Teaching via Best-Case Counterexamples in the Learning-with-Equivalence-Queries Paradigm

Anonymous Author(s)
Affiliation
Address
email

Abstract

We study the sample complexity of teaching, termed as "teaching dimension" (TD) in the literature, for the learning-with-equivalence-queries (LwEQ) paradigm. More concretely, we consider a learner who asks equivalence queries (i.e., "is the queried hypothesis the target hypothesis?"), and a teacher responds either "yes" or "no" along with a counterexample to the queried hypothesis. This learning paradigm has been extensively studied when the learner receives worst-case or random counterexamples; in this paper, we consider the optimal teacher who picks best-case counterexamples to teach the target hypothesis within a hypothesis class. For this optimal teacher, we introduce LwEQ-TD, a notion of TD capturing the teaching complexity (i.e., the number of queries made) in this paradigm. We show that a significant reduction in queries can be achieved with best-case counterexamples, in contrast to worst-case or random counterexamples, for different hypothesis classes. Furthermore, we establish new connections of LwEQ-TD to the well-studied notions of TD in the learning-from-samples paradigm.

1 Introduction

Learning-with-queries paradigm involves a learner who asks structured queries to a teacher in order to locate a target hypothesis. This paradigm has been extensively studied in machine learning and formal methods research, including automata learning [1–4], model checking [5], oracle-guided synthesis (OGIS) [6], model learning [7], among others. Classical literature involves different kinds of queries that a learner could ask [1, 8, 9], such as *membership* queries (i.e., "is the queried instance consistent with the target hypothesis?") and *equivalence* queries (i.e., "is the queried hypothesis the target hypothesis?"). In this paper, we consider the learning-with-equivalence-queries (LwEQ) paradigm where the learner is only asking equivalence queries and the teacher responds to a query either "yes" or "no" along with a counterexample on which the current hypothesis disagrees with the target hypothesis. LwEQ paradigm captures a variety of important problem settings, such as counterexample-guided synthesis (CEGIS) [10], data augmentation (CEGAR) [11], and learning regular languages from counterexamples [12]. The focus of our work is to understand the query complexity for the LwEQ paradigm, i.e., the number of equivalence queries needed by the learner for the exact identification of a target hypothesis [9].

The query complexity clearly depends on the learner model (i.e., the query function deciding the next hypothesis to query), as well as the informativeness of the counterexamples provided by the teacher. In the literature, the query complexity for the LwEQ paradigm has been extensively studied when the learner receives worst-case or random counterexamples [1, 9, 13–15]. To show worst-case bounds [1, 13], classical works have studied a teacher who responds with worst-case counterexamples to maximize the number of equivalence queries asked by the learner. Recent work [14] has studied this query complexity when facing a less benign teacher who provides counterexamples selected at

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

random from a known probability distribution. In particular, [14] proposed a learning algorithm (i.e., a query function for the learner) that achieves $\mathcal{O}(\log n)$ query complexity on the expected number of random counterexamples for any hypothesis class of size n. When contrasting this bound with worst-case bounds, it shows that random counterexamples could lead to exponential improvement in the query complexity when compared with worst-case counterexamples [13, 14]. Along these lines, an important research question is to understand the query complexity where the counterexamples are picked by a more informed and helpful teacher, instead of random or worst-case counterexamples.

In this paper, we consider a more powerful teaching setting, where the optimal teacher picks best-case counterexamples to steer the learner towards a target hypothesis. Our goal is to characterize the query complexity for this optimal teacher (also, referred to as *teaching complexity* in the paper), for the LwEQ paradigm. This teaching complexity has been extensively studied for binary classification in the learning-from-samples (LfS) paradigm [16, 17] and is termed as "teaching dimension" (TD) [18–30]. Beyond the LfS paradigm, various notions of teaching complexity have also been studied in other learning paradigms (see Section 1.1 for further details). We introduce a new notion of teaching complexity for the LwEQ paradigm, namely "LwEQ teaching dimension" (LwEQ-TD), capturing the number of queries needed by the learner when best-case counterexamples are provided by the optimal teacher. Our study of LwEQ-TD for some prominent hypothesis classes reveals the power of teaching via best-case counterexamples as we show a significant reduction in query complexity when compared to that for worst-case and random counterexamples. Furthermore, we establish several new connections of LwEQ-TD to the existing notions of TD in the LfS paradigm[1]. Table 1 provides a summary of these different teaching settings and learning paradigms; our main results and contributions are summarized below:

I. We characterize the query complexity for the optimal teacher in the LwEQ paradigm, termed as *learning-with-equivalence-queries teaching dimension* (LwEQ-TD). (see Section 3)

II. We study the query complexity in the LwEQ paradigm under different teaching settings: worst-case, random, and best-case, distinguished by the informativeness of counterexamples. We showcase the power of best-case counterexamples picked by the optimal teacher, in contrast to worst-case or random counterexamples, for different hypothesis classes, including Axes-aligned hyperplanes, Monotone monomials, and Orthogonal rectangles. (see Section 4)

III. We establish new connections between LwEQ-TD and LfS-TD by studying LwEQ-TD for different learner models based on the richness of their query functions. We show that LwEQ-TD is the same as wc-TD [18], RTD [22, 24], and NCTD [27] for a hypothesis class when restricting query functions to specific families. In general, LwEQ-TD is weaker than LfS-TD, e.g., LwEQ-TD is lower-bounded by local-PBTD [26, 29] of the hypothesis class when the learner's next query depends on the previous query. (see Section 5)

Learning \ Teaching	Worst-case Teacher	Random-case Teacher	Best-case Teacher
Learning-with-equivalence-queries (LwEQ)	Worst-case counterexamples [1, 9, 13, 31, 32]	Random counterexamples [14, 33]	LwEQ-TD **This work**
Learning-from-samples (LfS)	Worst-case examples (i.e., least informative)	i.i.d learning [16, 17]	LfS-TD / classical TD [18, 22, 25, 27, 29]

Table 1: An overview of different teaching settings in the context of LwEQ and LfS paradigms.

1.1 Background and Related Work

Learning-with-queries paradigm and equivalence queries. Learning-with-queries paradigm was introduced in [1] which proposed L^* algorithm for exact identification of DFAs (deterministic finite automaton) when the learner is allowed to ask membership queries and equivalence queries. Classical work has studied different types of queries (subset, membership, equivalence, correction, among others) [1, 8, 9]. Among these works, learning with membership queries has been explored in a variety of problem settings, such as PAC learning [16], active learning [34–36], and agnostic learning [37]. In the learning-with-equivalence-queries (LwEQ) paradigm, the learner can ask only equivalence queries; furthermore, in our work, we consider *proper* queries, i.e. the queried hypothesis

[1] We will collectively refer to these different notions of TD in the LfS paradigm as LfS-TD.

is within the hypothesis class (see Section 6 for a discussion on *improper* queries). In this LwEQ paradigm, [13] studied query complexity when the teacher picks worst-case counterexamples for some key hypothesis classes (e.g. DFA, NFA, Context-free Grammars), and showed an exponential lower bound. [14] studied random counterexamples, and proposed a learning algorithm, namely Max-Min, which achieves a substantially improved bound on query complexity. We will investigate the query complexity of Max-Min learner for the hypothesis class of Axes-aligned hyperplanes in Section 4.

