Understanding the Power and Limitations of Teaching with Imperfect Knowledge

Rati Devidze[1*] , **Farnam Mansouri**[1*] , **Luis Haug**[2] , **Yuxin Chen**[3] and **Adish Singla**[1]
[1]Max Planck Institute for Software Systems (MPI-SWS)
[2]ETH Zurich
[3]University of Chicago
{rdevidze,mfarnam,adishs}@mpi-sws.org, lhaug@inf.ethz.ch, chenyuxin@uchicago.edu

Abstract

Machine teaching studies the interaction between a teacher and a student/learner where the teacher selects training examples for the learner to learn a specific task. The typical assumption is that the teacher has perfect knowledge of the task—this knowledge comprises knowing the desired learning target, having the exact task representation used by the learner, and knowing the parameters capturing the learning dynamics of the learner. Inspired by real-world applications of machine teaching in education, we consider the setting where teacher's knowledge is limited and noisy, and the key research question we study is the following: When does a teacher succeed or fail in effectively teaching a learner using its imperfect knowledge? We answer this question by showing connections to how imperfect knowledge affects the teacher's solution of the corresponding machine teaching problem when constructing optimal teaching sets. Our results have important implications for designing robust teaching algorithms for real-world applications.

1 Introduction

The field of machine teaching studies the interaction between a teacher and a student/learner where the teacher's objective is to select a short sequence of examples for the learner to learn a specific task [Goldman and Kearns, 1995; Zhu *et al.*, 2018]. An important application is in education where the *learner* is a human student, and the *teacher* is a computerized intelligent tutoring system (ITS) that selects a curriculum of learning material for the student [Zhu, 2015; Rafferty *et al.*, 2016; Sen *et al.*, 2018; Hunziker *et al.*, 2019]. Another concrete application is the data poisoning (training-time) adversarial attacks where the *learner* is a machine learning (ML) system, and the *teacher* is a hacking algorithm that poisons the training data to maliciously change the learner's output to a desired target [Mei and Zhu, 2015; Zhu, 2018]. Regardless of the application and the teacher's intentions, machine teaching provides a formal model of quantifying the teaching effort and an algorithmic framework for deriving an optimized curriculum of material to have maximum influence on the learner with minimal effort. Considering applications in educational settings, the problem of designing an optimized curriculum is of utmost importance because it leads to more effective learning, increased engagement, and reduced drop-out of students [Archambault *et al.*, 2009].

The key issue in applying machine teaching algorithms to real-world applications is that these algorithms (and the corresponding theoretical guarantees) often make unrealistic assumptions about the teacher's knowledge of the learner and the task. It is typically assumed that the teacher has perfect knowledge of the following: (i) the learner, e.g., a computational model of the learning dynamics, and parameters capturing initial knowledge and learning rate, (ii) task specification, e.g., a complete ground truth data and representation of the task as used by the learner. Assuming such a powerful teacher might be meaningful for deriving theoretical guarantees (e.g., computing information-theoretic lower bounds of teaching complexity [Goldman and Kearns, 1995; Zilles *et al.*, 2011; Chen *et al.*, 2018b; Mansouri *et al.*, 2019]) or for understanding the vulnerability of an ML system (e.g., against a white-box poisoning attack [Zhang *et al.*, 2018; Ma *et al.*, 2019]). However, for applications in education where the student is a human learner, this assumption is clearly unrealistic: learning dynamics and task specifications are usually obtained from domain expertise or inferred from historical student data (see [Singla *et al.*, 2014; Piech *et al.*, 2015; Settles and Meeder, 2016; Sen *et al.*, 2018; Hunziker *et al.*, 2019]), and this information is often incomplete and noisy.

1.1 Our Approach and Contributions

Ironically, while the promise of machine teaching algorithms lies in providing a near-optimal teaching curriculum with guaranteed performance, the fundamental assumptions required by these algorithms are clearly violated in practice. The main research question we study in this paper is the following: *When does a teacher succeed or fail in effectively teaching a learner using its imperfect knowledge?*

To answer this question, we require a formal specification of the task, a learner model, and a concrete teaching algorithm. In our work, we study a classification task in the context of teaching human students the rules to identify animal species—an important skill required for biodiversity monitoring related

[*] Authors contributed equally to this work.

citizen-science projects [Sullivan *et al.*, 2009; Van Horn *et al.*, 2018]. This is one of the few real-world applications for which machine teaching algorithms with guarantees have been applied to teaching human learners in educational settings (see [Singla *et al.*, 2013; Singla *et al.*, 2014; Chen *et al.*, 2018a; Mac Aodha *et al.*, 2018]) and hence is well-suited for our work. We highlight some of the key contributions and results below:

- We formally study the problem of robust machine teaching. To quantify the effectiveness of teaching, we introduce two metrics that measure teacher's success in terms of the learner's eventual error, and the size of the teaching set as constructed by a teacher with imperfect knowledge (Section 2).
- We show that teaching is much more brittle w.r.t noise in learning rate and less so when considering noise in prior knowledge of the learner. This theoretical result aligns with a similar observation recently made in the context of a very different learning model (Section 3).
- When studying robustness w.r.t. noise in task specification, we provide natural regularity conditions on the data distributions and then use these conditions when specifying the guarantees. This allows us to take a less pessimistic view in comparison to contemporary works that study the worst-case setting (Section 4).

Remarks on proofs and reproducibility. Detailed proofs of theorems are provided in the longer version of the paper [Devidze *et al.*, 2020]. For the reproducibility of experimental results and facilitating research in this area, the code and dataset are publicly available.

1.2 Related Work on Robust Machine Teaching

A growing body of contemporary works has tackled the problem of robust machine teaching in different forms, however, with a very different focus compared to ours. For instance, [Liu *et al.*, 2018; Melo *et al.*, 2018; Dasgupta *et al.*, 2019; Kamalaruban *et al.*, 2019] have studied the problem of teaching a "blackbox" learner where the teacher has very limited or no knowledge of the learner. The focus of these papers has been on designing an *online* teaching algorithm that infers the learner model in an online fashion. These works often conclude that an *offline* teaching algorithm that operates with limited knowledge can perform arbitrarily bad by considering a worst-case setting. However, designing and deploying online algorithms is a lot more challenging in practice—the results in contemporary works have mostly been theoretical and might not be directly applicable in practice given the high sample complexity of online inference. The focus of our work is primarily on *offline* teaching algorithms, where knowledge about the task is usually obtained from domain expertise or inferred from historical student data. We aim at developing a fundamental understanding of how the performance guarantees of a teaching algorithm degrade w.r.t. the noise in teacher's knowledge when considering natural data distributions.

