves finite mistake bound (Section 6.1). A potential avenue for strengthening this result is suggested in Section 6.2, where we explore the relationship between c-online and improper CPAC learning.

The paper is structured as follows. Section 2 provides the general background and notation needed from online learning and computability theory. Section 3 introduces our main definitions of a-optimal online learning, optimally significant inputs, and c-online learning. The last three sections analyze c-online learning under increasingly looser notions of "optimality"—Section 4 considers a-optimal c-online learning, Section 5 optimal c-online learning, and Section 6 c-online learning.

2. General Background

This section provides the required background from online learning (Section 2.1) and computability theory (Section 2.2).

2.1. Online Learning

We first give an informal description of the online learning model and then introduce the formal notation that will be used throughout the paper. The definitions in this section are based on those given in Shalev-Shwartz and Ben-David (2014, Chapter 21).

Introduced in Littlestone (1988)'s seminal work, online learning takes place in rounds. Informally, at each round t, an adversary presents the learner with some point x_t, the learner makes a prediction p_t, and the adversary reveals the true label y_t. The goal of the learner is to minimize the number of mistakes it makes. Clearly, with no further restrictions, the adversary could contradict the learner at each time step and cause an unbounded number of mistakes. It is therefore assumed that the learner has access to a class of hypotheses and that the sequence of examples presented by the adversary is consistent with some hypothesis from this class.

Formally, let $\mathcal{X}$ be the *domain set* and $\mathcal{Y} = \{0, 1\}$ be the *label set*. A *hypothesis* is a function $h : \mathcal{X} \to \mathcal{Y}$ and a *hypothesis class* is a set of hypotheses $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$. The *support* of a hypothesis h is $h^{-1}(1) = \{x : h(x) = 1\}$. Given a set $E \subseteq \mathcal{X}$, the *characteristic function* of E is $\chi_E : x \mapsto \mathbb{1}_{[x \in E]}$. A *sample* $S \in \mathbb{S} = \cup_{T \in \mathbb{N}} (\mathcal{X} \times \mathcal{Y})^T$ is a finite sequence of labeled domain instances, where the *empty*

sample is denoted by ε. Given a sample $S = ((x_i, y_i))_{i=1}^T$, let $S_n = ((x_i, y_i))_{i=1}^n$ be the *length-n prefix* of S, where $0 \leq n \leq T$. Denote by $S^\frown S'$ the concatenation of two samples $S, S' \in \mathbb{S}$. The *empirical loss of a hypothesis h* with respect to a sample S is defined as $L_S(h) = \sum_{t=1}^T \mathbb{1}_{[h(x_t) \neq y_t]}$. The *empirical loss of a hypothesis class $\mathcal{H}$* is $L_S(\mathcal{H}) = \inf_{h \in \mathcal{H}} L_S(h)$. The set of all samples that are *$\mathcal{H}$-realizable* is denoted by $\mathbb{S}_\mathcal{H} = \{S \in \mathbb{S} : L_S(\mathcal{H}) = 0\}$. Given a sample S, define $\mathcal{H}_S = \{h \in \mathcal{H} : L_S(h) = 0\}$ as the set of all hypotheses from $\mathcal{H}$ that are *consistent* with S. For some labeled instance $(x, y) \in \mathcal{X} \times \mathcal{Y}$, let $\mathcal{H}^{(x,y)} = \{h \in \mathcal{H} : h(x) = y\}$. Furthermore, define $[n] = \{x \in \mathbb{N} : 1 \leq x \leq n\}$, where $n \in \mathbb{N}$.

Definition 1 (online learner) *An* online learner *is a function $A \in \mathcal{Y}^{\mathbb{S} \times \mathcal{X}}$ that takes an* input history *$S \in \mathbb{S}$ and a domain instance $x \in \mathcal{X}$ as input and predicts $A(S, x) \in \{0, 1\}$. Given a sample $S = ((x_t, y_t))_{t=1}^T$, representing one run of the online learning process, at time step $t \in [T]$, A's* prediction *is $A(S_{t-1}, x_t)$, its* output hypothesis *is $A(S_{t-1}, \cdot) \in \mathcal{Y}^\mathcal{X}$, and its* version space *is $\mathcal{H}_{S_{t-1}}$.*

Definition 2 (mistake bound) *The number of mistakes made by an online learner A on a sample $S = ((x_1, y_1), \ldots, (x_T, y_T))$ is $M_A(S) = \sum_{t=1}^T \mathbb{1}_{[A(S_{t-1}, x_t) \neq y_t]}$. The* mistake bound *of A with respect to a hypothesis class $\mathcal{H}$ is $M_A(\mathcal{H}) = \sup_{S \in \mathbb{S}_\mathcal{H}} M_A(S)$—that is, the most that A errs on any $\mathcal{H}$-realizable sample. The* optimal mistake bound *of $\mathcal{H}$ is $M(\mathcal{H}) = \inf_{A \in \mathcal{Y}^{\mathbb{S} \times \mathcal{X}}} M_A(\mathcal{H})$.*

Definition 3 (online learnable class) *A hypothesis class $\mathcal{H}$ is* online learnable *if $M(\mathcal{H}) < \infty$.*

Definition 4 (optimal online learner) *An online learner A is an* optimal online learner *for a hypothesis class $\mathcal{H}$ if $M_A(\mathcal{H}) = M(\mathcal{H})$.*

Definition 5 ($\mathcal{H}$-shattered tree) *Let $\mathcal{H} \subseteq \{0, 1\}^\mathcal{X}$ and $d \in \mathbb{N}$. We say that $(x_1, \ldots, x_{2^d-1}) \in \mathcal{X}^{2^d-1}$ is an* $\mathcal{H}$-shattered tree *of depth d if, for every $(y_1, \ldots, y_d) \in \{0, 1\}^d$, there exists $h \in \mathcal{H}$ such that for all $j \in [d]$ we have that $h(x_{i_j}) = y_j$, where $i_j = 2^{j-1} + \sum_{k=1}^{j-1} y_k 2^{j-1-k}$. Let $\mathcal{T}_\mathcal{H}^d$ denote the set of all $\mathcal{H}$-shattered trees of depth d.*

Remark 6 *Intuitively, in the definition above, $(x_1, \ldots, x_{2^d-1}) \in \mathcal{T}_\mathcal{H}^d$ represents a labeling of the nodes of a complete binary tree of depth d, with x_i labeling i^{th} node. Each $(y_1, \ldots, y_d)$ represents a different path through the tree starting from the root node $i_1 = 1$. If i_j is the current node in the path, we go to the left child of i_j if $y_j = 0$ and go to the right child if $y_j = 1$.*

Definition 7 (Littlestone dimension) *The* Littlestone dimension *of a hypothesis class $\mathcal{H}$ is the depth of the largest $\mathcal{H}$-shattered tree. Formally, if $\mathcal{H} \neq \emptyset$, Ldim$(\mathcal{H}) = \sup\{d \in \mathbb{N} : \mathcal{T}_\mathcal{H}^d \neq \emptyset\}$ and Ldim$(\emptyset) = -1$.*

Remark 8 *Note that, for any hypothesis class $\mathcal{H}$, if Ldim$(\mathcal{H}^{(x,r)}) =$ Ldim$(\mathcal{H})$ for some $x \in \mathcal{X}$ and $r \in \{0, 1\}$, we must have that Ldim$(\mathcal{H}^{(x,1-r)}) <$ Ldim$(\mathcal{H})$.*

Definition 9 (Standard Optimal Learner) *The* Standard Optimal Learner *for a hypothesis class $\mathcal{H}$ is defined as $SOL_\mathcal{H} : (S, x) \mapsto \mathbb{1}_{\left[\text{Ldim}\left(\mathcal{H}_S^{(x,1)}\right) \geq \text{Ldim}\left(\mathcal{H}_S^{(x,0)}\right)\right]}$.*

Theorem 10 (Littlestone, 1988, Theorem 3) *Given any hypothesis class $\mathcal{H}$, $M(\mathcal{H}) =$ Ldim$(\mathcal{H})$. In particular, for every online learner A, $M_A(\mathcal{H}) \geq$ Ldim$(\mathcal{H})$ and $M_{SOL_\mathcal{H}}(\mathcal{H}) =$ Ldim$(\mathcal{H})$.*[1]

1. Although Littlestone (1988, Theorem 3) only considers finite classes, the result is easily extendable to infinite classes if the learners are not required to be computable (see Shalev-Shwartz and Ben-David, 2014, Corollary 21.8)

2.2. Computability

We use notation given by Soare (2016). Let $\{P_e\}_{e\in\mathbb{N}}$ and $\{\varphi_e\}_{e\in\mathbb{N}}$ be effective numberings of all Turing machines and all *partial computable (p.c.) functions*, respectively. If P_e halts on input x and outputs y, we write $\varphi_e(x) = y$ and say that $\varphi_e(x)$ *converges* (denoted $\varphi_e(x)\downarrow$). Otherwise, $\varphi_e(x)$ *diverges* (denoted $\varphi_e(x)\uparrow$). The domain of φ_e is $\text{dom}(\varphi_e) = \{x : \varphi_e(x)\downarrow\}$ and its range is $\text{rng}(\varphi_e) = \{\varphi_e(x) : \varphi_e(x)\downarrow\}$. If $\text{dom}(\varphi_e) = \mathbb{N}$, φ_e is a *total computable (t.c.) function* (abbreviated *computable function*). We also extend this notation to n-place p.c. functions, where $\varphi_e^{(n)}$ is the p.c. function of n variables computed by P_e and φ_e denotes $\varphi_e^{(1)}$. A set E is *recursively enumerable (r.e.)* if it can be effectively enumerated—that is, if it is the domain of some p.c. function. E is *decidable* if its characteristic function, $\chi_E : x \mapsto \mathbb{1}_{[x\in E]}$, is computable. The *restriction* of φ_e to an r.e. set X is the p.c. function $\varphi_e|_X$, where $\varphi_e|_X(x)$ equals $\varphi_e(x)$ if $x \in X \cap \text{dom}(\varphi_e)$ and is undefined otherwise. We say φ_{e_2} is a *p.c. extension* of φ_{e_1} if $\varphi_{e_2}|_{\text{dom}(\varphi_{e_1})} = \varphi_{e_1}$.

The *canonical index* of a finite set $F \subset \mathbb{N}$ is an integer y that explicitly specifies all elements of F, and D_y denotes the finite set with canonical index y.[2] Furthermore, given a sequence $Z \in \cup_{n\in\mathbb{N}}\mathbb{N}^n$, we let $\langle Z\rangle$ denote the encoding of Z by a standard 1:1 computable function from $\cup_{n\in\mathbb{N}}\mathbb{N}^n$ to $\mathbb{N}$. In a slight abuse of notation, we extend this notation to apply when $Z \in \mathbb{S}$.[3] Additionally, for a set X of such integer sequences, we define $\langle X\rangle = \{\langle Z\rangle : Z \in X\}$.