Teaching in the LfS paradigm for binary classification. Algorithmic machine teaching, first introduced by [18, 38], studies the interaction between a teacher and a learner where the teacher's goal is to find an optimal sequence of training samples to teach a target hypothesis. [18] introduced a measure of teaching complexity, named teaching dimension (TD) of the hypothesis class, in the learning-from-samples (LfS) paradigm. The classical notion of TD in [18] characterized the minimum number of samples (i.e., examples) needed to teach a target hypothesis to a version-space learner who picks hypothesis within the version space *arbitrarily* (in an adversarial way). In the past two decades, several new teaching settings have been studied, driven by the motivation to lower teaching complexity and to find settings for which TD has better connections with Vapnik–Chervonenkis dimension (VCD) [39]. In particular, several new teaching models and complexity measures have been proposed for both the batch teaching settings (e.g., worst-case [18], recursive [22, 24], preference-based [25], and non-clashing models [27]) and the sequential settings (e.g., local preference-based model [26, 29]). These teaching settings, in turn, lead to different notions of TD, that we collectively refer to as LfS-TD. In recent work, [29, 30] has characterized these different notions of TD through a unified framework of modeling learners with preference/ranking functions. In Section 5, we will build on this framework to model the learner's query functions in the LwEQ paradigm through ranking functions, allowing us to connect LwEQ-TD with LfS-TD.

Teaching in other learning settings. Within binary classification, teaching complexity results have been extended beyond version space learners, including models for gradient learners [40, 41], models inspired by control theory [42, 43], and models for human-centered applications [44–47]. Furthermore, a recent line of research has studied robust notions of teaching in settings where the teacher has limited information about the learner's dynamics [48–50]. Given the importance of teacher-learner interactions in many real-world applications, teaching has also been studied in richer domains. In particular, teaching complexity has been investigated for imitation learning settings where the teacher provides demonstrations [51–55], and for reinforcement learning settings where the teacher provides reward feedback [56, 57]. We see these works as complementary to ours, and we refer the reader to see [28, 58] for an overview.

2 Problem Setup

Teaching framework. Let $\mathcal{X}$ be a ground set of unlabeled instances and $\mathcal{Y}$ the set of labels. Let $\mathcal{H}$ be a finite class of hypotheses; each element $h \in \mathcal{H}$ is a function $h : \mathcal{X} \to \mathcal{Y}$. Here, we only consider boolean functions, and hence $\mathcal{Y} = \{0, 1\}$. Let $\mathcal{Z} \subseteq \mathcal{X} \times \mathcal{Y}$ be the ground set of labeled examples. Each element $z = (x_z, y_z) \in \mathcal{Z}$ represents a labeled example where the label is given by the target hypothesis h^*, i.e., $y_z = h^*(x_z)$. Furthermore, for any $Z \subseteq \mathcal{Z}$, we define *version space* induced by the examples Z as the subset of hypotheses $\mathcal{H}(Z) \subseteq \mathcal{H}$ that is consistent with the labels of all the examples, i.e.,

$$\mathcal{H}(Z) := \{h \in \mathcal{H} \,|\, \forall z = (x_z, y_z) \in Z,\ h(x_z) = y_z\}\,. \tag{1}$$

Equivalence queries. We consider the LwEQ paradigm where a learner seeks to identify a target hypothesis from the hypothesis class via equivalence queries. In an equivalence query, the learner asks if the current hypothesis, say $h' \in \mathcal{H}$, is equivalent to the target hypothesis h^* or not. The teacher provides a response $\mathbf{r}$ where $\mathbf{r}$ is either "yes" if $h' \equiv h^*$ or "no" along with a counterexample $z := (x_z, y_z) \in \mathcal{Z}$, such that $h'(x_z) \neq y_z$.

Learner model and query protocol. We consider a generic model of the learner that captures our assumptions about how the learner conjectures its hypothesis for equivalence queries based on the responses received from the teacher. A key aspect of this model is the learner's query function ℓ over the hypotheses. The learner, based on the information encoded in the inputs of this query function—which include the current hypothesis and the history of counterexamples—will choose one hypothesis

in $\mathcal{H}$. In the beginning, the learner starts with an initial hypothesis $h_0 \in \mathcal{H}$, the history is $Z_0 = \emptyset$, and the version space is $H_0 = \mathcal{H}$. At a time step $t \geq 1$, the learner first picks the hypothesis h_t as follows:

$$\ell(Z_{t-1}, h_{t-1}) \longrightarrow h_t \in \mathcal{H}(Z_{t-1}). \tag{2}$$

where Z_{t-1} is the history of counterexamples seen up until time t and $\mathcal{H}(Z_{t-1})$ is the corresponding version space. The learner then queries h_t for equivalence to h^* and receives a response $\mathbf{r}_t$. Then, the query protocol proceeds as follows: (i) if $\mathbf{r}_t$ is "yes", the learner has identified h^* and stops; (ii) otherwise $\mathbf{r}_t$ is "no" along with a counterexample z_t using which the learner updates $Z_t = Z_{t-1} \cup \{z_t\}$, and continues. We summarize this query protocol in Algorithm 1.

Algorithm 1: Query protocol between the learner and the teacher

Learner's initial hypothesis is $h_0 \in \mathcal{H}$, history is $Z_0 = \emptyset$, and version space is $\mathcal{H}_0 = \mathcal{H}$;
for $t = 1, 2, 3, \cdots$ **do**
  learner picks $h_t \in \mathcal{H}(Z_{t-1})$ based on Z_{t-1} and h_{t-1} as per Eq. (2);
  learner performs an equivalence query with h_t;
  teacher provides a response r_t that is either a "yes" or a 'no" along with a counterexample z_t;
  if r_t *is "yes"* **then**
    learner identifies h^* and stops;
  else
    learner updates $Z_t = Z_{t-1} \cup \{z_t\}$;

We assume that both the learner and the teacher have full knowledge of $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{H}$; in addition, the teacher knows the target h^* as well as the learner's query function ℓ. In this work, we consider learner models which could be characterized by a specific query function ℓ as discussed above. These include well-known learners studied in the literature, such as a *constant* query learner (denoted as ℓ_{const}) who picks the next hypothesis h_t arbitrarily in $\mathcal{H}(Z_{t-1})$ without any preference [1, 18, 29], a *global* query learner (denoted as ℓ_{global}) who uses a global ranking over $\mathcal{H}$ to pick the next hypothesis h_t in $\mathcal{H}(Z_{t-1})$ as per Eq. (2) [25, 29], and the Max-Min learning algorithm (denoted as $\ell_{\text{Max-Min}}$) introduced in recent work on LwEQ paradigm [14].