In another line of contemporary work on teaching a reinforcement learning agent, [Haug *et al.*, 2018; Tschiatschek *et al.*, 2019] have considered the setting where teacher and learner have different worldview and preferences—the focus of these works is on designing a teaching algorithm to account for these mismatches, and do not directly tackle the question we study in this paper. There has also been some empirical work on understanding the robustness and effect of different model components as part of the popular Bayesian Knowledge Tracing (BKT) teaching model used in ITS [Klingler *et al.*, 2015; Khajah *et al.*, 2016]—we see this work as complementary to ours as we take a more formal approach towards understanding the robustness of theoretical guarantees provided by machine teaching algorithms.

2 Problem Formulation

In this section, we first introduce the task representation, the learner's model, and the teacher's optimization problem. Then, we formulate the problem of teaching with imperfect knowledge, and discuss the notions of successful teaching.

2.1 Teaching Task and Representation

We consider the problem of teaching a binary classification task. Let $\mathcal{X}$ denote a ground set of instances (e.g., images), and the learner uses a feature representation of $\phi : \mathcal{X} \to \mathbb{R}^d$. Let $\mathcal{H}$ be a finite class of hypotheses considered by the learner where each hypothesis $h \in \mathcal{H}$ is a function $h : \mathcal{X} \to \{-1, +1\}$. As a concrete setting, $\mathcal{H}$ could be the set of hypotheses of the form $h(x) = \text{sign}(\langle \theta_h, \phi(x) \rangle)$ where $\theta_h \in \mathbb{R}^d$ is the weight vector associated with hypothesis h.

Each instance $x \in \mathcal{X}$ is associated with a ground truth label given by the function $y^* : \mathcal{X} \to \{-1, 1\}$, and we denote the ground truth label of an instance x as $y^*(x)$. The ground truth labels given by y^* are not known to the leaner. We use $\mathcal{Z}$ to denote instances with their labels where a labeled example $z \in \mathcal{Z}$ is given by $z = (x, y^*(x))$.

As typically studied in machine teaching literature, we consider a realizable setting where there exists a hypothesis $h^* \in \mathcal{H}$ such that $\forall x, h^*(x) = y^*(x)$.[1] The teacher's goal can then be stated as that of teaching the hypothesis h^* to the learner by providing a minimal number of labeled examples to the learner. Before formulating the teacher's optimization problem of selecting labeled examples, we state the learning dynamics of the learner below.

2.2 Learner Model

We consider a probabilistic learner model that generalizes the well-studied *version space* models in classical machine teaching literature (see [Goldman and Kearns, 1995]). At a high-level, the learner works as follows: During the learning process, the learner maintains a score for each hypothesis given by $Q(h)$ capturing learner's belief of how good the hypothesis h is. Given $Q(h)$, the learner acts probabilistically by drawing a hypothesis with probability $\frac{Q(h)}{\sum_{h' \in \mathcal{H}} Q(h')}$. Next, we discuss how the scores are updated.

[1] This assumption is w.l.o.g.: In a non-realizable setting, the teacher could consider $h^* \in \mathcal{H}$ as a hypothesis with minimal error in terms of disagreement of labels w.r.t. the labels given by y^* and the results presented in this paper can be extended to this general setting.

Before teaching starts, the learner's prior knowledge about the task is captured by initial scores given by $Q_0(h)$. For simplicity and as considered in [Singla *et al.*, 2014; Chen *et al.*, 2018a], we will assume that $Q_0(h)$ is a probability distribution over $\mathcal{H}$. After receiving a set of labeled examples $S = \{(x_s, y_s)\}_{s=1,2,\dots}$ from the teacher, we denote the learner's score as $Q(h|S)$ which are updated as follows:

$$Q(h|S) = Q_0(h) \cdot \Pi_{s=1,2,\dots,|S|} J(y_s|h, x_s, \eta) \tag{1}$$

where J is a likelihood function parameterized by $\eta \in (0, 1]$. In this paper, we consider the following likelihood function given by:

$$J(y_s|h, x_s, \eta) = \begin{cases} 1 - \eta & \text{for } h(x_s) \neq y_s \\ 1 & \text{o.w.} \end{cases} \tag{2}$$

Here, the quantity η captures a notion of the learning rate. This model reduces to a randomized variant of the classical learner model [Goldman and Kearns, 1995]) for $\eta = 1$. The main results and findings in the paper also generalize to more complex likelihood functions such as the logistic functions considered by [Singla *et al.*, 2014; Mac Aodha *et al.*, 2018].

An important quantity of interest in this paper is the learner's expected error after receiving a set of examples S. Let $\mathrm{err}(h) = \frac{\sum_{z=(x,y)\in\mathcal{Z}} 1_{h(x)\neq y}}{|\mathcal{Z}|}$ be the expected error of h. The learner's expected error is given by the following:

$$\textsc{Err}(S) = \sum_{h\in\mathcal{H}} \frac{Q(h|S)}{\sum_{h'\in\mathcal{H}} Q(h'|S)} \mathrm{err}(h) \tag{3}$$

2.3 Teaching with Perfect Knowledge

We first consider the optimal teaching problem when the teacher has perfect knowledge of the teaching task represented as $(Q_0, \eta, \mathcal{Z}, h^*, \phi, \mathcal{H})$. In particular, the teacher's knowledge comprises: (i) learning dynamics captured by learner's initial knowledge Q_0 and learning rate η, (ii) task specification captured by the target hypothesis h^*, the ground set of labeled examples $\mathcal{Z}$, the feature map ϕ, and hypothesis class $\mathcal{H}$.

Teacher's primary goal is to find a smallest set of labeled examples to teach so that learner's error is below a certain desirable threshold ϵ. To construct the optimal teaching set, instead of directly optimizing for a reduction in error, it is common in the literature to construct surrogate objective functions which capture learner's progress towards learning h^* (also, see [Goldman and Kearns, 1995; Singla *et al.*, 2014; Chen *et al.*, 2018a; Mac Aodha *et al.*, 2018].).