3. Setup and definitions

This section introduces our main definitions of anytime optimal online learning (Section 3.1), optimally significant inputs (Section 3.2), and c-online learning (Section 3.3).

3.1. Anytime optimal online learning

We present a notion of anytime optimal online learning, which we claim is a more natural conceptualization of "optimality" when referring to online learning.

As a motivating example, consider the class $\mathcal{H}_d = \{\chi_{[n]}\}_{n=1}^{2^d}$ over the domain $\mathcal{X} = \mathbb{N}$, where $2 < d < \infty$ (recall that $[n] = \{1, 2, \ldots, n\}$ and χ_A is the characteristic function of the set $A \subseteq \mathbb{N}$). Further define $E = \{2^d + i\}_{i=1}^{d-1}$ and let $\mathcal{H}'_d = \mathcal{H}_d \cup \{\chi_E\}$. That is, $\mathcal{H}_d$ is a set of 2^d thresholds over the natural numbers and E is a set of $d-1$ distinct domain instances that are not given the label 1 by any $h \in \mathcal{H}_d$. It is easy to verify that $\text{Ldim}(\mathcal{H}'_d) = \text{Ldim}(\mathcal{H}_d) = d$. Now, let A be the learner that behaves as follows: for all inputs $(S, x) \in \mathbb{S} \times \mathcal{X}$, $A(S, x) = SOL_{\mathcal{H}'_d}(S, x)$ if S is $\mathcal{H}_d$-realizable and $A(S, x) = 0$ otherwise. Note that A is still an optimal online learner for $\mathcal{H}'_d$ as it errs no more than d times on any $\mathcal{H}'_d$-realizable sample; however, on the $\mathcal{H}'_d$-realizable sample $((x, 1))_{x\in E}$, A errs $d-1$ times while $SOL_{\mathcal{H}'_d}$ only errs at time step 1. It is clear that any $\mathcal{H}'_d$-realizable sample that contains some $x \in E$ with the label 1 can only be realized by χ_E; hence, a "truly optimal" learner should incur no mistakes after seeing any $x \in E$ with the label 1.

2. Specifically, the *canonical index* of a finite set $F \subset \mathbb{N}$ is the integer $y = \sum_{x\in F} 2^x$. The elements of the finite set with canonical index y, D_y, are the positions of the "on" bits in y's binary expansion.
3. To be explicit, given an n-tuple of integers $Z = (z_1, \ldots, z_n)$, we have that $\langle Z\rangle = \Pi_{i=1}^{n} p_i^{z_i+1}$, where p_i is the ith prime number. Similarly, given a sample $S = ((x_1, y_1), \ldots, (x_n, y_n))$, we define $\langle S\rangle = \langle(x_1, y_1, \ldots, x_n, y_n)\rangle$. Note that any 1:1 partially computable function is computably invertible on its range, so Z and S are computably recoverable given $\langle Z\rangle$ and $\langle S\rangle$ respectively.

The above example illustrates a gap between the commonly accepted definition of optimal online learning and the stricter optimality displayed by the Standard Optimal Learner. We define our notion of anytime optimal online learning below, where the learner makes the optimal number of mistakes even after conditioning on a given input sample. The properties of anytime optimal online learning are further explored in Section 4.1.

Definition 11 (post-S mistake bound) *Given a hypothesis class $\mathcal{H}$, an online learner A, and an $\mathcal{H}$-realizable sample $S \in \mathbb{S}_{\mathcal{H}}$, we define the* post-$S$ mistake bound *of A with respect to $\mathcal{H}$ as*

$$M_A^S(\mathcal{H}) = \sup_{\substack{S' \in \mathbb{S}: \\ S^\frown S' \in \mathbb{S}_{\mathcal{H}}}} M_A(S^\frown S') - M_A(S).$$

That is, $M_A^S(\mathcal{H})$ is the most that A can be made to err after witnessing S. The optimal post-S mistake bound *of $\mathcal{H}$ is defined as $M^S(\mathcal{H}) = \inf_{A \in \mathcal{Y}^{\mathbb{S} \times \mathcal{X}}} M_A^S(\mathcal{H})$.*

Definition 12 (anytime optimal (a-optimal) online learner) *An online learner A is* anytime optimal (a-optimal) *for a hypothesis class $\mathcal{H}$ if $M_A^S(\mathcal{H}) = M^S(\mathcal{H})$ for all $S \in \mathbb{S}_{\mathcal{H}}$.*

3.2. Significant inputs for optimal and a-optimal online learning

Frances and Litman (1998) introduced the concept of *significant points*, points on which all optimal online learners agree on in the first time step of online learning. Formally, we say that $x \in \mathcal{X}$ is an *optimally significant point* for online learning a class $\mathcal{H}$ if $A(\varepsilon, x) = A'(\varepsilon, x)$ for any two optimal online learners A and A'. Furthermore, Lemma 3 from Frances and Litman (1998) characterizes all optimally significant points as follows: x is an optimally significant point for $\mathcal{H}$ iff there exists $r \in \{0, 1\}$ such that $\text{Ldim}(\mathcal{H}^{(x,r)}) = \text{Ldim}(\mathcal{H})$. Moreover, $A(\varepsilon, x) = r$ for all online learners A that are optimal w.r.t. $\mathcal{H}$. Below, we extend this definition to apply beyond the first time step.

Definition 13 (optimally significant input) *Let $\mathcal{H}$ be any hypothesis class. We say that $(S, x) \in \mathbb{S}_{\mathcal{H}} \times \mathcal{X}$ is an* optimally significant input *for online learning $\mathcal{H}$ if $A(S, x) = A'(S, x)$ for any two optimal online learners A and A' for $\mathcal{H}$. Let $\mathcal{I}_{\mathcal{H}}$ be the set of all optimally significant inputs for $\mathcal{H}$.*

Definition 14 (anytime optimally (a-optimally) significant input) *Let $\mathcal{H}$ be any hypothesis class. We say that $(S, x) \in \mathbb{S}_{\mathcal{H}} \times \mathcal{X}$ is an* anytime optimally (a-optimally) significant input *for online learning $\mathcal{H}$ if $A(S, x) = A'(S, x)$ for any two a-optimal online learners A and A' for $\mathcal{H}$.*

3.3. Computable online learning

When defining a computably online learnable hypothesis class, we require both the class and the learner to conform to some notion of "computability."[4] Following the computable PAC (CPAC) setting (Agarwal et al., 2020), we let $\mathcal{X} = \mathbb{N}$, and assume, as a minimum, that the class consists of computable hypotheses. It is also desirable to assume an effective enumeration of (the encodings of) the hypotheses. A class $\mathcal{H} \subset \{0, 1\}^{\mathbb{N}}$ of computable hypotheses is *recursively enumerably representable (RER)* if there exists an r.e. set $E \subset \mathbb{N}$ such that $\mathcal{H} = \{\varphi_e : e \in E\}$. A class $\mathcal{H}$ is *decidably representable (DR)* if each $h \in \mathcal{H}$ has finite support and $\{y : \exists h \in \mathcal{H}\ (D_y = h^{-1}(1))\}$ is a decidable set. Next, we define what it means for the learner itself to be computable.

4. The reader is referred to Section 2.2 for the relevant notation from computability theory.

Definition 15 (computable online (c-online) learner) *Let $\mathcal{H} \subset \{0,1\}^{\mathbb{N}}$ be any class of computable hypotheses. A two-place p.c. function $A : \mathbb{N}^2 \to \mathbb{N}$ is a* computable online (c-online) learner *for $\mathcal{H}$, if, for every $\mathcal{H}$-realizable sample $S \in \mathbb{S}_{\mathcal{H}}$ and every domain instance $x \in \mathcal{X}$, $A(\langle S \rangle, x) \downarrow = y$ for some $y \in \{0,1\}$. That is, $dom(A) \supseteq \langle \mathbb{S}_{\mathcal{H}} \rangle \times \mathcal{X}$ and $rng(A|_{\langle \mathbb{S}_{\mathcal{H}} \rangle \times \mathcal{X}}) \subseteq \{0,1\}$.*

Definition 16 (computable optimal online learner) *A* computable optimal online learner *A for a class $\mathcal{H} \subset \{0,1\}^{\mathbb{N}}$ of computable hypotheses is a c-online learner for $\mathcal{H}$ with $M_A(\mathcal{H}) = M(\mathcal{H})$.*[5]

Definition 17 (computable a-optimal online learner) *A* computable anytime optimal (a-optimal) online learner *A for a class $\mathcal{H} \subset \{0,1\}^{\mathbb{N}}$ of computable hypotheses is a c-online learner for $\mathcal{H}$ with $M_A^S(\mathcal{H}) = M^S(\mathcal{H})$ for all $S \in \mathbb{S}_{\mathcal{H}}$.*

Definition 18 (computably online (c-online) learnable class) *A class $\mathcal{H} \subset \{0,1\}^{\mathbb{N}}$ of computable hypotheses is* computably online (c-online) learnable *if there exists a c-online learner A for $\mathcal{H}$ with $M_A(\mathcal{H}) < \infty$.*

Definition 19 (optimally c-online learnable class) *A class $\mathcal{H} \subset \{0,1\}^{\mathbb{N}}$ of computable hypotheses is* optimally c-online learnable *if there exists a computable optimal online learner for $\mathcal{H}$.*

Definition 20 (a-optimally c-online learnable class) *A class $\mathcal{H} \subset \{0,1\}^{\mathbb{N}}$ of computable hypotheses is* anytime optimally (a-optimally) c-online learnable *if there exists a computable a-optimal online learner for $\mathcal{H}$.*

4. Anytime optimal c-online learnability

We start our analysis by considering the computability of a-optimal online learners. In Section 4.2, we show the existence of a computational separation between a-optimal and optimal online learning, proving that a-optimal online learning is computationally more difficult. Our proof relies on properties of a-optimal online learners presented in section 4.1 below.

4.1. Properties of anytime optimal online learners

The following lemma gives a characterization of the optimal post-S mistake bound of anytime optimal online learning in terms of the Littlestone dimension of the version space. The proof is implicit in the proof of Theorem 3 from Littlestone (1988).

Lemma 21 (characterizing the mistake bound of a-optimal online learning) *Let $\mathcal{H}$ be any hypothesis class. Then, for any $\mathcal{H}$-realizable sample $S \in \mathbb{S}_{\mathcal{H}}$, we have that $M^S(\mathcal{H}) = Ldim(\mathcal{H}_S)$. In particular, for every online learner A, $M_A^S(\mathcal{H}) \geq Ldim(\mathcal{H}_S)$ and $M_{SOL_{\mathcal{H}}}^S(\mathcal{H}) = Ldim(\mathcal{H}_S)$.*

Informally, the next lemma states that an input is a-optimally significant iff it causes an "imbalance" in the Littlestone tree of the version space. Again, the proof is implicit in the proof of Theorem 3 from Littlestone (1988).