Complexity of teaching (i.e., the number of queries made). In this paper, we study the number of equivalence queries asked by the learner to the teacher to identify a target hypothesis in Algorithm 1, and we call it the *query complexity* or *teaching complexity for LwEQ* paradigm interchangeably. Clearly, this query complexity depends on the learner's query function ℓ and the choice of counterexamples by the teacher. In the following sections, we study this complexity for different teacher types depending on the informativeness of the provided counterexamples, as well as for different families of learner types.

3 The Query Complexity with Best-Case Teacher: LwEQ-TD

In this section, we consider the optimal teacher who picks best-case counterexamples with the objective of minimizing the learner's queries for identifying h^*. For this optimal teacher, we provide a formal characterization of teaching complexity, namely *teaching dimension for learning-with-equivalence queries* (LwEQ-TD) paradigm, inspired by different notions of teaching dimension for learning-from-samples (LfS-TD) paradigm [18, 24, 26, 29].

Notation. We denote by $\mathcal{L}$ a family of learner models—alternatively, we can think of them as a family of query functions. To begin, we fix a query function $\ell \in \mathcal{L}$ that the learner uses to pick the next hypotheses for equivalence queries. Our characterization below will be based on understanding the minimal "cost" (i.e., the number of queries needed) in steering the learner from a hypothesis h with the history of counterexamples Z to the target hypothesis h^*—in Algorithm 1, h refers to h_{t-1} and Z refers to Z_{t-1} at the beginning of time t.

Minimal cost of steering. We begin by providing a recursive function that captures this cost of steering and will be key to formalize teaching complexity in the LwEQ paradigm. As is typically considered in the LfS paradigm [18, 24, 26, 29], we consider the "adversarial" perspective in how the learner breaks ties in picking the hypothesis w.r.t. its query function in Eq. (2). The optimal cost[2] for steering

[2]Note that D_ℓ in Eq. (3) has a $\max$ operator w.r.t. the learner's choice and $\min$ operator with the teacher's choice [29]. Furthermore, this function does not count the last query when the learner's $h_t = h^*$.

the learner from the current history (Z, h) to h^* in the query protocol of Algorithm 1 is then given by:

$$D_\ell(Z, h, h^*) = \begin{cases} 0, & \text{if } \ell(Z,h) = \{h^*\} \\ 1 + \max\limits_{h' \in \ell(Z,h)} \min\limits_{z : h'(x_z) \neq y_z} & D_\ell(Z \cup \{z\}, h', h^*), \quad \text{otherwise} \end{cases} \tag{3}$$

LwEQ-TD. Given a fixed query function $\ell \in \mathcal{L}$, an initial hypothesis h_0, and a target hypothesis h^*, LwEQ-TD w.r.t. ℓ and h^* is the optimal cost for teaching the target hypothesis h^*:

$$\text{LwEQ-TD}_{\mathcal{X},\mathcal{H},h_0}(\ell, h^*) := D_\ell(\emptyset, h_0, h^*). \tag{4}$$

To characterize the teaching complexity for the hypothesis class, we consider the worst-case target hypothesis in $\mathcal{H}$, given by:

$$\text{LwEQ-TD}_{\mathcal{X},\mathcal{H},h_0}(\ell) = \max_{h^*} D_\ell(\emptyset, h_0, h^*). \tag{5}$$

Finally, to compare LwEQ-TD with the existing notions of TD in the LfS paradigm, we define LwEQ-TD for a given family of query functions $\mathcal{L}$. Based on [24, 27, 29], we define LwEQ-TD w.r.t the family $\mathcal{L}$ as the teaching complexity w.r.t. the best ℓ in that family:

$$\text{LwEQ-TD}_{\mathcal{X},\mathcal{H},h_0}(\mathcal{L}) = \min_{\ell \in \mathcal{L}} \max_{h^*} D_\ell(\emptyset, h_0, h^*). \tag{6}$$

In the following two sections, we will investigate how LwEQ-TD (the complexity of the optimal teacher in the LwEQ paradigm) compares with various other complexity notions. In Section 4, we investigate different teachers in the LwEQ paradigm, showcasing the power of the optimal teacher. In Section 5, we establish new connections of LwEQ-TD to the well-studied notions of LfS-TD in the LfS paradigm.

4 The Query Complexity for Different Teachers in the LwEQ Paradigm

In this section, we study the query complexity in the LwEQ paradigm for different types of teachers. In Section 3, we characterized the query complexity for the optimal teacher who provides best-case counterexamples. The goal of this section is to compare the query complexity for the optimal teacher with other variants of teachers as discussed below. In this section, we will denote the teacher as TEQ, and we consider the following four variants of teachers:

- best-TEQ responds "yes" or "no" along with best-case counterexamples (see Section 3).
- random-TEQ responds "yes" or "no" along with random counterexamples (picked uniformly at random). As discussed in [14], we characterize the query complexity for random-TEQ as the expected number of counterexamples provided by random-TEQ in Algorithm 1.
- worst-TEQ responds "yes" or "no" along with worst-case (least informative) counterexamples. We characterize query complexity for worst-TEQ by replacing the $\min$ over the choice of counterexamples with a $\max$ in the cost function D_ℓ (as discussed for LwEQ-TD in Eq. (3)).
- binary-TEQ responds "yes" or "no" without any counterexamples.

In the rest of the section, we compare the query complexity for these teachers when interacting with different learners and we characterize this complexity by the richness of their query functions.

4.1 Warm-up: Query Complexity Bounds When Teaching Different Types of Learners

In this section, we study the query complexity bounds for different teachers when teaching various learner models characterized by the richness of the underlying query functions. In particular, we consider the query functions ℓ_{const}, ℓ_{global}, and $\ell_{\text{Max-Min}}$ (see Section 2). ℓ_{const} picks the next hypothesis h_t arbitrarily in $\mathcal{H}(Z_{t-1})$ without any preference [29]. $\ell_{\text{Max-Min}}$, introduced in recent work on LwEQ paradigm [14], uses a more rich history-dependent query function and was used specifically for achieving a better query complexity for the case of randomly chosen counterexamples. ℓ_{global} uses a global ranking over $\mathcal{H}$ to pick the next hypothesis h_t in $\mathcal{H}(Z_{t-1})$ as per Eq. (2) and is popularly studied in the LfS paradigm [25, 29]. To study the query complexity for different settings (corresponding to four teachers and three learners), we consider the hypothesis class of Axes-aligned hyperplanes, a simple yet rich hypothesis class, that generalizes the canonical hypothesis class of threshold boolean functions to higher dimensions. Next, we introduce this hypothesis class.