Let us define a set function $F : 2^{\mathcal{Z}} \to \mathbb{R}_{\geq 0}$ as follows:

$$F(S) = \sum_{h\in\mathcal{H}} \Big(Q_0(h) - Q(h|S)\Big) \cdot \mathrm{err}(h) \tag{4}$$

Here, the quantity $\big(Q_0(h) - Q(h|S)\big)$ captures the reduction in the score for hypothesis h after learner receives examples set S. In particular, the surrogate objective function F is a soft variant of set cover, and allows one to design greedy algorithms to find near-optimal teaching sets.

For a given ϵ, one can find a corresponding (sufficient) stopping value C_ϵ such that $F(S) \geq C_\epsilon$ implies that $\textsc{Err}(S) \leq \epsilon$. As used in the optimization frameworks of [Singla *et al.*, 2014; Mac Aodha *et al.*, 2018], we use $C_\epsilon = \sum_{h\in\mathcal{H}} Q_0(h) \cdot \mathrm{err}(h) - \epsilon \cdot Q_0(h^*)$. This leads to the following optimization problem:

$$\min_{S\subseteq\mathcal{Z}} |S| \text{ s.t. } F(S) \geq \sum_{h\in\mathcal{H}} Q_0(h) \cdot \mathrm{err}(h) - \epsilon \cdot Q_0(h^*) \tag{5}$$

where $F(S)$ is given in Eq. 4. We use $\textsc{Opt}_\epsilon$ to denote the optimal teaching set as a solution to the problem (5).

2.4 Teaching with Imperfect Knowledge

We now consider a teacher with imperfect knowledge and study the following different settings:

- having noise on learner's initial knowledge Q_0 (Section 3.1)
- having noise on learner's learning rate η (Section 3.2).
- having access to ground truth labels for only a subset of instances instead of the whole ground set $\mathcal{X}$ (Section 4.1).
- having a noisy feature map, i.e., teacher's assumed feature map does not match with ϕ used by the learner (Section 4.2).

We denote the teacher's view of the imperfect knowledge as $(\widetilde{Q}_0, \widetilde{\eta}, \widetilde{\mathcal{Z}}, \widetilde{h^*}, \widetilde{\phi}, \widetilde{\mathcal{H}})$. Given this knowledge, the teacher has its own view of quantities such as $\widetilde{Q}(h|S)$ (cf., Eq. 1), $\widetilde{\mathrm{err}}(h)$ (cf., err(.) used in Eq. 3), and $\widetilde{F}$ (cf., Eq. 4) as counterparts to those of a teacher with perfect knowledge. The optimization problem from the viewpoint of the teacher with imperfect knowledge can be written as follows:

$$\min_{S\subseteq\widetilde{\mathcal{Z}}} |S| \text{ s.t. } \widetilde{F}(S) \geq \sum_{h\in\widetilde{\mathcal{H}}} \widetilde{Q}_0(h) \cdot \widetilde{\mathrm{err}}(h) - \epsilon \cdot \widetilde{Q_0}(\widetilde{h^*}) \tag{6}$$

In the subsequent sections, we will introduce notions of Δ-imperfect knowledge depending on a set/tuple of parameters Δ. Let us denote by $\widetilde{\textsc{Opt}}_{\epsilon,\Delta}$ the teaching set found by Δ-imperfect teacher as a solution to the problem (6). The following definitions quantify the success of a teacher with imperfect knowledge w.r.t. to measure $M.1$ (related to learner's error) and measure $M.2$ (related to teaching set size).

Definition 1 ($M.1$-successful)**.** *We say a teacher is $M.1$-successful if the learner's eventual error upon receiving the set $\widetilde{\textsc{Opt}}_{\epsilon,\Delta}$ is $O(\epsilon)$ (here we treat the parameters as constant).*

Definition 2 ($M.2$-successful)**.** *We say a teacher is $M.2$-successful if $|\widetilde{\textsc{Opt}}_{\epsilon,\Delta}| \leq |\textsc{Opt}_{\hat{\epsilon}}|$, where $\hat{\epsilon} = \Theta(\epsilon)$ (here we treat the parameters as constant). In other words, the size of the teacher's teaching set is competitive w.r.t. that of a teacher with perfect knowledge which constructs an optimal teaching set for an $\Theta(\epsilon)$ error threshold.*[2]

3 Imperfect Knowledge about the Dynamics

In this section, we explore the effectiveness of teaching when the teacher has imperfect knowledge of the learner's prior knowledge Q_0 and learning rate η. In fact, these

[2] This is the style of bound often considered in literature when taking an optimization perspective on teaching [Singla *et al.*, 2014; Chen *et al.*, 2018a]. One might be tempted to directly bound the size $|\widetilde{\textsc{Opt}}_{\epsilon,\Delta}|$ as a function of $|\textsc{Opt}_\epsilon|$, however, this is usually not possible without making further assumptions about the data distribution.

two parameters are key to many popular learner models (e.g., Bayesian Knowledge Tracing (BKT) models in educational applications [Piech *et al.*, 2015; Klingler *et al.*, 2015; Khajah *et al.*, 2016], spaced-repetition models used in vocabulary applications [Settles and Meeder, 2016; Hunziker *et al.*, 2019], or gradient learner models studied for data-poisoning attacks [Liu *et al.*, 2018]).

3.1 Noise in the Learning Prior

Here, we consider the setting where the teacher has a noisy estimate $\widetilde{Q_0}$ of learner's initial distribution Q_0, i.e., $(\widetilde{Q_0}, \widetilde{\eta}, \widetilde{\mathcal{Z}}, \widetilde{h^*}, \widetilde{\phi}, \widetilde{\mathcal{H}}) := (\widetilde{Q_0}, \eta, \mathcal{Z}, h^*, \phi, \mathcal{H})$. The following definition quantifies the noise in $\widetilde{Q_0}$ w.r.t. the true Q_0.

Definition 3 (Δ_{Q_0}-imperfect)**.** *Let* $\Delta_{Q_0} = (\delta_1, \delta_2)$ *for* $\delta_1, \delta_2 \geq 0$. *We say that teacher's estimated distribution* $\widetilde{Q_0}$ *is* Δ_{Q_0}*-imperfect if the following holds:*

$$\forall\, h \in \mathcal{H}, (1 - \delta_1) \cdot Q_0(h) \leq \widetilde{Q_0}(h) \leq Q_0(h) \cdot (1 + \delta_2).$$

The following theorem quantifies the effectiveness of teaching w.r.t. measures $M.1$ and $M.2$ (see Definitions 1, 2).