Lemma 22 (characterizing a-optimally significant inputs) *Let $\mathcal{H}$ be any hypothesis class. Then, an input $(S, x) \in \mathbb{S} \times \mathcal{X}$ is a-optimally significant for $\mathcal{H}$ iff $Ldim(\mathcal{H}_S^{(x,1)}) \neq Ldim(\mathcal{H}_S^{(x,0)})$. Furthermore, $A(S, x) = \arg\max_{r \in \{0,1\}} Ldim(\mathcal{H}_S^{(x,r)})$ for all a-optimal online learners A for $\mathcal{H}$.*

5. Note that when A is a c-online learner for a class $\mathcal{H}$ of computable hypotheses, $M_A(S) = \sum_{t=1}^{T} \mathbb{1}_{[A(\langle S_{t-1} \rangle, x_t) \neq y_t]}$ is well-defined for any $\mathcal{H}$-realizable $S = ((x_t, y_t))_{t=1}^{T}$. We can extend the notation for $M_A(\mathcal{H})$ and $M_A^S(\mathcal{H})$ similarly.

4.2. Computational gap between optimal and a-optimal online learning

In this section, we show that a-optimal online learning is computationally more difficult than optimal online learning. In particular, we construct an RER class that is optimally but not a-optimally c-online learnable. This result is extended to the DR case in Appendix C.

Theorem 23 *There exists an RER class $\mathcal{H} \subset \{0,1\}^{\mathbb{N}}$ of computable hypotheses with finite Littlestone dimension such that $\mathcal{H}$ is optimally c-online learnable but not a-optimally c-online learnable.*

Proof Consider the following class:

$$\mathcal{H}^{RER}_{halt} = \bigcup_{e\in\mathbb{N}} \left\{\chi_{\{3e\}}\right\} \cup \bigcup_{\substack{e\in\mathbb{N}:\\ \varphi_e(e)\downarrow}} \left\{\chi_{\{3e,\ 3e+1\}}, \chi_{\{3e,\ 3e+1,\ 3e+2\}}\right\}.$$

For simplicity, let $\mathcal{H} = \mathcal{H}^{RER}_{halt}$. Note that $\mathcal{H}$ is RER, each $h \in \mathcal{H}$ is computable, and $\mathrm{Ldim}(\mathcal{H}) < \infty$.

Assume, by way of contradiction, that there exists a computable a-optimal online learner A for $\mathcal{H}$. For each $e \in \mathbb{N}$, let $S^e = ((3e, 1))$ and $x^e = 3e + 1$. Further define $f : e \mapsto A(\langle S^e\rangle, x^e)$. First, note that f is computable, since for each $e \in \mathbb{N}$ the sample S^e is $\mathcal{H}$-realizable and $A(\langle S^e\rangle, x^e)\downarrow$. Next, we show by Lemma 22 that each (S^e, x^e) is an a-optimally significant input. Note that for any $e \in \mathbb{N}$,

$$\mathcal{H}^{(x^e,0)}_{S^e} = \{\chi_{\{3e\}}\} \qquad \text{and} \qquad \mathcal{H}^{(x^e,1)}_{S^e} = \begin{cases} \{\chi_{\{3e,\ 3e+1\}}, \chi_{\{3e,\ 3e+1,\ 3e+2\}}\} & \text{if } \varphi_e(e)\downarrow \\ \emptyset & \text{otherwise.} \end{cases}$$

Therefore, if $\varphi_e(e)\downarrow$, $\mathrm{Ldim}(\mathcal{H}^{(x^e,1)}_{S^e}) = 1 > \mathrm{Ldim}(\mathcal{H}^{(x^e,0)}_{S^e}) = 0$ and $A(\langle S^e\rangle, x^e) = 1$. On the other hand, if $\varphi_e(e)\uparrow$, $\mathrm{Ldim}(\mathcal{H}^{(x^e,0)}_{S^e}) = 0 > \mathrm{Ldim}(\mathcal{H}^{(x^e,1)}_{S^e}) = -1$ and $A(\langle S^e\rangle, x^e) = 0$. Hence, f is computable and equals $\chi_{\{e\in\mathbb{N}:\ \varphi_e(e)\downarrow\}}$, contradicting the undecidability of the halting problem.

Although $\mathcal{H}$ is not a-optimally c-online learnable, we show the existence of a computable optimal online learner B for $\mathcal{H}$. It is easy to verify that $\mathrm{Ldim}(\mathcal{H}) \geq 2$; hence, it suffices to show that $M_B(\mathcal{H}) = 2$. B predicts 0 until, for some $e \in \mathbb{N}$, a mistake is made on $x_1 \in \{3e,\ 3e+1,\ 3e+2\}$, at which point it matches $\chi_{\{3e,\ 3e+1,\ x_1\}}$. If $x_1 = 3e + 2$, B will not err again. Otherwise, it could possibly err on $x_2 \in \{3e+1,\ 3e+2\}$. If $x_2 = 3e+2$, the target function must be $\chi_{\{3e,\ 3e+1,\ 3e+2\}}$; otherwise, if $x_2 = 3e + 1$, the target function must be $\chi_{\{3e\}}$. In either case, B errs no more than $\mathrm{Ldim}(\mathcal{H}) = 2$ times on any $\mathcal{H}$-realizable sample. ■

5. Optimal c-online learnability

In this section, we loosen the requirement of a-optimality, turning our focus to all optimal online learners instead. We give a necessary and sufficient condition for when optimal c-online learning is possible (Section 5.2) and show that the Littlestone dimension no longer characterizes the mistake bound of optimal c-online learning (Section 5.3). We also give a complete characterization of all optimally significant inputs (Section 5.1), a result which is used in our main proofs.

5.1. Characterizing optimally significant inputs

The following lemma gives a complete characterization of all optimally significant inputs.

Lemma 24 (characterizing optimally significant inputs) *Let $\mathcal{H}$ be any hypothesis class satisfying Ldim$(\mathcal{H}) = d < \infty$. Let $S = ((x_1, y_1), \ldots, (x_T, y_T))$ be any $\mathcal{H}$-realizable sample and $x_{T+1} \in \mathcal{X}$ be any domain instance, where $T \in \mathbb{N}$. Then, (S, x_{T+1}) is a significant input w.r.t. optimal online learning $\mathcal{H}$ iff the following conditions both hold:*

1. *for each $t \in [T+1]$, Ldim$(\mathcal{H}_{S_{t-1}}) = \max_{r \in \{0,1\}}$ Ldim$(\mathcal{H}_{S_{t-1}}^{(x_t, r)})$, and*
2. *for each $t \in [T]$, Ldim$(\mathcal{H}_{S_{t-1}}^{(x_t, y_t)}) \geq$ Ldim$(\mathcal{H}_{S_{t-1}}) - 1$.*

Furthermore, $A(S_{t-1}, x_t) = \arg\max_{r \in \{0,1\}}$ Ldim$(\mathcal{H}_{S_{t-1}}^{(x_t, r)})$ for all $t \in [T+1]$ and all optimal online learners A.

Proof It follows from Lemma 35 (Appendix A) that conditions 1 and 2 above are equivalent to the following two conditions:

I. $\text{Ldim}(\mathcal{H}_S) = \max_{r \in \{0,1\}} \text{Ldim}(\mathcal{H}_S^{(x_{T+1}, r)})$, and

II. $M_A(S) = \text{Ldim}(\mathcal{H}) - \text{Ldim}(\mathcal{H}_S)$ for every online learner A that is optimal for $\mathcal{H}$.

It remains to show that (S, x_{T+1}) is optimally significant iff conditions I and II hold. First, assume for the sake of contradiction that the two conditions hold but there exists an optimal online learner A that predicts $A(S, x_{T+1}) = 1 - r^*$, where $r^* = \arg\max_{r \in \{0,1\}} \text{Ldim}(\mathcal{H}_S^{(x_{T+1}, r)})$. Then, on the sample $S^* = S^\frown((x_{T+1}, r^*))$, A makes $\text{Ldim}(\mathcal{H}) - \text{Ldim}(\mathcal{H}_S) + 1$ mistakes and, by Lemma 21, can be made to err at least $\text{Ldim}(\mathcal{H}_S^{(x_{T+1}, r^*)}) = \text{Ldim}(\mathcal{H}_S)$ times after witnessing S^*, a contradiction. Furthermore, it follows from Lemma 35 that $A(S_{t-1}, x_t) = \arg\max_{r \in \{0,1\}} \text{Ldim}(\mathcal{H}_{S_{t-1}}^{(x_t, r)})$ for all $t \in [T+1]$ and all optimal online learners A.

Conversely, if (S, x_{T+1}) is optimally significant, there exists $r^* \in \{0, 1\}$ such that for all online learners A, if A is optimal then $A(S, x_{T+1}) = r^*$. For an a-optimal online learner A^*, let A' be the learner that agrees with A^* on all inputs except (S, x_{T+1}). Since this single change in prediction causes A' to no longer be optimal, we must have that $M_{A'}(S^*) + M_{A'}^{S^*}(\mathcal{H}) \geq d + 1$, where $S^* = S^\frown((x_{T+1}, r^*))$. Note that $M_{A'}(S^*) + M_{A'}^{S^*}(\mathcal{H}) = M_{A^*}(S) + 1 + M_{A^*}^{S^*}(\mathcal{H})$, so we must have that $M_{A^*}(S) + M_{A^*}^{S^*}(\mathcal{H}) \geq d$. However, since A^* is a-optimal, the maximum values for $M_{A^*}(S)$ and $M_{A^*}^{S^*}(\mathcal{H})$ are $d - \text{Ldim}(\mathcal{H}_S)$ and $\text{Ldim}(\mathcal{H}_S)$ respectively. Hence, the inequality is only satisfied when both conditions I and II hold. ■

Corollary 25 (version space of optimally significant inputs) *Let $\mathcal{H}$ be a hypothesis class with Ldim$(\mathcal{H}) = d < \infty$ and let $(S, x) \in \mathcal{I}_\mathcal{H}$ be any optimally significant input for $\mathcal{H}$. Then, there exists $m \in \mathbb{N}$ such that $M_A(S) = m$ for all optimal online learners A and Ldim$(\mathcal{H}_S) = d - m$.*

5.2. Characterizing optimal c-online learning

In this section, we give a necessary and sufficient condition for optimal c-online learning in the RER setting (Corollary 27). The condition follows from Theorem 26, which shows that the predictions of all optimal online learners are computable on inputs that are optimally significant. Corollary 28 shows that any infinite RER class of Littlestone dimension 1 is optimally c-online learnable.