Definition 4.1 (Axes-aligned hyperplanes)**.** *Fix an input space $\mathcal{X} = \{1, \cdots, n\}^d$ in $\mathbb{R}^d$ with labels set $\mathcal{Y} = \{0, 1\}$, where an input $\mathbf{x} \in \mathcal{X}$ is a d-dimensional point in $\mathbb{R}^d$, i.e., $\mathbf{x} := (x_1, x_2, \cdots, x_d)$ such that all $x_i \in [1, n+1]$. We define the hypothesis class of Axes-aligned hyperplanes as:*

$$\mathcal{H}_{\text{axh}} := \{h \mid \exists\, i \in [1, d+1],\ j \in [n+1],\ \text{s.t. } \forall \mathbf{x} \in \mathcal{X},\ \ h(\mathbf{x}) = 1 \text{ if } x_i \leq j,\ \text{otherwise } 0\}\,. \quad (7)$$

We summarize the query complexity bounds for ℓ_{const}, ℓ_{global}, and $\ell_{\text{Max-Min}}$ under different teaching scenarios in Table 2. We note that $|\mathcal{H}_{\text{axh}}| = d \cdot (n+1)$. To capture the global ranking of ℓ_{global} we use the function $g : \mathcal{H}_{\text{axh}} \to [|\mathcal{H}_{\text{axh}}|]$. For a target hypothesis $h^* \in \mathcal{H}_{\text{axh}}$ and ranking function g, the bounds for ℓ_{global} are described using the quantities—$t_g^* := |\{h \in \mathcal{H}_{\text{axh}} \mid g(h) \leq g(h^*)\}|$ (note $0 \leq t_g^* \leq d \cdot (n+1)$) and n_g^* ($0 \leq n_g^* \leq n+1$, where n_g^* is the highest number of hypotheses aligned along an axis (which could be at most $n+1$) preferred over h^* as per the ranking g). In the Appendix of the supplementary, we state the bounds in Table 2 as theorems and provide detailed proof.

Learner \ Teacher	binary-TEQ (Yes/No)	worst-TEQ	random-TEQ	best-TEQ
ℓ_{const}	$d \cdot (n+1)$	$\Omega\left(d \cdot (n+1)\right)$	$\mathcal{O}\left(d \cdot \log(n+1)\right)$	2
ℓ_{global}	t_g^*	$\Omega\left(t_g^*\right)$	$\mathcal{O}\left(d \cdot \log n_g^*\right)$	1
$\ell_{\text{Max-Min}}$	(Not-Applicable)	$\Omega\left(d \cdot \log(n+1)\right)$	$\mathcal{O}\left(\log(d \cdot (n+1))\right)$	2

Table 2: Query complexity for the Axes-aligned hyperplanes considering different teachers and learners; see Section 4.1 for details.

4.2 Lower and Upper Bounds on Query Complexity

Now, we study lower bounds on the query complexity for the setting of three teachers (binary-TEQ, worst-TEQ, random-TEQ) teaching learner models characterized with optimal query functions. In contrast to these lower bounds, we establish upper bounds on the query complexity for the setting of the best-case teacher (best-TEQ) teaching a much weaker learner model characterized with a global query function (see Section 2). To study the bounds in these different teaching settings, we consider two more hypothesis classes beyond Axes-aligned hyperplanes (see Definition 4.1): (i) Monotone monomials [18, 38], and (ii) Orthogonal rectangles [18, 32, 59, 60], which have been extensively studied in the LfS paradigm.

We summarize the results in Table 3 with detailed proofs deferred to the Appendix of the supplementary. First, we state the results on query complexity for Axes-aligned hyperplanes in the following:

Theorem 1. *Consider the hypothesis class of Axes-aligned hyperplanes $\mathcal{H}_{\text{axh}}$ (see Eq. (7)). There exists a global learner ℓ_{global} such that best-*TEQ *achieves* LwEQ-TD(ℓ_{global}) *of exactly 1, whereas for any optimal learner random-*TEQ *provides at least $\Omega\left(\log d\right)$ counterexamples "in expectation", and worst-*TEQ *provides at least $\Omega\left(d + \log(n+1)\right)$ counterexamples "in the worst-case".*

In Section 4.1, we showed that a specific choice of ℓ_{global} upper bounds the query complexity for best-TEQ by 1. On the other hand, as per Theorem 25 (see [14]), the query complexity for a random teacher is connected to the VC dimension of a hypothesis class independent of the learner model. We show that VCD$(\mathcal{H}_{\text{axh}})$ is $\Omega\left(\log d\right)$, which entails a direct lower bound on the query complexity for random-TEQ for any learner model. Furthermore, we note that the worst-TEQ could force any learner to query a hypothesis in every axis in addition to $\Omega\left(\log(n+1)\right)$ queries in the axis of the target hypothesis.

Now, we define the hypothesis class of Monotone monomials before we state the aforementioned query complexity bounds for this class in Theorem 2.

Definition 4.2 (Monotone monomials)**.** *Fix a set of literals $\{v_1, v_2, \cdots, v_n\}$, an input space $\mathcal{X} = \{0, 1\}^n$, and labels set $\mathcal{Y} = \{0, 1\}$. A monotone monomial is a negation-free conjunction of literals. If $mono(i_1, i_2, \cdots, i_k)$ denotes a monotone monomial over the set of k literals $\{v_{i_1}, v_{i_2}, \cdots, v_{i_k}\}$, then it canonically represents a hypothesis $h(\mathbf{x}) := x_{i_1} \wedge x_{i_2} \wedge \cdots \wedge x_{i_k}$ over the input space. With this understanding, we define the hypothesis class of Monotone monomials as:*

$$\mathcal{H}_{\text{mono}} := \{h \mid \exists\ \{i_1, i_2, \cdots, i_k\} \subseteq \{1, 2, \cdots, n\},\ \text{s.t. } \forall \mathbf{x} \in \mathcal{X},\ \ h(\mathbf{x}) = x_{i_1} \wedge \cdots \wedge x_{i_k}\}\,. \quad (8)$$

Theorem 2. *Consider the hypothesis class of Monotone monomials $\mathcal{H}_{\text{mono}}$ (see Eq. (8)). There exists a global learner ℓ_{global} such that best-*TEQ *achieves* LwEQ-TD(ℓ_{global}) *of exactly 1, whereas for any optimal learner model random-*TEQ *provides at least $\Omega\left(n\right)$ counterexamples "in expectation", and worst-*TEQ *provides at least $\Omega\left(n\right)$ counterexamples "in the worst-case".*

Hypothesis class \ Teacher	binary-TEQ (Yes/No)	worst-TEQ	random-TEQ	best-TEQ
Axes-aligned hyperplanes	$\Omega\left(d \cdot (n+1)\right)$	$\Omega\left(d + \log(n+1)\right)$	$\Omega\left(\log d\right)$	1
Monotone monomials	$\Omega\left(2^n\right)$	$\Omega\left(n\right)$	$\Omega\left(n\right)$	1
Orthogonal rectangles	$\Omega\left((n \cdot (n+1))^d\right)$	$\Omega\left(d \cdot \log(n+1)\right)$	$\Omega\left(d\right)$	2

Table 3: Lower and upper bounds on query complexity for the Axes-aligned hyperplanes, Monotone monomials, Orthogonal rectangles considering different teachers and learners.

Now, we define the hypothesis class of Orthogonal rectangles before we state Theorem 3 establishing the aforementioned query complexity bounds.