Theorem 1. *Fix* $\epsilon \geq 0$, $\delta_1 \geq 0$, *and* $\delta_2 \geq 0$. *Consider a teacher with knowledge* $(\widetilde{Q_0}, \eta, \mathcal{Z}, h^*, \phi, \mathcal{H})$, *where* $\widetilde{Q_0}$ *is* Δ_{Q_0}*-imperfect w.r.t. true* Q_0 *for* $\Delta_{Q_0} = (\delta_1, \delta_2)$. *Then, in the worst-case for any problem setting and any* Δ_{Q_0}*-imperfect* $\widetilde{Q_0}$, *the teacher is successful w.r.t. measures* $M.1$ *and* $M.2$ *with the following bounds:*

1. *The learner's error is* $O(\epsilon)$ *and is bounded as* $\textsc{Err}(\widetilde{\textsc{Opt}}_{\epsilon, \Delta_{Q_0}}) \leq \frac{\epsilon \cdot (1+\delta_2)}{(1-\delta_1)}$.
2. *The size of the teaching set is bounded as* $|\widetilde{\textsc{Opt}}_{\epsilon, \Delta_{Q_0}}| \leq |\textsc{Opt}_{\hat{\epsilon}}|$ where $\hat{\epsilon} = \frac{\epsilon \cdot (1-\delta_1)}{(1+\delta_2)}$.

The proof is provided in the longer version of the paper [Devidze *et al.*, 2020].

3.2 Noise in the Learning Rate

Next, we consider a setting where the teacher has an imperfect estimate of the learner's learning rate η while having perfect knowledge about the rest of the parameters, i.e., the teacher's knowledge is $(Q_0, \widetilde{\eta}, \mathcal{Z}, h^*, \phi, \mathcal{H})$. The following definition quantifies the noise in $\widetilde{\eta}$ w.r.t. true η.

Definition 4 (Δ_η-imperfect)**.** *Let* $\Delta_\eta = (\delta)$ *for* $\delta \geq 0$. *We say that a teacher's estimate* $\widetilde{\eta}$ *is* Δ_η*-imperfect if* $|\widetilde{\eta} - \eta| \leq \delta$, *where both* $\widetilde{\eta} \in (0, 1]$ *and* $\eta \in (0, 1]$.

The following two worst-case scenarios are of interest: (i) a teacher who overestimates the learning rate with $\widetilde{\eta} = \min\{\eta + \delta, 1\}$ and (ii) a teacher who underestimates the learning rate with $\widetilde{\eta} = \max\{\eta - \delta, 0\}$. The following theorem quantifies the challenges in teaching successfully in this setting.

Theorem 2. *Fix* $\epsilon \geq 0$ *and* $\delta > 0$. *Consider a teacher with knowledge* $(Q_0, \widetilde{\eta}, \mathcal{Z}, h^*, \phi, \mathcal{H})$ *where* $\widetilde{\eta}$ *is* Δ_η*-imperfect w.r.t. true* η *for* $\Delta_\eta = (\delta)$. *Then, for any* Δ_η*-imperfect* $\widetilde{\eta}$, *there exists a problem setting such that the teacher is unsuccessful w.r.t. measures* $M.1$ *and* $M.2$*:*

1. *For any fixed* ϵ *and* Δ_η, *there exist problem settings where* $\textsc{Err}(\widetilde{\textsc{Opt}}_{\epsilon, \Delta_\eta}) \geq \frac{1}{2}$.
2. *For any fixed* ϵ *and* Δ_η, *and any* $\hat{\epsilon}$ *arbitrarily close to* 0, *there exist problem settings where* $|\widetilde{\textsc{Opt}}_{\epsilon, \Delta_\eta}| \geq |\textsc{Opt}_{\hat{\epsilon}}|$.

The proof, provided in the longer version [Devidze *et al.*, 2020], is given by creating two problem settings: (i) a setting for the teacher who overestimates η that leads to the first statement about the learner's error, and (ii) a setting for the teacher who underestimates η that leads to the second statement about the size of the teaching set. Comparing Theorem 1 and Theorem 2, these results suggest that noise in the teacher's assumption about the learning rate is a lot more hazardous compared to noise about the learner's initial distribution. While we derived these results by focusing on a very specific task and learner model, similar observations were made in the context of a different type of teaching setting when teaching a gradient learner [Yeo *et al.*, 2019].

Theorem 2 only provides a pessimistic view that the teacher can fail badly. On closer inspection, the negative results arise from two separate issues: (i) teacher computing wrong utility of examples in (6), and (ii) teacher having a wrong estimate of stopping criteria in (6) which in turn depends on learner's progress. Empirically, we found that the second reason seems to be the dominant one for the teacher's failure. One practical way to fix this issue is to develop an interactive teaching strategy where the teacher's stopping criteria is determined by the learner's true progress measured in an online fashion instead of the progress as estimated by the teacher using its offline model (also, see discussions in Section 1.2).

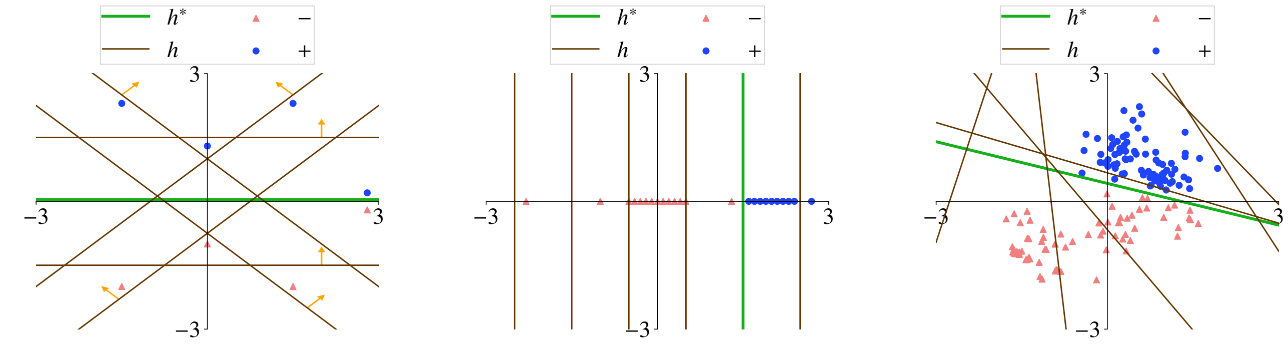
4 Imperfect Knowledge about Representation

In this section, we explore the effect of teaching when the teacher has imperfect knowledge of the task specification, in particular, limited ground truth data and noisy representation of the task used by the learner.