Theorem 26 (computability of optimally significant predictions) *Let $\mathcal{H} \subset \{0,1\}^{\mathbb{N}}$ be any RER class of computable hypotheses with finite Littlestone dimension. Then, there exists a partial computable function $p^{sig}_{\mathcal{H}}$ such that $p^{sig}_{\mathcal{H}}(\langle S \rangle, x) = A(S,x)$ for any optimally significant input $(S,x) \in \mathcal{I}_{\mathcal{H}}$ and any optimal online learner A for $\mathcal{H}$.*

Proof Let $\mathcal{X} = \mathbb{N}$ and $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be any RER class of computable hypotheses with $\text{Ldim}(\mathcal{H}) = d < \infty$. First, we show the existence of a Turing machine $M_{\mathcal{H}}$ that behaves as follows: for any $S \in \mathbb{S}$, $x \in \mathcal{X}$, and $d' \in \mathbb{N}$, if there exists $r \in \{0,1\}$ for which $\text{Ldim}(\mathcal{H}^{(x,r)}_S) = d'$ and $\text{Ldim}(\mathcal{H}^{(x,1-r)}_S) < d'$, $M_{\mathcal{H}}$ halts on input $(\langle S \rangle, x, d')$ and outputs r. Note that for any RER class $\mathcal{H}'$ of computable hypotheses, the set $\langle \mathcal{T}^{d'}_{\mathcal{H}'} \rangle$ of (the encodings of) all $\mathcal{H}'$-shattered trees of depth d' is r.e.. Therefore, since both $\mathcal{H}^{(x,1)}_S$ and $\mathcal{H}^{(x,0)}_S$ are RER, $M_{\mathcal{H}}$ simultaneously runs the enumerators for $\mathcal{T}^{d'}_{\mathcal{H}^{(x,1)}_S}$ and $\mathcal{T}^{d'}_{\mathcal{H}^{(x,0)}_S}$ until one yields an output. If the enumerator for $\mathcal{T}^{d'}_{\mathcal{H}^{(x,y)}_S}$ yields an output first, $M_{\mathcal{H}}$ halts and outputs y. Now, if there exists r for which $\text{Ldim}(\mathcal{H}^{(x,r)}_S) = d'$ and $\text{Ldim}(\mathcal{H}^{(x,1-r)}_S) < d'$, we must have that $\mathcal{T}^{d'}_{\mathcal{H}^{(x,r)}_S} \neq \emptyset$ and $\mathcal{T}^{d'}_{\mathcal{H}^{(x,1-r)}_S} = \emptyset$; hence, $M_{\mathcal{H}}$ will eventually halt and output r.

Now, consider the Turing machine $P^{sig}_{\mathcal{H}}$ that behaves as follows on any input $(\langle S \rangle, x_{T+1})$, where $S = ((x_1,y_1),\ldots,(x_T,y_T))$ for some $T \in \mathbb{N}$: 1) initialize $m = 0$; 2) for each $t \in [T+1]$, let p_t be the result of running $M_{\mathcal{H}}$ on input $(\langle S_{t-1} \rangle, x_t, d-m)$ and increment m if $p_t \neq y_t$; 3) output p_{T+1}.

We will show that if $(S, x_{T+1}) \in \mathcal{I}_{\mathcal{H}}$, $p_t = \arg\max_{r \in \{0,1\}} \text{Ldim}(\mathcal{H}^{(x_t,r)}_{S_{t-1}})$ for each $t \in [T+1]$; hence, by Lemma 24, $p^{sig}_{\mathcal{H}}$ is computed by $P^{sig}_{\mathcal{H}}$. We proceed by induction on $t \in [T+1]$. If $t = 1$, by lemma 24, there exists $r_1 \in \{0,1\}$ such that $\text{Ldim}(\mathcal{H}^{(x_1,r_1)}_{S_0}) = d$ and $\text{Ldim}(\mathcal{H}^{(x_1,1-r_1)}_{S_0}) < d$. Therefore, $M_{\mathcal{H}}$ halts on input $(\langle S_0 \rangle, x_1, d)$ and outputs r_1. Now, consider any $\tau \in [T+1]$ such that the condition holds for all $t < \tau$. Then, $m_{\tau-1} = \sum_{t=1}^{\tau-1} \mathbb{1}_{[p_t \neq y_t]}$ is the number of mistakes that all optimal online learners make on $S_{\tau-1}$. Hence, by Corollary 25, $\text{Ldim}(\mathcal{H}_{S_{\tau-1}}) = d - m_{\tau-1}$ and, by Lemma 24, there exists $r_\tau \in \{0,1\}$ such that $\text{Ldim}(\mathcal{H}^{(x_\tau,r_\tau)}_{S_{\tau-1}}) = d - m_{\tau-1}$ and $\text{Ldim}(\mathcal{H}^{(x_\tau,1-r_\tau)}_{S_{\tau-1}}) < d - m_{\tau-1}$. Therefore, $M_{\mathcal{H}}$ halts on $(\langle S_{\tau-1} \rangle, x_\tau, d - m_{\tau-1})$ and outputs $p_\tau = r_\tau$, as required. ■

Corollary 27 (characterizing optimal c-online learning) *Let $\mathcal{H} \subset \{0,1\}^{\mathbb{N}}$ be any RER class of computable hypotheses with finite Littlestone dimension and let $p^{sig}_{\mathcal{H}}$ be the partial computable function defined in Theorem 26. Then, $\mathcal{H}$ is optimally c-online learnable iff there exists a p.c. extension $p^{real}_{\mathcal{H}}$ of $p^{sig}_{\mathcal{H}}$ such that $\text{dom}(p^{real}_{\mathcal{H}}) \supseteq \langle \mathbb{S}_{\mathcal{H}} \rangle \times \mathcal{X}$ and $\text{rng}(p^{real}_{\mathcal{H}}|_{\langle \mathbb{S}_{\mathcal{H}} \rangle \times \mathcal{X}}) \subseteq \{0,1\}$.*

Corollary 28 (optimal c-online learnability of classes with Littlestone dimension 1) *Let $\mathcal{H} \subset \{0,1\}^{\mathbb{N}}$ be any infinite RER class of computable hypotheses with $\text{Ldim}(\mathcal{H}) = 1$. Then, $\mathcal{H}$ is optimally c-online learnable.*

Proof By Corollary 27, it suffices to show that $\mathbb{S}_{\mathcal{H}} \times \mathcal{X} \subseteq \mathcal{I}_{\mathcal{H}}$. Let $T \in \mathbb{N}$ and consider any $\mathcal{H}$-realizable sample $S = ((x_1,y_1),\ldots,(x_T,y_T)) \in \mathbb{S}_{\mathcal{H}}$ and any $x_{T+1} \in \mathcal{X}$. We will show that

(S, x_{T+1}) satisfies Lemma 24 and is hence an optimally significant input for $\mathcal{H}$. Let $\tau \in [T]$ be the earliest time step such that $\text{Ldim}(\mathcal{H}_{S_{\tau-1}}) \neq \text{Ldim}(\mathcal{H}^{(x_\tau, y_\tau)}_{S_{\tau-1}})$. If no such time step exists, let $\tau = T+1$. Then, $\text{Ldim}(\mathcal{H}_{S_{t-1}}) = 1$ for all $t \leq \tau$ and, since S is $\mathcal{H}$-realizable, $\text{Ldim}(\mathcal{H}_{S_{t-1}}) = 0$ for all $\tau < t \leq T+1$. Therefore, condition 2 of Lemma 24 is satisfied for all $t \in [T]$ and condition 1 is satisfied for all $t \neq \tau$. Now, since $\mathcal{H}$ is infinite and at most one hypothesis is removed from the version space at each time step before τ, $\mathcal{H}_{S_{\tau-1}}$ is also infinite and there exists $r \in \{0,1\}$ such that $\mathcal{H}^{(x_\tau, r)}_{S_{\tau-1}}$ is infinite. Hence, $\text{Ldim}(\mathcal{H}_{S_{\tau-1}}) = \text{Ldim}(\mathcal{H}^{(x_\tau, r)}_{S_{\tau-1}}) = 1$ and condition 1 holds for $t = \tau$. ■

5.3. Littlestone dimension fails to characterize optimal mistake bound of online learning

In this section, we show that the Littlestone dimension no longer characterizes the mistake bound of optimal c-online learning. Specifically, we construct a DR class of computable hypotheses that has finite Littlestone dimension but is not optimally c-online learnable. Without the RER requirement, constructing such a class is not too difficult. In fact, the class $\mathcal{H}_{halting} = \bigcup_{e \in \mathbb{N}: \varphi_e(e)\downarrow} \{\chi_{\{2e,\, 2e+1\}}\} \cup \bigcup_{e \in \mathbb{N}: \varphi_e(e)\uparrow} \{\chi_{\{2e\}}\}$, presented by Agarwal et al. (2020, Theorem 9), has Littlestone dimension 1 but any computable optimal online learner for this class would decide the halting problem.

Theorem 29 *There exists a DR class $\mathcal{H} \subset \{0,1\}^{\mathbb{N}}$ of computable hypotheses such that $\text{Ldim}(\mathcal{H}) < \infty$ but $\mathcal{H}$ is not optimally c-online learnable.*

Proof For each $x \in \mathbb{N}$, let $\{C^{(x)}_i\}_{i \in \mathbb{N}: i > 0}$ be an effective enumeration of all halting computations starting from input x (see Soare, 2016, Section 1.5.2). Further define, for each $x \in \mathbb{N}$, the p.c. function c_x such that if P_e halts on input x, $C^{(x)}_{c_x(e)}$ is the halting certificate. That is, for each $e \in \mathbb{N}$,

$$c_x(e) = \begin{cases} i & \text{if there exists } i \text{ s.t. } C^{(x)}_i \text{ is a halting computation for } P_e \text{ on input } x \\ \text{undefined} & \text{otherwise.} \end{cases}$$

Now, consider the following class:

$$\begin{aligned} \mathcal{H}^{DR}_{ext} = & \bigcup_{e \in \mathbb{N}:\, \varphi_e(0)\downarrow} \left\{ \chi_{\{2^e,\, 2^e 3^{c_0(e)}\}} \right\} \\ & \cup \bigcup_{e \in \mathbb{N}:\, \varphi_e(0)\downarrow \text{ and } \varphi_e(e)\downarrow = 1} \left\{ \chi_{\{2^e,\, 2^e 5^{c_0(e)},\, 2^e 7^{c_e(e)}\}}, \chi_{\{2^e,\, 2^e 5^{c_0(e)},\, 2^e 11^{c_e(e)}\}} \right\} \\ & \cup \bigcup_{e \in \mathbb{N}:\, \varphi_e(0)\downarrow \text{ and } \varphi_e(e)\downarrow = 0} \left\{ \chi_{\{2^e,\, 2^e 5^{c_0(e)},\, 2^e 13^{c_e(e)}\}}, \chi_{\{2^e,\, 2^e 3^{c_0(e)},\, 2^e 13^{c_e(e)}\}} \right\}. \end{aligned}$$

For simplicity, let $\mathcal{H} = \mathcal{H}^{DR}_{ext}$. Note that each $h \in \mathcal{H}$ is computable since $c_x(e)$ is evaluated only if $\varphi_e(x)\downarrow$. Furthermore, $\text{Ldim}(\mathcal{H}) = 2$ (Appendix B.1) and $\mathcal{H}$ is DR (Appendix B.2).