Definition 4.3 (Orthogonal rectangles)**.** *Fix an input space $\mathcal{X} = \{1, \cdots, n\}^d$ in $\mathbb{R}^d$ with labels set $\mathcal{Y} = \{0, 1\}$, where an input $\mathbf{x} \in \mathcal{X}$ is a d-dimensional point in $\mathbb{R}^d$, i.e, $\mathbf{x} := (x_1, x_2, \cdots, x_d)$ such that all $x_i \in [1, n+1]$. We define the class of Orthogonal rectangles as:*

$$\mathcal{H}_{\text{rec}} := \left\{ h \,\middle|\, \exists \left\{a_j, b_j\right\}_{j \in [1,d+1]} \subset [n+1]\,,\ \text{s.t. } \forall \mathbf{x} \in \mathcal{X},\ \ h(\mathbf{x}) = \begin{cases} 1 & \text{if } \forall j,\ a_j < x_j \leq b_j \\ 0 & \text{otherwise.} \end{cases} \right\}. \tag{9}$$

Theorem 3. *Consider the hypothesis class of Orthogonal rectangles $\mathcal{H}_{\text{rec}}$ (see Eq. (9)). There exists a global learner ℓ_{global} such that best-*TEQ *achieves LwEQ-TD(ℓ_{global}) of exactly 2, whereas for any optimal learner model random-*TEQ *provides at least $\Omega\left(d\right)$ counterexamples "in expectation", and worst-*TEQ *provides at least $\Omega\left(d \cdot \log(n+1)\right)$ counterexamples "in the worst-case".*

As noted for Axes-aligned hyperplanes (see Theorem 1), even for rich classes of Monotone monomials (see Theorem 2) and Orthogonal rectangles (see Theorem 3) the lower bound results are significantly higher than the upper bound results, thereby establishing the power of the optimal teacher.

5 Teaching Dimensions for the LwEQ and LfS Paradigms

In this section, we draw new comparisons of the notion of LwEQ-TD in the LwEQ paradigm to existing notions of TD in the LfS paradigm. Based on the teaching setting and the learner models, one gets different notions of teaching complexity and we collectively refer to these notions as LfS-TD (see Footnote 1) [18, 21, 22, 25, 27, 29, 61]. For comparisons with LfS-TD below, we consider the framework of [29] that provides a unified view of teaching settings by modeling the learners through preference/ranking functions. In the following, we first introduce a notation for these ranking functions σ and then discuss how a ranking function σ, in turn, induces a learner in the LwEQ paradigm.

Learner's query function ℓ using a framework of a ranking σ Consider a hypothesis class $\mathcal{H}$, a version space $H \subseteq \mathcal{H}$, and hypotheses $h', h'', h \in \mathcal{H}$ such that $h', h'' \in H$. Building on the notation for a preference function σ (see [29]), we define a ranking function $\sigma : \mathcal{H} \times 2^{\mathcal{H}} \times \mathcal{H} \to \mathbb{R}$ where $\sigma(h'; H, h)$ signifies how h' is ranked in the version space H from h. Thus, we say h' is *ranked* (or preferred) over h'' in the current version space H from the current hypothesis h if $\sigma(h'; H, h) \leq \sigma(h''; H, h)$, and vice versa. Similar to preference-based learners [29], a learner's query function (see Section 2) could use this ranking to pick the most preferred hypothesis for an equivalence query. In this section, we consider query functions for our learner model ℓ (see Section 2) which use a framework of a ranking function σ to pick the next hypothesis h_t based on the current history of counterexamples seen Z_{t-1} and the current hypothesis h_{t-1} as follows:

$$\ell(Z_{t-1}, h_{t-1}) \longrightarrow h_t \in \underset{h' \in \mathcal{H}(Z_{t-1})}{\arg\min}\ \sigma(h'; \mathcal{H}(Z_{t-1}), h_{t-1}). \tag{10}$$

In this section, we identify a learner model ℓ_σ[3] with the corresponding ranking σ and consequently use the notation LwEQ-TD$_{\mathcal{X},\mathcal{H},h_0}(\sigma)$ (for fixed $\mathcal{X}, \mathcal{H}, h_0$) for LwEQ-TD (see Eq. (5), Section 3) for the learner model ℓ_σ. We denote a family of ranking functions σ as Σ. For a family of ranking functions Σ, the corresponding LwEQ teaching dimension is denoted as LwEQ-TD$_{\mathcal{X},\mathcal{H},h_0}(\Sigma)$ (see Eq. (6), Section 2).

[3]We use the notation ℓ_σ for a learner's query function ℓ using a ranking σ.

For families of learner models induced by specific types of ranking functions, we connect LwEQ-TD to existing notions of LfS-TD. In the following, we consider ranking functions broadly categorized into two classes : (i) ranking functions independent of Z_{t-1} and h_{t-1}, and (ii) ranking functions dependent on Z_{t-1} and/or h_{t-1}.

5.1 LwEQ Learners with Ranking Functions Independent of Z_{t-1} and h_{t-1}

These ranking functions induce learners whose next equivalence query at time t (i.e., choice of the hypothesis h_t) is independent of the history of counterexamples Z_{t-1} and independent of the hypothesis h_{t-1} (see Algorithm 1, Section 2). In Section 4.1, we discussed these learner families under the name of *contant* and *global* learners, and we formalize these families below using the ranking functions framework of Eq. (10). In particular, we introduce two families of ranking functions: (i) Σ_{const} is a family of *constant* ranking functions where $\sigma \in \Sigma_{\text{const}}$ ranks every hypothesis arbitrarily without any preference, and (ii) Σ_{global} is a family of *global* ranking functions where $\sigma \in \Sigma_{\text{global}}$ ranks hypothesis based on a global preference. These two families are given below:

$$\Sigma_{\text{const}} = \{\sigma \mid \exists\, c \in [|\mathcal{H}|]\,,\ \text{s.t.}\, \forall h', Z, h,\ \sigma(h'; \mathcal{H}(Z), h) = c\}\,. \tag{11}$$

$$\Sigma_{\text{global}} = \{\sigma \mid \exists\, g : \mathcal{H} \to [|\mathcal{H}|]\,,\ \text{s.t.}\, \forall h', Z, h,\ \sigma(h'; \mathcal{H}(Z), h) = g(h')\}\,. \tag{12}$$

Given these ranking functions, the following theorem establishes the connection of LwEQ-TD$_{\mathcal{X},\mathcal{H},h_0}(\Sigma_{\text{const}})$ and LwEQ-TD$_{\mathcal{X},\mathcal{H},h_0}(\Sigma_{\text{global}})$ with existing notions of LfS-TD (see (6), Section 3).