4.1 Limited Ground Truth Labels

Here, we consider the setting where the teacher has ground truth labels for only a subset of examples $\widetilde{\mathcal{Z}} \subseteq \mathcal{Z}$. The typical process followed when applying machine teaching algorithms is to first sample a small set of instances $\widetilde{\mathcal{X}} \subseteq \mathcal{X}$ and then get expert annotations to obtain $\widetilde{\mathcal{Z}} \subseteq \mathcal{Z}$ (e.g., see [Singla *et al.*, 2014; Mac Aodha *et al.*, 2018]. Then, the teacher selects a hypothesis $\widetilde{h^*}$ as the one with minimal empirical error given by $\widetilde{h^*} \in \arg\min_{h \in \mathcal{H}} \widetilde{\mathrm{err}}(h)$. For this setting, we represent the knowledge of the teacher as $(Q_0, \eta, \widetilde{\mathcal{Z}}, \widetilde{h^*}, \phi, \mathcal{H})$.

As long as the set $\widetilde{\mathcal{Z}}$ is constructed *i.i.d.*, the teacher can construct teaching sets to ensure that the learner's error would be low (i.e., teaching is successful w.r.t. measure $M.1$). This argument follows from the standard concentration inequalities which ensures that with high probability, the teacher has a good estimate of $\widetilde{\mathrm{err}}(\cdot)$, i.e., $\forall\, h \in \mathcal{H}, |\widetilde{\mathrm{err}}(h) - \mathrm{err}(h)|$ is small (see Theorem 3). However, regarding the teacher's performance on measure $M.2$, without any additional assumptions about data distribution, it is easy to construct a pessimistic scenario where the data distribution is skewed and the teaching set



(a) Setting with a few extreme data points

(b) Setting with skewed data distribution

(c) Real-world problem setting

Figure 1: **(a)** shows a problem setting with a few extreme points that are important for teaching—assuming $\eta = 1$ and $\epsilon = 0$ for simplicity of arguments, the optimal teaching set consists of only two examples lying close to coordinates $(3, 0)$. However, if these two examples are not present in $\widetilde{\mathcal{Z}}$, the teaching set can be arbitrarily large (i.e., of size 6 in the illustration). **(b)** shows another problem setting with skewed data distribution where a small perturbation of data could lead to big changes in the prediction of hypotheses. **(c)** shows a real-world problem setting to distinguish animal species. As can be seen, the data distribution here is more "well-behaved" and does not suffer from issues present in the other two problem settings. Note that only a few hypotheses are shown in the illustration, see Section 5 for more details.

$\widetilde{\text{OPT}}_{\epsilon,\Delta}$ constructed by a teacher with imperfect knowledge is arbitrarily large w.r.t. the optimal teaching set OPT_ϵ (see Figure 1).

Building on insights from the problem settings discussed in Figure 1, we consider additional structural assumptions on the problem setting as discussed below. First, we introduce the notion of δ-perturbed set of examples.

Definition 5 (δ-perturbed)**.** *Consider a set of labeled examples $S \subseteq \mathcal{Z}$. We call S' a δ-perturbed version of S, if there exists a bijective map $S \mapsto S', (x, y) \mapsto (x', y)$ such that $\|\phi(x) - \phi(x')\|_2 \le \delta$.*

We will also need the following smoothness notion for proving robustness guarantees (for bounding the size in Theorem 3 and for bounding both the error/size in Theorem 4).

Definition 6 (λ-smoothness)**.** *Let $\delta \ge 0$, $\lambda \ge 0$. Consider any set $S \subseteq \mathcal{Z}$, and let S' be any δ-perturbed version of S. Then, we call the problem setting λ-smooth when the following holds: for any $h \in \mathcal{H}$, the mismatch in labels assigned by h to examples S and S' is upper-bounded by $\lambda \cdot \delta$.*

Definition 7 below quantifies the imperfection in teacher's knowledge arising from the sampling process coupled with additional structural conditions.

Definition 7 ($\Delta_{\mathcal{Z}}$-imperfect)**.** *Let $\Delta_{\mathcal{Z}} = (\delta_1, \delta_2, \delta_3)$ for $\delta_1, \delta_2, \delta_3 \ge 0$. We say that a teacher's knowledge is $\Delta_{\mathcal{Z}}$-imperfect if the following statements hold with probability at least $(1 - \delta_1)$:*

- $\forall\, h, |\widetilde{\text{err}}(h) - \text{err}(h)| \le \delta_2,$
- *for any set of labeled examples $S \subseteq \mathcal{Z}$ with $|S| \le |\widetilde{\mathcal{Z}}|$, there exists a δ_3-perturbed version of S in $\widetilde{\mathcal{Z}}$.*

Note that in the above definition, the bound on error is satisfied from the *i.i.d.* sampling process and doesn't require any further structural assumption. The second condition implicitly adds regularity conditions on the underlying data distribution ensuring that it does not have characteristics as seen in Figure 1a and Figure 1b. The following theorem quantifies the effectiveness of a $\Delta_{\mathcal{Z}}$-imperfect teacher.