By Theorem 26, since $\mathcal{H}$ is RER, there exists a p.c. function $p^{sig}_{\mathcal{H}}$ such that $p^{sig}_{\mathcal{H}}(\langle S \rangle, x) = A(S, x)$ for any optimally significant input $(S, x) \in \mathcal{I}_{\mathcal{H}}$ and any optimal online learner A for $\mathcal{H}$. For each $e \in \mathbb{N}$, let $S^e = ((2^e, 1))$ and define the p.c. functions $x : e \mapsto 2^e 3^{c_0(e)}$ and $f : e \mapsto p^{sig}_{\mathcal{H}}(\langle S^e \rangle, x(e))$. In Appendix B.3, we show using Lemma 24 that $(S^e, x(e))$ is an optimally

significant input for $\mathcal{H}$ iff $\varphi_e(0)\downarrow$ and $\varphi_e(e)\downarrow\in\{0,1\}$. Furthermore, for any $e\in\mathbb{N}$ such that $\varphi_e(0)\downarrow$ and $\varphi_e(e)\downarrow\in\{0,1\}$, we have that

$$f(e)=p_{\mathcal{H}}^{sig}(\langle S^e\rangle,x(e))=\begin{cases}1 & \text{if } \varphi_e(0)\downarrow \text{ and } \varphi_e(e)\downarrow=0\\ 0 & \text{if } \varphi_e(0)\downarrow \text{ and } \varphi_e(e)\downarrow=1.\end{cases}$$

Now, assume for the sake of contradiction that $\mathcal{H}$ is optimally c-online learnable. Then, by Corollary 27, there exists a p.c. extension $p_{\mathcal{H}}^{real}$ of $p_{\mathcal{H}}^{sig}$ such that $\mathrm{dom}(p_{\mathcal{H}}^{real})\supseteq\langle\mathbb{S}_{\mathcal{H}}\rangle\times\mathcal{X}$ and $\mathrm{rng}(p_{\mathcal{H}}^{real}|_{\langle\mathbb{S}_{\mathcal{H}}\rangle\times\mathcal{X}})=\{0,1\}$. It follows that the following function is also partial computable:

$$g(e)=\begin{cases}0 & \text{if } e=0\\ p_{\mathcal{H}}^{real}(\langle S^e\rangle,x(e)) & \text{otherwise.}\end{cases}$$

We will show that for any $e>0$ such that $\varphi_e(0)\downarrow$, we have that $g(e)\neq\varphi_e(e)$. First, if $\varphi_e(e)\downarrow\in\{0,1\}$, $(S^e,x(e))$ is optimally significant for $\mathcal{H}$ and $g(e)=f(e)=1-\varphi_e(e)$. Otherwise, if $\varphi_e(e)\uparrow$ or $\varphi_e(e)\downarrow\notin\{0,1\}$, we must have that $g(e)\downarrow\in\{0,1\}$ since S^e is $\mathcal{H}$-realizable for any e satisfying $\varphi_e(0)\downarrow$. Now, since g is p.c. and each p.c. function has infinitely many indices, there exists $e'>0$ such that $g=\varphi_{e'}$. However, since $g(0)\downarrow$, this would imply the existence of some $e'>0$ such that $\varphi_{e'}(0)\downarrow$ and $g(e')=\varphi_{e'}(e')$, a contradiction. ■

6. C-online learnability

A corollary of Theorem 10 is that the finiteness of the Littlestone dimension characterizes whether a class is online learnable at all—that is, whether it is online learnable with finite mistake bound. Although the class $\mathcal{H}_{ext}^{DR}$ presented in Theorem 29 is not optimally c-online learnable, it is still c-online learnable by the learner that predicts 0 except on instances it has seen labeled 1. In this section, we analyze c-online learning when there is no requirement for optimality. As a first step, we construct a non-RER class of computable hypotheses that has finite Littlestone dimension but is not c-online learnable (Section 6.1). Next, we explore the connection between c-online and CPAC learning and suggest a potential avenue for strengthening the result to the RER setting (Section 6.2).

6.1. Finite Littlestone dimension fails to characterize c-online learning

The following theorem shows that, in the non-RER setting, the finiteness of the Littlestone dimension no longer characterizes c-online learnability.

Theorem 30 *There exists a class $\mathcal{H}\subset\{0,1\}^{\mathbb{N}}$ of computable hypotheses such that $Ldim(\mathcal{H})<\infty$ but $\mathcal{H}$ is not c-online learnable.*

Proof Recall that any c-online learner is a two-place partial computable function. The idea is to construct a class $\mathcal{H}$ such that for any two-place p.c. function A and for any input length T there exists a hypothesis $h\in\mathcal{H}$ and T consecutive domain instances $x_1,\ldots,x_T\in\mathbb{N}$ such that, on the sample $S=((x_t,h(x_t)))_{t=1}^T$, we have that $A(\langle S_{t-1}\rangle,x_t)\neq h(x_t)$ for all time steps $t\in[T]$. Hence, any c-online learner for $\mathcal{H}$ will have an infinite mistake bound.

Formally, define the functions $s_1:n\mapsto\sum_{i=0}^{n}i$ and $s_2:n\mapsto\sum_{i=0}^{n}s_1(i)$. For each $i\in\mathbb{N}$ and $j\leq i$, let $N_i=\{n:s_2(i)\leq n<s_2(i{+}1)\}$ and $N_{i,j}=\{n:s_2(i){+}s_1(j)\leq n<s_2(i){+}s_1(j{+}1)\}$.

Note that the natural numbers can be partitioned into disjoint sets $\mathbb{N} = \sqcup_{i\in\mathbb{N}} N_i$ and each N_i can be further partitioned as $N_i = \sqcup_{j=0}^{i} N_{i,j}$. Let I_1, I_2, I, and m be functions defined as follows: for each $i \in \mathbb{N}$, $j \leq i$, and $n \in N_{i,j}$, $I_1(n) = i$, $I_2(n) = j$, $I(n) = I_1(n) - I_2(n)$, and $m(n) = \min N_{i,j}$.

Let $\{A_e\}_{e\in\mathbb{N}}$ be an effective numbering of all two-place p.c. functions and define the function $L : n \mapsto \mathbb{1}_{[A_{I(n)}(\langle S^n\rangle,\, n)\downarrow=0]}$, where $S^n = ((n', L(n')))_{n'=m(n)}^{n-1}$. Now, let $\mathcal{H}_{split} = \{h_i\}_{i\in\mathbb{N}}$, where

$$h_i(n) = \begin{cases} L(n) & \text{if } I_1(n) = i \\ 0 & \text{otherwise.} \end{cases}$$

For simplicity, let $\mathcal{H} = \mathcal{H}_{split}$. Note that each h_i is computable since $|h_i^{-1}(1)| \leq s_2(i) < \infty$. However, $\mathcal{H}$ is not RER, since otherwise a Turing machine for computing L would exist. Furthermore, $\text{Ldim}(\mathcal{H}) = 1$ since each domain instance is given the label 1 by at most one $h \in \mathcal{H}$.

Now, assume for the sake of contradiction that $\mathcal{H}$ is c-online learnable and let A_e be a c-online learner for $\mathcal{H}$. Since A_e has finite mistake bound, there exists $M \in \mathbb{N}$ such that $M_{A_e}(\mathcal{H}) \leq M$. However, we will show the existence of an $\mathcal{H}$-realizable sample on which A_e errs $M + 1$ times. Let $i = M + e$, $j = M$, and $S = ((n, h_i(n)))_{n=\min N_{i,j}}^{\max N_{i,j}}$. We will show that for each $t \in [|S|] = [M + 1]$, we have that $A_e(\langle S_{t-1}\rangle, n_t) = 1 - h_i(n_t)$, where $n_t = \min N_{i,j} + t - 1$ is the t^{th} domain instance in S. By definition, since $n_t \in N_{i,j}$, we have that $I_1(n_t) = i$; hence, $h_i(n_t) = L(n_t) = \mathbb{1}_{[A_{I(n_t)}(\langle S^{n_t}\rangle,\, n_t)\downarrow=0]}$, where $S^{n_t} = ((n', L(n')))_{n'=m(n_t)}^{n_t-1}$. Note that $I(n_t) = e$ and $S^{n_t} = S_{t-1}$. Therefore, $h_i(t) = \mathbb{1}_{[A_e(\langle S_{t-1}\rangle, n_t)\downarrow=0]}$. Now, since A_e is a c-online learner for $\mathcal{H}$ and S_{t-1} is an $\mathcal{H}$-realizable sample, we will always have that $A_e(\langle S_{t-1}\rangle, n_t) \downarrow\in \{0, 1\}$. Therefore, $h_i(n_t) = 1 - A_e(\langle S_{t-1}\rangle, n_t)$ for each $t \in [M + 1]$ and $M_{A_e}(S) = M + 1$, as required. ■

6.2. Connection between c-online and CPAC learning

It is natural to ask whether Theorem 30 can be extended to the RER setting. That is, does there exist an RER class $\mathcal{H}$ of computable hypotheses such that $\text{Ldim}(\mathcal{H}) < \infty$ but no c-online learner for $\mathcal{H}$ achieves $M_A(\mathcal{H}) < \infty$? In this section, we propose a potential avenue for addressing this question.

Recently, Sterkenburg (2022) proved a necessary condition for agnostic improper CPAC learnability and constructed an RER class of finite VC-dimension not satisfying this condition. In Lemma 34, we show that this condition is also necessary for agnostic c-online learnability. In particular, we show that any class that is agnostically c-online learnable is also agnostically improperly CPAC learnable but by a probabilistic learner (Lemma 33).