Theorem 4. *Fix $\mathcal{X}, \mathcal{H}, h_0$. For learners whose query function is induced by a ranking function independent of Z_{t-1} and h_{t-1}, the corresponding LwEQ-TD is equivalent to the notions of LfS-TD as follows:*

$$\text{LwEQ-TD}_{\mathcal{X},\mathcal{H},h_0}(\Sigma_{\text{const}}) = \text{wc-TD}(\mathcal{H}) \tag{13}$$

$$\text{LwEQ-TD}_{\mathcal{X},\mathcal{H},h_0}(\Sigma_{\text{global}}) = \text{RTD}(\mathcal{H}) \tag{14}$$

When the rankings are independent of Z_{t-1} and h_{t-1}, Eq. (13) and Eq. (14) connects LwEQ-TD to the notions of wc-TD [18] and RTD [22, 24] in batch teaching settings (the optimal teaching sequence is invariant to its permutation) in the LfS paradigm. To show the equalities we note that if $Z' \subseteq \mathcal{Z}$ is a teaching sequence for fixed h^* in the LfS paradigm, then there exists a permutation of Z' which forms a teaching sequence of counterexamples for the learner in the LwEQ paradigm. Furthermore, this theorem allows us to connect LwEQ-TD$_{\mathcal{X},\mathcal{H},h_0}(\Sigma_{\text{global}})$ with the VCD$(\mathcal{H})$ (as RTD$(\mathcal{H}) = \mathcal{O}\left(\text{VCD}^2(\mathcal{H})\right)$, see [29]). A detailed proof of the theorem is in the Appendix of the supplementary.

5.2 LwEQ Learners with Ranking Functions dependent on Z_{t-1} and/or h_{t-1}

These ranking functions induce learners whose next equivalence query at time t (i.e., choice of the hypothesis h_t) is dependent on the history of counterexamples Z_{t-1} and/or dependent on the hypothesis h_{t-1} (see Algorithm 1, Section 2). We formalize these families below using the ranking functions framework of Eq. (10). In particular, we introduce three families of ranking functions: (i) Σ_{gvs} is a family of *global version space* ranking functions where $\sigma \in \Sigma_{\text{gvs}}$ ranks every hypothesis based on a global preference dependent on Z_{t-1} but independent of h_{t-1}, (ii) Σ_{local} is a family of *local* ranking functions where $\sigma \in \Sigma_{\text{local}}$ ranks hypothesis based on a local preference dependent only on h_{t-1}, and (iii) Σ_{lvs} is a family of *local version space* ranking functions where $\sigma \in \Sigma_{\text{lvs}}$ ranks hypothesis based on a local preference dependent on Z_{t-1} and h_{t-1}. These three families are given below:

$$\Sigma_{\text{gvs}} = \left\{\sigma \,\middle|\, \exists\, g : \mathcal{H} \times 2^{\mathcal{H}} \to \mathbb{R},\ \text{s.t.}\, \forall h', Z, h,\ \sigma(h'; \mathcal{H}(Z), h) = g(h', \mathcal{H}(Z))\right\}. \tag{15}$$

$$\Sigma_{\text{local}} = \{\sigma \mid \exists\, g : \mathcal{H} \times \mathcal{H} \to \mathbb{R},\ \text{s.t.}\, \forall h', Z, h,\ \sigma(h'; \mathcal{H}(Z), h) = g(h', h)\}\,. \tag{16}$$

$$\Sigma_{\text{lvs}} = \left\{\sigma \,\middle|\, \exists\, g : \mathcal{H} \times 2^{\mathcal{H}} \times \mathcal{H} \to \mathbb{R},\ \text{s.t.}\, \forall h', Z, h,\ \sigma(h'; \mathcal{H}(Z), h) = g(h', \mathcal{H}(Z), h)\right\}. \tag{17}$$

As discussed in the LfS paradigm, these ranking functions could lead to the learner and the teacher collude to achieve arbitrarily low teaching complexity [29]. To avoid this, we consider a specific collusion-free behavior where the ranking is consistent with its choice of hypothesis. More formally,

Definition 5.1 (**Collusion-free ranking** [29]). *Consider a time t where the learner's current hypothesis is h_{t-1} and the history of inputs seen is Z_{t-1}. Further assume that the learner's preferred hypothesis for time t is uniquely given by $\arg\min_{h' \in \mathcal{H}(Z_{t-1})} \sigma(h'; \mathcal{H}(Z_{t-1}), h_{t-1}) = \{\hat{h}\}$. Let S be additional examples provided by an adversary from time t onwards. We call a ranking function σ collusion-free if for any S consistent with $\hat{h}$, it holds that $\arg\min_{h' \in \mathcal{H}(Z_{t-1} \cup S)} \sigma(h'; \mathcal{H}(Z_{t-1} \cup S), \hat{h}) = \{\hat{h}\}$.*

We assume that the families of rankings studied in Eq. (15)-(17) satisfy Definition 5.1, and consequently the corresponding collusion-free families are denoted as $\Sigma^{\text{CF}}_{\text{gvs}}$, $\Sigma^{\text{CF}}_{\text{local}}$, and $\Sigma^{\text{CF}}_{\text{lvs}}$. For these ranking functions, we study the connection of LwEQ-TD$_{\mathcal{X},\mathcal{H},h_0}(\Sigma^{\text{CF}}_{\text{gvs}})$, LwEQ-TD$_{\mathcal{X},\mathcal{H},h_0}(\Sigma^{\text{CF}}_{\text{local}})$, and LwEQ-TD$_{\mathcal{X},\mathcal{H},h_0}(\Sigma^{\text{CF}}_{\text{lvs}})$ with existing notions of LfS-TD, in particular non-clashing TD [27], local-PBTD [29, 61], and wc-TD [18, 29] (see (6), Section 3). We state these connections in the following theorem with detailed proof in the Appendix of the supplementary.

Theorem 5. *Fix $\mathcal{X}, \mathcal{H}, h_0$. For learners whose query function is induced by a ranking function dependent on Z_{t-1} and/or h_{t-1}, the corresponding LwEQ-TD is connected to the notions of LfS-TD as:*

$$\text{LwEQ-TD}_{\mathcal{X},\mathcal{H},h_0}\left(\Sigma^{\text{CF}}_{\text{gvs}}\right) = \text{NCTD}(\mathcal{H}) \tag{18}$$

$$\text{local-PBTD}(\mathcal{H}) \leq \text{LwEQ-TD}_{\mathcal{X},\mathcal{H},h_0}(\Sigma^{\text{CF}}_{\text{local}}) \leq \text{wc-TD}(\mathcal{H}) \tag{19}$$

$$\text{lvs-PBTD}(\mathcal{H}) \leq \text{LwEQ-TD}_{\mathcal{X},\mathcal{H},h_0}(\Sigma^{\text{CF}}_{\text{lvs}}) \leq \text{wc-TD}(\mathcal{H}) \tag{20}$$

To achieve equality in Eq. (18), we observe that the family of rankings $\Sigma^{\text{CF}}_{\text{gvs}}$ leads to a batched teaching setting in the LfS paradigm. To show this, we establish the permutation invariance of a teaching sequence $Z' \subseteq \mathcal{Z}$ in the LfS paradigm to form a teaching sequence of counterexamples in the LwEQ paradigm using the collusion-freeness of the underlying ranking function of the learner. To show the lower bound in Eq. (19) and Eq. (20), we note that a teaching set of counterexamples $Z \subseteq \mathcal{Z}$ in the LwEQ paradigm forms a teaching set in the LfS paradigm. On the other hand, for the upper bound of wc-TD$(\mathcal{H})$ (see Theorem 4) we note that a *constant* query learner (see Section 2) could pick any consistent hypothesis in a version space to maximize LwEQ-TD$_{\mathcal{X},\mathcal{H},h_0}(\Sigma_{\text{const}})$ (see (3) in Section 3).