Theorem 3. *Fix $\epsilon \ge 0$ and $\Delta_{\mathcal{Z}} = (\delta_1, \delta_2, \delta_3)$ with $\delta_1, \delta_2, \delta_3 \ge 0$. Consider a $\Delta_{\mathcal{Z}}$-imperfect teacher with knowledge $(Q_0, \eta, \widetilde{\mathcal{Z}}, \widetilde{h^*}, \phi, \mathcal{H})$. Assume the problem setting is λ-smooth for some $\lambda \ge 0$, $\eta < 1$, and $|\widetilde{\mathcal{Z}}|$ is sufficiently large. Then, for any sample $\widetilde{\mathcal{Z}}$ and selection of $\widetilde{h^*}$, with probability at least $(1 - \delta_1)$, the teacher is successful with the following bounds:*

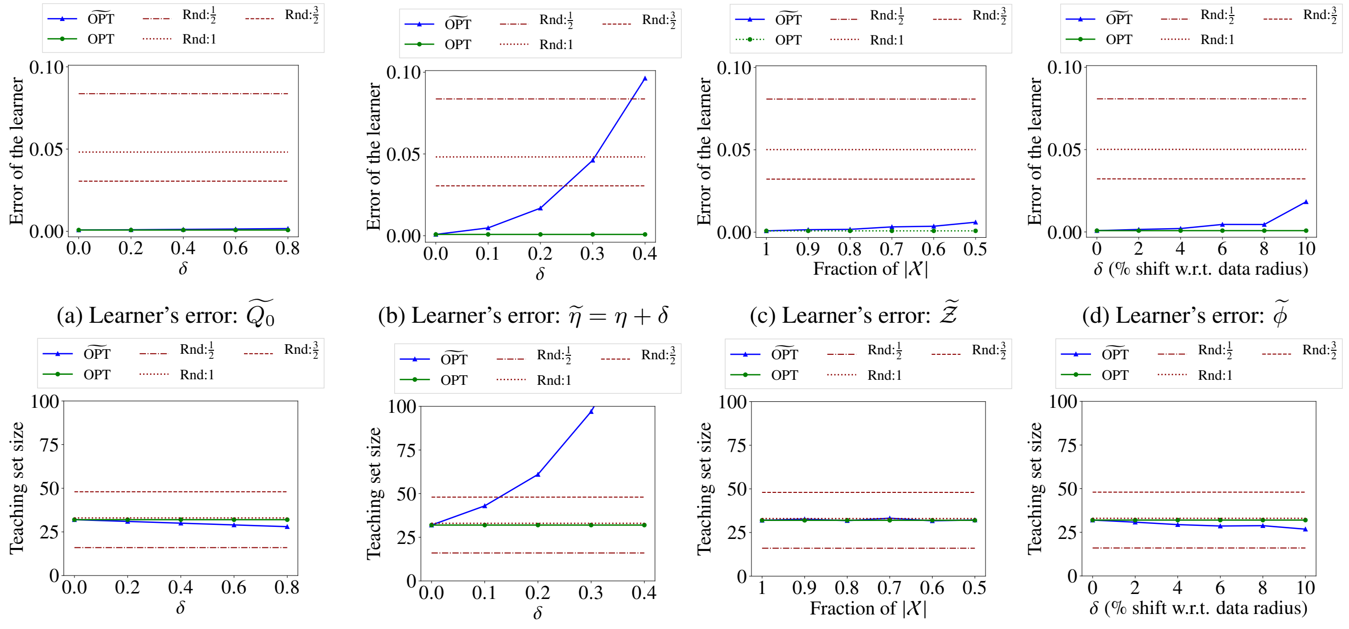
1. *The learner's error is $O(\epsilon)$ and is bounded as $\text{ERR}(\widetilde{\text{OPT}}_{\epsilon,\Delta_{\mathcal{Z}}}) \le \frac{(\epsilon \cdot Q_{\max} + \delta_2)}{Q(h^*)}$.*
2. *The size of the teaching set is bounded as $|\widetilde{\text{OPT}}_{\epsilon,\Delta_{\mathcal{Z}}}| \le |\text{OPT}_{\hat{\epsilon}}|$ where $\hat{\epsilon} = \frac{(\epsilon \cdot Q_{\min} - \delta_2) \cdot (1-\eta)^{\lambda \cdot \delta_3}}{Q(h^*)}$*

where $Q_{\max} = \max_h Q_0(h)$ and $Q_{\min} = \min_h Q_0(h)$.

The proof of the theorem is provided in the longer version of the paper [Devidze *et al.*, 2020]. Note that the bound is only valid for $\eta < 1$. When η approaches 1, and for the extreme case of $\eta = 1$, the learner reduces to a noise-free version space learner who eliminates all inconsistent hypothesis immediately. For this setting, bounding the teaching set size requires more combinatorial assumptions on the dataset (e.g., based on the separability of data from the hyperplanes)—however, for practical applications, η bounded away from 1 is a more natural setting as analyzed in this theorem.

4.2 Noise in Feature Embedding

Here, we consider imperfect knowledge in terms of noisy feature map $\widetilde{\phi}$. This is a challenging setting as noise in ϕ means error in the predictions of hypotheses $h \in \mathcal{H}$ which in turn leads to noise in error of hypotheses err(.) and in the likelihood function J. As noted earlier, the teacher will select a hypothesis $\widetilde{h^*}$ as the one with minimal error given by $\widetilde{h^*} \in$



(a) Learner's error: $\widetilde{Q_0}$

(b) Learner's error: $\widetilde{\eta} = \eta + \delta$

(c) Learner's error: $\widetilde{\mathcal{Z}}$

(d) Learner's error: $\widetilde{\phi}$

(e) Teaching set size: $\widetilde{Q_0}$

(f) Teaching set size: $\widetilde{\eta} = \eta - \delta$

(g) Teaching set size: $\widetilde{\mathcal{Z}}$

(h) Teaching set size: $\widetilde{\phi}$

Figure 2: Experimental results for the problem in Figure 1c. **(a, e)**: robustness of teaching for Δ_{Q_0}-imperfect teacher. **(b, f)**: for Δ_η-imperfect teacher, the learner's error could be high when the teacher overestimates η or the teaching set size could be arbitrary large when the teacher underestimates η. **(c, g)**: robustness of teaching for $\Delta_{\mathcal{Z}}$-imperfect teacher. **(d, h)**: robustness of teaching for Δ_ϕ-imperfect teacher.

$\arg\min_{h \in \mathcal{H}} \widetilde{\text{err}}(h)$. The following definition quantifies the imperfection in the teacher's knowledge $(Q_0, \eta, \mathcal{Z}, \widetilde{h^*}, \widetilde{\phi}, \mathcal{H})$.

Definition 8 (Δ_ϕ-imperfect)**.** *Let $\Delta_\phi = (\delta_1, \delta_2)$ for $\delta_1, \delta_2 \geq 0$. We say that a teacher's knowledge is Δ_ϕ-imperfect if the following holds:*

- $\forall\, x \in \mathcal{X}, ||\phi(x) - \widetilde{\phi}(x)||_2 \leq \delta_1,$
- $\forall\, h, |\widetilde{\text{err}}(h) - \text{err}(h)| \leq \delta_2.$

The following theorem quantifies the effectiveness of teaching of a Δ_ϕ-imperfect teacher.