Thus far, we have been concerned with *realizable c-online learners*—learners whose predictions are only guaranteed to be computable on realizable samples. We therefore extend the definition of agnostic online learning introduced by Ben-David et al. (2009) to the computable setting. Let $\mathcal{X} = \mathbb{N}$ and $\mathcal{H} \subset \{0, 1\}^{\mathcal{X}}$ be any class of computable hypotheses. An *agnostic c-online learner* $A : \mathbb{N}^2 \to \mathbb{Q} \cap [0, 1]$ is a two-place total computable function, where for any sample $S \in \mathbb{S}$ and any domain instance $x \in \mathcal{X}$, $A(\langle S\rangle, x)$ is the probability of predicting the label 1 on the given input.[6] The *loss* of a hypothesis $h : \mathcal{X} \to [0, 1]$ on a labeled instance (x, y) is $\ell(h, (x, y)) = \mathbb{P}_{p\sim\text{Bernoulli}(h(x))}[p \neq y] = |h(x) - y|$. The *expected regret* of an agnostic c-online learner A with respect to $\mathcal{H}$ and a sample size T is $\mathbb{E}[R_A(\mathcal{H}, T)] = \sup_{S=((x_t,y_t))_{t=1}^T} \left[\sum_{t=1}^T \ell(A_t, (x_t, y_t)) - \inf_{h\in\mathcal{H}} \sum_{t=1}^T \ell(h, (x_t, y_t))\right]$, where

6. Since there exists a computable bijection between $\mathbb{N}$ and $\mathbb{Q} \cap [0, 1]$, we can assume, without loss of generality, that A is a valid computable function.

$A_t = A(\langle S_{t-1}\rangle, \cdot)$. The *error* of $h : \mathcal{X} \to [0,1]$ w.r.t. a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ is $L_{\mathcal{D}}(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\, \ell(h,(x,y))$ and the error of a hypothesis class $\mathcal{H}$ w.r.t. D is $L_{\mathcal{D}}(\mathcal{H}) = \inf_{h\in\mathcal{H}} L_{\mathcal{D}}(h)$.

Definition 31 (agnostic c-online learnable) *A class $\mathcal{H} \subset \{0,1\}^{\mathbb{N}}$ of computable hypotheses is* agnostically c-online learnable *if there exists an agnostic c-online learner A whose expected regret grows sublinearly in the length of the input sample. That is,* $\lim_{T\to\infty} \frac{\mathbb{E}[R_A(\mathcal{H},T)]}{T} = 0$.

Definition 32 ((agnostic) improper CPAC learnable by a probabilistic learner) *A class $\mathcal{H}$ of computable hypotheses is* improperly CPAC learnable by a probabilistic learner (in the realizable setting) *if there exists a partial computable function $A : \mathbb{N}^2 \to \mathbb{Q}\cap[0,1]$ and a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ such that dom$(A) \supseteq \langle \mathbb{S}_{\mathcal{H}}\rangle \times \mathcal{X}$ and for all $\epsilon,\delta \in (0,1)$, all $m \geq m_{\mathcal{H}}(\epsilon,\delta)$, and all distributions $\mathcal{D}$ over $\mathcal{X}\times\mathcal{Y}$ that satisfy $L_{\mathcal{D}}(\mathcal{H}) = 0$, we have that with probability at least $1-\delta$ over $S \sim \mathcal{D}^m$, $L_{\mathcal{D}}(A_S) \leq L_{\mathcal{D}}(\mathcal{H}) + \epsilon$, where $A_S = A(\langle S\rangle, \cdot)$. We say that $\mathcal{H}$ is* agnostically improperly CPAC learnable by a probabilistic learner *if A is a total computable function and the above condition holds for any distributions $\mathcal{D}$ over $\mathcal{X}\times\mathcal{Y}$.*

Lemma 33 (computable online-to-batch conversion) *Let $\mathcal{H} \subset \{0,1\}^{\mathbb{N}}$ be any class of computable hypotheses that is (agnostically) c-online learnable. Then, $\mathcal{H}$ is (agnostically) improperly CPAC learnable by a probabilistic learner.*

Proof Let A be an agnostic c-online learner for $\mathcal{H}$. We use A to construct an agnostic improper CPAC learner B for $\mathcal{H}$ that is probabilistic. For any $S = ((x_t,y_t))_{t=1}^T$ and $x \in \mathcal{X}$, define $B(\langle S\rangle, x) = \frac{1}{T}\sum_{t=1}^T A(\langle S_{t-1}\rangle, x)$. We can think of B as representing an algorithm that uniformly at random picks some $t \in [T]$ and outputs $A(\langle S_{t-1}\rangle,\cdot)$ as its hypothesis. As required, B is a computable function from $\mathbb{N}^2$ into $\mathbb{Q}\cap[0,1]$. The proof that B is a PAC learner for $\mathcal{H}$ follows from the standard online-to-batch conversion argument (see Kakade and Tewari, 2008; Shalev-Shwartz and Ben-David, 2014, Exercise 21.7.5). The proof can also be extended to the realizable setting. ■

Lemma 34 (necessary condition for agnostic c-online learnability) *Let $\mathcal{H} \subset \{0,1\}^{\mathbb{N}}$ be any class of computable hypotheses that is agnostically c-online learnable. Then, $\mathcal{H}$ satisfies the following two conditions: (1) Ldim$(\mathcal{H}) < \infty$ and (2) for sufficiently large n, there exists an algorithm C_n that on any input $X \subset \mathcal{X}$ of size n, outputs a labeling $g : X \to \{0,1\}$ for which $((x,g(x)))_{x\in X}$ is not $\mathcal{H}$-realizable.*

Proof The first condition follows from Ben-David et al. (2009), who showed that $\mathcal{H}$ is agnostically online learnable in the standard setting iff Ldim$(\mathcal{H}) < \infty$. The second condition follows almost directly form Sterkenburg (2022, Lemma 9), who showed that if $\mathcal{H}$ is agnostically improperly CPAC learnable, for sufficiently large n, there exists an algorithm C_n satisfying the stated property. Their proof, which follows from the Computable No-Free-Lunch theorem (Agarwal et al., 2020, Lemma 19), can also be extended to probabilistic learners. Hence, the result follows from Lemma 33. ■

Open Question Is there an RER class of computable hypotheses with finite Littlestone dimension that is not c-online learnable? Lemma 33 suggests one approach to addressing this question: constructing a class with finite Littlestone dimension that is not improperly CPAC learnable (by a probabilistic learner). Similarly, Lemma 34 could be applied to construct a class that is not c-online learnable in the *agnostic* setting.

In Appendix D, we show that the class $\mathcal{H}_{init}$ presented by Sterkenburg (2022, Theorem 10)—the only known RER class of finite VC-dimension that is not improperly CPAC learnable—has infinite Littlestone dimension. Hence, this class cannot be used to address the question stated above. It remains open whether there exists an RER class of computable functions that has finite Littlestone dimension but is not improperly CPAC learnable.

7. Conclusion and Future Work

In this paper, we investigate computable online learning under three different settings. First, we formalize anytime optimal (a-optimal) online learning, a natural conceptualization of "optimality," and show that it is computationally more difficult than optimal online learning. Second, we give a necessary and sufficient condition for optimal c-online learning and prove that the Littlestone dimension no longer characterizes the optimal mistake bound of c-online learning. Finally, we demonstrate that, in the non-RER setting, the finiteness of the Littlestone dimension no longer determines whether a class is c-online learnable with finite mistake bound. Although this last result remains open in the RER setting, we show that it is equivalent to asking whether there exists an RER class of computable functions that has finite Littlestone dimension but is not improperly CPAC learnable.

As we have shown that some very fundamental results from online learning fail in the computable setting, it would be interesting for future work to explore computable online learning in various related settings—for example, agnostic online learning, proper online learning, and differentially private PAC learning.

Furthermore, similar to Sterkenburg (2022)'s characterization of proper CPAC learning, our characterization of optimal c-online learning relies on computability-theoretic concepts. A major remaining open problem is to find purely combinatorial characterizations of computable learnability.

Acknowledgments

We would like to thank CIFAR and the Vector Institute for their support: CIFAR for supporting Shai as a Canada AI CIFAR chair and the Vector Institute for supporting Niki through a research grant and Shai through a faculty appointment. We would also like to thank Alex Bie, Tosca Lechner, and Matt Regehr for interesting and helpful discussions.

References

Sushant Agarwal, Nivasini Ananthakrishnan, Shai Ben-David, Tosca Lechner, and Ruth Urner. On learnability wih computable learners. In *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, volume 117 of *Proceedings of Machine Learning Research*, pages 48–60, 2020.

Sushant Agarwal, Nivasini Ananthakrishnan, Shai Ben-David, Tosca Lechner, and Ruth Urner. Open problem: Are all VC-classes CPAC learnable? In *Proceedings of 34th Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 4636–4641, 2021.

Noga Alon, Mark Bun, Roi Livni, Maryanthe Malliaris, and Shay Moran. Private and online learnability are equivalent. *Journal of the ACM*, 69(4), 2022.

Shai Ben-David, Dávid Pál, and Shai Shalev-Shwartz. Agnostic online learning. In *Proceedings of 22nd Conference on Learning Theory*, 2009.

Shai Ben-David, Pavel Hrubeš, Shay Moran, Amir Shpilka, and Amir Yehudayoff. Learnability can be undecidable. *Nature Machine Intelligence*, 1(1):44–48, 2019.

Hunter Chase and James Freitag. Bounds in query learning. In *Proceedings of 33rd Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 1142–1160, 2020.

Moti Frances and Ami Litman. Optimal mistake bound learning is hard. *Information and Computation*, 144(1):66–82, 1998.

Steve Hanneke, Roi Livni, and Shay Moran. Online learning with simple predictors and a combinatorial characterization of minimax in 0/1 games. In *Proceedings of 34th Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 2289–2314, 2021.

Sham Kakade and Ambuj Tewari. CMSC 35900 lecture 13: Online to batch conversions. *Toyota Technical Institute at Chicago*, 2008. URL https://home.ttic.edu/~tewari/lectures/lecture13.pdf.

Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine learning*, 2(4):285–318, 1988.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: from theory to algorithms*. Cambridge University Press, 2014.

Robert I. Soare. *Turing computability: theory and applications*, volume 4 of *Theory and Applications of Computability*. Springer Berlin Heidelberg, 2016.

Tom F. Sterkenburg. On characterizations of learnability with computable learners. In *Proceedings of 35th Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 3365–3379, 2022.

Appendix A. Proof of Lemma 24

Lemma 35 *Let $\mathcal{H}$ be a hypothesis class such that $\mathit{Ldim}(\mathcal{H}) = d < \infty$. Let $S = ((x_t, y_t))_{t=1}^{T}$ be any $\mathcal{H}$-realizable sample and $x_{T+1} \in \mathcal{X}$ be any domain instance, where $T \in \mathbb{N}$. Then, the following conditions are equivalent:*

A. For each $t \in [T]$, $\mathit{Ldim}(\mathcal{H}_{S_{t-1}}) = \max\limits_{r \in \{0,1\}} \mathit{Ldim}(\mathcal{H}_{S_{t-1}}^{(x_t,r)})$ and $\mathit{Ldim}(\mathcal{H}_{S_t}) \geq \mathit{Ldim}(\mathcal{H}_{S_{t-1}}) - 1$

B. $M_A(S) = \mathit{Ldim}(\mathcal{H}) - \mathit{Ldim}(\mathcal{H}_S)$ for every online learner A that is optimal for $\mathcal{H}$.