6 Discussion and Conclusion

We investigated the query complexity for the learning-with-equivalence-queries (LwEQ) paradigm when the counterexamples are provided by the optimal teacher. We introduced LwEQ-TD, a notion of teaching dimension, to characterize the complexity of teaching (i.e., the number of queries made) for the optimal teacher. We showed the power of best-case counterexamples picked by the optimal teacher, in contrast to worst-case or random counterexamples, for different hypothesis classes, including Axes-aligned hyperplanes, Monotone monomials, and Orthogonal rectangles. We further established new connections of LwEQ-TD with existing notions of TD in the learning-with-samples paradigm, including wc-TD, RTD, NCTD, and local-PBTD.

In the learning-with-queries paradigm, several works have analyzed the query complexity for a combination of different query types [1, 8, 15], such as membership queries, equivalence queries, among others. Building on our characterization of LwEQ-TD, an important research direction of future work is to investigate similar notions of TD when a learner can ask a combination of different queries. Alternatively, one could study LwEQ in the setting of *improper* equivalence queries where the learner could pick the queried hypothesis outside the hypothesis class. [32] considered improper equivalence queries in the LwEQ paradigm, leading to a reduction in query complexity for various hypothesis classes. It would be important to characterize the teaching complexity for this more generic setting of the equivalence queries.

Another important research direction is to further investigate LwEQ-TD for more complex hypothesis classes, including DFA (deterministic finite automaton), NFA (nondeterministic finite-state acceptors), CFG (context-free grammars), among others. As a concrete hypothesis class, consider DFA^2_s (i.e., DFA with alphabet size 2 and state size s). For this hypothesis class, the query complexity is exponential in s when considering worst-case counterexamples [13]. [62] computed VCD for a variety of hypothesis classes including DFA^2_s; this result along with query complexity bounds of the Max-Min algorithm in [14] establishes the bound of $\Theta(s \log s)$ on the expected number of random counterexamples. We conjecture that LwEQ-TD for DFA^2_s is of the order $\mathcal{O}(s)$, and leave this investigation as future work.

References

[1] Dana Angluin. Learning Regular Sets from Queries and Counterexamples. *Information and Computation*, 75(2):87–106, 1987.

[2] Leonard Pitt. Inductive inference, dfas, and computational complexity. In *Proceedings of the International Workshop on Analogical and Inductive Inference*, AII '89, page 18–44, 1989.

[3] J. Balcázar, J. Díaz, and R. Gavaldà. Algorithms for learning finite automata from queries: A unified view. In *Advances in Algorithms, Languages, and Complexity*, 1997.

[4] B. Bollig, P. Habermehl, C. Kern, and M. Leucker. Angluin-style learning of nfa. In *IJCAI*, 2009.

[5] Doron Peled, Moshe Y. Vardi, and Mihalis Yannakakis. *Black Box Checking*, pages 225–240. 1999.

[6] Susmit Jha and Sanjit Seshia. A Theory of Formal Synthesis via Inductive Learning. *Acta Informatica*, 54, 2017.

[7] Frits Vaandrager. Model Learning. *Communications of the ACM*, 60(2):86–95, 2017.

[8] Dana Angluin. Queries Revisited. *Theoretical Computer Science*, 313(2):175 – 194, 2004.

[9] Dana Angluin. Queries and Concept Learning. *Machine Learning*, 2(4):319–342, 1988.

[10] Armando Solar-Lezama, Liviu Tancau, Rastislav Bodik, Sanjit Seshia, and Vijay Saraswat. Combinatorial Sketching for Finite Programs. *SIGARCH*, 34(5):404–415, 2006.

[11] Tommaso Dreossi, Shromona Ghosh, Xiangyu Yue, Kurt Keutzer, Alberto Sangiovanni-Vincentelli, and Sanjit A. Seshia. Counterexample-Guided Data Augmentation. In *IJCAI*, page 2071–2078, 2018.

[12] Oscar Habarra and Tao Jiang. Learning Regular Languages from Counterexamples. *Journal of Computer and System Sciences*, 43(2):299–316, 1991.

[13] Dana Angluin. Negative Results for Equivalence Queries. *Machine Learning*, 5(2):121–150, 1990.

[14] Dana Angluin and Tyler Dohrn. The Power of Random Counterexamples. In *ALT*, pages 452–465, 2017.

[15] Hunter Chase and James Freitag. Bounds in Query Learning. In *COLT*, pages 1142–1160, 2020.

[16] L. G. Valiant. A Theory of the Learnable. *Communications of ACM*, 27(11):1134–1142, 1984.

[17] V.N. Vapnik. An Overview of Statistical Learning Theory. *IEEE Transactions on Neural Networks*, 10(5):988–999, 1999.

[18] Sally A. Goldman and Michael J. Kearns. On the complexity of teaching. *Journal of Computer and System Sciences*, 50(1):20–31, 1995.

[19] Christian Kuhlmann. On Teaching and Learning Intersection-Closed Concept Classes. In *European Conference on Computational Learning Theory*, pages 168–182, 1999.

[20] Frank J Balbach. Measuring Teachability using Variants of the Teaching Dimension. *Theoretical Computer Science*, 397(1-3):94–113, 2008.

[21] Sandra Zilles, Steffen Lange, Robert Holte, and Martin Zinkevich. Teaching Dimensions Based on Cooperative Learning. In *COLT*, pages 135–146, 2008.

[22] Sandra Zilles, Steffen Lange, Robert Holte, and Martin Zinkevich. Models of Cooperative Teaching and Learning. *JMLR*, 12:349–384, 2011.

[23] Xiaojin Zhu. Machine Teaching for Bayesian Learners in the Exponential Family. In *NeurIPS*, pages 1905–1913, 2013.

[24] Thorsten Doliwa, Gaojian Fan, Hans Ulrich Simon, and Sandra Zilles. Recursive Teaching Dimension, VC-Dimension and Sample Compression. *JMLR*, 15(1):3107–3131, 2014.

[25] Ziyuan Gao, Christoph Ries, Hans U Simon, and Sandra Zilles. Preference-Based Teaching. *JMLR*, 18(31):1–32, 2017.

[26] Yuxin Chen, Adish Singla, Oisin Mac Aodha, Pietro Perona, and Yisong Yue. Understanding the Role of Adaptivity in Machine Teaching: The Case of Version Space Learners. In *NeurIPS*, 2018.