Theorem 4. *Fix $\epsilon \geq 0$ and $\Delta_\phi = (\delta_1, \delta_2)$ with $\delta_1, \delta_2 \geq 0$. Consider a Δ_ϕ-imperfect teacher with knowledge $(Q_0, \eta, \mathcal{Z}, \widetilde{h^*}, \widetilde{\phi}, \mathcal{H})$. Assume the problem setting is λ-smooth for some $\lambda \geq 0$, that $\eta < 1$, and assume that the error $\widetilde{err}(\widetilde{h^*}) = 0$. Then, in the worst-case for any observed $\widetilde{\phi}$ and selection of $\widetilde{h^*}$, the teacher is successful with the following bounds:*

1. *The learner's error is $O(\epsilon)$ and is bounded as $\text{E}\small{\text{RR}}(\widetilde{\text{O}\small{\text{PT}}}_{\epsilon,\Delta_\phi}) \leq \frac{(\epsilon \cdot Q_{\max} + \delta_2)}{Q(h^*) \cdot (1-\eta)^{\lambda \cdot \delta_1}}$.*
2. *The size of the teaching set is bounded as $|\widetilde{\text{O}\small{\text{PT}}}_{\epsilon,\Delta_\phi}| \leq |\text{O}\small{\text{PT}}_{\hat{\epsilon}}|$ where $\hat{\epsilon} = \frac{(\epsilon \cdot Q_{\min} - \delta_2) \cdot (1-\eta)^{\lambda \cdot \delta_1}}{Q(h^*)}$.*

The proof is provided in the longer version of the paper [Devidze *et al.*, 2020]. In comparison to the error bound in Theorem 3, the error bound here with noise in ϕ is much worse—this is a lot more challenging setting given that hypotheses predictions on examples can be wrong in this setting. Here, for simplicity of the proof and presentation of results, we assumed that there exists some $\widetilde{h^*}$ for which error in teacher's representation is 0, i.e., $\widetilde{\text{err}}(\widetilde{h^*}) = 0$, see discussion in Footnote 1. The theorem suggests that when considering additional structural/smoothness assumptions on the problem, the teaching with imperfect knowledge about representations is robust w.r.t. both $M.1$ and $M.2$ success criteria. As we shall see in experiments, these robustness guarantees indeed hold in practice given that the real-world problem settings often respect these regularity assumptions.

5 Experimental Evaluation

In this section, we perform empirical studies to validate the guarantees provided by our theorems, and to showcase that the data regularity assumptions we made in the previous section are satisfied in real-world problem settings.

Teaching task. We consider a binary image classification task for identifying animal species. This specific task has been studied extensively in the machine teaching literature (see [Singla *et al.*, 2014; Chen *et al.*, 2018a; Mac Aodha *et al.*, 2018; Yeo *et al.*, 2019]). First, we state the problem setup from the viewpoint of a teacher with full knowledge represented as $(Q_0, \eta, \mathcal{Z}, h^*, \phi, \mathcal{H})$. Our problem setup is based on the task and dataset that is used in the works of [Singla *et al.*, 2014; Yeo *et al.*, 2019]. The task is to distinguish "moths" (− labeled class) from "butterflies" (+ labeled class). We have a total of $|\mathcal{Z}| = 160$ labeled images, and the embedding of instances is shown in Figure 1c. We have $|\mathcal{H}| = 67$ hypotheses, and a subset of these hypotheses along with h^* are shown in Figure 1c.

We consider Q_0 to be uniform distribution over $\mathcal{H}$, $\eta = 0.5$, and have desired $\epsilon = 0.001$.

Metrics and baselines. All the results corresponding to four different notions of imperfect teacher are shown in Figure 2, averaged over 10 runs. For performance metrics, we plot the eventual error of the learner and the size of the teaching set. In addition to $\textsc{Opt}_\epsilon$ (simply denoted as $\textsc{Opt}$ in plots) and $\widetilde{\textsc{Opt}}_{\epsilon,\Delta}$ (simply denoted as $\widetilde{\textsc{Opt}}$ in plots), we also have three more baselines denoted as Rnd:$\frac{1}{2}$, Rnd:1, and Rnd:$\frac{3}{2}$. These three baselines correspond to teachers who select examples randomly, with set sizes being $\frac{1}{2}$, 1, and $\frac{3}{2}$ times that of $|\textsc{Opt}|$.

Empirical results. We consider Δ_{Q_0}-imperfect teacher with $\Delta_{Q_0} = (\delta, \delta)$ (i.e., $\delta_1 = \delta_2 = \delta$) with $\delta \in [0, 0.8]$; results are shown in Figures 2a,2e. For Δ_η-imperfect teacher, we vary $\delta \in [0, 0.4]$ considering a teacher who overestimates or underestimates the learning rate; results are shown in Figures 2b,2f. For $\Delta_{\mathcal{Z}}$-imperfect teacher, we vary the fraction of instances $\mathcal{X}$ from 1 to 0.5 that we sample to construct $\widetilde{\mathcal{Z}}$, and sampling is done *i.i.d.*; the performance of this teacher is shown in Figures 2c,2g. For Δ_ϕ-imperfect teacher, we computed noisy representation $\widetilde{\phi}$ by adding a random vector in $\mathbb{R}^2$ of norm δ as noise to $\phi(x)\ \forall x \in \mathcal{X}$; results are shown in Figures 2d,2h. Note that in Figures 2d,2h, the norm δ is shown as a relative % shift w.r.t. data radius, where the radius is $\max_{x \in \mathcal{X}} ||\phi(x)||_2$ (see Figure 1c).

The results in these plots validate the performance guarantees that we proved in previous sections. It is important to note that for $\Delta_{\mathcal{Z}}$-imperfect and Δ_ϕ-imperfect teacher, any additional structural assumptions as were needed by Definitions 7,8 and Theorems 3,4 are naturally satisfied in real-world problem settings, as is evident in the performance plots.