Furthermore, for all $t \in [T]$ and all optimal online learners A, we have that $A(S_{t-1}, x_t) = \arg\max_{r \in \{0,1\}} \mathit{Ldim}(\mathcal{H}_{S_{t-1}}^{(x_t,r)})$.

Proof (A $\implies$ B) Assume that condition A holds and let A^* be an a-optimal online learner for $\mathcal{H}$. Note that, by Lemma 22, each (S_{t-1}, x_t) is an a-optimally significant input and $A^*(S_{t-1}, x_t) = r_t^* = \arg\max_{r\in\{0,1\}} \text{Ldim}(\mathcal{H}_{S_{t-1}}^{(x_t,r)})$. Hence, it follows from condition A that the Littlestone dimension of the version space decreases iff A^* errs and decreases by at most one at each time step. Therefore, $M_{A^*}(S_t) = d - \text{Ldim}(\mathcal{H}_{S_t})$ for any $t \in [T]$.

We will show that condition B holds by showing that, for each $t \in [T]$, every optimal online learner must agree with A^* on (S_{t-1}, x_t). Assume for the sake of contradiction that there exists an optimal online learner A such that for some $t \in [T]$, $A(S_{t-1}, x_t) = 1 - r_t^*$. Let τ be the earliest such time step. Then, on the sample $S_{\tau-1}{}^\frown((x_\tau, r_\tau^*))$, A errs $M_{A^*}(S_{\tau-1}) + 1$ times. However, by Lemma 21, A can be made to err at least $\text{Ldim}(\mathcal{H}_{S_{\tau-1}}^{(x_\tau,r_\tau^*)}) = \text{Ldim}(\mathcal{H}_{S_{\tau-1}}) = d - M_{A^*}(S_{\tau-1})$ more times, a contradiction.

(B $\implies$ A) Let A^* be an a-optimal online learner such that $A^*(S_{t-1}, x_t) = y_t$ for all (S_{t-1}, x_t) that are not a-optimally significant. That is, A^* errs iff $\text{Ldim}(\mathcal{H}_{S_{t-1}}^{(x_t,y_t)}) < \text{Ldim}(\mathcal{H}_{S_{t-1}}^{(x_t,1-y_t)})$. Furthermore, $M_{A^*}(S) \leq d - \text{Ldim}(\mathcal{H}_S)$, as every time A^* errs the Littlestone dimension of the version space decreases by at least one. We will show that if condition A does not hold, this inequality is strict.

First, if there exists $t \in [T]$ such that $\text{Ldim}(\mathcal{H}_{S_{t-1}}^{(x_t,y_t)}) < \text{Ldim}(\mathcal{H}_{S_{t-1}})$ and $\text{Ldim}(\mathcal{H}_{S_{t-1}}^{(x_t,1-y_t)}) < \text{Ldim}(\mathcal{H}_{S_{t-1}})$, there are two cases. Either A^* does not err at time step t and the Littlestone dimension of the version space decreases by at least one, or A^* errs and the Littlestone dimension of the version space decreases by at least two. Similarly, if there exists $t \in [T]$ such that $\text{Ldim}(\mathcal{H}_{S_t}) \leq \text{Ldim}(\mathcal{H}_{S_{t-1}}) - 2$, the Littlestone dimension of the version space goes down by at least one more than the number of mistakes made. In either case, $M_{A^*}(S) < d - \text{Ldim}(\mathcal{H}_S)$. ■

Appendix B. Proof of Theorem 29

B.1. Littlestone dimension of $\mathcal{H}_{ext}^{DR}$

Lemma 36 *Ldim*$(\mathcal{H}_{ext}^{DR}) = 2$.

Proof For simplicity, let $\mathcal{H} = \mathcal{H}_{ext}^{DR}$. First, we will show that $\text{Ldim}(\mathcal{H}) \geq 2$. Consider any three distinct indices $e_1, e_2, e_3 \in \mathbb{N}$ such that $\varphi_{e_i}(0) \downarrow$ for all $i \in [3]$ and $\varphi_{e_1}(e_1) \downarrow = 1$. Then, the $\mathbb{N}$-labeled tree of depth 2 given by $2^{e_2} \leftarrow 2^{e_1} \rightarrow 2^{e_1}3^{c_0(e_1)}$ is shattered by $\chi_{\{2^{e_3}, 2^{e_3}3^{c_0(e_3)}\}}$, $\chi_{\{2^{e_2}, 2^{e_2}3^{c_0(e_2)}\}}$, $\chi_{\{2^{e_1}, 2^{e_1}5^{c_0(e_1)}, 2^{e_1}7^{c_{e_1}(e_1)}\}}$, $\chi_{\{2^{e_1}, 2^{e_1}3^{c_0(e_1)}\}} \in \mathcal{H}$.

Next, we will show that $\text{Ldim}(\mathcal{H}) \leq 2$ by showing the existence of a learner B (not necessarily computable) which errs at most twice on any $\mathcal{H}$-realizable sample. B predicts 0 until (possibly) a mistake is made on x_1. There are two cases for x_1. If $x_1 = 2^e y^i$ for some $e, i \in \mathbb{N}$ s.t. $i > 0$ and $y \in \{3, 5, 7, 11, 13\}$, B matches $\chi_{\{2^e, 2^e y^i\}}$ until a mistake is potentially made on x_2, at which point it matches the target function $\chi_{\{2^e, 2^e y^i, x_2\}}$ and does not err again. If $x_1 = 2^e$ for some $e \in \mathbb{N}$, there are three cases. If $\varphi_e(e) \downarrow = 1$, B matches $\chi_{\{2^e, 2^e 5^{c_0(e)}\}}$, if $\varphi_e(e) \downarrow = 0$, B matches $\chi_{\{2^e, 2^e 13^{c_e(e)}\}}$, and otherwise B matches $\chi_{\{2^e, 2^e 3^{c_0(e)}\}}$. In either case, B can be made to err at most once more. ■

B.2. Proof that $\mathcal{H}_{ext}^{DR}$ is DR

Lemma 37 *$\mathcal{H}_{ext}^{DR}$ is decidably representable.*

Proof First, note that the set $\{(e,i,x) : C_i^{(x)}$ is a halting certificate for P_e on input $x\}$ is decidable by the following Turing machine P_{cert}. On any input (e,i,x), after ensuring that $i > 0$, P_{cert} simulates running P_e on input x and checks each configuration that P_e goes through against the corresponding one in $C_i^{(x)}$. If at any point the configurations are not the same or if there are no more configurations left to check from $C_i^{(x)}$, P_{cert} halts and outputs 0. Otherwise, if P_e halts on input x and all the configurations match, P_{cert} halts and outputs 1. P_{cert} is guaranteed to halt since $C_i^{(x)}$ is a finite sequence of configurations.

Now, we will show that the set $\{y : \exists h \in \mathcal{H}\ (D_y = h^{-1}(1))\}$ is decidable by the following Turing machine P. Given the canonical index y of any finite set as input, P first decodes y into its associated set D_y and checks if D_y equals any of the sets $\{2^e, 2^e3^i\}$, $\{2^e, 2^e5^i, 2^e7^j\}$, $\{2^e, 2^e5^i, 2^e11^j\}$, $\{2^e, 2^e5^i, 2^e13^j\}$, $\{2^e, 2^e3^i, 2^e13^j\}$ for some $e,i,j \in \mathbb{N}$ such that $i,j > 0$. If not, P halts and outputs 0. Otherwise, if $D_y = \{2^e, 2^e3^i\}$, P halts and outputs the result of running P_{cert} on $(e,i,0)$. Otherwise, P evaluates P_{cert} on $(e,i,0)$ and (e,j,e) and, if either result is 0, halts and outputs 0. If both invocations of P_{cert} yield 1, let r be the result of evaluating P_e on input e. P outputs 1 if $r = 0$ and $2^e13^j \in D_y$ or if $r = 1$ and $2^e13^j \notin D_y$. Otherwise, it outputs 0. ■

B.3. Optimally significant inputs for $\mathcal{H}_{ext}^{DR}$

Lemma 38 *For each $e \in \mathbb{N}$, let $S^e = ((2^e, 1))$ and define the p.c. function $x : e \mapsto 2^e3^{c_0(e)}$. $(S^e, x(e))$ is a significant input w.r.t. optimal online learning $\mathcal{H}_{ext}^{DR}$ iff $\varphi_e(0)\downarrow$ and $\varphi_e(e)\downarrow \in \{0,1\}$. Furthermore, for any optimal online learner A for $\mathcal{H}_{ext}^{DR}$, if $\varphi_e(0)\downarrow$ and $\varphi_e(e)\downarrow = r$ for some $r \in \{0,1\}$, $A(S^e, x(e)) = 1 - r$*

Proof Let $\mathcal{H} = \mathcal{H}_{ext}^{DR}$. First, consider any $e \in \mathbb{N}$ such that $\varphi_e(0)\downarrow$ and $\varphi_e(e)\downarrow \in \{0,1\}$. We will show that $(S^e, x(e))$ is an optimally significant input by showing that it satisfies Lemma 24. That is, we need to show that $\mathrm{Ldim}(\mathcal{H}_{S^e}) \geq \mathrm{Ldim}(\mathcal{H}) - 1$, $\mathrm{Ldim}(\mathcal{H}) = \max_{r\in\{0,1\}} \mathrm{Ldim}(\mathcal{H}^{(2^e,r)})$, and $\mathrm{Ldim}(\mathcal{H}_{S^e}) = \max_{r\in\{0,1\}} \mathrm{Ldim}(\mathcal{H}_{S^e}^{(x(e),r)})$.

By Lemma 36, $\mathrm{Ldim}(\mathcal{H}) = 2$, and it is easy to verify that $\mathrm{Ldim}(\mathcal{H}^{(2^e,0)}) = 2$ and $\mathrm{Ldim}(\mathcal{H}_{S^e}) = 1$. Hence, the first two conditions are satisfied. For the third condition there are two cases. Note that for $r \in \{0,1\}$,

$$\mathrm{Ldim}(\mathcal{H}_{S^e}^{(x(e),r)}) = \begin{cases} \{\chi_{\{2^e, 2^e5^{c_0(e)}, 2^e7^{c_e(e)}\}}, \chi_{\{2^e, 2^e5^{c_0(e)}, 2^e11^{c_e(e)}\}}\} & \text{if } r = 0 \text{ and } \varphi_e(e)\downarrow = 1 \\ \{\chi_{\{2^e, 2^e3^{c_0(e)}\}}, \chi_{\{2^e, 2^e3^{c_0(e)}, 2^e13^{c_e(e)}\}}\} & \text{if } r = 1 \text{ and } \varphi_e(e)\downarrow = 0. \end{cases}$$

Hence, $\mathrm{Ldim}(\mathcal{H}_{S^e}^{(x(e),1-\varphi_e(e))}) = \mathrm{Ldim}(\mathcal{H}_{S^e}) = 1$ and by Lemma 24, $(S^e, x(e))$ is an optimally significant input and $A(S^e, x(e)) = 1 - \varphi_e(e)$ for any optimal online learner A, as required.