[27] David Kirkpatrick, Hans U. Simon, and Sandra Zilles. Optimal Collusion-Free Teaching. In *ALT*, volume 98, pages 506–528, 2019.

[28] Xiaojin Zhu, Adish Singla, Sandra Zilles, and Anna N. Rafferty. An Overview of Machine Teaching. *CoRR*, abs/1801.05927, 2018.

[29] Farnam Mansouri, Yuxin Chen, Ara Vartanian, Xiaojin Zhu, and Adish Singla. Preference-based Batch and Sequential Teaching: Towards a Unified View of Models. In *NeurIPS*, pages 9195–9205, 2019.

[30] Farnam Mansouri, Yuxin Chen, Ara Vartanian, Xiaojin Zhu, and Adish Singla. Preference-Based Batch and Sequential Teaching. *CoRR*, abs/2010.10012, 2020.

[31] Dana Angluin, Lisa Hellerstein, and Marek Karpinski. Learning read-once formulas with queries. Technical report, USA, 1989.

[32] G. Turan and W. Maass. On the Complexity of Learning from Counterexamples. In *FOCS*, pages 262–267, 1989.

[33] Jagdeep Singh Bhatia. Simple and Fast Algorithms for Interactive Machine Learning with Random Counter-examples. *JMLR*, 22(15):1–30, 2021.

[34] Y. Freund, H. Seung, E. Shamir, and Naftali Tishby. Selective Sampling Using the Query by Committee Algorithm. *Machine Learning*, 28:133–168, 2004.

[35] D. Cohn, L. Atlas, and R. Ladner. Improving Generalization with Active Learning. *Machine Learning*, 15:201–221, 1994.

[36] Sanjoy Dasgupta. Coarse Sample Complexity Bounds for Active Learning. In *NeurIPS*, page 235–242, 2005.

[37] Vitaly Feldman. On the Power of Membership Queries in Agnostic Learning. *JMLR*, 10:163–182, 2009.

[38] Ayumi Shinohara and Satoru Miyano. Teachability in Computational Learning. *New Gen. Comput.*, 8(4):337–347, 1991.

[39] VN Vapnik and A Ya Chervonenkis. On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities. *Theory of Probability and its Applications*, 16(2):264, 1971.

[40] Weiyang Liu, Bo Dai, Ahmad Humayun, Charlene Tay, Chen Yu, Linda B. Smith, James M. Rehg, and Le Song. Iterative Machine Teaching. In *ICML*, pages 2149–2158, 2017.

[41] Weiyang Liu, Bo Dai, Xingguo Li, Zhen Liu, James M. Rehg, and Le Song. Towards Black-box Iterative Machine Teaching. In *ICML*, pages 3147–3155, 2018.

[42] Xiaojin Zhu. An Optimal Control View of Adversarial Machine Learning. *CoRR*, abs/1811.04422, 2018.

[43] Laurent Lessard, Xuezhou Zhang, and Xiaojin Zhu. An Optimal Control Approach to Sequential Machine Teaching. In *AISTATS*, pages 2495–2503, 2019.

[44] Adish Singla, Ilija Bogunovic, G Bartók, A Karbasi, and A Krause. On Actively Teaching the Crowd to Classify. In *NIPS Workshop on Data Driven Education*, 2013.

[45] Adish Singla, Ilija Bogunovic, Gábor Bartók, Amin Karbasi, and Andreas Krause. Near-Optimally Teaching the Crowd to Classify. In *ICML*, volume 32, pages 154–162, 2014.

[46] Anette Hunziker, Yuxin Chen, Oisin Mac Aodha, Manuel Gomez Rodriguez, Andreas Krause, Pietro Perona, Yisong Yue, and Adish Singla. Teaching Multiple Concepts to a Forgetful Learner. In *NeurIPS*, pages 4050–4060, 2019.

[47] Oisin Mac Aodha, Shihan Su, Yuxin Chen, Pietro Perona, and Yisong Yue. Teaching Categories to Human Learners with Visual Explanations. In *CVPR*, pages 3820–3828, 2018.

[48] Sanjoy Dasgupta, Daniel Hsu, Stefanos Poulis, and Xiaojin Zhu. Teaching a Black-Box Learner. In *ICML*, pages 1547–1555, 2019.

[49] Rati Devidze, Farnam Mansouri, Luis Haug, Yuxin Chen, and Adish Singla. Understanding the Power and Limitations of Teaching with Imperfect Knowledge. In *IJCAI*, pages 2647–2654.

[50] Ferdinando Cicalese, Sergio Filho, Eduardo Laber, and Marco Molinaro. Teaching with Limited Information on the Learner's Behaviour. In *ICML*, 2020.

[51] Maya Cakmak and Manuel Lopes. Algorithmic and Human Teaching of Sequential Decision Tasks. In *AAAI*, volume 26, 2012.

[52] Thomas J. Walsh and Sergiu Goschin. Dynamic teaching in sequential decision making environments. In *UAI*, page 863–872, 2012.

[53] Parameswaran Kamalaruban, Rati Devidze, Volkan Cevher, and Adish Singla. Interactive Teaching Algorithms for Inverse Reinforcement Learning. In *IJCAI*, pages 2692–2700, 2019.

[54] Daniel S Brown and Scott Niekum. Machine Teaching for Inverse Reinforcement Learning: Algorithms and Applications. In *AAAI*, 2019.

[55] Sebastian Tschiatschek, Ahana Ghosh, Luis Haug, Rati Devidze, and Adish Singla. Learner-aware Teaching: Inverse Reinforcement Learning with Preferences and Constraints. In *NeurIPS*, 2019.

[56] Amin Rakhsha, Goran Radanovic, Rati Devidze, Xiaojin Zhu, and Adish Singla. Policy Teaching via Environment Poisoning: Training-time Adversarial Attacks against Reinforcement Learning. In *ICML*, 2020.

[57] Xuezhou Zhang, Shubham Kumar Bharti, Yuzhe Ma, Adish Singla, and Xiaojin Zhu. The Sample Complexity of Teaching by Reinforcement on Q-Learning. In *AAAI*, 2021.

[58] Xiaojin Zhu. Machine Teaching: An Inverse Problem to Machine Learning and an Approach Toward Optimal Education. In *AAAI*, pages 4083–4087, 2015.

[59] Kathleen Romanik. Approximate Testing and Learnability. page 327–332, 1992.

[60] Kathleen Romantik and Carl Smith. Testing Geometric Objects. *Computational Geometry*, 4(3):157–176, 1994.

[61] Yuxin Chen, Oisin Mac Aodha, Shihan Su, Pietro Perona, and Yisong Yue. Near-Optimal Machine Teaching via Explanatory Teaching Sets. In *AISTATS*, 2018.

[62] Y. Ishigami and Sei'ichi Tani. The VC-Dimensions of Finite Automata with n States. In *ALT*, 1993.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Section 4 and Section 5.
   (b) Did you describe the limitations of your work? [Yes] We discuss how our results can be extended further as future work directions in Section 6.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] All proofs can be found in appendices.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]