6 Conclusions

We studied the problem of machine teaching when the teacher's knowledge is imperfect. We focused on understanding the robustness of a teacher who constructs teaching sets based on its imperfect knowledge. When having imperfect knowledge about the learner model, our results suggest that having a good estimate of the learning rate is a lot more important than the learner's prior knowledge. In terms of imperfect knowledge about the task specification, we introduced some regularity assumptions under which the teacher is robust. Our empirical experiments on a real-world teaching problem further validate our theoretical results. Our findings have important implications in designing teaching algorithms for real-world applications in education.

References

[Archambault *et al.*, 2009] Isabelle Archambault, Michel Janosz, Jean-Sébastien Fallu, and Linda S Pagani. Student engagement and its relationship with early high school dropout. *Journal of adolescence*, 32(3):651–670, 2009.

[Chen *et al.*, 2018a] Yuxin Chen, Oisin Mac Aodha, Shihan Su, Pietro Perona, and Yisong Yue. Near-optimal machine teaching via explanatory teaching sets. In *AISTATS*, 2018.

[Chen *et al.*, 2018b] 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*, pages 1483–1493, 2018.

[Dasgupta *et al.*, 2019] Sanjoy Dasgupta, Daniel Hsu, Stefanos Poulis, and Xiaojin Zhu. Teaching a black-box learner. In *ICML*, pages 1547–1555, 2019.

[Devidze *et al.*, 2020] Rati Devidze, Farnam Mansouri, Luis Haug, Yuxin Chen, and Adish Singla. Understanding the power and limitations of teaching with imperfect knowledge. *CoRR*, abs/2003.09712, 2020.

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

[Haug *et al.*, 2018] Luis Haug, Sebastian Tschiatschek, and Adish Singla. Teaching inverse reinforcement learners via features and demonstrations. In *NeurIPS*, 2018.

[Hunziker *et al.*, 2019] 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*, 2019.

[Kamalaruban *et al.*, 2019] Parameswaran Kamalaruban, Rati Devidze, Volkan Cevher, and Adish Singla. Interactive teaching algorithms for inverse reinforcement learning. In *IJCAI*, pages 2692–2700, 2019.

[Khajah *et al.*, 2016] Mohammad Khajah, Robert V. Lindsey, and Michael Mozer. How deep is knowledge tracing? In *EDM*, 2016.

[Klingler *et al.*, 2015] Severin Klingler, Tanja Käser, Barbara Solenthaler, and Markus Gross. On the performance characteristics of latent-factor and knowledge tracing models. *International Educational Data Mining Society*, 2015.

[Liu *et al.*, 2018] Weiyang Liu, Bo Dai, Xingguo Li, Zhen Liu, James M. Rehg, and Le Song. Towards black-box iterative machine teaching. In *ICML*, 2018.

[Ma *et al.*, 2019] Yuzhe Ma, Xiaojin Zhu, and Justin Hsu. Data poisoning against differentially-private learners: Attacks and defenses. In *IJCAI*, pages 4732–4738, 2019.

[Mac Aodha *et al.*, 2018] 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.

[Mansouri *et al.*, 2019] Farnam Mansouri, Yuxin Chen, Ara Vartanian, Jerry Zhu, and Adish Singla. Preference-based batch and sequential teaching: Towards a unified view of models. In *NeurIPS*, pages 9195–9205, 2019.

[Mei and Zhu, 2015] Shike Mei and Xiaojin Zhu. Using machine teaching to identify optimal training-set attacks on machine learners. In *AAAI*, pages 2871–2877, 2015.

[Melo *et al.*, 2018] Francisco S. Melo, Carla Guerra, and Manuel Lopes. Interactive optimal teaching with unknown learners. In *IJCAI*, pages 2567–2573, 2018.

[Piech *et al.*, 2015] Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein. Deep knowledge tracing. In *Advances in neural information processing systems*, 2015.

[Rafferty *et al.*, 2016] Anna N Rafferty, Emma Brunskill, Thomas L Griffiths, and Patrick Shafto. Faster teaching via pomdp planning. *Cognitive science*, 2016.

[Sen *et al.*, 2018] Ayon Sen, Purav Patel, Martina A. Rau, Blake Mason, Robert Nowak, Timothy T. Rogers, and Xiaojin Zhu. Machine beats human at sequencing visuals for perceptual-fluency practice. In *EDM*, 2018.

[Settles and Meeder, 2016] Burr Settles and Brendan Meeder. A trainable spaced repetition model for language learning. In *ACL*, pages 1848–1858, 2016.

[Singla *et al.*, 2013] 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.

[Singla *et al.*, 2014] Adish Singla, Ilija Bogunovic, Gábor Bartók, Amin Karbasi, and Andreas Krause. Near-optimally teaching the crowd to classify. In *ICML*, 2014.

[Sullivan *et al.*, 2009] Brian Sullivan, Christopher Wood, Marshall Iliff, Rick Bonney, Daniel Fink, and Steve Kelling. eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation*, 2009.

[Tschiatschek *et al.*, 2019] Sebastian Tschiatschek, Ahana Ghosh, Luis Haug, Rati Devidze, and Adish Singla. Learner-aware teaching: Inverse reinforcement learning with preferences and constraints. In *NeurIPS*, 2019.

[Van Horn *et al.*, 2018] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *CVPR*, 2018.

[Yeo *et al.*, 2019] Teresa Yeo, Parameswaran Kamalaruban, Adish Singla, Arpit Merchant, Thibault Asselborn, Louis Faucon, Pierre Dillenbourg, and Volkan Cevher. Iterative classroom teaching. In *AAAI*, pages 5684–5692, 2019.

[Zhang *et al.*, 2018] Xuezhou Zhang, Xiaojin Zhu, and Stephen J. Wright. Training set debugging using trusted items. In *AAAI*, pages 4482–4489, 2018.

[Zhu *et al.*, 2018] Xiaojin Zhu, Adish Singla, Sandra Zilles, and Anna N. Rafferty. An overview of machine teaching. *CoRR*, abs/1801.05927, 2018.

[Zhu, 2015] Xiaojin Zhu. Machine teaching: An inverse problem to machine learning and an approach toward optimal education. In *AAAI*, pages 4083–4087, 2015.

[Zhu, 2018] Xiaojin Zhu. An optimal control view of adversarial machine learning. *CoRR*, abs/1811.04422, 2018.

[Zilles *et al.*, 2011] Sandra Zilles, Steffen Lange, Robert Holte, and Martin Zinkevich. Models of cooperative teaching and learning. *JMLR*, 12(Feb):349–384, 2011.