Conversely, for any $e \in \mathbb{N}$ such that $\varphi_e(0)\uparrow$, S^e is not $\mathcal{H}$-realizable and $(S^e, x(e))$ cannot be an optimally significant input. Now, for any $e \in \mathbb{N}$ such that $\varphi_e(0)\downarrow$ but $\varphi_e(e) \notin \{0,1\}$, $\mathcal{H}_{S^e} = \{\chi_{\{2^e, 2^e3^{c_0(e)}\}}\}$ and $\mathrm{Ldim}(\mathcal{H}_{S^e}) = 0 < \mathrm{Ldim}(\mathcal{H}) - 1$. Hence, Lemma 24 is not satisfied and $(S^e, x(e))$ is not an optimally significant input. ■

Appendix C. Extending Theorem 23 to the DR setting

In this section, we extend Theorem 23 to the DR setting. The technique is similar to that used in the proof of Theorem 29.

Theorem 39 *There exists a DR class $\mathcal{H} \subset \{0,1\}^{\mathbb{N}}$ of computable hypotheses with finite Littlestone dimension such that $\mathcal{H}$ is optimally c-online learnable but not a-optimally c-online learnable.*

Proof For each $x \in \mathbb{N}$, let the p.c. function c_x be defined as in Theorem 29 and consider the following class:

$$\mathcal{H}^{DR}_{halt} = \bigcup_{e\in\mathbb{N}:\ \varphi_e(0)\downarrow} \left\{\chi_{\left\{2^e,\ 2^e 3^{c_0(e)}\right\}}\right\} \cup \bigcup_{e\in\mathbb{N}:\ \varphi_e(0)\downarrow \text{ and } \varphi_e(e)\downarrow} \left\{\chi_{\left\{2^e,\ 2^e 5^{c_0(e)},\ 2^e 7^{c_e(e)}\right\}}, \chi_{\left\{2^e,\ 2^e 5^{c_0(e)},\ 2^e 11^{c_e(e)}\right\}}\right\}.$$

For simplicity, let $\mathcal{H} = \mathcal{H}^{DR}_{halt}$. Since $|h^{-1}(1)| \leq 3$ for each $h \in \mathcal{H}$, we have that $\mathrm{Ldim}(\mathcal{H}) < \infty$. Furthermore, each $h \in \mathcal{H}$ is computable since $c_x(e)$ is evaluated only if $\varphi_e(x)\downarrow$. To show that $\mathcal{H}$ is DR, the same proof technique presented in Appendix B.2 can be applied.

Now, assume for the sake of contradiction that there exists a computable a-optimal online learner A for $\mathcal{H}$. For each $e \in \mathbb{N}$, let $S^e = ((2^e, 1))$ and define the p.c. functions $x : e \mapsto 2^e 3^{c_0(e)}$ and $f : e \mapsto A(\langle S^e\rangle, x(e))$. We will show that

$$f(e) = A(\langle S^e\rangle, x(e)) = \begin{cases} 1 & \text{if } \varphi_e(0)\downarrow \text{ and } \varphi_e(e)\uparrow \\ 0 & \text{if } \varphi_e(0)\downarrow \text{ and } \varphi_e(e)\downarrow \\ \text{undefined} & \text{if } \varphi_e(0)\uparrow. \end{cases}$$

First, note that $f(e)\downarrow$ iff $\varphi_e(0)\downarrow$: if $\varphi_e(0)\downarrow$, S^e is $\mathcal{H}$-realizable and $c_0(e)\downarrow$; otherwise, S^e is not $\mathcal{H}$-realizable and $c_0(e)\uparrow$. Next, we show by Lemma 22 that if $\varphi_e(0)\downarrow$, we must have that $(S^e, x(e))$ is a-optimally significant for $\mathcal{H}$. Note that for any e such that $\varphi_e(0)\downarrow$ we must have that

$$\mathcal{H}^{(x(e),1)}_{S^e} = \left\{\chi_{\left\{2^e, 2^e 3^{c_0(e)}\right\}}\right\}$$

and

$$\mathcal{H}^{(x(e),0)}_{S^e} = \begin{cases} \emptyset & \text{if } \varphi_e(e)\uparrow \\ \left\{\chi_{\left\{2^e,\ 2^e 5^{c_0(e)},\ 2^e 7^{c_e(e)}\right\}}, \chi_{\left\{2^e,\ 2^e 5^{c_0(e)},\ 2^e 11^{c_e(e)}\right\}}\right\} & \text{if } \varphi_e(e)\downarrow \end{cases}$$

Therefore, if $\varphi_e(0)\downarrow$ and $\varphi_e(e)\uparrow$, $\mathrm{Ldim}(\mathcal{H}^{(x(e),1)}_{S^e}) = 0 > \mathrm{Ldim}(\mathcal{H}^{(x(e),0)}_{S^e}) = -1$ and $f(e) = 1$. On the other hand, if $\varphi_e(0)\downarrow$ and $\varphi_e(e)\downarrow$, $\mathrm{Ldim}(\mathcal{H}^{(x(e),0)}_{S^e}) = 1 > \mathrm{Ldim}(\mathcal{H}^{(x(e),1)}_{S^e}) = 0$ and $f(e) = 0$. Next, we can use f to construct the following p.c. function:

$$g(e) = \begin{cases} 1 & \text{if } e = 0 \\ 1 & \text{if } e > 0 \text{ and } f(e) = 1 \\ \text{undefined} & \text{if } e > 0 \text{ and } f(e) = 0 \text{ or } f(e)\uparrow. \end{cases}$$

Since g is a p.c. function, there exists e such that $\varphi_e = g$. Furthermore, since each p.c. function has infinitely many indices, we can assume that $e > 0$. Now, by definition of g, since $e > 0$,

$$g(e)\downarrow \iff f(e) = 1 \iff \varphi_e(0)\downarrow \ \wedge\ \varphi_e(e)\uparrow \iff g(e)\uparrow,$$

contradicting the existence of an a-optimal c-online learner for $\mathcal{H}$.

Although $\mathcal{H}$ is not a-optimally c-online learnable, we can show that there exists a computable optimal online learner B for $\mathcal{H}$. It is easy to verify that $\text{Ldim}(\mathcal{H}) \geq 2$; hence, it suffices to show that $M_B(\mathcal{H}) = 2 = \text{Ldim}(\mathcal{H})$. B predicts 0 until a mistake is made on $(x_1, 1)$. There are three cases for x_1. If $x_1 = 2^e y^i$ for some $e, i \in \mathbb{N}$ such that $i > 0$ and $y \in \{5, 7, 11\}$, B will match the function $\chi_{\{2^e, 2^e 5^{c_0(e)}, 2^e y^i\}}$. Since $(x_1, 1)$ is realizable iff $\varphi_e(0)\downarrow$ and $\varphi_e(e)\downarrow$, B's hypothesis is computable and can be made to err at most once before the target function is determined. If $x_1 = 2^e 3^i$ for some $e, i \in \mathbb{N}$ such that $i > 0$, B will match the target function $\chi_{\{2^e, 2^e 3^i\}}$ and make no further mistakes. Finally, if $x_1 = 2^e$ for some $e \in \mathbb{N}$, B matches $\chi_{\{2^e, 2^e 5^{c_0(e)}\}}$, which is computable since $\varphi_e(0)\downarrow$. B can only be made to err on $(2^e 3^{c_0(e)}, 1)$, $(2^e 5^{c_0(e)}, 0)$, $(2^e 7^{c_e(e)}, 1)$, or $(2^e 11^{c_e(e)}, 1)$ (the last two only if $\varphi_e(e)\downarrow$), after which it will match the target function and not err again. ■

Appendix D. Littlestone dimension of $\mathcal{H}_{init}$

In this section, we show that the class $\mathcal{H}_{init}$ presented by Sterkenburg (2022, Theorem 10) has infinite Littlestone dimension.

Proposition 40 *Define $\mathcal{H}_{init} = \{h_s\}_{s\in\mathbb{N}}$, where, for each $s, x \in \mathbb{N}$,*

$$h_s(x) = \begin{cases} 1 & \text{if } \varphi_{x,s}(x)\downarrow \\ 0 & \text{otherwise,} \end{cases}$$

and $\varphi_{i,s}(x)\downarrow$ denotes that φ_i halts on input x within s computation steps. Then, $\text{Ldim}(\mathcal{H}_{init}) = \infty$.

Proof We say that a hypothesis class $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ contains k thresholds if there are $x_1, \ldots, x_k \in \mathcal{X}$ and $h_1, \ldots, h_k \in \mathcal{H}$ such that for all $i, j \in [k]$, $h_i(x_j) = \mathbb{1}_{[i\geq j]}$. It is not difficult to show that if $\mathcal{H}$ contains 2^n thresholds, then $\text{Ldim}(\mathcal{H}) \geq n$ (see Alon et al., 2022, Appendix A). We will show that $\text{Ldim}(\mathcal{H}_{init}) = \infty$ by showing that for each $k \in \mathbb{N}$, $\mathcal{H}$ contains k thresholds.

Define $H = \{z \in \mathbb{N} : \varphi_z(z)\downarrow\}$ and for any $z \in H$, let $s_z = \arg\min_{s\in\mathbb{N}} \varphi_{z,s}(z)\downarrow$. That is, s_z is the earliest time step at which $\varphi_z(z)\downarrow$. First, we will show that for each $z_1 \in H$, there exists $z_2 \in H$ such that $s_{z_2} > s_{z_1}$. That is, $\varphi_{z_2}(z_2)$ converges strictly after $\varphi_{z_1}(z_1)$. Assume by way of contradiction that there exists some $z_1 \in H$ such that for all $z_2 \in H$, $s_{z_2} \leq s_{z_1}$. Then, $H = \{z \in \mathbb{N} : \varphi_{z,s_{z_1}}(z)\downarrow\}$ and $\overline{H} = \{z \in \mathbb{N} : \varphi_{z,s_{z_1}}(z)\uparrow\}$. However, this would imply that $\overline{H}$ is recursively enumerable, which contradicts the undecidability of H.

Therefore, for any $k \in \mathbb{N}$, there exist $x_1, \ldots, x_k \in H$ such that $s_{x_1} < \ldots < s_{x_k}$. Note that $h_{s_{x_1}}, \ldots, h_{s_{x_k}}$ form k thresholds over these instances, since for each $i, j \in [k]$, $h_{s_{x_i}}(x_j) = \mathbb{1}_{[\varphi_{x_j,s_{x_i}}(x_j)\downarrow]} = \mathbb{1}_{[s_{x_i} \geq s_{x_j}]} = \mathbb{1}_{[i \geq j]}$. ■